# Foreword

It is a great pleasure for the RBLA Editorial Committee to present this first RBLA international special issue, *Corpus studies: future directions*, which focuses on one of the cutting-edge research areas in contemporary linguistics: corpus linguistics and its applications, to the scientific community.

We heartily thank Prof. Stefan Th. Gries, from the University of California at Santa Barbara, for having generously accepted our invitation to be the guest-editor for this special issue which, as the readers will see shortly, reflects his future-oriented thinking for the area.

We are very thankful to our colleagues who submitted their contributions and applied themselves to formulate their insightful directions for the further growth of corpus linguistics. RBLA is honored to be able to offer the international community a special issue that gathers this level of excellence.

We hope our readers will embark in a very inspiring and motivating experience as they explore the pages ahead.

Heliana Mello

# Introduction to this special issue

Stefan Th. Gries
University of California, Santa Barbara

## 1 Introduction

If one asks a corpus linguist how long the field has been around, two answers are heard most often. One would say that corpus linguistic methods have been around for quite some time, would point to early Bible concordances or Käding's (1897) work, would adduce European comparative linguists and American structuralists from the first half of the 20th century as additional examples, etc. The other would say that corpus linguistics really only began to take shape with, on the European stage, Firth's (1951) work on collocation or the work on the Survey of English Usage and/or, on the American stage, Fries's (1952) work on spoken American English, etc.

Regardless of which of these points of view one holds – they are probably both correct from some points of view and corpus linguists might adopt either one over where necessary to make a particular rhetorical move – it is probably no exaggeration to say that it is only over the last 20 years or so, that corpus linguistics has really taken off and developed into one of the most widely-used methods in linguistics. This is visible on many different levels:

- on the level of **resources**: technological developments took place that facilitated the creation of the first mega corpora of the kind exemplified by the British National Corpus;
- on the level of the role that corpus data play in the development and refinement of more comprehensive **theories of language** i.e. in work going beyond mere description. While such developments are still resisted by some – as is the view of corpus linguistics as a 'mere' methodology – (cf. Worlock Pope's (2010) the special issue of the *International Journal of Corpus Linguistics* on the so-called bootcamp discourse) the ways in which corpus linguistics on the one hand and cognitive linguistics and psycholinguistics on the other hand feed into each other is hard to ignore or resist;

- on the level of **statistical methodology**: the overall developmental trend in linguistics towards more quantitative methods can – finally! – also be seen in corpus linguistics. In fact, I have argued elsewhere that, since corpus linguistics is essentially based on nothing but distributional and quantitative data, the field should have been the one to lead the current quantitative revolution in linguistics rather than leaving this honor to, mainly, psycholinguistics …;
- on the level of **competences by practitioners** of the field: many practitioners in the field have long been constrained by a few commercial corpus analysis tools, which limited researchers' ability to think outside of the (software tool) box, the field is now shaping up and many researchers turn to more versatile, powerful, and elegant tools such as the Natural Language Toolkit (cf. <http://www.nltk.org>) or programming languages (cf. Gries 2009 for one example), which finally allows the field to handle the complex types of data in more appropriate ways than was possible before.

By now, corpus linguistics is well established: the field has several international peer-reviewed journals, its own book series with international publishers, a lively conference circuit, and corpus-based methods have contributed to research in most sub-disciplines of linguistics. This also means that researchers don't have to include in their papers justifications or even defenses of why they are using corpus data anymore – corpus linguistics has succeeded to become many of its methods are now mainstream (in a positive sense).

## 2 This special issue

In spite of its impressive success story, corpus linguistics is still in need of maturation and further evolution, and this special issue is devoted to this topic. When I was invited to guest-edit a special issue of the *Brazilian Journal of Applied Linguistics* (*BJAL*) on corpus linguistics, I quickly decided to *not* edit the typical kind of issue in which 'standard' research articles present nice and significant results – my goal became to edit a special issue that outlines where the field of corpus linguistics should go next, an issue that, so to speak, provides direction to the field just as good plenary addresses would do. I thought it was particularly fitting that such a special issue would appear in an open-access journal, which makes the contributions more accessible than copyright restrictions of some commercial journals often allow for so I was delighted that the editorial team of BJAL accepted this plan.

The next step consisted of identifying a range of fields which I considered benefited much from, and contributed much to, corpus linguistics as well as persuading a range of prominent scholars in these fields to contribute to this special issue a paper that answered the following question:

> In your area of research and in your work with corpora – and I am writing to you because of your work in _____ – where do you think the field of corpus linguistics has to go and/or mature, and why? What are developments in terms of resources, standards, technology, methods, etc. that you think are essential and/or at least desirable, and why, or what can we do then?

I was very lucky to receive affirmative and encouraging responses from high-profile colleagues for a number of linguistic areas or sub-disciplines, which are listed in Table 1. Each of the papers outlines answers to the above guiding questions in its own way, usually providing a short state-of-the-art overview, followed by perspectives, recommendations, lists of desiderata, case studies, and much more that should give the field food for thought for the foreseeable future – they certainly did that for me.

As a final note, a heartfelt 'thank you!' is due to my associate editor at *BJAL*, Heliana Ribeiro de Mello, without whom this special issue would not have materialized. And, I would of course also like to express my sincere thanks to the contributors, who agreed to contribute to a special issue with a somewhat unusual focus and who sent in thoughtful and inspiring papers that clearly outline how corpus linguistics can evolve further in ways that no single author ever could. If this special issue gets you thinking and planning, they deserve all the credit for that.

TABLE
Overview of this special issue

| Area/sub-discipline: corpora and… | Authores |
|---|---|
| … quatitative research/methods | R Harald Baayen (University of Tübingen) |
| … metaphor research | Tony Berber Sardinha (Catholic University of São Paulo) |
| … sociolinguistics | Tyler Kendall (University of Oregon, Eugene) |
| … multi-modal data | Dawn Knight (University of Nottingham) |
| … historical linguistics | Merja Kytö (Uppsala University) |
| … second/foreign language learning | Fanny Meunier (Catholic University of Louvain) |
| … discourse pragmatics | Massimo Moneglia (University of Florença) |
| … cognitive linguistics | John Newman (University of Alberta, Edmonton) |
| … dialectology | Benedikt Szmrecsanyi & Christoph Wolk<br>(Freiburg Institute for Advanced Studies ) |

## References

FIRTH, J.R. *Papers in linguistics*, 1934-1951. Oxford: Oxford University Press, 1951.

FRIES, C. C. *The structure of English*: an introduction to the construction of English sentences. New York: Harcourt Brace, 1952.

GRIES, St.Th. *Quantitative corpus linguistics with R*: a practical introduction. London / New York: Routledge, Taylor & Francis Group, 2009.

KÄDING, F.W. *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: no publ., 1897.

WORLOCK POPE, C. (Ed.). The bootcamp discourse and beyond. Special issue of the *International Journal of Corpus Linguistics,* v. 15, n. 2, 2010.

# *Corpus linguistics and naive discriminative learning*
# *A linguística de corpus e a aprendizagem discriminativa ingênua*

R. Harald Baayen
Universität Tübingen
Tübingen / Germany

ABSTRACT: Three classifiers from machine learning (the generalized linear mixed model, memory based learning, and support vector machines) are compared with a naive discriminative learning classifier, derived from basic principles of error-driven learning characterizing animal and human learning. Tested on the dative alternation in English, using the Switchboard data from (BRESNAN; CUENI; NIKITINA; BAAYEN, 2007), naive discriminative learning emerges with state-of-the-art predictive accuracy. Naive discriminative learning offers a united framework for understanding the learning of probabilistic distributional patterns, for classification, and for a cognitive grounding of distinctive collexeme analysis.

KEYWORDS: machine learning; dative alternation; Switchboard; probabilistic distributional patterns; collexeme analysis.

RESUMO: Três classificadores de aprendizagem de máquina (modelos mistos lineares generalizados, aprendizagem baseada na memória e máquinas de apoio a vetores) são comparados com o classificador da aprendizagem discriminativa ingênua, derivada de princípios básicos da aprendizagem guiada por erros de humanos e animais. Testada na alternância dativa do inglês, usando os dados do Switchboard (BRESNAN; CUENI; NIKITINA; BAAYEN, 2007), a aprendizagem discriminativa ingênua emerge com uma acurácia predicativa no estado da arte. A aprendizagem discriminativa ingênua oferece um arcabouço unificado para a compreensão da aprendizagem de padrões distribucionais probabilísticos, para a classificação, e para um embasamento cognitivo para a análise de colexemas distintivos.

PALAVRAS-CHAVE: aprendizagem de máquinas; alternância; dativa; Switchboard; padrões de distribuição probabilística; análise de colexemas.

* baayen@ualberta.ca

According to Gries (2011), linguistics is a distributional science exploring the distribution of elements at all levels of linguistic structure. He describes corpus linguistics as investigating the frequencies of occurrence of such elements in corpora, their dispersion, and their co-occurrence properties. Although this characterization of present-day corpus linguistics is factually correct, the aim of the present paper is to argue that corpus linguistics should be more ambitious, and that for a better understanding of the data its current descriptive approach may profit from complementation with cognitive computational modeling.

Consider the dative alternation in English. Bresnan *et al.* (2007) presented an analysis of the dative alternation in which the choice between the double object construction (*Mary gave John the book*) and the prepositional object construction (*Mary gave the book to John*) was modeled as a function of a wide range of predictors, including the accessibility, definiteness, length, and animacy of theme and recipient (see also FORD; BRESNAN, 2010). A mixed-effects logistic regression model indicated that their variables were highly successful in predicting which construction is most likely to be used, with approximately 94% accuracy.

The statistical technique used by Bresnan and colleagues, logistic regression modeling, is but one of many excellent statistical classifiers currently available to the corpus linguist, such as memory based learning (MBL, DAELEMANS; BOSCH, 2005), analogical modeling of language (AML, SKOUSEN, 1989), support vector machines (SVM, VAPNIK, 1995), and random forests (RF, STROBL; MALLEY; TUTZ, 2009; TAGLIAMONTE; BAAYEN, 2010). The mathematics underlying these techniques varies widely, from iterative optimization of the model fit (regression), nearest-neighbor similarity-based inference (memory based learning), kernel methods (support vector machines), and recursive conditioning with subsampling (random forests). All these statistical techniques tend to provide a good description of the speaker-listener's knowledge, but it is unlikely that they provide a good characterization of how speaker-listeners actually acquire and use this knowledge. Of these four techniques, only memory-based learning, as a computational implementation of an exemplar-based model, may arguably reflect human performance.

A first question addressed in the present study is whether these different statistical models provide a correct characterization of the knowledge that a speaker has of how to choose between these two dative constructions. A statistical model may faithfully reflect a speaker's knowledge, but it is also conceivable that it underestimates or overestimates what native speakers of English actually have internalized. This question will be addressed by comparing statistical models with a model based on principles of human learning.

A second question concerns how frequency of occurrence and co-occurrence frequencies come into play in human classification behavior as compared to machine classification. For machine classification, we can easily count how often a linguistic element occurs, and how often it co-occurs with other elements. The success of machine classification in reproducing linguistic choice behavior suggests that probabilities of occurrence are somehow available to the human classifier. But is frequency of (co-)occurrence available to the human classifier in the same way as to the machine classifier? Simple frequency of occurrence information is often modeled by means of some 'counter in the head', implemented in cognitive models in the form of 'resting activation levels', as in the interactive activation models of McClelland and Rumelhart (1981); Coltheart, Rastle, Perry, Langdon, and Ziegler (2001); Van Heuven, Dijsktra, and Grainger (1998), in the form of frequency based rankings (MURRAY; FORSTER, 2004), as a unit's verification time (LEVELT, ROELOFS; MEYER, 1999), or in the Bayesian approach of Norris, straightforwardly as a unit's long-term a-priori probability (NORRIS, 2006; NORRIS; McQUEEN, 2008). A potential problem that arises in this context is that large numbers of such 'counters in the head' are required, not only for simple or complex words, but also for hundreds of millions of word *n*-grams, given recent experimental results indicating human sensitivity to *n*-gram frequency (ARNON; SNIDER, 2010; TREMBLAY; BAAYEN, 2010). Moreover, given the tendency of human memory to merge, or blend, previous experiences, it is rather unlikely that the human classifier has at its disposal exactly the same frequency information that we make available to our machine classifiers.

To address these questions, the present study explores what a general model of human learning may offer corpus linguistics as a computational theory of human classification.

TABLE 1
Example instance base for discriminative learning with the Rescorla-Wagner equations, with as cues the definiteness and pronominality of the theme, and as outcome the construction (double object, NP NP, versus prepositional object, NP PP)

| Frequency | Definiteness of Theme | Pronominality of Theme | Construction |
|---|---|---|---|
| 7 | definite | non-pronominal | NP NP |
| 1 | definite | pronominal | NP NP |
| 28 | indefinite | non-pronominal | NP NP |
| 1 | indefinite | pronominal | NP NP |
| 3 | definite | non-pronominal | NP PP |
| 4 | definite | pronominal | NP PP |
| 6 | indefinite | non-pronominal | NP PP |
| 0 | indefinite | pronominal | NP PP |

## 1. Naive Discriminative Learning

In psychology, the model of Wagner and Rescorla (1972) is one of the most influential and fruitful theories of animal and human learning (MILLER; BARNET; GRAHAME, 1995; SIEGEL; ALLAN, 1996). Its learning algorithm is closely related to the connectionist delta-rule (cf. GLUCK; BOWER, 1988; ANDERSON, 2000) and to the Kalman filter (cf. DAYAN; KAKADE, 2001), and can be viewed as an instantiation of a general probabilistic learning mechanism (see, e.g., CHATER; TANENBAUM; YUILLE, 2006; HSU, CHATER; VITÁNYI, 2010).

### 1.1. The Rescorla-Wagner equations

Rescorla and Wagner formulated a set of equations that specify how the strength of association of a cue in the input to a given outcome is modified by experience. By way of example, consider the instance base in Table 1, which specifies for the four combinations of the pronominality and definiteness of the theme (*the book* in *John gave the book to Mary*) which construction is used (the double object construction, NP NP, or the prepositional object construction, NP PP). The eight possible combinations occur with different frequencies, modeled on the data of Bresnan *et al.* (2007). The cues in this example are the values for definiteness and pronominality. The outcomes are the two constructions. There are in all 50 learning trials, more than half of which pair an indefinite non-pronominal theme with the double object construction (e.g., *John gave a book to Mary*).

The Rescorla-Wagner equations implement a form of supervised learning. It is assumed that the learner predicts an outcome given the available cues. Depending on whether this prediction is correct, the weights (association strengts) from the cues to the outcomes are adjusted such that at subsequent trials, prediction accuracy will improve.

Let PRESENT $(C, t)$ denote the presence of a cue $C$ (definiteness, pronominality) and PRESENT $(O, t)$ the presence of outcome $O$ (construction) at time $t$, and let ABSENT $(C, t)$ and ABSENT $(O, t)$ denote their absence at time $t$. The Rescorla-Wagner equations specify the association strength $V_i^{t+1}$ of cue $C_i$ with outcome $O$ at time $t+1$ by means of the recurrence relation

(1) $$V_i^{t+1} = V_i^{t} + \Delta V_i^{t},$$

which simply states that the association strength at time $t + 1$ is equal to its previous association strength at time $t$ modified by some change change in association strength $\Delta V_i^t$, defined as

(2) $$\Delta V_i^t = \begin{cases} 0 & \text{if } \textsc{absent}\ (C_i, t), \\ \alpha_i \beta_1 \left(\lambda - \sum_{\mathit{PRESENT}\ (C_j,\ t)} Vj\right) & \text{if } \textsc{present}\ (C_j, t)\ \&\ \textsc{present}\ (O, t), \\ \alpha_i \beta_2 \left(0 - \sum_{\mathit{PRESENT}\ (C_j,\ t)} Vj\right) & \text{if } \textsc{present}\ (C_j, t)\ \&\ \textsc{absent}\ (O, t). \end{cases}$$

Standard settings for the parameters are $\lambda = 1$, $\alpha_1 = \alpha_2 = 0.1$, $\beta_1 = \beta_2 = 0.1$. If a cue is not present in the input, its association strength is not changed. When the cue is present, the change in association strength depends on whether or not the outcome is present. Association strengths are increased when cue and outcome co-occur, and decreased when the cue occurs without the outcome. Furthermore, when more cues are present simultaneously, adjustments are more conservative. In this case, we can speak of cue competition.

Figure 1 illustrates, for a random presentation of the 50 learning trials, how the association strengths (or weights) from cues to outcomes develop over time. As indefinite nonpronominal themes dominate the instance base, and strongly favor the double object construction, the weights from the cues **indefinite** and **non-pronominal** to the construction NP NP increase steadily during the learning process.

## 2. The equilibrium equations for the Rescorla-Wagner equations

The Rescorla-Wagner equations have recently turned out to be of considerable interest for understanding child language acquisition, see, for instance, Ramscar and Yarlett (2007); Ramscar, Yarlett, Dye, Denny, and Thorpe (2010); Ramscar, Dye, Popick, and O'Donnell-McCarthy (2011). For corpus linguistics, the equilibrium equations for the Rescorla-Wagner equations developed by Danks (2003) are of key interest. Danks was able to derive a set of equations that define the association strengths (weights) from cues to outcomes for the situation in which these strengths no longer change, i.e., for the adult state of the learner. It can be shown that when

(3) $$V_i^{t+1} = V_i^t, \text{ or, equivalently,}$$

(4) $$V_i^{t+1} - V_i^t = 0,$$

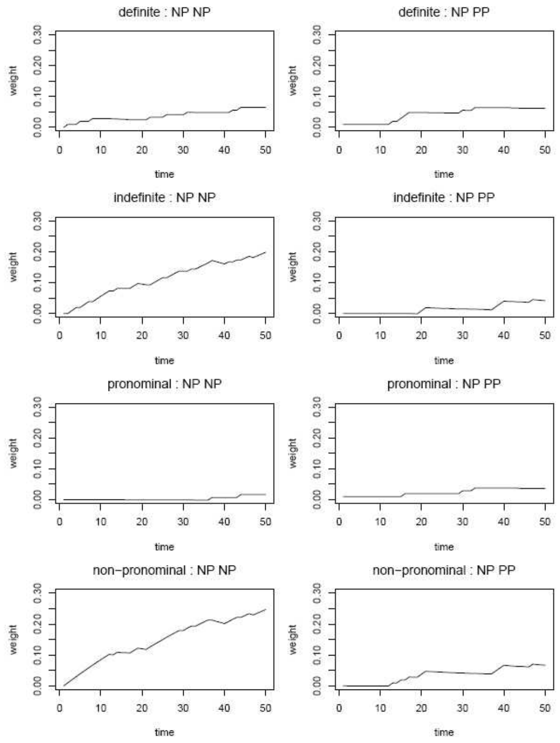


FIGURE 1 – Development of the weights from cues (definite/indefinite/pronominal/non-pronominal) to outcomes (NP NP/NP PP) given the instance base summarized in Table 1. The 50 instance tokens were presented for learning once, in random order.

the weights to the outcomes can be estimated by solving the following set of equations, with *W* the matrix of unknown weights:[1]

(5) $$CW = O.$$

In (5), *C* is the matrix of conditional probabilities of the outcomes. It is obtained by first calculating the matrix *M* listing the frequencies with which cues co-occur:

(6)
$$M = \left(\begin{array}{lcccc} & \text{indefinite} & \text{pronominal} & \text{nonpronominal} & \text{definite} \\ \text{indefinite} & 35 & 1 & 34 & 0 \\ \text{pronominal} & 1 & 6 & 0 & 5 \\ \text{nonpronominal} & 34 & 0 & 44 & 10 \\ \text{definite} & 0 & 5 & 10 & 15 \end{array}\right)$$

As can be verified by inspecting Table 1, the cue **indefinite** occurs 35 times, the combination of **indefinite** and **pronominal** occurs once, **indefinite** co-occurs 34 times with **non-pronominal**, and so on. From this matrix, we derive the matrix of conditional probabilies of cue *j* given cue *i*:

(7)
$$C = \left(\begin{array}{lcccc} & \text{indefinite} & \text{pronominal} & \text{nonpronominal} & \text{definite} \\ \text{indefinite} & 0.50 & 0.01 & 0.49 & 0.00 \\ \text{pronominal} & 0.08 & 0.50 & 0.00 & 0.42 \\ \text{nonpronominal} & 0.39 & 0.00 & 0.50 & 0.11 \\ \text{definite} & 0.00 & 0.17 & 0.33 & 0.50 \end{array}\right)$$

The probability of **indefinite** given **indefinite** is 35/(35+1+34+0)=0.5, that of **indefinite** given **pronominal** is 1/(1+6+0+5)=0.083, and so on.

The matrix *W* is the matrix of association strengths from cues (rows) to outcomes (columns) that we want to estimate. Finally, the matrix ***O***,

(8)
$$O = \left(\begin{array}{lcc} & \text{NP NP} & \text{NP PP} \\ \text{indefinite} & 0.41 & 0.09 \\ \text{pronominal} & 0.17 & 0.33 \\ \text{nonpronominal} & 0.40 & 0.10 \\ \text{definite} & 0.27 & 0.23 \end{array}\right)$$

[1] Equation (5) is formulated using notation from matrix algebra. The following example ilustrates the principle of the calculation involved.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} v & w \\ x & y \end{pmatrix} = \begin{pmatrix} av + bx & aw + by \\ cv + dx & bw + dy \end{pmatrix}$$

lists the conditional probabilities of the constructions (columns) given the cues (rows). It is obtained from the co-occurrence matrix of cues (*M*) and the co-occurrence matrix of cues and constructions *N*,

(9)

$$N = \left(\begin{array}{lrr} & \text{NP NP} & \text{NP PP} \\ \text{indefinite} & 29 & 6 \\ \text{pronominal} & 2 & 4 \\ \text{nonpronominal} & 35 & 9 \\ \text{definite} & 8 & 7 \end{array}\right)$$

For instance, the probability of the double object construction given (i) the **indefinite** cue is 29/(35+1+34+0)=0.414, and given (ii) the **pronominal** cue it is 2/(1+6+0+5)=0.167. The set of equations (5) can be solved using the generalized inverse, which will yield a solution that is optimal in the least-squares sense, resulting in the weight matrix

TABLE 2
Probabilities of the two constructions following from the equilibrium equations for the Rescorla-Wagner model

| | indefinite | indefinite | definite | definite |
|---|---|---|---|---|
| | non-pronominal | pronominal | non-pronominal | pronominal |
| NP NP | 0.84 | 0.49 | 0.65 | 0.3 |
| NP PP | 0.16 | 0.51 | 0.35 | 0.7 |

(10)

$$W = \left(\begin{array}{lrr} & \text{NP NP} & \text{NP PP} \\ \text{indefinite} & 0.38 & 0.12 \\ \text{definite} & 0.19 & 0.31 \\ \text{nonpronominal} & 0.46 & 0.04 \\ \text{pronominal} & 0.11 & 0.39 \end{array}\right)$$

The support for the two constructions given a set of input cues is obtained by summation over the association strengths (weights) of the active cues in the input. For instance, for indefinite non-pronominal themes, the summed support for the NP NP construction is 0.38+0.46=0.84, while the support for the NP PP construction is 0.12+0.04=0.16. Hence, the probability of the double object construction equals 0.84/(0.84+0.16)= 0.84, and that for the prepositional object construction is 0.16. (In this example, the two measures of support sum up to one, but this is not generally the case for more complex data sets.) One can think of the weights being chosen in such a way that, given the co-occurrences of cues and outcomes, the probability of a construction given the different cues in the input is optimized.

We can view this model as providing a re-representation of the data: Eight frequencies (see Table 1) have been replaced by eight weights, representing 50 trials of learning. The model does not work with exemplars, nevertheless, its weights do reflect exemplar frequencies. For instance, the probabilities of the double object construction in Table 2 are correlated with the original frequencies ($r_s$=0.94, $p$=0.051). It is worth noting that the probabilities in Table 2 are obtained with a model that is completely driven by the input, and that is devoid of free parameters – the learning parameters of the Rescorla-Wagner equations (2) drop out of the equilibrium equations.

Baayen, Millin, Filipovic Durdjevic, Hendrix, and Marelli (2011) made use of discriminative learning to model visual lexical decision and self-paced reading latencies in Serbian and English. They obtained excellent fits to empirical latencies, both in terms of good correlations at the item level, as well as in terms of the relative importance and effect sizes of a wide range of lexical distributional predictors. Simulated latencies correctly reflected morphological family size effects as well as whole-word frequency effects for complex words, without any complex words being represented in the model as individual units. Their model also predicts word n-gram frequency effects (see also BAAYEN; HENDRIX, 2011). It provides a highly parsimonious account of morphological processing, both in terms of the representations it assumes, and in terms of the extremely limited number of free parameters that it requires to fit the data. For monomorphemic words, the model is essentially parameter free, as in the present example for the dative alternation.

Baayen *et al.* (2011) refer to the present approach as *naive* discriminative learning, because the probability of a given outcome is estimated independently from all other outcomes. This is a simplification, but thus far it seems that this simplification does not affect performance much, just as often observed for *naive Bayes* classifiers, while making it possible to obtain model predictions without having to simulate the learning process itself.

The question to which we now turn is to what extent naive discriminative learning provides a good fit to corpus data. If the model provides decent fits, then, given that it is grounded in well-established principles of human learning, and given that it performs well in simulations of human processing costs at the lexical level, we can compare discriminative learing with well-established statistical methods in order to answer the question of whether human learning is comparable, superior, or inferior to machine learning. We explore this issue by a more comprehensive analysis of the dative alternation data.

## 3. Predicting the dative alternation

From the **dative** dataset in the **languageR** package (BAAYEN, 2000), the subset of data points extracted from the Switchboard corpus were selected for further analysis. For this subset of the data, information about the speaker is available. In what follows, the probability of the prepositional object construction is taken as the response variable. Software for naive discriminative classification is available in the **ndl** package for R, available at **www.r-project.org**. Example code is provided in the appendix.

### 3.1. Prediction accuracy

A discriminative learning model predicting construction (double object versus prepositional object) was fitted with the predictors Verb, Semantic Class, and the Animacy, Definiteness, Pronominality, and Length of recipient and theme. As the model currently requires discrete cues, as a workaround, the length of recipient and theme were split into three ranges: length 1, lengths 2-4, and lengths exceeding 4. These three length levels were used as cues, instead of the original numerical values. As Brenan *et al.* (2007) did not observe significant by-speaker variability, speaker is not included as a predictor in our initial model. (Models including speaker as predictor will be introduced below.)

To evaluate goodness of fit, we used two measures, the index of concordance *C* and the model's accuracy. The index of concordance *C* is also known as the receiver operating characteristic curve area 'C' (see, e.g. HARRELL, 2001). Values of *C* exceeding 0.8 are generally regarded as indicative of a succesful classifier. Accuracy was defined here as the proportion of correctly predicted constructions, with as cut-off criterion for a correct prediction that the probability for the correct prediction exceed 0.5. According to these measures, the naive discriminative learning model performed well, with *C*=0.97 and an accuracy of 0.92.

To place the performance of naive discriminative learning (NDL) in perspective, we compared it with memory based learning (MBL), logistic mixed-effects regression (GLMM), and a support vector machine with a linear kernel (SVM). The index of concordance obtained with MBL, using TiMBL version 6.3 (DAELEMANS; ZAVREL; SLOOT; BOSCH, 2010), was *C*=0.89. Its accuracy was 0.92. TiMBL was supplied with speaker information.

A logistic mixed-effects regression model, fitted with the LME4 package for R (BATES, D.; MAECHLER, 2009), with both Speaker and Verb as random-effect factors did not converge. As the GLMM did not detect significant speaker-bound variance, we therefore fitted a model with verb as only random-effect factor, including length of theme and recipient as (numerical) covariates. The index of concordance for this model was *C*=0.97, accuracy was at 0.93. The regression model required 18 parameters (one random-effect standard deviation, an intercept, and 16 coefficients for slopes and contrasts) to achieve this fit. A support vector machine, provided with access to Speaker information, and fitted with the **svm** function in the E1017 package for R (DIMITRIADOU; HORNICK; LEISCH; MEYER; WEINGESSEL, 2009), yielded *C*=0.97 with accuracy at 0.93, requiring 524 support vectors.

From this comparison, naive discriminative learning emerges as more or less comparable in classificatory accuracy to existing state-of-the-art classifiers. It is outperformed in both *C* and accuracy only by the support vector machine, the currently best-performing classifier available. We note here that the NDL classifier used here is completely parameter-free. The weights are fully determined, and only determined, by the corpus input. There are no choices that the user could make to influence the results.

Since speaker information was available to TiMBL and to the SVM, we fitted a second naive discriminative learning model to the data, this time including speaker as a predictor. The index of concordance increased slightly to 0.98, and accuracy to 0.95. Further improvement can be obtained by allowing pairs of predictor values to function as cues, following the naive discriminative reader model of Baayen *et al.* (2011). They included both letters and letter bigrams as cues, the former representing static knowledge of which letters are present in the input, the latter representing information about sequences of letters. Analogously, pairs of features, e.g., semantic class **p** combined with a **given theme**, can be brought into the learning process. This amounts to considering (when calculating the conditional co-occurrence matrix ***C***

TABLE 3
Index of concordance *C* and accuracy for all data (left) and average across 10-fold cross-validation

| | all data | | 10-fold cross-validation | |
|---|---|---|---|---|
| | C | Accuracy | C | Accuracy |
| SVM | 0.98 | 0.95 | 0.95 | 0.91 |
| TiMBL | 0.89 | 0.92 | 0.89 | 0.92 |
| GLMM | 0.97 | 0.93 | 0.96 | 0.92 |
| NDL (verb) | 0.97 | 0.92 | 0.89 | 0.85 |
| NDL (verb+speaker) | 0.98 | 0.95 | 0.93 | 0.89 |
| NDL-2 (verb+speaker) | 0.99 | 0.96 | 0.94 | 0.91 |

not only pairwise co-occurrences of cues, but also the co-occurrences of triplets and quadruplets of cues. Within the framework of naive discriminative learning, this is the functional equivalent of interactions in a regression model. In what follows, NDL-2 refers to a model which includes pairs of features for all predictors, excluding however pairs involving Verb or Speaker. With this richer representation of the input, the index of concordance increased to 0.99 and accuracy to 0.96.

However, we now need to assess whether naive discriminative learning achieves this good performance at the cost of overfitting. To assess this possibility, we made use of 10-fold cross-validation, using exactly the same folds for each of the classifiers. The right half of Table 3 summarizes the results. In cross-validation, naive discriminative learning performs less well than the SVM and the GLMM, but similar to TiMBL. Fortunately, concordance and accuracy remain high.

We are now in the position to tentatively answer our first question, of whether machine learning outperforms human learning. If naive discriminative learning is indeed a reasonable approximation of human learning, then the answer is that human learning builds a representation of past experience comparable to that of other machine learning techniques. However, for generalization to unseen, new data, human classification seems thus far to be outperformed, albeit only slightly, by some of the best machine classifiers currently available.

### 3.2. Effect sizes and variable importance

One of the advantages of regression models for linguistic analysis is that the estimated coefficients offer the researcher insight into what forces shape the probabilities of a construction. For instance, a pronominal theme is assigned a *b* weight of 2.2398 on the log odds scale, indicating that pronominal themes are much more likely to be expressed in a prepositional object construction than in a double object construction. This kind of information is more difficult to extract from a support vector machine or from a memory based model, for which one has to inspect the support vectors or the similarity neighborhoods respectively. Interestingly, the weights of the naive discriminative learner provide the same kind of information as the coefficients of the regression model. For instance, in the model with verb (and without speaker), a non-pronominal theme has a negative weight equal to -0.046 for the prepositional object construction, whereas a pronominal theme has a positive weight of 0.203. The difference between the two, henceforth the NDL treatment contrast, is 0.248. This difference should be similar to the

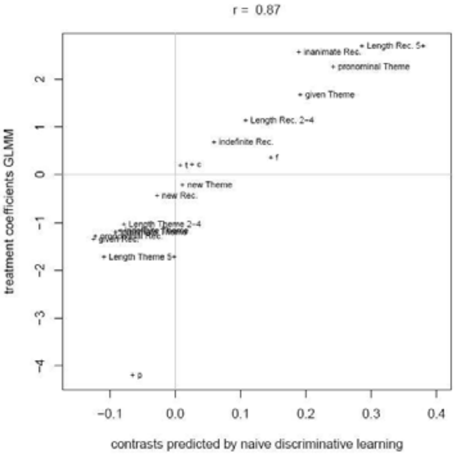


FIGURE 2 – Treatment contrasts generated from the association strengths of the naive discriminative learner (horizontal axis) and the treatment constrasts of a generalized linear mixed-effects model (vertical axis). Semantic class: reference level **abstract** (*give it some thought*); c: communication (*tell, give me your name*); f: future transfer of possession (*owe, promise*); p: prevention of possession (*cost, deny*); t: transfer of possession (*give an armband, send*). Reference levels for the other predictors are **animate**, **definite**, **accessible**, **'non-pronominal**, and **Length 1**.

treatment contrast for the pronominality of the theme, which is defined as the difference (on the logit scale) between a pronominal theme and the reference level of the non-pronominal theme. When we plot the NDL treatment contrast together with the treatment coefficients of the logistic regression model, we find that the two enter into a strong correlation, r = 0.87 (*t*(16) = 7.18, *p* = 0), as can be seen in Figure 2.

For sparse data, the naive discriminative learner tends to be more conservative than the GLMM. The +p data point in the lower left of Figure 2 represents the 'prevention of possession' semantic class, which supports 182 instances with the double object construction and only one case with the prepositional object construction. The logistic regression model concludes that a prepositional object construction is extremely unlikely, assigning +p verbs a negative weight of no less than -4. The naive discriminative learner is assigning this type of verb a larger, though still small, probability.

In order to assess the importance of a predictor for classification accuracy across the very different classifiers considered above, we permute the values of the predictor in order to break its potential relation with the dependent variable. We then inspect to what extent classification accuracy decreases. The greater the decrease in classification accuracy, the greater the importance of the predictor. This non-parametric approach is inspired by how variable importance is assessed for random forests, which are also non-parametric classifiers (see, e.g., STROBL *et al*, 2009). Figure 3 summarizes the results for the regression model, the support vector machine, and for naive discriminative learning.

First consider variable importance for the regression model, summarized in the upper left panel. The pronominality of the theme emerges as the most important predictor for regression accuracy, followed by verb, and at a distance, by the definiteness of the theme. Semantic class has a negative score, indicating that random permutation of its values resulted in slightly improved accuracy. By chance, the random reordering of values resulted in a configuration that affords a slightly better model to be fitted. This may arise when the values of an irrelevant predictor are reshuffled. By mirroring the minimal score to the right of zero, we obtain an interval that characterizes irrelevant predictors (see, e.g., STROBL *et al.*, 2009, for this logic in the context of random forests). For the regression model, this interval contains, in addition to Semantic Class, the predictors Animacy of Theme and Definiteness of Recipient.

The support vector machine comes to rather different conclusions. Its classification accuracy is very sensitive to having access to verb and speaker information. Accuracy is also affected negatively by removal of the predictors specifying the pronominality of theme and recipient, as well as the length of the recipient.

Predictors marked as important by naive discriminative learning are the animacy of the recipient, the length of the theme, the pronominality of the theme, the identity of the verb, and the accessibility of the theme. Speaker is characterized as having no importance, in accordance with the GLMM but contrary to the results obtained with the SVM.

For all models, overall accuracy (which is in the nineties) is hardly affected by permuting the values of a single predictor. This especially striking for the naive discriminative learning model with cue pairs (lower right panel), for which the reductions in accuracy are an order of magnitude smaller than those for the other models (note the different scale on the horizontal axis in the lower right panel of Figure 3). Apparently, this model is exceptionally robust against noise predictors.

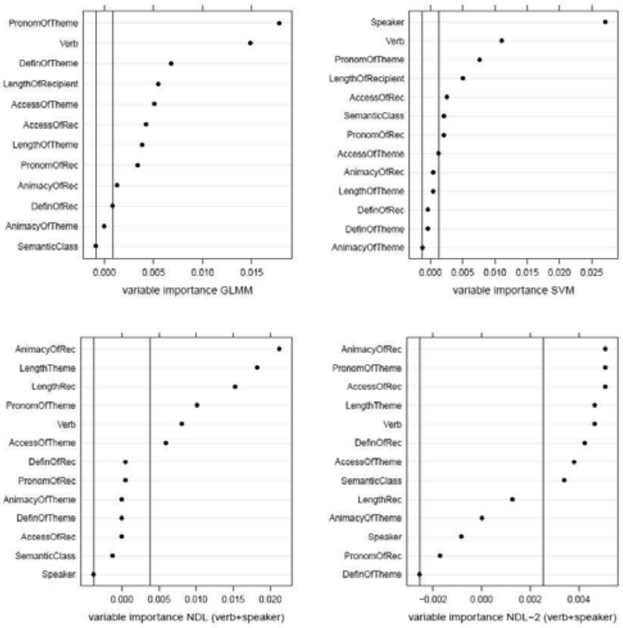


FIGURE 3 – Permutation accuracy importance: the reduction in accuracy for predicting the prepositional object construction when a predictor is randomly permuted, for mixed-effects logistic regression (upper left), a support vector machine (upper right), naive discriminative learning (lower left), and naive discriminative learning with feature pairs (lower right).

The minor effect of variable permutation also indicates that, apparently, individual predictors are not that important. This is in all likelihood a consequence of the correlational structure characterizing the predictor space. For the dative set, each of the predictors listed in Table 4 can be predicted from the other predictors, with 2 up to 6 of the other predictors having significant coefficients ($p$ <0.05), and with prediction accuracies up to 95%. Although this kind of rampant collinearity can pose serious problems for statistical analysis (in fact, a conditional variable importance measure (STROBL; BOULESTEIX; KNEIB; AUGUSTIN; ZEILEIS, 2008) for random forests would be a better choice than the straightforward permutation measure used above), it probably provides exactly the redundancy that makes human learning of language data robust. The improvement in classification accuracy of the naive discriminative learner when provided with feature pairs instead of single features as cues provides further support for the importance of redundancy. By making richer co-occurrence information available to the model, classification accuracy increases. The other side of the same coin is that permuting one predictor's values leaves prediction accuracy virtually unchanged: The 'functional' burden of individual predictors is small.

TABLE 4
Prediction accuracy and number of significant predictors for (logistic) regression models predicting one predictor from the remaining other predictors

| | Accuracy | Number of Significant Predictors |
|---|---|---|
| Animacy of Recipient | 0.93 | 3 |
| Definiteness of Recipient | 0.91 | 2 |
| Pronominality of Recipient | 0.91 | 5 |
| Accessibility of Recipient | 0.95 | 4 |
| Length of Recipient | 0.33 | 3 |
| Animacy of Theme | 0.03 | 4 |
| Definitiness of Theme | 0.79 | 4 |
| Pronominality of Theme | 0.90 | 4 |
| Accessibility of Theme | 0.76 | 6 |
| Length of Theme | 0.04 | 2 |

### 3.3. Non-normal speaker variability

From a methodological perspective, it is noteworthy that Figure 3 clarifies that the importance of individual predictors is evaluated rather differently by the different models. The information gain ratios used by TiMBL to evaluate exemplar similarity, not shown here, provide yet another, and again different, ranking of variable importance. In the light of this diversity, one would hope that the variable importance suggested by models that are cognitively more realistic is closer to the truth. Whether this is indeed the case for naive discriminative learning awaits further validation, perhaps through psycholinguistic experimentation.

In what follows, we focus on one particularly salient difference, the discrepancy between the SVM and the other models when it comes to the importance of Speaker. Figure 4 visualizes the distributions of the contributions of the verb and speaker weights to the probability of the prepositional object construction in NDL-2, as well as the random intercepts for the verbs in the generalized linear mixed model. The left panels show estimated probability density functions, the right panels the corresponding quantile-quantile plots.

The top panels present the NDL-2 weights for the associations of verbs to the prepositional object construction in the naive discriminative learning model. These weights follow, approximately, a normal distribution. The central panels graph the distribution of the random intercepts for the verbs in the GLMM, these also roughly follow a normal distribution. The NDL-2 verb weights and the GLMM random intercepts for verbs correlate well, r = 0.77 ($t(36) = 7.18, p = 0$), indicating that the two models are representing the same variation in a very similar way.

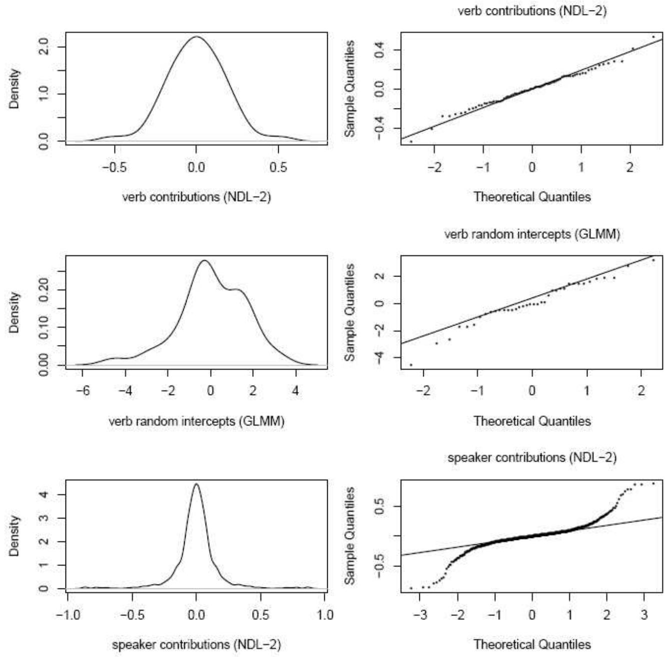


FIGURE 4 – Distributions of the contributions of the individual verbs (top) and speakers (bottom) to the likelihood of the prepositional object construction, and the by-verb random intercepts in the generalized linear mixed model (center panels)

The bottom panels summarize the distribution of the association strengths from speakers to the prepositional object construction in the NDL. These weights are characterized by a symmetrical distribution that, however, deviates markedly from normality. There are too many very small weights close to zero, combined with long but slim tails with outliers. This is, at least in part, due to the sparsity of information on individual speakers (the median number of observations for Speaker is 4, less than half of the median for Verb, 10.4).

The generalized linear mixed model builds on the assumption that random intercepts follow a normal distribution. For the speakers, this assumption is clearly violated. The mixed-effects model either fails to detect non-normally-distributed speaker variability, or infers that including speaker as random-effect factor does not lead to improved prediction. As the GLMM slightly outperforms the SVM under cross-validation, it seems likely that the SVM may be overfitting the data. The permutation variable importance for speaker in the naive discriminative learning models points in the same direction.

Returning to the difference between machine learning and human learning, the performance of naive discriminative learning suggests that human learning might be sensitive to variation (such as variation coming with individual speakers) that machine learning would back off from. However, for the human learner, thanks to the highly redundant nature of the set of predictors, the consequences of human overfitting seem negligible.

## 4. Naive discriminative learning and distinctive collexeme analysis

We have seen that naive discriminative learning provides a statistical tool for classification that, at least for the present data set, performs comparably to other state-of-the-art statistical classifiers. Crucially, naive discriminative classification is theoretically motivated as the end state of human discriminative learning. Over time, very simple adjustments to the association strengths of verbs to constructions result in excellent classification performance. The aim of this section is to show that within this new approach, a measure for distinctive collexeme analysis can be straightforwardly formulated.

Distinctive collexeme analysis (GRIES; STEFANOWITSCH, 2004) quantifies to what extent a word is attracted to a particular construction. For instance, for the verb *take*, a contingency table (Table 5) serves as the input to a Fisher exact test of independence. The p-value produced by this test is log-transformed. The absolute value of the resulting measure is used to gauge attraction to or repulsion from a given construction. For *take*, distinctive collexeme strength is 35.7, indicating extremely strong attraction to the prepositional object construction. (Here, and in what follows, the focus is on the prepositional object construction.)

From a statistical perspective, it is somewhat odd to derive a measure from a p-value. An alternative approach is to make use of a measure from information theory, the Kullback-Leibler divergence, also known as relative entropy. Relative entropy specifies the difference between two probability distributions. The first probability distribution, *p*, concerns the probabilities of the two constructions for the verb *take*. The second probability distribution, *q*, specifies the probabilities of the two constructions in general.

TABLE 5
Contingency table for distinctive collexeme analysis of *take*

|  | NP NP | NP PP |
|---|---|---|
| *take* | 2 | 56 |
| other verbs | 1857 | 445 |

TABLE 6
The probability distribution *p* and *q* required for the calculation of the relative entropy for *take*

|  | *p* | *q* |
|---|---|---|
| double object construction | 2/(2+56) | (2+1857)/(2+56+1857+445) |
| prepositional object construction | 56/(2+56) | (56+445)/(2+56+1857+445) |

From Table 5 these probabilities can be obtained straightforwardly, as shown in Table 6. Given the two distributions *p* and *q*, their relative entropy is defined as

(11) $$\mathrm{RE}\,(\mathrm{p}, \mathrm{q}) = \sum_i p_i \log_2 \frac{p_i}{q_i},$$

which for *take* evaluates to 1.95.

Alternatively, the $\Delta P$ measure (ALLAN, 1980; ELLIS, 2006) can be used. This measure comes from learning and conditioning theory in psychology, where it has been found to be useful to probe cue learnability. Given a contingency table *m* cross-tabulating for the presence and absence of a given cue *C* and outcome *O*,

(12) $$m = \left( \begin{array}{ccc} & O & -O \\ C & a & b \\ -C & c & d \end{array} \right)$$

this one-way dependency statistic is defined as

(13) $$\begin{aligned} \Delta P &= \mathrm{Pr}\,(O|C) - P\,(O\,|-C) \\ &= a\,/(a+b) - c\,/(c+d) \\ &= (ad - bc)\,/\,[(a+b)\;(c+d)] \end{aligned}$$

$\Delta P$ ranges between -1 and 1, and represents the difference between two conditional probabilities, the probability of the outcome given the cue, and the probability of the outcome in the absence of the cue. For the data in Table 5, $\Delta P$ for the cue *take* and the outcome NP NP is -0.77, indicating that the cue *take* decreases the probability of the double object construction. Conversely, $\Delta P$ for the cue *take* and the prepositional object construction is 0.77, indicating that *take* is a reliable cue for this construction.

Yet another option for quantifying a verb's preference for a construction is to use the random intercepts of the generalized linear mixed model. For *take*, this random intercept (the adjustment of the baseline log-odds for the prepositional object construction) is 2.75, again indicating that the use of this verb is biased towards the prepositional object construction.

Finally, we can also use the association strength of a verb to a construction as estimated by naive discriminative learning as a measure for distinctive collexeme strength. In the model with both Verb and Speaker, the association strength (weight) to the prepositional object construction for *take* is 0.13. The verb *promise* has the largest negative association strength for the prepositional object construction (-0.28), and the verb *read* the largest (0.54).

Figure 5 presents a scatterplot matrix for the five measures for distinctive collexeme analysis, calculated across all verbs. First note that all measures enter into positive correlations that are consistently significant according to the non-parametric Spearman correlation test. The standard measure of Collexeme Strength is most clearly correlated with the relative entropy measure. $\Delta P$ correlates well with Relative Entropy, with the Random Intercepts, and with the Cue Strengths. Furthermore, the random intercepts of the GLMM and the verb-to-construction association strengths of the NDL are strongly correlated. The Random Intercepts and the Cue Strengths emerge as less prone to generate extreme outliers. For instance, whereas *take* is an extreme outlier on the scale of the Collexeme Strength and Relative Entropy measures, it is well integrated within the cloud of data points for the Random Intercepts and Cue Strengths.

What this survey of measures suggests is that corpus linguistics has a range of measures for verb-specific constructional preferences at its disposal that probably all do a decent job of highlighting verbs with strong constructional biases. The Cue-to-Construction Strength measure, however, is particularly interesting and promising, in that it is derived from the Rescorla-Wagner equations, described by Ellis (2006) as "the most influential formula in the history of conditioning theory". As a speaker/listener becomes more and more proficient in a language, the association strengths of words to constructions become more and more fine-tuned to the distributional properties of the language. For a verb such as *take*, the speaker/listener comes to expect the prepositional object construction. Sentences such as *We hope he took his mother the ingredients to bake a Simnel Mothering Cake* (**stpauls-healdsburg.org/wp-content/uploads/.../2010/201004-stpauls.pdf**) then come as a surprise, violating the expectation of a prepositional object construction, but at the same time constituting a learning experience with concomitant adjustments of the association strengths of this verb to the double object construction.

## 5. General Discussion

Corpus linguistics is generally conceived of as a descriptive subdiscipline of linguistics. As increasingly powerful and realistic models of human learning and cognition become available, however, corpus linguistics can begin to take on the challenge of not only describing distributional patterns in corpora, but also of explaining the consequences of the observed distributional patterns for human learning and linguistic choice behavior.

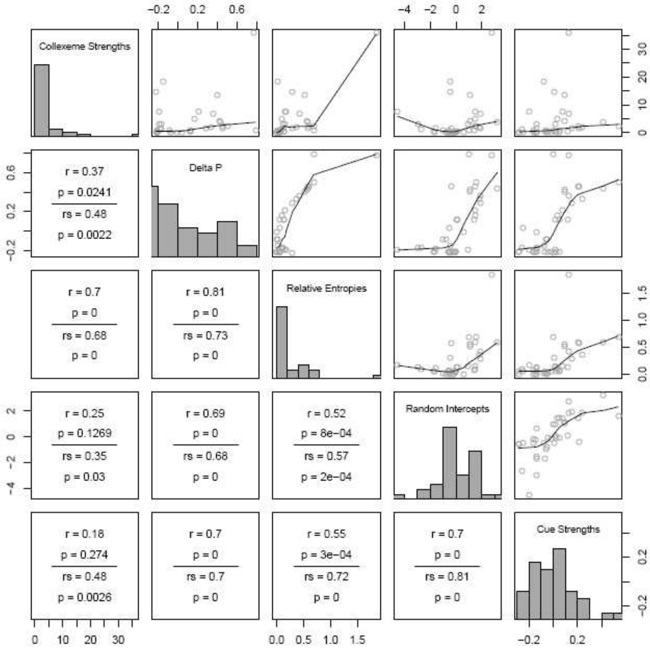


FIGURE 5 – Different measures of collexeme strength and their pairwise correlations (Pearson and Spearman)

Over the last decades, the statistical evaluation of distributional patterns has become increasingly important in corpus linguistics. Statistical models provide excellent insight into the quantitative structure of distributional patterns, but it is unclear to what extent such models provide an adequate characterization of the speaker-listener's actual knowledge. Moreover, the way in which statistical models derive a quantitative characterization of distributional patterns will, in general, be very different from how the speaker-listener acquires this knowledge.

As a first step towards a better cognitive grounding of quantitative analysis in corpus linguistics, the present study introduces a classifier grounded in naive discriminative learning. Using the data of Bresnan *et al.* (2007) on the dative alternation in spoken English as a case study, we have been able to show that, in theory, human classification can achieve nearly the same high level of accuracy as current state-of-the-art machine-learning techniques.

We have to be careful with this conclusion, however. First, this study has examined only one data set. Naive discriminative learning may not perform as well on other more complex data sets. Second, the validity of naive discriminative learning as a model for how speaker-listeners acquire and represent probabilistic knowledge depends on the validity of the Rescorla-Wagner equations. These equations specify learning under optimal conditions, without noise factors such as lack of attention, incomplete assessment of relevant cues, and incomplete knowledge of the targeted outcomes. The present results for naive discriminative learning therefore probably represent an upper bound for human performance. Third, although it is well known that dopamine neurons display a short-latency, phasic reward signal that indicates the difference between actual and predicted rewards (SCHULTZ, 2002; DAW; SHOHAMY, 2008), providing a neuro-biological justification for the hypothesis that learning is error-driven, it is well-known that the Rescorla-Wagner equations, however fruitful, do not cover all aspects of learning (MILLER *et al.*, 1995; SIEGEL; ALLAN, 1996).

Although naive discriminative classification performs well for the present data set, the conclusion that machine classification and human classification would be equivalent is not warranted. An examination of variable importance across models suggests that although statistical models can achieve comparable performance, they may do so by assigning predictors rather different explanatory relevance. There is a surprising and from a statistical perspective disquieting lack of convergence in the variable importance assigned to the predictors for the dative constructions across the support vector machine model, the logistic regression model, and the naive discriminative learner (Figure 3). In the face of such diversity, one would hope that that a statistical classifier derived from principles of human learning may provide superior estimates of variable importance for human-produced quantitative data. Without experimental support, unfortunately, this remains a conjecture at best.

The association strengths from verbs to constructions emerge in the naive discriminative learning approach as a natural alternative for quantifying distinctive collexeme strength. Although the five measures considered in this study (Collexeme strength, Δ*P*, Relative Entropy, Random Intercepts, and Cue Strength) are all correlated and useful as measures of collexeme strength, it is only the Cue Strength measure that is fully grounded in learning theory. It offers an important advantage compared to the other measure originating in psychology, Δ*P*. While Δ*P* is appropriate for 2 by 2 contingency tables (ALLAN, 1980), the Cue Strength measure handles *n* by 2 contingency tables appropriately. Crucially, the Cue Strength measure takes into account that many different cues may compete for a given outcome. Consider, for instance, the expression in the second row of equation (2) above,

$$\lambda - \sum_{\text{present } (C_j,\, t)} V_j.$$

When many cues are present simultaneously, the sum over cues will be larger, hence a larger number is subtracted from λ, and as a consequence, the cue-to-outcome association strength will increase with a smaller amount. Furthermore, when estimating the equilibrium association strength, the co-occurrence frequencies of the individual cues and outcomes are taken into account. By contrast, the Δ*P* measure ignores all other cues that can co-occur with a given outcome.

The naive discriminative learning model compresses experience with 2360 verb tokens, each characterized by 14 values (construction, verb, speaker, and 11 predictors) into a matrix of cue-to-construction association strengths with dimensions 865 by 2, a reduction from 2360 x 14 = 33040 values to only 1730 values, which amounts to a reduction by almost a factor 20. This reduced representation of past experience in terms of cue-to-construction strengths is reminiscent of connectionist models. The discriminative learning approach shares with the connectionist models of Seindeberg and McClelland

(1989) and Harm and Seindeberg (2004), as well as with the Competition Model (BATES, E.; MacWHINNEY, 1987; MacWHINNEY, 2005) the axiom that learning and generalization is driven by the distributional properties of the input. The discriminative learning model differs from the abovementioned connectionist models in terms of its architecture, which is much simpler. It does not make use of subsymbolic, distributed, representations, and it dispenses with hidden layers of all kinds. As a consequence, it is extremely parsimonious in free parameters: The only free parameter in the present study is whether to make use of single features or of feature-pairs. Computation of the weight matrix is also computationally much more efficient than in connectionist models. Computational efficiency also compares very favorably with random forests (BREIMAN, 2001; STROBL *et al.*, 2009), a high-performance non-parametric classifier that, unfortunately, is extremely slow for data with factors such as speaker and verb that have very large numbers of levels.

Note that the discriminative learner approach offers the possibility of gauging not only verb-related constructional preferences, but also speaker-related constructional preferences, by means of the weights on the connections from speakers to constructions.

The way in which knowledge is represented in naive discriminative learning differs from other (non-connectionist) computational models for linguistic generalization. In exemplar-based approaches, it is assumed that in the course of experience, exemplars are stored in memory. Prediction is based on similarity neighborhoods in exemplar space. Data Oriented Parsing (BOD, 2006), Analogical Modeling of Language (SKOUSEN, 1989), and Memory Based Learning (DAELEMANS; BOSCH, 2005) provide examples of this general approach.

An important advantage of exemplar-based approaches is that the generalization process is simple and remarkably accurate in its predictions, as witnessed for the present data set by the classification results obtained with TiMBL, using its out-of-the-box default settings of parameters. An important disadvantage is that exemplars must be assumed to be available in memory, which may be unrealistic for human language processing. For example, recent studies suggest that the frequency with which a given sequence of words occurs in the language is predictive for how quickly such a sequence is processed (ARNON; SNIDER, 2010; TREMBLAY; BAAYEN, 2010). This frequency effect persists for non-idiomatic sequences and for sequences that are incomplete phrases (as, e.g., *the president of the*). The assumption that shorter *n*-grams are stored in memory implies that hundreds of millions of exemplars would be remembered. This seems unrealistic. While naive discriminative learning shares with memory based learning the premise that each exemplar is important and contributes to learning, unlike memory-based learning, it does not need to posit that individual exemplars 'exist' independently in memory: Exemplar information is merged into the weights.

Instead of calculating predictions over an acquired instance space at run time, as in memory-based learning, one can instead seek to construct rule systems or constraint systems that capture the quantitative forces shaping behavior without having to store exemplars. The Gradual Learning Algorithm of Stochastic Optimality Theory (BOERSMA; HAYES, 2001) and the Mimimum Generalization Learner (ALBRIGHT; HAYES, 2003) are examples of this approach. These rule-based approaches do not run into the problem that the instance base can become extremely voluminous, but they are challenged by frequency effects documented for linguistic units such as regular complex words and *n*-grams. Rule-based approaches tend to ignore these frequency effects, leaving them aside as an unsolved issue supposedly irrelevant to understanding the nature of generalization in human cognition. Rule-based approaches are also challenged by a proliferation of rules necessary to capture the fine details of the quantitative patterns in the data. Naive discriminative learning, by contrast, dispenses with the necessity of positing that the speaker-listener deduces large and complex rule sets from the input. Excellent classification accuracy can be obtained without storing exemplars and without rule induction or deduction.

In summary, the potential importance of naive discriminative learning for corpus linguistics is that it offers a unified framework for learning, for classification, and for distinctive collexeme (and distinctive collocutor) analysis. It is conceivable that variable importance is more adequately assessed by means of discriminative learning. Furthermore, naive discriminative learning may detect non-normally distributed variability where classic mixed models cannot do so. Finally, in discriminative learning, single cues make only modest contributions to classification accuracy. The present case study suggests that cues for outcomes tend to be highly interdependent and to a considerable extent predictable from each other. As such, they constitute a rich and redundant feature space in which a highly context-sensitive error-driven learning algorithm such as defined by the Rescorla-Wagner equations functions well, unhampered by issues of collinearity that plague (parametric) regression models.

Assuming that naive discriminative learning is on the right track as a characterization of human learning and categorization, many important questions remain unanswered. One such question is how speakers/listeners become knowledgeable about the cues and outcomes on which naive discriminative classification is based. Another question is why the language input to the model typically displays high-dimensional correlational structure, as exemplified by the dative alternation data. Although intercorrelated, redundant feature spaces are apparently relatively easy to learn, at least under ideal conditions, it remains unclear why the data take the distributional forms typically attested in corpora. Furthermore, our use of the equilibrium equations for the Rescorla-Wagner equations assumes that the adult system would be completely stable and not subject to further change, which is only approximately correct.

The Rescorla-Wagner characterization of discriminative learning is in all likelihood incomplete, in that it does not do justice to tiny biases favoring outcomes that are cognitively easier to process (such as given information preceding new information). Within a speech community, such small biases would, under favorable circumstances, gain momentum, leading to locally optimal, 'functional' distributional patterns. Under this scenario, predictors such as animacy, definiteness, and information status would not shape an individual speaker's production as, for instance, in the variable rule approach of Cedergren and Sankoff (1974). Instead of a given utterance being governed by a probabilistic set of cognitive constraints operating at the level of an individual's brain, an utterance would be shaped by past experience under error-driven discriminative learning, much as described above for the dative alternation. However, tiny cognitive biases, neglected in the current formulation of the naive discriminative learner, would over time give rise to a speech community the utterances of which would then reflect, to some extent, varying from speech community to speech community, these very cognitive biases.

A challenge for corpus linguistics is to develop multi-agent computational models demonstrating that indeed tiny cognitive biases in discriminative learning can generate the kind of grammars and their trajectories of diachronic change that we find in human speech communities. With efficient algorithms such as provided by the equilibrium equations, realistic computational methods are coming within reach.

## References

ALBRIGHT, A.; HAYES, B. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, v. 90, p. 119-161, 2003.

ALLAN, L. G. A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, v. 15, p. 147-149, 1980.

ANDERSON, J. R. *Learning and memory*: An integrated approach. New York: Wiley, 2000.

ARNON, I.; SNIDER, N. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Journal of Memory and Language*, v. 62, p. 67-82, 2010.

ARPPE, A.; BAAYEN, R. H. ndl: Naive discriminative learning: an implementation in r [Computer software manual]. 2011. Available at: <http://CRAN.R-project.org/package=ndl (R package version 0.4)>.

BAAYEN, R. H. languageR: Data sets and functions with "analyzing linguistic data:A practical introduction to statistics". [Computer software manual]. 2009. Available at: <http://CRAN.R-project.org/package=languageR (R package version 0.955)>.

BAAYEN, R. H.; HENDRIX, P. Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. Empirically examining parsimony and redundancy in usage-based models, LSA workshop, January 2011.

BAAYEN, R. H.; MILIN, P.; FILIPOVIC DURDJEVIC, D.; HENDRIX, P.; MARELLI, M. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 2011. (in press).

BATES, D.; MAECHLER, M. lme4: Linear mixed-effects models using s4 classes [Computer software manual]. 2009. Available at: <http://CRAN.R-project.org/package=lme4 (R package version 0.999375-31)>.

BATES, E.; MacWHINNEY, B. Competition, variation, and language learning. In: MacWHINNEY, B. (Ed.). *Mechanisms of language acquisition.* Hillsdale, NJ: Lawrence Erlbalm Assoc., 1987.

BOD, R. Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, v. 23, n. 3, p. 291-320, 2006.

BOERSMA, P.; HAYES, B. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, v. 32, p. 45-86, 2001.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5-32, 2001.

BRESNAN, J.; CUENI, A.; NIKITINA, T.; BAAYEN, R. H. Predicting the dative alternation. In: Bouma, G.; KRAEMER, I.; ZWARTS, J. (Ed.). *Cognitive foundations of interpretation*. Royal Netherlands Academy of Arts and Sciences, 2007.

CEDERGREN, H.; SANKOFF, D. Variable rules: Performance as a statistical reflection of competence. *Language*, v. 50, n. 2, p. 333-355, 1974.

CHATER, N.; TENENBAUM, J. B.; YUILLE, A. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Science*, v. 10, n. 7, p. 287-291, 2006.

COLTHEART, M.; RASTLE, K.; PERRY, C.; LANGDON, R.; ZIEGLER, J. DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, v. 108, n. 1, p. 204-256, 2001.

DAELEMANS, W.; BOSCH, A. Van den. *Memory-based language processing*. Cambridge: Cambridge University Press, 2005.

DAELEMANS, W.; ZAVREL, J.; SLOOT, K. Van der; BOSCH, A. Van den. TiMBL: Tilburg Memory Based Learner Reference Guide. Version 6.3 (Technical Report No. ILK 10-01). 2010. Computational Linguistics Tilburg University.

DANKS, D. Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, v. 47, n. 2, p. 109-121, 2003.

DAW, N.; SHOHAMY, D. The cognitive neuroscience of motivation and learning. *Social Cognition*, v. 26, n. 5, p. 593-620, 2008.

DAYAN, P.; KAKADE, S. Explaining away in weight space. In: LEEN, T. K.; DIETTERICH, T. G.; TRESP, V. (Ed.). *Advances in neural information processing systems 13*. Cambridge, MA: MIT Press, 2001.

DIMITRIADOU, E.; HORNIK, K.; LEISCH, F.; MEYER, D.; WEINGESSEL, A. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien [Computer software manual]. 2009. (R package version 1.5-19)

ELLIS, N. C. Language acquisition as rational contingency learning. *Applied Linguistics*, v. 27, n. 1, p. 1-24, 2006.

FORD, M.; BRESNAN, J. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, v. 86, n. 1, p. 168-213, 2010.

GLUCK, M. A.; BOWER, G. H. From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology*: General, v. 117, n. 3, p. 227-247, 1988.

GRIES, St. Th. Frequency tables: tests, effect sizes, and explorations. In: GLYNN, D.; ROBINSON, J. (Ed.). *Polisemy and synonymy*: Corpus methods and applications in Cognitive Linguistics. Amsterdam / Philadelphia: John Benjamins, 2011.

GRIES, St. Th.; STEFANOWITSCH, A. Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics*, v. 9, n. 1, p. 97-129, 2004.

HARM, M. W.; SEIDENBERG, M. S. Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, v. 111, p. 662-720, 2004.

HARRELL, F. *Regression modeling strategies*. Berlin: Springer, 2001.

HSU, A. S.; CHATER, N.; VITÁNYI, P. The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis, 2010. Manuscript submitted for publication.

LEVELT, W. J. M.; ROELOFS, A.; MEYER, A. S. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, v. 22, p. 1-38, 1999.

MacWHINNEY, B. A united model of language acquisition. In: KROLL, J.; GROOT, A. de (Ed.). *Handbook of bilingualism*: Psycholinguistic approaches Oxford University Press, 2005.

McCLELLAND, J. L.; RUMELHART, D. E. An interactive activation model of context effects in letter perception: Part I. An account of the basic findings. *Psychological Review*, v. 88, p. 375-407, 1981.

MILLER, R. R.; BARNET, R. C.; GRAHAME, N. J. Assessment of the Rescorla-Wagner Model. *Psychological Bulletin*, v. 117, n. 3, p. 363-386, 1995.

MURRAY, W. S.; FORSTER, K. Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, v. 111, p. 721-756, 2004.

NORRIS, D. The Bayesian Reader: Explaining Word Recognition as an Optimal Bayesian Decision Process. *Psychological Review*, v. 113, n. 2, p. 327-357, 2006.

NORRIS, D.; McQUEEN, J. Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, v. 115, n. 2, p. 357-395, 2008.

RAMSCAR, M.; DYE, M.; POPICK, H. M.; O'DONNELL-McCARTHY, F. The Right Words or Les Mots Justes? Why Changing the Way We Speak to Children Can Help Them Learn Numbers Faster. PLoS ONE, 2011. (To appear)

RAMSCAR, M.; YARLETT, D. Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, v. 31, n. 6, p. 927-960, 2007.

RAMSCAR, M.; YARLETT, D.; DYE, M.; DENNY, K.; THORPE, K. The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, v. 34, n. 7, 2010. (In press).

SCHULTZ, W. Getting formal with dopamine and reward. *Neuron*, v. 36, n. 2, p. 241-263, 2002.

SEIDENBERG, M. S.; McCLELLAND, J. L. A distributed, developmental model of word recognition and naming. *Psychological Review*, v. 96, p. 523-568, 1989.

SIEGEL, S.; ALLAN, L. G. The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, v. 3, n. 3, p. 314-321, 1996.

SKOUSEN, R. *Analogical modeling of language*. Dordrecht: Kluwer, 1989.

STROBL, C.; BOULESTEIX, A.-L.; KNEIB, T.; AUGUSTIN, T.; ZEILEIS, A. Conditional variable importance for random forests. BMC Bioinformatics, 9. 2008. Available at:<http://www.biomedcentral.com/1471-2105/9/307>.

STROBL, C.; MALLEY, J.; TUTZ, G. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*, v. 14, n. 4, p. 323-348, 2009.

TAGLIAMONTE, S.; BAAYEN, R. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. 2010. Manuscript submitted for publication.

TREMBLAY, A.; BAAYEN, R. H. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In: WOOD, D. (Ed.). *Perspectives on Formulaic Language*. Acquisition and Communication, 2010.

Van HEUVEN, W. J. B.; DIJKSTRA, A.; GRAINGER, J. Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language*, v. 39, p. 458-483, 1998.

VAPNIK, V. *The nature of statistical learning theory*. Berlin: Springer-Verlag, 1995.

WAGNER, A.; RESCORLA, R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: BLACK, A. H.; PROKASY, W. F. (Ed.). *Classical conditioning ii*. Appleton-Century-Crofts, 1972.

## APPENDIX: naive discriminative classification with the ndl package

The ndl package (Arppe & Baayen, 2011), available in the CRAN archives at www.r-project.org, provides software for naive discriminative learning for the R statistical programming environment. The following provides an introduction to its basic functionality.

As a first step, we attach the package, extract the dative dataset, and remove the data for which no speaker information is available.

```
> library (ndl)
> data (dative)
> dative = dative [! is. na (dative$Speaker), – 2]
```

We fit a basic naive discriminative classifier to the data using the standard formula based interface, where the dot is expanded into all predictors in the dative data frame other than the dependent variable (*RealizationOfRecipient*):

```
> dative.nd1 = nd1Classify (RealizationOfRecipient ~., data = dative)
```

Numeric predictors are converted into factors, by default each factor has two levels. This default can be changed by the user, as explained in the documentation. Models with cue pairs can be specified using the interaction notation for R formulae. For instance,

```
> dative.nd12 = nd1Classify (RealizationOfRecipient  ~  (SemanticClass +
+     LengthOfRecipient + AnimacyOfRec  + DefinOfRec  + PronomOfRec +
+     LengthOfTheme + AnimacyOfTheme + DefinOfTheme + PronomOfTheme +
+     AccessOfRec + AccessOfTheme)  + Verb  + Speaker,  data = dative)
```

includes pairwise cues for all independent variables, except verb and speaker.

The weight matrix can be extracted from the model object, which is a list:

```
> names (dative.nd1)
[1] "activationMatrix"  "weight-Matrix"     "cuesOutcomes"     "frequency"
[5] "call"              "formula"           "data"
> head (dative.nd1$weightMatrix)
                                     NP              PP
AccessOfRecaccessible       –0.015468613     0.083503412
AccessOfRecgiven             0.094783250    –0.026748451
AccessOfRecnew              –0.009894431     0.077929230
AccessOfThemeaccessible      0.089768608    –0.021733809
AccessOfThemegiven          –0.093523058     0.161557856
AccessOfThemenew             0.073174656    –0.005139857
```

The association strengths of the individual verbs to the constructions can be accessed as follows:

```
> w = dative.nd1$SweightMatrix
> verbs = w [grep (“Verb”, rownames (w)), ]
> verbs = verbs [order (verbs [, “pp” ]),]
> head (verbs)
                          NP              PP
Verbaward            0.6194557      –0.6146320
 Verbbet             0.3843946      –0.3795708
Verbowe              0.3570426      –0.3522188
Verbpromise          0.3425307      –0.3377070
Verbtell             0.3036573      –0.2988335
 Verbteach           0.1962304      –0.1914066
> tail (verbs)
                          NP              PP
Verbhand            –0.2039228       0.2087466
 Verbbring          –0.2059774      0.2108011
Verbleave            –0.2593050     0.2641288
Verbowrite          –0.4221433       0.4269670
Verbread            –0.4432427       0.4480664
 Verbafford         –0.6125922       0.6174159
```

A summary method for ndl objects is available that provides a wide range of measures of goodness of fit, including

```
> summary (dative.nd1) $statistics$C
[1] 0.9820687
> summary (dative.nd1) $statistics$accuracy
[1] 0.9457627
```

A crosstabulation of observed and predicted values is available with

```
> summary (dative.nd1) $statistics$crosstable
     NP    PP
NP   1821  38
PP   90    411
```

The predicted probabilities of the double object and prepositional object constructions for each row of the dative data frame are obtained with

```
> p = acts2probs (dative.nd1$activationMatrix)$p
>read (p)
```

```
                  NP              PP
        [1,]      0.7582780       0.2417220
        [2,]      0.1872549       0.8127451
        [3,]      0.5710474       0.4289526
        [4,]      0.5707516       0.4292484
        [5,]      0.5190592       0.4809408
        [6,]      0.4767222       0.5232778
> tail (p)
                  NP              PP
      [2355,]     0.5009516       0.4990484
      [2356,]     0.6145346       0.3854654
      [2357,]     0.6999555       0.3000445
      [2358,]     0.4434956       0.5565044
      [2359,]     0.6017827       0.3982173
      [2360,]     0.6433302       0.3566698
```

Crossvalidation can be carried out as follows:

```
> dative.nd1.10 = nd1Crossvalidate (RealizationOfRecipient ~.,
+         data = dative)

> summary (dative.nd1.10)$statistics.summary [“Mean”, “C”]
[1] 0.9265221

> summary (dative.nd1.10)$statistics.summary [“Mean”, “accuracy”]
[1] 0.8889831
```

Permutation variable importance is assessed with

```
> dative.varimp = nd1Varimp (dative.ndl)

> library (lattice)
> dotplot (sort (summary (dative.nd1)$statistics$accuracy – dative.varimp$accuracy),
+      xlab = “permutation variable importance”)
```

Recebido em 25/12/2010. Aprovado em 01/06/2011.

# Metaphor and Corpus Linguistics[1]

## *Metáfora e linguística de corpus*

Tony Berber Sardinha*
Catholic University of São Paulo
São Paulo / Brasil

ABSTRACT: In this paper, I look at four different aspects of metaphor research from a corpus linguistic perspective, namely: (1) the lexicogrammar of metaphors, which refers to the patterning of linguistic metaphor revealed by corpus analysis; (2) metaphor probabilities, which is a facet of metaphor that emerges from frequency-based studies of metaphor; (3) dimensions of metaphor variation, or the search for systematic parameters of variation in metaphor use across different registers; and (4) automated metaphor retrieval, which relates to the development of software to help identify metaphors in corpora. I argue that these four aspects are interrelated, and that advances in one of them can drive changes in the others.

KEYWORDS: corpora; metaphor; metaphor identification; lexicogrammar; probabilities; Multi-Dimensional Analysis; metaphor retrieval software.

RESUMO: Neste artigo discuto quarto aspectos da pesquisa sobre metáfora do ponto de vista da linguística de corpus: (1) a lexicogramática das metáforas, que se refere aos padrões da metáfora linguística revelados pela análise de corpus; (2) probabilidades metafóricas, que é uma faceta da metáfora que emerge a partir dos estudos relacionados à freqüência de metáforas; (3) dimensões da variação de metáforas, ou a busca por parâmetros sistemáticos de variação de uso de metáfora em diferentes gêneros; e (4) captura automática de metáfora, que está relacionada ao desenvolvimento de softwares que auxiliam na identificação de metáforas em corpora. I defendo que esses quatro aspectos são interrelacionados, e que progressos em um deles podem acarretar mudanças nos outros.

PALAVRAS-CHAVE: corpora; metáfora; identificação de metáfora; lexicogramática; probabilidades; Análise Multidimensional; software de captura de metáforas.

* tony@corpuslg.org

[1] I want to thank CPNq (Brasília, Brazil) for supporting my research.

## 1. Introduction

The field of metaphor studies is vast and has, a long tradition that dates back to ancient Greece. Over time, numerous theories of metaphor and a range of different methods for metaphor identification have been proposed (see GIBBS, 2008a). Corpus Linguistics is a newcomer to the field, but its influence is already being felt:

> A related emerging concern for empirical studies of metaphor focuses on the true frequency of metaphors in language and other media. Claims about the importance or ubiquity of particular metaphorical patterns, in either language or thought, are often made without adequate empirical support, such as reporting the frequencies with which different metaphors are found in particular texts, or comparing the findings from one's own textual analysis of metaphor with those seen in large corpora. (GIBBS, 2008b, p. 12)

Notwithstanding the recognition of its role, Corpus Linguistics has only begun to make itself noticed in the vast field of metaphor scholarship. One reason is the fact that it is a relatively recent approach to metaphor analysis, with the first studies dating back to 1999 (DEIGNAN, 1999a; 1999b; 1999c). Another reason is that metaphor traditionally requires hand analysis, which is too time consuming to carry out in large corpora. A number of metaphor retrieval computer tools have been developed but they have not made an impact in the field, partly because they are not widely available and partly because their performance is still not particularly high (see section 5 below).

There is a growing body of research at the interface between metaphor and Corpus Linguistics. Deignan (2005) offers a detailed treatment of bottom-up approaches to metaphor analysis, with an emphasis on concordancing and how linguistic metaphor is signaled by recurring patterns of use. In Stefanowitsch and Gries (2006) several different approaches to metaphor identification in corpora are presented, which Stefanowitsch (2006, p. 2-6) classifies into seven distinct groups based on the kind of searching performed: manual, source domain vocabulary, target domain vocabulary, both source and target domains, metaphor markers, and extraction from corpora annotated for semantic fields or for conceptual mappings. Wikberg (2007) discusses central issues in using corpora for metaphor research, and concludes that close reading of text passages is necessary for determining metaphoricity. Berber Sardinha (2009) provides an overview of corpus-based and corpus-driven approaches to metaphor identification in corpora, showing how they can be retrieved by programs such as WordSmith Tools Keywords (SCOTT, 2004) and the Metaphor Candidate Identifier (see section 5 below).

At the same time, there is crucial work being done at setting criteria for metaphor identification by hand analysis. This includes MIP (Metaphor Identification Procedure) and MIV (Metaphor Identification Through Vehicle Terms). MIP (PRAGGLEJAZ GROUP, 2007) and its more recent version MIPVU (STEEN, DORST, HERRMANN *et al.*, 2010) both lay down guidelines for metaphor identification. MIP/MIPVU details steps for coding metaphors at the word level, showing how to determine metaphoricity by taking into account the basic and contextual meanings of each word. MIV (CREET, 2006) also presents detailed procedures for metaphor identification, but singles out Vehicle terms (metaphorically used language), which may or may not be single words. Other equally important work in this area includes Steen (2007) and Cameron and Maslen (2010). The former gives a thorough account of issues in metaphor identification and interpretation, as well as how these relate to language and thought. The latter focuses on discourse dynamics and takes a comprehensive look at systematicity, that is, how recurrent connections between Topic (what the metaphor refers to) and Vehicle can reveal aspects of discourse. Work on methods for metaphor identification and interpretation, even though not strictly from a corpus perspective, can provide valuable insights into issues that affect corpus research, such as the lexical patterning of metaphorical language, and criteria for determining metaphor use.

I would argue the following are particularly important findings from previous research:

1) Metaphor use seems to correlate with lexicogrammatical patterns. Patterns used to express metaphor are typically different from patterns employed to denote literal language. In other words, metaphorically used language selects particular patterns.
2) In particular genres (articles, reports, speeches, etc.) or registers (academic, fiction, business, etc.)[2] (see e.g. BIBER, CONRAD, 2009), metaphorically

[2] There are numerous definitions of genre and register in the literature. Here, genres are understood as 'recognizable communicative events, characterized by a set of communicative purposes identified and mutually understood by members [...] of [a] community where they regularly occurs.' (BHATIA, 2004, p. 23). And register is defined 'as a cover term for any language variety defined by its situational characteristics, including the speaker's purpose, the relationship between speaker and hearer, and the production circumstances.' (BIBER, 2009, p. 823)

used language has probabilities of use that are different from those in literal language. Also, probabilities of metaphor use for particular words or expressions in specialized varieties differ from those in general language.

3) Metaphor use varies systematically across different genres and registers and this may give rise to dimensions of metaphor variation.
4) Specialized systems for metaphor retrieval by machine have been developed to automate metaphor identification from corpora.

Based on these, the major areas that I think should mature in CL metaphor research are the following:

1) The lexicogrammar of metaphor;
2) The probabilistic nature of metaphor use;
3) Variation in metaphor use;
4) Automating metaphor identification.

In the following sections, I will focus on each of these points in turn. In order to make these points, I will report findings from previous research.

However, we must first distinguish between two basic types of CL metaphor research: whole corpus and concordance-based. In the former, researchers code all the metaphors in the whole corpus, usually by hand, and then retrieve the metaphors based on the hand analysis done ahead of time; in the latter, they run concordances for particular items and then analyze only those occurrences. Whole corpus analyses are affected by the amount of data that need to be coded. Concordance-based is not, because analysis is typically carried out on a sample (e.g. one thousand lines) of concordance lines extracted from the corpus. Concordance-based analyses are influenced by the choice of search terms, since these will define what will and will not be found in the corpus.

These points overlap and draw on each other; for instance, the more we know about the linguistic patterning underlying metaphor use, the better we can establish both the probabilities of use and the dimensions of variation of metaphor across registers, and vice-versa. And the more we know about the patterning of metaphor, its probabilities and variation, the better positioned we are to determine which aspects of metaphor the computer can be taught to recognize with reasonable degrees of accuracy.

One further point that has not deserved much attention in CL metaphor research is extending the scope of inquiry beyond English. The vast majority of the literature focuses on metaphors in English, and few other languages are reported at all. There are exceptions, notably Steen *et al.* (2010) analysis of 130,000 words of Dutch. A basic ingredient, the corpus, can be easily compiled for a large number of languages, given the wide availability of electronic texts on the Web. Other resources may be harder to find, which may hinder progress of this kind of research in other languages. I will present findings of analyses of Portuguese corpora below.

## 2. The lexicogrammar of metaphor

One of the ways in which a metaphor reveals itself in corpora is by its patterns of usage, which typically contrasts with the patterns of non-metaphorical language. This has proved valuable as a criterion for both metaphor manifestation and identification.

To illustrate, I will use data from my own analysis of autobiographical narratives in Brazilian Portuguese (BERBER SARDINHA, 2007B). These were recorded by the Museu da Pessoa (Museum of the Person), an organization that aims at preserving history by recording people telling a personal narrative about their lives. These recordings are then transcribed and many of them are made public on the institution's website. I collected a corpus of such narratives and used both hand and machine analyses to identify the metaphors in them.

One set of metaphorical patterns that emerged in the analysis was the following:

TABLE 1
Metaphorical patterns of PEGAR

| Pattern | Direct English translation |
|---|---|
| PEGAR$_{\textit{past tense}}$ + , + *Vb past*[3] | GRAB/CATCH/PICK UP $_{\textit{past tense}}$ + , + *Vb past* |
| PEGAR$_{\textit{past tense}}$ + *Determiner* + [Disease] | GRAB/CATCH/PICK UP $_{\textit{past tense}}$ + *Determiner* + [Disease] |
| PEGAR$_{\textit{past tense}}$ + e + *Vb past* | GRAB/CATCH/PICK UP $_{\textit{past tense}}$ + e + *Vb past* |

[3] I use the following convention to represent patterns: CAPITALS = lemmas; subscript italics = grammatical constraint on lemma; *Italics* = part of speech; [Square brackets] = Semantic fields; other formats: actual strings/words; the plus sign = followed by, up to five words away.

These are exemplified in the concordance below.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | , | o | pai | **pegou** | , | disse | que |
| 2 | nesse | trabalho | eu | **pegava** | , | eu | fiz |
| 3 | taquara | e | eu | **peguei** | , | fiz | por |
| 4 | é | que | me | **pegou** | a | depressão | estou |
| 5 | E | aí | ela | **pegou** | a | malária | . |
| 6 | e | a | mãe | **pegou** | e | disse | : |
| 7 | frente | dele | , | **pegou** | e | disse | : |
| 8 | que | a | gente | **pegava** | e | fazia | , |
| 9 | com | dó | de | **pegar** | e | levar | pra |
| 10 | , | aí | ela | **pegou** | e | falou | : |
| 11 | . | Aí | ele | **pegou** | e | danou | pra |
| 12 | Aí | o | velho | **pegou** | e | veio | mais |
| 13 | . | Aí | ele | **pegou** | e | casou | com |
| 14 | pitimbado | . | Ele | **pegou** | e | falou | : |
| 15 | . | Aí | ela | **pegou** | e | falou | : |
| 16 | , | aí | eu | **peguei** | e | falei | com |
| 17 | no | pé | , | **pegou** | e | falou | comigo |
| 18 | tá | bom | , | **peguei** | e | fui | lá |
| 19 | . | Aí | eu | **peguei** | e | disse | : |
| 20 | fiz | ? | Eu | **peguei** | e | fui | na |
| 21 | ? | Então | ela | **pegava** | e | mandava | eu |
| 22 | . | Aí | eu | **peguei** | e | falei | . |
| 23 | Então | , | eu | **peguei** | e | comprei | esse |
| 24 | “ | Aí | eu | **pegava** | e | fazia | a |
| 25 | de | carro | , | **peguei** | e | vendi | o |
| 26 | ! | Aí | eu | **peguei** | e | falei | para |
| 27 | , | aí | eu | **peguei** | e | liguei | para |
| 28 | Maria | do | Carmo | **pegou** | e | falou | “ |
| 29 | Cultura | , | então | **peguei** | e | pedi | a |
| 30 | um | dia | ele | **pegou** | e | começou | falando |
| 31 | tempo | que | ele | **pegou** | essa | doença | , |
| 32 | . | Lá | ela | **pegou** | essa | febre | . |
| 33 | panqueca | . | Denise | **pegou** | uma | anemia | profunda |
| 34 | ficou | doente | , | **pegou** | uma | doença | grave |
| 35 | ele | pegou | , | **pegou** | uma | hepatite | o |
| 36 | doente | . | Ela | **pegou** | uma | febre | que |
| 37 | logo | . | Ele | **pegou** | uma | infecção | intestinal |

FIGURE 1: Concordances for metaphorical patterns of PEGAR

The metaphors realized by these patterns appear in the table below.

TABLE 2
Metaphorical patterns of PEGAR

| Pattern | Metaphor |
|---|---|
| PEGAR$_{past\ tense}$ + , + *Vb past* | AN ACTION IS AN OBJECT |
| PEGAR$_{past\ tense}$ + *Determiner* + [Disease] | A DISEASE IS AN OBJECT |
| PEGAR$_{past\ tense}$ + e + *Vb past* | AN ACTION IS AN OBJECT |

Examples of each pattern are shown below:

TABLE 3
Examples of metaphorical patterns of PEGAR

| Pattern and Metaphor | Concordance lines in Figure 1 | Example |
|---|---|---|
| PEGAR$_{past\ tense}$ + , + *Vb past*<br>AN ACTION IS AN OBJECT | 1 – 3 | O pai pegou, disse que uma coisa que a gente tinha que ter era estudo, entende?<br>My father turned and said that one thing we needed to have was schooling, see what I mean? |
| PEGAR$_{past\ tense}$ + *Det* + [Disease]<br>A DISEASE IS AN OBJECT | 4 – 5<br>31 – 32<br>33 – 37 | Ela pegou a malária. E foi indo, foi a causa da morte dela.<br>She caught malaria. She carried on, this was the cause of her death.<br>E ela estava em Marajó. Lá ela pegou essa febre.<br>She was in Marajó. She caught this fever over there.<br>Pegou uma hepatite, o cabelo foi tudo para o beleléu, tudo num mês.<br>*He caught hepatitis, his hair fell out, all in a month's time.* |
| PEGAR$_{past\ tense}$ + e + *Vb past*<br>AN ACTION IS AN OBJECT | 6 – 30 | Subindo um viaduto, assim, e a mãe pegou e disse: “Cardoso, caiu uma, caiu um papelão.”<br>*Going up an overpass, like that, and my mom turned around and said: “Cardoso, a piece of cardboard dropped out”.* |

As can be seen in the translated examples, in the metaphor AN ACTION IS AN OBJECT PEGAR means something equivalent to the English 'turn (a)round and'. In the other metaphor, A DISEASE IS AN OBJECT, it means its direct equivalent, 'to catch'.

I labeled the instances of 'turn (a)round and' as AN ACTION IS AN OBJECT, but these might as well have been named in other ways, for instance as AN IDEA IS AN OBJECT, since they might imply 'grabbing an idea' and expressing it in words or actions. Labeling metaphors, especially conceptual ones, is tricky, and there are no specific guidelines. This is certainly an area where more clarity is needed; this will become more pressing as research that resorts to metaphor categorization intensifies.

Semantic categories are very useful in formulating patterns. In this particular case, they were applied after the fact, by looking at and grouping citations in a concordance. They can be more useful, though, if applied as search terms to query a corpus, because researchers need only to specify the semantic grouping and not each individual word in it. The problem is of course that it requires a semantically annotated corpus. Increasing the availability of semantically annotated corpora (in several languages) is another front that needs development both in Corpus Linguistics in general and in CL metaphor research in particular.

By contrast, the basic patterns for literal uses of PEGAR are:

TABLE 4
Non-metaphorical patterns of PEGAR

| Pattern<br>Example | Concordance lines in Figure 2 | Direct English (translation)<br>Examples |
|---|---|---|
| PEGAR + *Det* + [Concrete] | | GRAB + *Det* + [Concrete] |
| Pegou a bagagem | 1 | Picked up the baggage |
| Pegou a bola | 7 | Grab that ball |
| Pegava o saco | 20 | Picked up a bag |
| Pegou uns pincéis | 30 | Picked up a few brushes |
| PEGAR + *Prep* + *Det* + [Concrete] | | SIT/HOLD/PICK UP + *Prep* + *Det* + [Concrete] |
| Me pegava no colo | 12 | Sit on someone's (lap) |
| Pegar no lixo | 13 | Pick something up from the trash |

These are illustrated in the following concordance:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | ? | Conclusão | ? | **Pegou** | a | bagagem | e |
| 2 | levantava | cedo | e | **pegava** | a | enchadinha | . |
| 3 | falava | assim | , | **pegava** | a | identidade | da |
| 4 | menina | saiu | , | **pegou** | a | lata | . |
| 5 | , | a | gente | **pegando** | a | meia | com |
| 6 | lá | fora | , | **peguei** | a | receita | joguei |
| 7 | que | tinha | que | **pegar** | aquela | bola | . |
| 8 | eu | gostava | de | **pegar** | aquela | coisinha | de |
| 9 | aqui | ! | Vou | **pegar** | as | cobertas | “ |
| 10 | segunda | época | , | **pegou** | as | provas | e |
| 11 | no | coisa | e | **pegava** | as | terras | , |
| 12 | , | que | me | **pegava** | no | colo | e |
| 13 | de | pegar | . | **Pegar** | no | lixo | . |
| 14 | bisavó | , | aí | **pegou** | no | mato | , |
| 15 | ele | atravessava | , | **pegava** | o | bonde | , |
| 16 | a | mais | e | **pegava** | o | dinheiro | pra |
| 17 | pagar | pra | ele | **pegar** | o | diploma | . |
| 18 | que | a | gente | **pegava** | o | gibi | , |
| 19 | ele | foi | lá | **pegar** | o | prêmio | , |
| 20 | , | né | ? | **Pegava** | o | saco | , |
| 21 | Aí | um | dia | **peguei** | um | anúncio | e |
| 22 | notável | , | ele | **pegava** | um | livro | . |
| 23 | em | quando | ele | **pegava** | um | passarinho | esse |
| 24 | , | pra | ela | **pegar** | um | pedaço | a |
| 25 | , | mas | nunca | **peguei** | um | peixe | o |
| 26 | favor | do | cara | **pegar** | um | pente | para |
| 27 | às | vezes | vocês | **pegavam** | uma | quantidade | de |
| 28 | empregad̯ | não | podia | **pegar** | uma | revista | , |
| 29 | Eu | gosto | , | **pegar** | uma | tesoura | . |
| 30 | de | cor | , | **pegou** | uns | pincéis | , |
| 31 | teve | interesse | em | **pegar** | uns | quadros | meus |

FIGURE 2: Concordance with non-metaphorical uses of PEGAR (GRAB)

This illustrated the existence of a lexicogrammar of individual metaphors, which patterns the way metaphor choices are made in texts. Patterns may signal metaphor (or non-metaphor) with a certain likelihood. The next point looks at the cumulative effects of the presence of metaphor in corpora, from a probabilistic point of view.

## 3. Metaphor probabilities

Patterns of metaphor use occur in language with particular probabilities of occurrence attached to them. There is little research in this aspect of metaphor use, even though this is an important characteristic of metaphor, because it may reveal how likely it is that we encounter metaphors in written and spoken texts. Theory emphasizes that a metaphor is a frequent linguistic feature, and that all language users are likely to come across or employ metaphors to express various meanings. Empirical research also makes similar claims. For instance, according to Deignan and Potter (2004,p. 1236) 'non-literal language is extremely common, often accounting for a substantial proportion of the corpus citations of a word.' Gibbs and Franks (2002, p. 151) likewise note that their data 'show just how prominent metaphor was.' And Moules *et al.* (2004) observe that they were 'struck with how often metaphors arise in the language of grief'. Such claims imply that the probability of metaphor use is high in language, and so in order to verify whether they are true, we must look at the probability of use of metaphor in corpora.

I did research by looking at metaphor probability in 2007, and this involved determining the metaphor status (metaphor versus non-metaphor) of each individual word in an 85,000 word corpus of teleconferences held at investment banks in Portuguese in Brazil; these meetings were attended by bank staff, investors and journalists, and were broadcast over the phone. I then searched a large general corpus of Portuguese (Banco de Português, +220 million words) for the same words found to be used metaphorically in the teleconference corpus. Finally, I compared the frequency of metaphor versus non-metaphor across the two corpora.

In that study, probabilities were calculated in three different ways.

First, all metaphorically used words (MUW) tokens as a proportion of all word tokens in the specialized corpus. This can answer the question of how likely it is that any one word token is a MUW:

4311 MUW tokens / 85438 word tokens in the corpus = .05 (5%)

This indicates that a small share of the words in the corpus are MUWs. The likelihood of word tokens being an MUW is therefore approximately 1 in 20. Literal is the default status for words in the corpus.

Second, all MUW tokens as a proportion of their joint frequency (including both metaphors and non-metaphors) in the specialized corpus. This can provide an answer to the question of how likely it is that an MUW selects a metaphorical meaning:

4311 MUW tokens / 5021 sum of frequency of all MUW types = .86 (86%)

This suggests that MUW types tend to be re-used metaphorically in the same corpus. That is, of all the words in the corpus, those that had taken on a metaphorical meaning tend to so more often than otherwise (that is, be used literally). Metaphor is the default status for MUWs.

Third, the frequency in the reference corpus of all MUW types found in the specialized corpus as a proportion of their joint frequency (including both metaphors and non-metaphors) in the reference corpus. This can help answer how likely it is that MUWs in the specialized corpus are metaphors in language in general:

15220 MUW tokens / 21854 sum of frequency of all MUW types = .7 (70%)

This shows that MUWs in the specialized corpus tend to be MUWs in general language as well, albeit to a lesser degree.

However, when I looked at each word individually and compared their probability of metaphor use in the specialized corpus against the general corpus, I noticed that the vast majority showed 'upward resetting' (HALLIDAY, 1991), that is, their probability of metaphor use was higher in the specialized corpus:

TABLE 5
Resetting of probability of metaphor use

| | |
|---|---|
| Upward resetting (general corpus < specialized corpus) | 323 |
| No resetting (general corpus = specialized corpus) | 63 |
| Downward resetting (general corpus > specialized corpus) | 37 |
| Total | 423 |

Examples of upward resetting MUWs are shown in Table 6.

TABLE 6
Examples of upward resetting

| Word (literal translation) | MUW Prob. in General Corpus | MUW Prob. in Specialized Corpus | Example (adapted translation) |
|---|---|---|---|
| Aliança<br>(*alliance*) | .01 | 1.00 | *o (banco) firmou uma aliança estratégica com (companhia x)*<br>(the bank formed a strategic alliance with (company x)) |
| Parada<br>(*stopped*) | .03 | 1.00 | *a economia parada*<br>(slow economy) |
| Bola<br>(*ball*) | .04 | 1.00 | *acho que precisa de uma bola de cristal para saber o que vai acontecer*<br>(I think we need a crystal ball to know what's going to happen) |
| Jogamos<br>(*throw*) | .05 | 1.00 | *é isso que nós jogamos na projeção*<br>(that's what we build into our projection) |
| Atingidos<br>*(hit)* | .05 | 1.00 | *bases essenciais para que estes objetivos sejam atingidos*<br>(essential basis for us to meet our objectives) |
| Fotografia<br>(*snapshot*) | .05 | 1.00 | *aí tem uma fotografia do que é a transação*<br>(there's a snapshot of what a transaction is) |
| Depositado<br>(*deposited*) | .05 | 1.00 | *que mostra a confiança que o mercado tem depositado no (banco)*<br>(that shows the trust that the market has placed in the bank) |
| Bala<br>(*bullet*) | .06 | 1.00 | *eu entenderia que eles estariam guardando bala*<br>(I would understand that they were holding fire) |
| Loteria<br>(*lottery*) | .07 | 1.00 | *mas isso é loteria, não temos idéia do que vai ocorrer*<br>(but this is a lottery, we have no idea what's going to happen) |
| Travada<br>(*locked*) | .07 | 1.00 | *nós temos a moeda dólar travada para a aquisição*<br>(we have the dollar locked in for the acquisition) |
| Empatar<br>(*tie*) | .08 | 1.00 | *como isso poderia empatar o balanço em reais*<br>(how that might balance the books in Brazilian reais) |
| Canal<br>(*channel*) | .09 | 1.00 | *entre outras estratégias que o banco pode adotar, um canal alternativo*<br>(among other strategies that the bank may adopt, an alternative channel) |
| Chute<br>(*kick*) | .10 | 1.00 | i*sso não foi um chute, foi sim uma análise metodológica*<br>(that was not a guess, it was a methodological analysis) |

Characteristically, these are words of the financial domain. Their metaphoricity is strengthened in the specialized corpus.

Taken together, these findings seem to suggest that metaphors are not evenly distributed across texts; rather, they are typical of certain words/patterns and not others. On the basis of this evidence, metaphors might be seen as a matter of degree (more/less probable) rather than of category (yes/no). In addition, certain metaphors seem to be typical of particular genres or registers rather than of 'language as a whole'. The next section will explore the consequences of that from the point of view of variation.

## 4. Dimensions of metaphor variation

In the previous section, I presented evidence to suggest that the frequency of use of metaphor varies between specialized and general language. The question that arises is whether there is variation across different genres and registers as well. If the answer is affirmative, then this may suggest that metaphor use is patterned at the level of both lexicogrammar and register.

One way in which language use at the level of register may be seen to be systematically patterned is through dimensions of variation. This concept was introduced by Biber (1985; 1988) to refer to underlying parameters of variation, where 'each dimension represents a different set of co-occurring linguistic features' (BIBER, 2009, p. 829). He has developed a method for identifying these dimensions which was termed Multi-Dimensional Analysis of Variation (MDA), which can be defined as:

> a corpus-based methodological approach to, (i) identify the salient linguistic co-occurrence patterns in a language, in empirical/quantitative terms, and (ii) compare registers in the linguistic space defined by those co-occurrence patterns. (BIBER; DAVIES; JONES *et al.*, 2006, p. 5).

To carry out an MDA, the following steps need to be taken:

(a) "An appropriate corpus is designed based on previous research and analysis. Texts are collected, transcribed (in the case of spoken texts), and input into the computer. (In many cases, pre-existing corpora can be used.)

(b) Research is conducted to identify the linguistic features to be included in the analysis, together with functional associations of the linguistic features.

(c) Computer programs are developed for automated grammatical analysis, to identify or 'tag' all relevant linguistic features in texts.

(d) The entire corpus of texts is tagged automatically by computer, and all texts are edited interactively to ensure that the linguistic features are accurately identified.

(e) Additional computer programs are developed and run to compute normed frequency counts of each linguistic feature in each text of the corpus.

(f) The co-occurrence patterns among linguistic features are identified through a factor analysis of the frequency counts.

(g) The 'factors' from the factor analysis are interpreted functionally as underlying dimensions of variation.

(h) Dimension scores for each text are computed; the mean dimension scores for each register are then compared to analyze the salient linguistic similarities and differences among registers." (BIBER, 2009, p. 825-826).

MDA research has  identified dimensions of variation for a number of different languages and varieties. The first MDA description is that of English, which consists of six dimensions, namely: (1) Involved vs. informational production; (2) Narrative vs. Non-narrative concerns; (3) Explicit vs. Situation-dependent reference; (4) Overt expression of persuasion; (5) Abstract vs. Non-abstract information; and (6) On-line informational elaboration.

There were no previous studies that focused explicitly on metaphor variation. Nor were there MDA studies that included variables relating to metaphor use. Nevertheless, there is mounting evidence that metaphor use varies across registers. For instance, Cameron's (2003) study of metaphor in classroom discourse found a rate of metaphor use of 1 out of  every 37 words. My own study of conference calls (referred to in the previous section, BERBER SARDINHA, 2008) showed that metaphor was used at a rate of 1 out of  every 20 words. My research into metaphor use in autobiographical narratives (BERBER SARDINHA, 2010b) indicated the rate of metaphor use to be at 1 out of  every 115 words. And Krennmayr's (personal communication) study of several registers indicated that metaphor use varied from 18.4% of word tokens in academic discourse, to 16.6% in news, to 11.8% in fiction, to 7.8% in conversation. Different identification methods were used in these studies, as well as different definitions of what is counted as a metaphorical unit, therefore these figures are not directly comparable. This is confirmed by my own analyses of the same autobiographical narrative corpus; an early analysis showed a rate of 1 metaphor every 364 words, but more recently this changed to 1 in every 115, due to changes in the procedures for metaphor identification.

Despite these problems, this combined evidence may suggest that different registers use metaphors at different rates, and that perhaps casual spoken non-scripted registers such as conversation and personal narratives employ fewer metaphors than information-laden written registers such as academic or news.

To verify that, I decided to conduct an MDA of three major registers of Brazilian Portuguese, and include in the variable set a number of metaphor-related variables.

The corpus used for this study consisted of a small subset of the Brazilian MDA Corpus, which in turn is taken from the much larger Brazilian Corpus (1 billion words; http://corpusbrasileiro.pucsp.br):

TABLE 7
MDA of metaphor variation corpus

| Register | Tokens | Texts |
|---|---|---|
| Conversation | 17,042 | 8 |
| Academic | 16,915 | 8 |
| Newspaper | 18,165 | 67 |
| Total | 52,122 | 83 |

The corpus was compiled to meet a target of around 50 thousand words, distributed roughly equally among its registers. The target was chosen because it did not seem too large for manual analysis. Previous studies that involved close reading of entire corpora have used less data, such as Cameron (2003), who took a corpus of 27,000 words of classroom talk, Cameron (2010), with a 27,000 word corpus of reconciliation discourse, and Charteris-Black (2004), whose corpus of American political speeches was 33,000 words long. There is no consensus on corpus size for such research projects, and other studies used larger data sets, such as Steen *et al.* (2010), which is based on a corpus of 190,000 of English data and 130,000 of Dutch.

Initially, the variable pool included 57 variables. The corpus was tagged for part of speech by the Tree-Tagger (trained for Portuguese), and for metaphor features by hand. After that, variable frequencies were taken and examined, and a number of low frequency variables were dropped. An initial factor analysis was run (in SPSS) and communalities were examined. Some variables were dropped based on their communalities, either because they were too low (<.4 according to STEVENS, 2002, p. 410) or too high (1 or higher). The result was a final set of 25 variables, shown below.

To code metaphors, I drew on the concepts of metaphor Topic and Vehicle. A metaphor Topic is that which is being referred to metaphorically. A Vehicle, in turn, is that which is used metaphorically. For instance, in the metaphor 'waste of time', 'time' is the Topic, and 'waste' the Vehicle. Time is being metaphorized in terms of a precious resource that should not be wasted.

Metaphor variables

| Variable | Example<br>*Translation* |
|---|---|
| 1) metaphor density | Metaphorically used types / total word tokens in the text |
| 2) metaphor topic: people | Fiz amigos<br>*I made friends* |
| 3) metaphor topic: social | Pressão de diversos países<br>*Pressure by different countries* |
| 4) metaphor topic: abstract | Queda da participação<br>*Drop in participation* |
| 5) metaphor vehicle: movement/position | Elevadas taxas<br>*High taxes* |
| 6) metaphor vehicle: object/buildings | Relacionamento construtivo<br>*Constructive relationship* |
| 7) metaphor vehicle: other | Campo de tensões<br>*Field of tension* |
| 8) vehicle word POS: verb | Aumentar o tempo<br>*Increase the time* |

Linguistic variables

1) adjectives
2) adverbs
3) demonstratives
4) future tense
5) nouns
6) past participles
7) past tense verbs
8) possessives
9) prepositions
10) pronouns: 1st person
11) pronouns: 2nd person

12) pronouns: 3rd person
13) proper nouns
14) public verbs
15) be as main verb (ser, estar)
16) subordinate clauses
17) verbs

In order to determine how many factors are present in the data, a graph known as 'scree plot' is normally used in MDA. It plots the eigenvalues, or variances of the factors. Researchers look at the line searching for points where it breaks, indicating major differences in factor variances. The scree plot for the initial factor solution seemed to indicate a three-factor solution, as shown in the figure below.

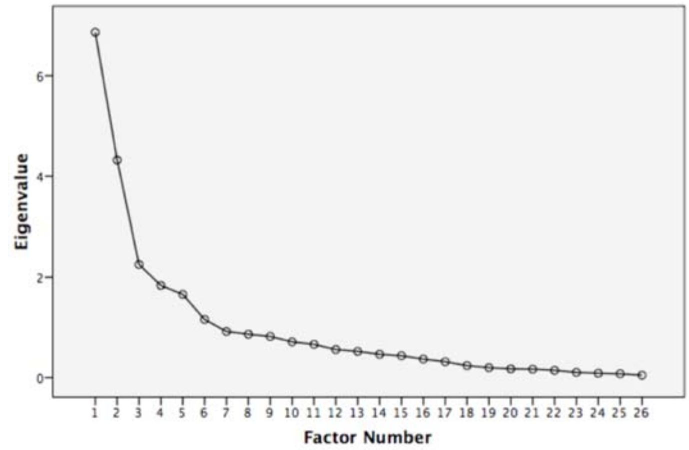


FIGURE 3 - Scree plot for metaphor MDA data

A three-factor Promax rotated analysis was then run on the data. The total variance captured was 47%, which is close to Biber's (1988) final 6-factor solution (at 49%). This suggested that the factor analysis seemed to have tapped into a good portion of the variation present in the data. Factor intercorrelations were small, at -.088 (between factors 1 and 2), -.405 (between 1 and 3), and .29 (2 and 3). This is again similar to Biber (1988), where they ranged from -.49 to .3.

The structure of the first factor is shown below.

TABLE 8
Factor 1 structure

| Variable | Loadings |
|---|---|
| adverbs | .84 |
| subordinate clause | .76 |
| pron 3rd person | .75 |
| tokens | .69 |
| demonstratives | .65 |
| past tense | .61 |
| adjectives | .58 |
| pron 1st person | .57 |
| pron 2nd person | .53 |
| possessives | .42 |
| verbs (other) | .41 |
| be as main verb (ser, estar) | .39 |
| public verbs | .36 |
| Metaphor density | -.58 |
| proper nouns | -.56 |

This factor encompassed a large number of linguistic features and only one metaphor variable (density). Adverbs, subordination, be as main verb, first and second person pronouns are all features occurring on Biber's Dimension 1 (BIBER, 1988, p. 105-107), signaling involved production. Public verbs and past tense are present on Biber's 1988 Dimension 2, indicating narratives. Adjectives appear on his Dimensions 1, 2, and 5, associated with informational, non-narrative and abstract discourse. And demonstratives occur on his Dimension 6, linked to online informational elaboration. In all, these features seem to indicate verbal, narrative, involved discourse produced under real-time conditions. The proper nouns at the bottom of the factor suggest an informational focus.

The distribution of registers along this dimension is shown below.

TABLE 9

Dimension 1 'Involved narrative production vs metaphor use'.

Mean factor scores for each register

DIMENSION 1

| | |
|---|---|
| + | |
| 22 | |
| 21 | conversation |
| 20 | |
| 19 | |
| … | |
| 0 | |
| -1 | academic |
| -2 | newspaper |
| -3 | |
| … | |
| -20 | |
| - | |

F=53.051, p=.000, $R^2$=.617

The first factor is generally the one that captures most of the variation. This is reflected in the distance between conversation, at the top, and the other two registers at the bottom. The register with the highest score on this dimension was conversation, which means conversations have high quantities of the positive features (mostly verbal features, as indicated above), and low quantities of negative ones (proper nouns and metaphors). I labeled this dimension 'Involved narrative production versus metaphor use', because the positive features seem to highlight the involved nature of conversation, while at the same time revealing that involvement seems to be achieved with very little need for metaphors. Proper nouns are missing in conversation because they are generally replaced by pronouns. This appears to confirm the earlier hunch that in casual spoken registers, metaphor is not a frequent feature.

The structure of the second factor is shown below.

TABLE 10
Factor 2 structure

| FACTOR 2 | |
|---|---|
| Variable | Loadings |
| verbs as vehicles words | .92 |
| social topics | .77 |
| object/building vehicles | .73 |
| abstract topics | .69 |
| other vehicles | .60 |
| movement/position vehicles | .40 |
| (adjectives | -.32) |
| (Metaphor density | -.42) |

In this factor, a large number of metaphor features are clustered together, and there are only positive features, since the variables at the negative end of the scale have higher loadings in other factors and are therefore disregarded for the computation of factor scores (but they are considered during factor interpretation). This paints a non-specific picture of metaphor use, one that does not seem to differentiate between different kinds of topics and vehicles. It seems to suggest that those three registers appear to have no preferences for particular kinds of metaphorically used words. Abstract and social topics are linked to particular kinds of vehicles, but not to metaphor density (it is in brackets because it had a higher loading on factor 1). This suggests that there is some association between abstraction and metaphorical language, but not between abstraction and metaphor frequency.

The distribution of registers along dimension 2 is shown below.

TABLE 11
Dimension 2 'non-specific metaphor use'

Mean factor scores for each register
DIMENSION 2

| | |
|---|---|
| + | |
| 11 | |
| … | |
| 2 | |
| 1 | newspaper |
| 0 | |
| -1 | academic |
| -2 | |
| -3 | |
| -4 | |
| -5 | conversation |
| -6 | |
| … | |
| -9 | |
| - | |

F=5.775, p=.005, R2=.145

Unlike in the previous factor, in this one registers are not distributed far apart, suggesting there is not much difference between them. I called this dimension 'non-specific metaphor use' because of the lack of correlation between particular kinds of metaphor and registers. Newspaper is the less metaphor specific register, wich suggests that it will employ just about any kind of metaphorically used word or refer to about any topic metaphorically. Conversation, on the other hand, seems to be a little less non-selective, but not enough to have any noticeable preference (otherwise the variables in the factor would have broken differently across the positive and negative ends). The fact that conversation is also metaphorically sparse (as suggested in the previous factor) may also influence these results, since there may not be enough metaphors to go around in conversation to constitute some sort of solid preference for any particular topic or vehicle.

Finally, the structure of the third factor is shown below.

TABLE 12
Factor 3 structure

| FACTOR 3 | |
|---|---|
| Variable | Loadings |
| nouns | .89 |
| prepositions | .84 |
| past participle | .50 |
| (proper nouns | .49) |
| (movement/position vehicles | .34) |
| (abstract topics | .30) |
| | |
| (pron 1st person | -.41) |
| (pron 2nd person | -.40) |

None of the variables that entered in the calculation of factor scores for dimension 3 is metaphor-related, namely nouns, prepositions and past participles. The remaining variables (in brackets) have higher loadings in other factors. These three variables seem to suggest an information focus, since nouns and prepositions are used in nominal groups which can package information densely. And past participles can form part of passive voice, which is a common feature of elaborate informative and/or argumentative registers. Pronouns, which cluster together on the negative pole of the dimension, are indicative of an interactive focus. This distribution of variables resembles in part that of Biber's (1988) first dimension, 'Involved vs. informational production', and so our dimension was named after that.

Metaphors are often thought of as devices that can help express abstract ideas as more concrete ones. Thus, it is interesting that characteristics normally associated with abstraction and information, such as the ones in this factor, are not linked to higher metaphor use (metaphor density). There is some association to abstract topics and to movement and position vehicles (metaphors of things going up and down, in and out, etc.), though, but these have higher loadings on factor 2, shown previously.

The distribution of registers on dimension 3 appears below.

TABLE 13

Dimension 3 'informational versus involved production'

Mean factor scores for each register

DIMENSION 3

| | |
|---|---|
| + | |
| 4 | |
| … | |
| 2 | academic |
| 1 | |
| 0 | newspaper |
| -1 | |
| -2 | |
| -3 | |
| -4 | conversation |
| -5 | |
| … | |
| -10 | |
| - | |

F=17.582, p=.000, R2=.349

Academic is the most informational register; newspaper is at the center, suggesting that on average it is both informational and involved. Conversation is at the bottom end of the scale, representing involved production. Once again, metaphors do not seem to come into play in defining conversation. This again reflects the scarcity of metaphor in this register.

On this factor, the ordering of registers is different from that on the other factors. In the previous factors, it was conversation – academic – news (regardless of polarity), and here it is academic – news – conversation. The ordering in and of itself is not particularly revealing, since registers are aligned on the scale according to their scores. What has remained consistent across the factors is the larger difference between the scores for conversation, on the one hand, and for the remaining registers, on the other. This suggests that conversation is a more distinctive register, which in turn perhaps reflects the basic distinction between spoken and written language, with the written registers (academic and newspaper) sharing more characteristics between themselves than with the spoken register.

In this section, I looked at the extent to which variation in metaphor is systematic and whether it can give rise to dimensions. Statistically significant results suggest that there is systematic variation in metaphor use across registers, with conversation standing in contrast with both academic and newspaper as a more metaphor-scarce language variety. The type of metaphor used in registers was not a good predictor of variation, though. There was some evidence to suggest that abstract topics are often metaphorized in informational registers. Metaphor density, on the other hand, was a strong component in the factors, forming a pole in factor 1. Registers seemed to be distinguished in terms of the quantity of metaphors present in them, with written registers sharing most of the metaphors, and conversation the fewest.

It must be stressed that these dimensions are not final. Larger corpora must be analyzed before a definitive set of dimensions is agreed on. Biber himself carried out preliminary analyses (BIBER, 1985) before arriving at the six dimensions that are currently referred to. Problems such as the subjective nature of metaphor identification and the labor intensive nature of such work on large quantities of text surely impose limits on both the range of registers that can be investigated and the number of texts that are included to represent each register. Work on dimensions of variation has been made possible in large part by automatic taggers (especially the 'Biber tagger', which is a reference in MDA). Similarly, if research in metaphor dimensions of variation is to continue and expand, then software for metaphor identification must be developed. This is the topic of the next section.

## 5. Automated metaphor retrieval

I have been engaged in developing software for metaphor extraction for several years. This has led to several prototypes of the Metaphor Candidate Identifier (BERBER SARDINHA, 2006; 2007a; 2010b; 2010b), a program that is intended to find possible metaphors (i.e., candidates) in corpora. It has been made available online for several years under different names (Metaphor Tagger, Metaphor Identifier). Support for the online versions has stopped while development of a desktop version is underway. The version I will report on here is number 4 (desktop), and it works as follows:

(a) For each word token in a corpus (of Portuguese or English), grab its collocates from 5 words to the left to 5 words to the right.

(b) For each of these collocates, determine its part of speech and lemma.

(c) Build list of node and collocate pairs, including lemma and part of speech.

(d) Search for each node-collocate pair in a database of metaphor patterns (built during training).

(e) If match is found, consider that word token a potential metaphor; if not, consider it as not being a potential metaphor.

The basis of the program is a large metaphor pattern database, consisting of over 541 thousand patterns. An example of a pattern found in the database is:

> NL_CW2R varrer_mapa (translation: sweep map)
> NL: Node is a lemma
> CW: Collocate is a word (not lemma)
> 2R: Collocate is at two words to the right of the node

This pattern will capture the expression 'varrer do mapa' (sweep off the map).

Not all patterns have positional constraints such as this; others will capture occurrences within the whole width of the collocational span. Others will be formed by semantic fields (represented in square brackets), such as:

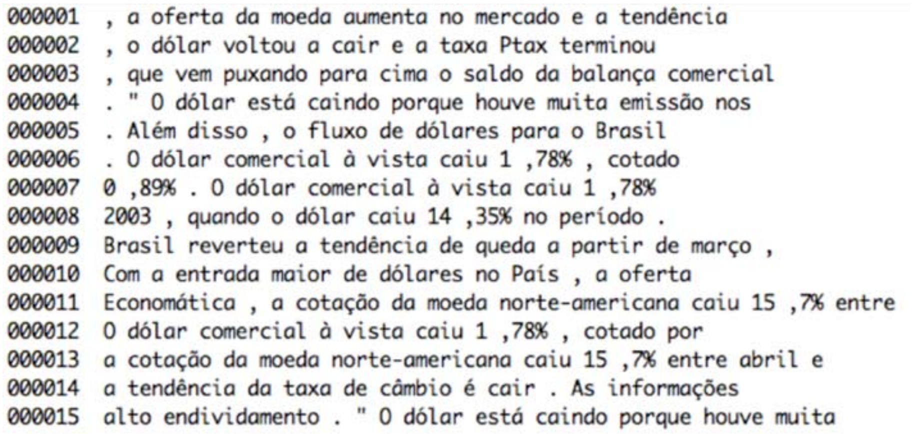
> abaixo [not concrete] (translation: under/below)

This pattern will match expressions such as 'abaixo das expectativas' (below expectations).

Semantic fields are entered in a separate database, in the form of word lists. Currently, the program will not do word disambiguation, and will simply match words in the lists to those in the corpus; errors may occur because of that, for instance, by treating 'meia' (sock) to be 'meia' (half). There are no word disambiguation programs for Portuguese available.

The program (written as a script in Unix shell and Perl) works reasonably fast, being able to process a million words in under five minutes in a standard desktop computer with 4 GB RAM.

The MCI outputs segments of text where it has found a metaphor pattern. The screen below shows the output of the analysis of a text on economics:



```
000001  , a oferta da moeda aumenta no mercado e a tendência
000002  , o dólar voltou a cair e a taxa Ptax terminou
000003  , que vem puxando para cima o saldo da balança comercial
000004  . " O dólar está caindo porque houve muita emissão nos
000005  . Além disso , o fluxo de dólares para o Brasil
000006  . O dólar comercial à vista caiu 1 ,78% , cotado
000007  0 ,89% . O dólar comercial à vista caiu 1 ,78%
000008  2003 , quando o dólar caiu 14 ,35% no período .
000009  Brasil reverteu a tendência de queda a partir de março ,
000010  Com a entrada maior de dólares no País , a oferta
000011  Economática , a cotação da moeda norte-americana caiu 15 ,7% entre
000012  O dólar comercial à vista caiu 1 ,78% , cotado por
000013  a cotação da moeda norte-americana caiu 15 ,7% entre abril e
000014  a tendência da taxa de câmbio é cair . As informações
000015  alto endividamento . " O dólar está caindo porque houve muita
```

In this particular case, all of the 15 lines were correctly picked up as they all have at least one metaphor:

1. offer of currency grows
2. the dollar has fallen again
3. balance of trade is being pulled upwards
4. the dollar is falling
5. the flow of dollars
6. the dollar has fallen
7. the dollar has fallen
8. when the dollar fell
9. downward trend
10. dollars went in
11. exchange rate fell
12. dollar fell
13. exchange rate fell
14. exchange rate trend is downward
15. high debt … dollar is falling

I tested the MCI on a small corpus that was then hand coded for metaphors, made up by the following texts:

TABLE 14
Test corpus for the MC

| Register | Texts | Tokens | Types |
|---|---|---|---|
| News on the economy | 15 | 5,510 | 1,562 |
| News on science | 15 | 10,321 | 2,947 |
| Political speeches | 20 | 24,712 | 3,865 |
| Total | 50 | 40,543 | 6,543 |

I computed the following metrics:

Precision: Metaphors correctly found divided by the total number of attempts (an attempt occurs when the program selects a metaphor candidate).

Recall: Metaphors correctly found divided by the total number of existing metaphors in the corpus according to manual analysis.

Results appear below.

TABLE 15
MCI precision

| Texts | Metaphors correctly found by MCI (True positives) | Attempts (True positives + False positives) | Precision % |
|---|---|---|---|
| Economy | 427 | 478 | 89.3 |
| Science | 364 | 606 | 60 |
| Politics | 1136 | 1591 | 71.4 |
| *Overall* | *1927* | *2675* | *72* |

TABLE 16
MCI recall

| Texts | Metaphors correctly found by MCI (True positives) | Existing metaphors | Recall % |
|---|---|---|---|
| Economy | 427 | 578 | 74 |
| Science | 364 | 535 | 68 |
| Politics | 1136 | 1563 | 73 |
| *Overall* | 1927 | 2676 | *72* |

Both precision and recall were 72% on average for the whole corpus. This means that 7 out of every 10 candidates MCI identified were really metaphors, and for every 10 metaphors in the corpus, 7 were correctly picked up. A performance at 70% is far from the ideal 100% that would be initially expected of a metaphor retriever, but this must be weighed against the difficulties involved in finding metaphors in texts by hand. This is demonstrated by several studies, such as Cameron (2003, p. 169), who reports an initial agreement of only 14% among analysts on a text in her corpus. Beigman Klebanov, Beigman and Diermeier (2008), in their study on newspaper metaphors, observe that agreement varied between 1.7% to 4%. And Steen *et al.* (2010) also show discrepancy between human analysts. At the same time, both Cameron and Steen et al. show that disagreement can be avoided by having very clear criteria for what counts as a metaphor, and it can also be resolved through discussion between the analysts. Such results underscore the difficulties involved in identifying metaphors, and imply that the gold standard must remain hand analysis, despite its shortcomings. I agree with that, but would further add that machine analysis must not be seen as substitute for manual analysis of metaphor. And that machine analysis should be considered as an extra rater in research teams.

This is because just as different people tend to find different but true metaphors, so does the computer when compared to people. In another study (BERBER SARDINHA, 2010a), I compared two independent analyses, by the MCI and by hand, and showed that the MCI correctly retrieved a large number of metaphors that were not noticed by hand and eye. Figure 4 shows the results of this study: the intersection between the two procedures (manual and MCI) is small. Inspection of the metaphors found revealed that the computer analysis was more consistent, never missing any one metaphor that it was taught to recognize, generally conventionalized ones. Human analysis, on the other hand, was better at finding metaphors that depended on context to be noticed, and also spotted innovative metaphors. The computer never gets 'distracted' or tired, while humans do, especially in activities that demand sustained attention such as metaphor identification in corpora.

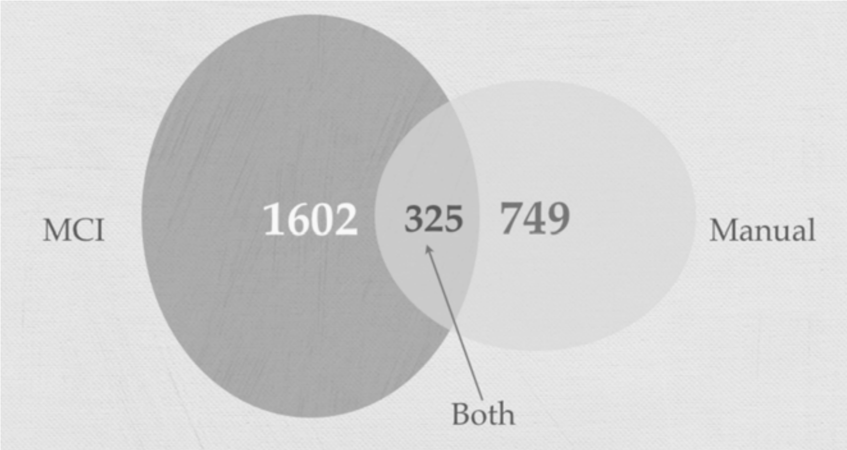


FIGURE 4: Comparison of true metaphors found by hand and by machine (MCI)

## 6. Final comments

In this paper, I argued that Corpus Linguistics has a great deal to contribute to metaphor studies, particularly with respect to research that shows:

1. The kinds of lexicogrammatical patterning that both arises from and signals metaphor in language use;
2. The extent to which metaphor use is patterned;
3. How metaphor varies across different genres and registers;
4. The extent to which such variation is systematic;
5. How research findings into linguistic patterning of metaphor can help develop tools to assist in automating at least in part the process of metaphor identification.

I also believe that these particular types of research can feed back on each other and support the development of resources to enable more CL metaphor research.

The development of resources such as metaphor identification assistance tools, semantically annotated corpora, and platforms for hand annotation of metaphors in corpora, among others, can all strengthen the important ties between metaphor and Corpus Linguistics. This way, the fields of metaphor and Corpus Linguistics can continue to mutually support and benefit from each other.

## References

BEIGMAN KLEBANOV, B. *et al.* Analyzing disagreements. In: *Workshop on Human Judgements in Computational Linguistics*, Coling 2008, Manchester, UK. 2008.

BERBER SARDINHA, T. A tagger for metaphors. Paper, RaAM - Researching and Applying Metaphor 6. Leeds, UK. 2006.

BERBER SARDINHA, T. Análise de metáfora em corpora. *Ilha do Desterro*, v. 52, p. 167-201, 2007a.

BERBER SARDINHA, T. Recontando a vida em narrativas pessoais: Um estudo de metáforas na perspectiva da Linguística de Corpus. *Organon*, v. 21, p. 143-160, 2007b.

BERBER SARDINHA, T. Metaphor probabilities in corpora. In: ZANOTTO, M. S. *et al.* (Ed.). *Confronting Metaphor in Use*: An Applied Linguistic Approach. Amsterdam/Atlanta, GA: Benjamins, 2008.

BERBER SARDINHA, T. Questões metodológicas de análise de corpora na perspectiva da Linguística de Corpus. *Gragoatá*, v. 26, p. 81-102, 2009.

BERBER SARDINHA, T. Improving and evaluating the Metaphor Candidate Identifier. Paper, RaAM – Researching and Applying Metaphor Conference. Amsterdam, the Netherlands. 2010a.

BERBER SARDINHA, T. MCI, um Identificador de Candidatos a Metáfora em corpora. In: SHEPHERD, T. *et al.* (Ed.). *Caminhos da Linguística de Corpus*. Campinas, SP: Mercado de Letras, 2010b. (Espaços da Linguística de Corpus).

BERBER SARDINHA, T. A program for finding metaphor candidates in corpora. *The ESPecialist*, v. 31, n. 1, p. 49-68, 2010c.

BHATIA, V. K. *Worlds of Written Discourse - A Genre-based View*. London, New York: Continuum, 2004. (Advances in Applied Linguistics).

BIBER, D. Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics*, v. 23, p. 337-360, 1985.

BIBER, D. *Variation across Speech and Writing*. Cambridge: Cambridge University Press, 1988.

BIBER, D. Multi-dimensional approaches. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus Linguistics* – An International Handbook. Berlin / New York: Walter de Gruyter, 2009.

BIBER, D.; CONRAD, S. *Register, genre, and style*. Cambridge; New York: Cambridge University Press, 2009. (Cambridge textbooks in linguistics).

BIBER, D. *et al.* Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora*, v. 1, n. 1, p. 1-37, 2006.

CAMERON, L. *Metaphor in Educational Discourse*. London: Continuum, 2003.

CAMERON, L. What is metaphor and why does it matter? In: CAMERON, L.; MASLEN, R. (Ed.). *Metaphor analysis*: Research practice in applied linguistics, social sciences and the humanities. London: Equinox, 2010.

CAMERON, L.; MASLEN, R. *Metaphor analysis*: Research practice in applied linguistics, social sciences and the humanities. London: Equinox, 2010. (Studies in applied linguistics).

CHARTERIS-BLACK, J. *Corpus Approaches to Critical Metaphor Analysis*. Basingstoke: Palgrave Macmillan, 2004.

CREET. *Metaphor Analysis Project*. 2006. Unpublished Work.

DEIGNAN, A. *Metaphor and Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins, 2005.

DEIGNAN, A.; POTTER, L. A corpus study of metaphors and metonyms in English and Italian. *Journal of Pragmatics*, v. 36, n. 7, p. 1231-1252, 2004.

GIBBS, R. W. (Ed.) *The Cambridge Handbook of Metaphor and Thought*. New York: Cambridge University Pressed. 2008a.

GIBBS, R. W. Metaphor and thought – The state of the art. In: GIBBS, R. W. (Ed.). *The Cambridge Handbook of Metaphor and Thought*. New York: Cambridge University Press, 2008b.

GIBBS, R. W.; FRANKS, H. Embodied metaphor in women's narratives about their experiences with cancer. *Health Communication*, v. 14, n. 2, p. 139-165, 2002.

HALLIDAY, M. A. K. Corpus studies and probabilistic grammar. In: AIJMER, K.; ALTENBERG, B. (Ed.). *English Corpus Linguistics*: Studies in Honour of Jan Svartvik. London: Longman, 1991.

MOULES, N. J. *et al.* Making room for grief: walking backwards and living forward. *Nursing Inquiry*, v. 11, n. 2, p. 99-107, 2004.

PRAGGLEJAZ GROUP. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, v. 22, n. 1, p. 1-39, 2007.

SCOTT, M. *WordSmith Tools, version 4*. Oxford: Oxford University Press, 2004.

STEEN, G. *Finding Metaphor in Grammar and Usage*. Amsterdam / Philadelphia: John Benjamins, 2007.

STEEN, G. *et al. A Method for Linguistic Metaphor Identification*: From MIP to MIPVU. Amsterdam: John Benjamins, 2010.

STEFANOWITSCH, A. Corpus-based Approaches to Metaphor and Metonymy. In: STEFANOWITSCH, A.; GRIES, St. Th. (Ed.). *Corpus-based Approaches to Metaphor and Metonymy*. Berlin; New York: M. de Gruyter, 2006.

STEFANOWITSCH, A.; GRIES, St. Th. *Corpus-based Approaches to Metaphor and Metonymy*. Berlin; New York: M. de Gruyter, 2006. (Trends in linguistics. Studies and monographs, 171).

STEVENS, J. *Applied multivariate statistics for the social sciences*. 4. ed. Mahwah, N.J.: Lawrence Erlbaum Associates, 2002.

WIKBERG, K. The role of corpus linguistics in metaphor research. In: JOHANNESON, N. L.; MINUGH, D. (Ed.). *Selected Papers from the 2006 and 2007 Stockholm Metaphor Festivals*. Stockholm: Department of English, Stockholm University, 2007.

Recebido em 08/04/2011. Aprovado em 08/05/2011.

# Corpora from a sociolinguistic perspective

# *Corpora sob uma perspectiva sociolinguística*

Tyler Kendall*
University of Oregon
Eugene / USA

ABSTRACT: In this paper, I consider the use of corpora in sociolinguistic research and, more broadly, the relationships between corpus linguistics and sociolinguistics. I consider the distinction between "conventional" and "unconventional" corpora (Beal et al. 2007a, b) and assess why conventional corpora have not had more traction in sociolinguistics. I then discuss the potential utility of corpora for sociolinguistic study in terms of the recent trajectory of sociolinguistic research interests (Eckert under review), acknowledging that, while many sociolinguists are increasingly using more advanced corpus-based techniques, many are, at the same time, moving away from corpus-like studies. I suggest two primary areas where corpus developers, both sociolinguistic and non-, could focus to develop more useful corpora: Corpora containing a wider range of non-standard (spoken) varieties and more flexible annotation and treatment of spoken language data.

KEYWORDS: Sociolinguistics; conventional and unconventional corpora; spoken language corpora; data management; annotation methods.

RESUMO: Neste artigo considero o uso de corpora na pesquisa sociolingüística e, de modo mais geral, a relação entre a linguística de corpus e a sociolinguística. Reflito sobre a distinção entre corpora "convencionais" e "não-convencionais" (BEAL ET AL. 2007 a, b) e avalio o porquê de corpora convencionais não terem atraído mais atenção no campo da sociolinguística. Na sequência, discuto a utilidade potencial de corpora para os estudos sociolingüísticos em termos da trajetória recente que tem sido adotada pela pesquisa nesta área (ECKHERT, em avaliação), reconhecendo que, se por um lado, muitos sociolinguistas têm ampliado o seu uso de técnicas avançadas da linguística de corpus, por outro, muitos estão, ao mesmo tempo, se afastando de estudos relaciados a corpora. Sugiro duas áreas principais nas quais compiladores de corpora, independentemente de serem sociolingüísticos ou não, poderiam enfocar para desenvolverem corpora mais úteis: corpora contendo uma amplitude maior de variedades (faladas )não-padrão e um esquema mais flexível de anotação e tratamento de dados orais.

PALAVRAS-CHAVE: Sociolinguística; corpora convencionais e não-convencionais; corpora orais; gerenciamento de dados; métodos de anotação.

* tsk@uoregon.edu

## 1. Introduction

Much work in sociolinguistics is firmly empirical and based on the analysis, whether quantitative or qualitative, of data of actual language use. As such, sociolinguistics is a field that has many natural connections to corpus linguistics, and these connections have not gone unnoticed. Several recent collections of papers (KRETZSCHMAR; ANDERSON; BEAL; CORRIGAN; OPAS-HÄNNINEN; PLICHTA, 2006; BEAL; CORRIGAN; MOISL, 2007a, 2007b; KENDALL; Van HERK, 2011), articles (BAUER, 2004; ANDERSON, 2008; ROMAINE, 2008), and a book (BAKER, 2010) have explicitly explored some of the relationships between sociolinguistics and corpus linguistics.[1] Despite these connections, however, there is often little direct interaction between scholars in these two fields. For instance, research undertaken on corpora like the British National Corpus (BNC) that might be described as "corpus sociolinguistic" (BAKER, 2010) does not appear to have caught on within mainstream sociolinguists to any large extent.[2]

Why is this the case? That is, why has corpus linguistics not had a larger influence on sociolinguistics? Beal *et al.* (2007a, b) made the useful clarification that much sociolinguistic work involves what they termed "unconventional" corpora, corpora that do not fit the standard mold of resources like the BNC. In fact, their volumes sought

> to establish whether or not annotation standards and guidelines of the kind already employed in the creation of more conventional corpora on standard spoken and written Englishes … should be extended to less conventional corpora so that they too may be 'tamed' in similar ways (BEAL *et al.*, 2007a: 1).

[1] Also, see Kretzschmar's (2009) *The Linguistics of Speech* for an interesting and helpful discussion of some historiographical connections between sociolinguistics and corpus linguistics.

[2] Before proceeding, it is worth commenting that "sociolinguistics" is a term that covers a diverse set of approaches to linguistics across several disciplines and encompasses many different traditions of research, ranging from, e.g., areas of linguistic anthropology, sociology, discourse studies, variationist linguistics, and so on. (Of course, "corpus linguistics" can also be considered a cover term for a number of different approaches to linguistics.) In this paper, I will often refer to "sociolinguistics" and "sociolinguists" in monolithic terms, but I realize that I am perhaps over-generalizing. It may help to explain that my own background is from the "variationist approach," the field-based, quantitative study of language variation and change pioneered by scholars like William Labov and Walt Wolfram. Some readers may find my point of view overly influenced by that flavor of sociolinguistics.

While the main question in this passage, about the applicability of standard corpora annotation to other kinds of corpus-like data, is a difficult question (and ultimately I will steer away from it, offering an alternative possibility in §5.2 of this paper), this notion of "conventional" and "unconventional" corpora is a useful one. Sociolinguistic research often focuses on non-standard varieties of language, and spoken language in particular, and the large "conventional" (i.e. standard language and often primarily written) corpora have simply not been of great use for pursuing sociolinguistic research on non- or less-standard varieties.[3]

In this paper, I consider the role of corpora and corpus linguistic methodologies in sociolinguistics and the division between conventional and unconventional corpora further. I begin, in section 2, by considering the role of corpora (broadly defined) in sociolinguistics. In section 3, I look at the ways that "traditional" corpora have been used to ask sociolinguistic questions, and consider why more work one might describe as "corpus sociolinguistic" (BAKER, 2010) has not been undertaken. I follow this up in section 4, by reviewing recent trajectories of research interest in sociolinguistics and considering how this impacts the relationship(s) between corpus linguistic and sociolinguistic work. In section 5, I consider what future directions corpus linguistics, and corpus development in particular, could take in order to facilitate corpus-based research on sociolinguistic questions. Finally, in section 6, I close with some summary comments.

[3] As a side note, the implementation of representativeness (McENERY; WILSON, 2001; McENERY; XIAO; TONO, 2006) in the construction of the major conventional corpora may be limiting from a sociolinguistic perspective. Corpora like the BNC, the Corpus of Contemporary American English (COCA), and those in the Brown family, in attempting to represent national varieties, are by necessity somewhat normative and exclusive – they downplay and/or normalize over the true diversity of language at a national scale. The notion of representativeness – what larger population of language use or users a corpus trustworthily samples – is crucial in all corpus-based work (see also GRIES, 2006). Yet, the sampled variability included in conventional corpora is often ordered along the dimension of register or genre, not the dimension(s) of social variation. This cannot be helped – as impressively large as, e.g., COCA is (at over 410 million words), it would be impossible for it to fully represent, say, the ethnic diversity of English in the U.S. For that matter, how does one even assess the full extent of ethnic diversity of English in the U.S.?

## 2. The place of corpora in sociolinguistics

Many approaches to sociolinguistics involve the analysis of bodies of naturally occurring talk. The size of these "bodies" of data range from quite small, such as a single conversation or story, to massive, e.g., hundreds of hours of recorded interviews collected over years of fieldwork. Increasingly, sociolinguists are calling these datasets "corpora".[4] Whether these corpora meet the "proper" definition maintained by corpus linguists (balanced, representative, machine-readable; cf. McENERY; WILSON, 2001; McENERY; XIAO; TONO, 2006) or not (they most often do not), they can still be usefully examined via corpus-based methodologies. Techniques such as examining concordances and collocational patterns, conducting keyword analyses, and the use of corpus analysis software tools themselves can shed useful light into even quite small datasets (cf. BAKER, 2010).

While it sometimes seems to be the case that sociolinguists' borrowings from corpus linguistics are shallow (and, as mentioned in footnote 4, possibly at times only name-deep), corpora certainly have a growing place in sociolinguistic research and some connections have already existed for decades. For instance, in an important paper, Poplack (1989) detailed the creation of her at-the-time "mega-corpus" of spoken Ottawa-Hull French, which contains 3.5 million words from 270 hours of recorded speech.[5] Other early

[4] Considering terminology further, "corpus" and "sociolinguistics" are both terms that are used variously by different groups of scholars. Within corpus linguistics, it often seems to be the case that "sociolinguistics" is used as a cover term for all kinds of corpus-based research that involves extra-linguistic factors. But among sociolinguists, much of this research fails to have traction; it is not always seen as sociolinguistic, or at least as "sociolinguistic enough". Meanwhile, there is also an increasing tendency for sociolinguistic researchers to consider and discuss their data as "corpora" in ways that over-generalize that term. More and more sociolinguistic field-based projects appear to outcome in collections of data that are named as corpora (e.g., hypothetically, "Smalltown USA Corpus"), when for corpus linguists they often have none of the characteristics of "corpus proper". It is hard to see where one draws the line between the "unconventional" corpora described in Beal *et al.* (2007a, 2007b) and data collections that simply are not appropriately considered "corpora". I do not want to dwell on terminological issues too much in this paper, but it is worth considering whether some of the most obvious current connections between sociolinguistics and corpus linguistics may only be name deep.

[5] While 3.5 million words may not seem like a "mega-corpus" to present day corpus researchers, it should be remembered that this word count reflects only natural, conversational spoken language, and, especially for a resource created over twenty years ago is an extremely impressive accomplishment. Within sociolinguistics, it remains one of the largest corpus-like corpora.

sociolinguistic work also focused extensively on describing aspects of their projects that in today's terms would likely be described as "corpus creation" (e.g., SHUY; WOLFRAM; RILEY, 1968; SANKOFF, D.; SANKOFF, G., 1973; cf. KENDALL, 2008). Tagliamonte's (2006) sociolinguistics textbook, *Analysing Sociolinguistic Variation*, outlines an approach to (variationist) sociolinguistics that many corpus linguistics would likely be comfortable with (and would, I think, find quite useful).

It seems clear that in coming years sociolinguistics will make increasing use of corpora and will increasingly interact with corpus-based approaches to linguistics from other areas. Endeavors like the Origins of New Zealand English (ONZE) project (GORDON; MACLAGAN; HAY, 2007), the Newcastle Electronic Corpus of Tyneside English (NECTE; ALLEN; BEAL; CORRIGAN; MAGUIRE; MOISL, 2007), and the Danish LANCHART project (LANguage CHAnge in Real Time; GREGERSEN, 2009), all of which involve the creation of impressive "unconventional" corpora, point to the fact that sociolinguists are paying more serious attention to corpus-based methodologies and the benefits of explicit corpus creation work.

There are also growing connections between sociolinguistics and corpus linguistics in terms of specific research. For instance, Torgersen, Gabrielatos, Hoffmann, and Fox (2011) provide an excellent example of how corpora and corpus linguistic methodologies can be used to pursue core sociolinguistic questions, such as the actuation of language change (WEINREICH; LABOV; HERZOG, 1968). In this work, they examine pragmatic markers (such as "innit" and "if you know what I mean") in two corpora of London speech, the Corpus of London Teenage Language (COLT) and the Linguistic Innovators Corpus (LIC). Through an analysis using the corpus-based heuristics of frequency and spread (i.e. dispersion among speakers) of pragmatic markers, Torgersen et al. shed light into the locus of language change in the highly complex and multi-ethnic urban center of London.

In sum, corpora and corpus-based methods have an important and still growing place in sociolinguistic research. Yet, the similarities are often approximate, and the connections often still indirect. Again, as Beal *et al.* (2007a, 2007b) pointed out, sociolinguists' corpora are typically "unconventional", not conforming to the models used in crafting major corpora like the BNC and the Corpus of Contemporary American English (COCA). In the next section, I change focus from corpora in sociolinguistic research, to sociolinguistics in corpus-based research to discuss the relationship between these two approaches to language more closely.

## 3. Corpus-based sociolinguistics

There have been several calls for scholars to mine traditional (i.e. "conventional") corpora for sociolinguistic research applications (BENDER, 2002; BAUER, 2004; ANDERSON, 2008; ROMAINE, 2008; BAKER, 2010). Baker's (2010) recent book, *Sociolinguistics and Corpus Linguistics*, points out that the previous literature in these two fields indicates "that some form of 'corpus sociolinguistics' is possible, although it might appear that corpus linguistics has made only a relatively small impact on sociolinguistics" (BAKER, 2010, p. 1). Baker provides many examples of research in support of a "corpus sociolinguistics", such as work on the BNC on sex-related language differences (SCHMID, 2003) and broader social differences (RAYSON; LEECH; HODGES, 1997). Research on sex-based language differences indeed seems an area of sociolinguistics well suited to corpus-based research.[6] One example, not reviewed by Baker, is recent work by Säily (2011; SÄILY; SUOMELA, 2009), who examined the relationship between speaker/ writer sex and morphological productivity in the Corpus of Early English Correspondence (CEEC) and the BNC and demonstrated ways in which the productivity of the *–ity* and *–ness* suffixes were similar or differed for males and females in the historical and present-day corpus data – findings of both sociolinguistic and morphological theoretical interest. A second, and perhaps better-known example (and one which is discussed at length in BAKER, 2010) is Biber's extensive research on genre- and register-based variation (e.g., BIBER; FINEGAN, 1989), which has clearly shown the value of corpora for understanding this important dimension of language variation.

Yet, for the most part (and as Baker points out), these kinds of standard corpora have had limited appeal and limited use by sociolinguists. I believe this is not entirely unexplainable. Many sociolinguists are interested first and foremost in spoken language, which is less available in conventional corpora than written language (cf. NEWMAN, 2008). Further, sociolinguistics is often about *fully* or at least *richly* situating language use in its social and interpersonal contexts and standard corpora, even those that are comprised of spoken language

[6] Arguably, the sex of an author or speaker is the easiest social/identity-related factor to annotate in a corpus, or to reconstruct post factum from corpus data. Note, however, that "gender," a socially constructed identity-related variable, is different from "sex," the biologically-based variable, and more difficult to evaluate without ethnographic inquiry (cf. CHESHIRE, 2004; BUCHOLTZ; HALL, 2006).

recordings, are often too divorced from social contextual information to be of use for in depth sociolinguistic study. Most spoken language corpora which currently do exist – such as many of those available from the Linguistics Data Consortium (LDC; http://ldc.upenn.edu/) – have been designed for speech technology and natural language processing research and simply do not capture the kinds of information that would be necessary for even basic sociolinguistic research (such as, the socio-economic class or ethnicity of the talkers).

It is clear from, for instance, work on the BNC (again, RAYSON *et al.*, 1997; SCHMID, 2003; SÄILY, 2011) that certain questions of a sociolinguistic nature, such as the differences in language use by sex, are fairly pursuable through standard and written corpora, but it is much less clear how one would look at more nuanced sociolinguistic patterns through these kind of corpora. For instance, it is difficult to see how researchers could examine social structures like "communities of practice" (WENGER, 2000; e.g., MALLINSON; CHILDS, 2007), groups built around a coordinated set of interests and activities, through a pre-existing corpus, and it is these sorts of questions which have become of most interest to a large number of sociolinguists in recent years. (I will return to this point in the next section.)

Further, much corpus linguistic work that does see itself as sociolinguistic focuses on lexis (i.e. examines socio-cultural, or, e.g., sex-based, differences through lexical patterns in corpora). Many of the studies reviewed by Baker (2010) are of this kind. For instance, Baker (2010, p. 70-73) gives examples of research on personal titles (such as "Mr." and "Mrs.") and gendered nouns ("boy(s)" and "girl(s)") and their changing use over time in the Brown family of corpora. Rayson *et al.* (1997) examine social differences in lexical frequencies in the BNC. These sorts of research endeavors are clearly sociolinguistic, in the sense that they inform us of changes in social structure and/or changes in the discourse on social structure, and have some similarities to work that is conducted in squarely sociolinguistic circles (e.g., D'ARCY; TAGLIAMONTE, 2010, who examine the complex sociolinguistic factors influencing the realization of relative pronouns in spoken English in Toronto). Nonetheless, many sociolinguists in recent years have focused more on morphosyntactic patterns, which are substantially harder to mine through corpus-based methods, and sociophonetic patterns, which seem even less analyzable through corpus methodologies. (Of course, SÄILY, 2011 examined morphological patterns, illustrating that "corpus sociolinguistic" work does not need to be limited to lexis.)

To some degree the lesser interest in lexis may have to do with sociolinguists' focus on spoken over written language. The primacy of words in corpus linguistics – e.g., the fact that we count a corpus' size in terms of word-count – does not perfectly fit research on conversational spoken language, which is characterized by a high occurrence of disfluencies, grammatical "errors," mispronunciations and the like.[7] It is hard to imagine how the sampling frame for the Brown family of corpora, for instance, 500 samples of about 2,000 words each (extracted from the start of a sentence to the completion of the first full sentence 2,000 words later; FRANCIS; KUčERA, 1979; see also BAKER, 2010, p. 59-68), could be appropriately applied to conversational spoken language.

Finally, it must be observed that many sociolinguists are interested in *specific* language varieties or language situations – such as the language practices of a particular (often small) community or social group – and, as a result, pre-existing corpora containing normative or standard varieties are simply not of great interest. Of course, corpora like the BNC and COCA are always valuable as benchmarks for assessing features in other varieties (and the Brown corpus is still a major source for word frequency information for a diverse range of research) but their use as the actual primary data of analysis is much less common among many sociolinguists.

In closing this section, I would agree that Baker's (2010) "corpus sociolinguistics" indeed appears possible and, I would argue, is being realized by some researchers, although I would also note that the uptake for this kind of work appears to be greater among researchers coming from corpus linguistic perspectives than among those coming from sociolinguistic backgrounds. Sociolinguists have been slower to adopt conventional corpora for research, for the reasons I outlined above, one of which, I revisit more fully now.

## 4. Convergence and divergence

The previous sections of this paper have illustrated, I think, a complex relationship between sociolinguistics and corpus linguistics. On the one hand, these fields have clear similarities. On the other, they also have clear differences.

[7] For instance, in Kendall, Bresnan, and Van Herk (forthcoming), a study of the variable pattern between *give* theme-NP recipient-PP and *give* recipient-NP theme-NP in African American English (discussed again below), we only approximate the word-count of our transcribed spoken data, finding it unrealistic to give the dataset an exact figure.

We observe that in some ways sociolinguistics and corpus linguistics have always been converging at the same time that we observe they have always been diverging in other ways. A consideration of Eckert's (2005, under review) paper, "Three waves of variation study", provides further light on this situation.

In this historiographical assessment of quantitative sociolinguistic work, Eckert classifies the study of sociolinguistic variation into three major categories, or "waves." The first wave is characterized by the study of broad correlational patterns between social features of talkers (and writers) and their use of variable language features. The second wave of study involves ethnography and studying smaller groups of speakers (and writers) to greater depth, focusing on more local patterns of language use. The third wave of study is about practice and agency, rather than social structures. Instead of searching for categories which correlate with language use, research in the third wave focuses more closely on understanding styles and the construction and negotiation of identity(/ies) rather than broad patterns of individual variable features. Eckert points out that these three waves are not necessarily chronologically ordered. Labov's (1963) ground-breaking first study – on Martha's Vineyard, with its deep ethnographic analysis of a small community – is seen as in the second wave, while his second (1966) foundational study – a large-scale survey of English in New York City – is seen as squarely first wave. Yet, despite there not being a direct chronology that corresponds to the three waves, current interest in sociolinguistics is moving increasingly towards third wave-like approaches (see also COUPLAND, 2007). First wave, and to a lesser extent second wave, sociolinguistic research would appear to fit comfortably within a corpus linguistics mold. It is in these "waves" of research that we can draw strong connections between sociolinguistic and corpus linguistic methods and practice where the quantitative large-scale analysis of corpora is most helpful. However, the focus and methodologies of third wave research appear to share less with corpus linguistics (and perhaps have more similarities with Conversation Analysis, cf. LIDDICOAT, 2007, than they do with large-scale corpus-based research).

Eckert provides a nice summary of a major way that third wave work differs from the earlier waves, especially the first wave – the kind of sociolinguistics most implementable through corpus methods.

> The survey method's primary virtues are coverage and replicability, both of which depend on the use of pre-determined social categories and fairly fleeting social contact with the speakers chosen to represent

> those categories. As a result, the social significance of variation can be surmised only on the basis of a general understanding of the categories that serve to select and classify speakers. There is no question that the broad demographic patterns of variation are important. But just as a map of New York City does not tell you what the streets are like, or what it's like to walk on them, the macro-sociological patterns of variation do not reveal what speakers at different places in the socioeconomic hierarchy are doing socially with those variables (Eckert under review: 6).

Further,

> [the] move from the study of structure to the study of practice, giving agency its place in the analysis, has defined the recent history of the social sciences and recent intellectual history more generally [...]. It does not negate the importance of structure, but emphasizes the role of structure in constraining practice and, in turn, the role of practice in producing and reproducing structure. In the study of variation, a focus on practice brings meaning into the foreground, as we try to get at what speakers are doing on the ground. At the same time, it moves us closer to the goal of studying the actual process of change (Eckert under review: 14).

These passages help explain the tension in actualizing a "corpus sociolinguistics." "Coverage and replicability" are two major tenets (and advantages) behind corpus-based work. Yet, it appears to be an impossible task to make replicable and generalizable, especially through corpus-based methods, the ethnographic and instance-specific knowledge a researcher must gain in order to understand the actual creation and negotiation of social meaning "on the ground."

From this, it seems that some sociolinguists will continue to be uninterested in corpora and corpus methods. Nonetheless, there are concrete steps that corpus developers could take to enhance the possibilities of a "corpus sociolinguistics" and to increase the utility of corpora for pursuing sociolinguistic research, and I turn to these now.

## 5. The future of corpora in sociolinguistic research

Technological advancements have been paramount in the development of sociolinguistics. The same is true of course for corpus linguistics. The current research in both approaches would be impossible without modern recording equipment and the ability to store, process, and analyze large amounts of text and audio data through computerized means. While it may be the case that sociolinguistics and corpus linguistics diverge in coming years in their research orientations and methodologies in some ways (as is indicated by Eckert's third wave), it seems likely that continued technological advancements in the development, annotation, and analysis of corpora will lead to increased opportunities for sociolinguistic engagement with corpora. This is especially true for research investigating aspects of language and social structure (i.e. work in the first and second waves), though I believe it is still the case for work that is less interested in large-scale quantitative study. All researchers working with recorded data can benefit from advancements in the treatment of these data.

Some of these advances will occur, I believe, without the need for an explicit "call to arms." Nonetheless, I here explicate two areas where I suggest corpus work could immediately benefit sociolinguistic research, and, conversely, insight from sociolinguistics could enrich broader corpus-based research: The creation of (spoken language) corpora for more diverse language varieties, and the implementation of annotation schemes that are more flexibly connected to data. I consider these in turn.

### 5.1. The need for large, publicly available corpora of more diverse (spoken) language varieties and increased sharing of existing data

One might argue that a primary benefit of corpus-based approaches to linguistic analysis is that the development, publication, and sharing of public corpora allows for the best possible advancement of empirical knowledge about language. By allowing (and, further, promoting) the repeated and repeatable analysis of the same publicly available datasets, corpus linguistics fosters an environment that more fully fits the "scientific method" mode of research than many other areas in linguistics. Scholars can question and refine previous findings by (re-)analyzing the original data; they can extend or modify the annotation schemes and data coding used in previous research; they can compare previously analyzed datasets directly to newly developed datasets; and so on. By working from a shared pool of data, researchers are best able to collectively develop agreed upon knowledge about language. This, I believe, is a major benefit of corpus-based work (which in my opinion has been under-boasted about by corpus linguists).

The vast bulk of sociolinguistic research, even that based on thoroughly balanced and representative linguistic databases, has been conducted on proprietary datasets that are not available for peer review or outside consideration. The common practice in sociolinguistics is for individual (groups of) researchers to develop highly specialized, but closed, databases, which are not made widely available to outsiders. This tendency is not ill intentioned, but rather is the outcome of historical processes in the field. A huge amount of effort, time, and money goes into the collection of sociolinguistic data (and the compilation of any spoken language dataset; NEWMAN, 2008) and within sociolinguistics (as, unfortunately, with many disciplines), academic "credit" has come from the analysis of the data and not its collection or compilation. Researchers traditionally have not wanted to get "scooped" (cf. CHILDS; Van HERK; THORBURN, 2011) on findings after doing the extensive and expensive work of data collection.[8] A second, and perhaps bigger, reason has related to rights management and informant privacy issues, since sociolinguistic fieldwork and interviewing often captures sensitive information that the informants may not want to make public or which fall under contracts with human subjects boards and have restricted access. These issues of anonymity and privacy are complex and difficult to answer when deciding to share fieldwork data (CHILDS *et al.* 2011). Finally, since sociolinguistic datasets have typically been developed in order to research a specific question or set of questions, it has often been assumed that once the original questions have been studied in depth there is not further interest in the datasets themselves. This trend of closed data appears to be changing and it is now the case that more groups of sociolinguistic researchers are making their data available to colleagues and to the public (cf. KENDALL, 2008; CHILDS *et al.* 2011), but it remains the case that sociolinguistics has so far not been able to benefit from the kind of peer review only possible when datasets are widely available for review and re-analysis.[9] This has also, of course, limited the ability of other (i.e. less sociolinguistically oriented) corpus linguists to draw from the vast amount of data collected in recent decades by

[8] As a reviewer aptly pointed out: One hoped for part of a solution would be greater recognition for corpus development work in career advancement, like promotion and tenure, so that corpus developers (sociolinguistic or not) were less incentivized to limit access to their data.

[9] The recent founding of journals like the *Journal of Experimental Linguistics*, <http://elanguage.net/journals/index.php/jel>, which focus on "reproducible research" and the publication of full datasets along with research articles, represents an exciting turn for areas of language research outside of corpus linguistics, most of which, like sociolinguistics, have heretofore, not made a general practice of working from shared data.

sociolinguists. One could imagine there being much richer corpora available, especially "conventional" corpora, if the developers of those corpora could draw on the spoken language data collections of sociolinguists.

To give a specific example, African American English (AAE) has been studied at exceptional length in North American sociolinguistics and has been the subject of a vast body of empirical and quantitative investigations (cf. SCHNEIDER, 1996, p. 3). This research has been driven by numerous exciting questions, from those involving diachrony – such as, how did AAE form in the first place? Is present day AAE the outcome of pidgin/creole forms or of a working class, slave-master variety of British English? Is present day AAE converging with, or diverging from, white varieties or regional varieties of American English? – to more applied sorts of questions about topics like education and social justice – such as, what are the educational positions and responsibilities of school systems towards AAE-speaking children? – and so forth. Many scholars have researched these questions (to list just a few: McDAVID, R.; McDAVID, V., 1951; WOLFRAM, 1969; LABOV, 1972a; DILLARD, 1972; FASOLD, 1972; SMITHERMAN, 1977; 1981; HEATH, 1983; RICKFORD, 1999; POPLACK; TAGLIAMONTE, 2001; WOLFRAM; THOMAS, 2002; CRAIG; WASHINGTON, 2006) and, while consensuses have emerged among sociolinguists in some areas, many questions are still quite actively pursued. However, one could argue that additional progress could be made if scholars had access to a large, shared pool of data against which they could test competing theories or could cite broadly available evidence in order to support or refute particular positions.

While some groups of sociolinguistic researchers have invested in developing thorough transcription and annotation schemes for their data (e.g., POPLACK, 1989; cf. TAGLIAMONTE, 2006), many other sociolinguists do not work with transcribed data, but rather code just the features of interest directly from the audio recordings (e.g., MILROY; GORDON, 2003, though one infers this point through the lack of discussion rather than an explicit statement about transcription). Thus, there are massive amounts of sociolinguistic recordings, which are simply not available in forms that avail themselves to corpus linguistic approaches. The costs of developing complete "corpus-like" data collections can unfortunately be too high, especially when the research questions at hand (often involving particular sociolinguistic variables, cf. WOLFRAM, 1993; MILROY; GORDON, 2003) are more quickly pursued by extracting just the tokens of interest from the audio recording rather than transcribing and annotating everything available.

In recent work investigating the dative alternation in African American English, Kendall, Bresnan, and Van Herk (forthcoming) attempted to take stock of the amount of transcribed sociolinguistic AAE data that was available if one pooled data from across several research groups. All told, we obtained only about a quarter million words of transcribed AAE speech, even though many scholars were extremely generous in making data available to us for analysis. This is not to say that a quarter million words is all that exists, but rather that these data (i.e. accurate transcripts of AAE speech) are scattered throughout the field and not available in any aggregateable form for corpus-based research. It seems clear that doing corpus-based analysis on AAE will require further corpus compilation and creation work.

In sum, countless researchers would be greatly aided by the availability of a large, publicly available corpus of African American English. And this is just one example of a non-standard variety of English. We can readily imagine how many language researchers would benefit from corpora developed for other varieties and varieties of other languages. We need more large-scale publically available corpora of non-standard language varieties.

## 5.2. Connecting "data" to data and the question of "taming"

A second area from which sociolinguistic research could benefit would be a greater focus on the kinds of annotation available in corpora. Corpus linguists – and also documentary linguists, natural language processing researchers, and others for that matter (e.g., BIRD; LIBERMAN, 2001; SIMONS; BIRD; SPANNE, 2008) – have developed extensive annotation frameworks, but often these annotation frameworks have not focused on capturing some of the information that sociolinguists are most interested in, such as a fuller range of social and demographic information about the speakers/writers and audiences in corpora, as well as the full interactional context and setting of the data.[10] In his "ethnography of speaking" approach

[10] In some cases, it would be more accurate to say that it is the entry of the annotation for particular corpora that fail to capture enough information to be widely useful for sociolinguistic queries rather than failings in the annotation *schemes* themselves. The annotation framework for the 4.2 million word demographically sample spoken portion of the BNC, for example, was designed to capture quite a range of demographic features for the speakers – including speaker sex, age group, education level, occupation, social class, and dialect background. For the recruited participants (those who agreed to carry the recording device) the information for many of these

to language, Hymes (e.g., 1974), for instance, proposed the S-P-E-A-K-I-N-G model, which in present day terms could be understood as an annotation framework. Many of the S-P-E-A-K-I-N-G model's components are included in standard corpora (such as information about the "Genre", and often "Participants"), but many are not (such as the "Act sequence" or "Key"). This is not to propose that Hymes' model in particular be adopted by corpus developers, but more simply to highlight some of the kinds of annotation that would further sociolinguistic research possibilities through corpora and, more generally, might lead to richer annotation frameworks than are most often currently used.

Of course, there are huge difficulties in implementing these kinds of ethnographically informed annotation systems in a general way. They are often not readily applicable on a wide-scale, or individual annotation schemes are too bound up with a specific project, or a specific researcher's agenda, to be of use beyond a specific corpus or a specific research project. Even social measures that may seem straightforward at first glance, like socio-economic class or education level, must often be contextualized for the particulars of the group under study (cf. SANKOFF; LABERGE, 1978). For instance, when studying the language use of non-mainstream populations, such as rural African Americans in the U.S. South, mainstream conceptions of socio-economic class or social status simply do not seem to be relevant and local understandings of social structure are necessary for in depth sociolinguistic research (KENDALL; WOLFRAM, 2009). How to best achieve the kind of annotation necessary to make cross-group comparisons in these sorts of situations, or whether such annotation is possible in the first place, is a difficult question to answer.

This question, however, returns us to the quote on the first page of this paper (BEAL *et al.* 2007a, p. 1). In their two edited volumes about "unconventional" corpora, Beal *et al.* (2007a, 2007b) discuss the difficulties of "taming" these unconventional corpora. Poplack, in her foreword to the volumes, explains,

fields is available (since the recruits were solicited based on their region, sex, age group, and social class), but for the talkers with whom the recruits interacted much less information is known. There are also additional problems that must be considered (such as, inaccurate or even false information) when relying on the self-disclosure of social information from recruits such as those in COLT and the BNC (cf. STENSTRÖM; ANDERSEN; HASUND, 2002, p. 18-19).

> Taming, as understood here, is largely a question of representation: How to represent forms for which there is no standard orthography, what to represent, how much to annotate, how much analysis to impose on the materials, how to represent ambiguities and indeterminacies, how to represent the finished product to the end-user (POPLACK, 2007, p. ix-x).

While I find Beal *et al.*'s discussion helpful (and Poplack's foreward particularly insightful), and while the papers in their volumes provide an excellent overview of current work on unconventional corpus development, I am not sure that I like the term "taming" for what needs to happen for less standard language datasets to be usefully developed into corpora. It seems to me that one reason traditional corpora have not been used as extensively for sociolinguistic research is precisely because they have been extensively "tamed," and this "taming" has rendered them less sociolinguistically "real" or useful.[11] A more preferable model might be one which embraces the multi-dimensionality of spoken language data and attempts to maintain the full richness of those dimensions through the corpus development process. (I resist the temptation to label this something like "data left in the wild".)

In Kendall (2008), I proposed a model for considering data within sociolinguistics that attempts to maintain close connections between layers of annotation or metadata. Crucially, this involves being explicit about layers of abstraction (steps away from the original source data) in our annotation and metadata creation processes. Figure 1, from that paper, contrasts what I consider to be a traditional approach to sociolinguistic analysis and data management with an approach that I believe has greater benefits. The basic premise is that sociolinguists are interested in understanding patterns of language in their social contexts, but that all quantitative work (or in fact any work based on records of speech, including audio-only and even video recordings, since recordings never capture the entirety of a real-world event)

[11] For example, the BNC's demographically sampled spoken component was built following social survey research practices (BURNARD, 2007; see also RAYSON *et al.*, 1997) and, at face value, appears to be quite similar to the sort of large-scale dialect survey sociolinguists might undertake. Rayson *et al.* (1997), in their examination of social factors in differences in lexical frequency in the spoken component, note however that work on the social differentiation of language in this part of the corpus is limited by the simplified transcription system. One might argue that it is primarily the extent of its "taming" that makes this part of the BNC less sociolinguistically useful than it otherwise would be.

involves abstractions away from the true, contextualized language data, the actual real-world speech event. In the "traditional" model, layers upon layers of annotation are developed, many of which increase the distance between the "data" (in quotes, indicating some level of abstraction from the actual or ideal data) and the real-world speech events that are ultimately the objects of interest, the true data (no quotes).

For example, if I am interested in studying variable realizations of the English past tense (like unmarking or non-standard past tense marking), I might audio record a speech event and from that recording develop a transcript, which, for sake of the example, we will assume accurately captures the variable realizations of the English past tense morpheme. I then extract the frequencies of the various realizations of the past tense morpheme along with other contextual information and then compile this as a spreadsheet, which I add to compiled data from other speakers and other speech events. In the end, I have a data file ready for quantitative analysis, but I have also moved several steps away from the original speech event. My language data has become a spreadsheet of frequencies or data tokens with very little available matrix talk, perhaps a concordance-like "keyword in context" amount of surrounding context. It is no longer quite "language," having been separated from its full communicative context. This likely does not matter as far as the success of my quantitative analysis goes, but the closer examination of individual tokens has become difficult, as has my ability to question the original coding of the morphemes.

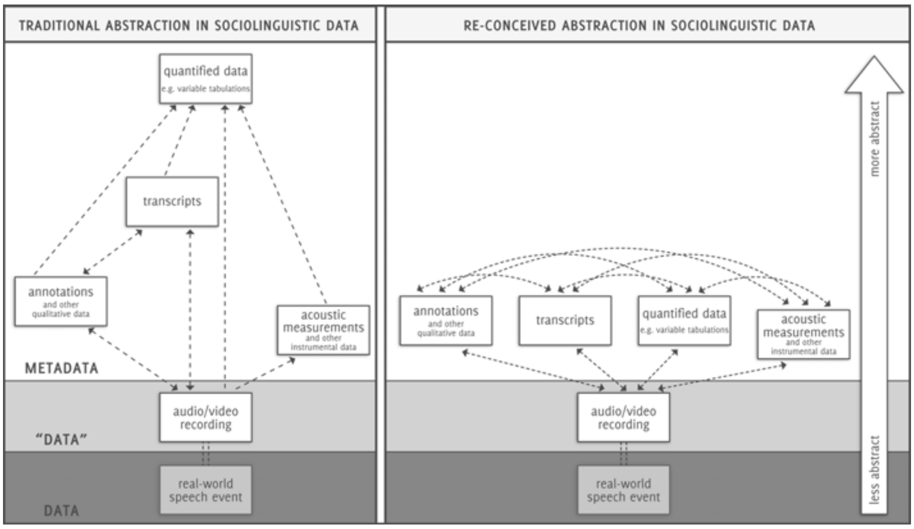
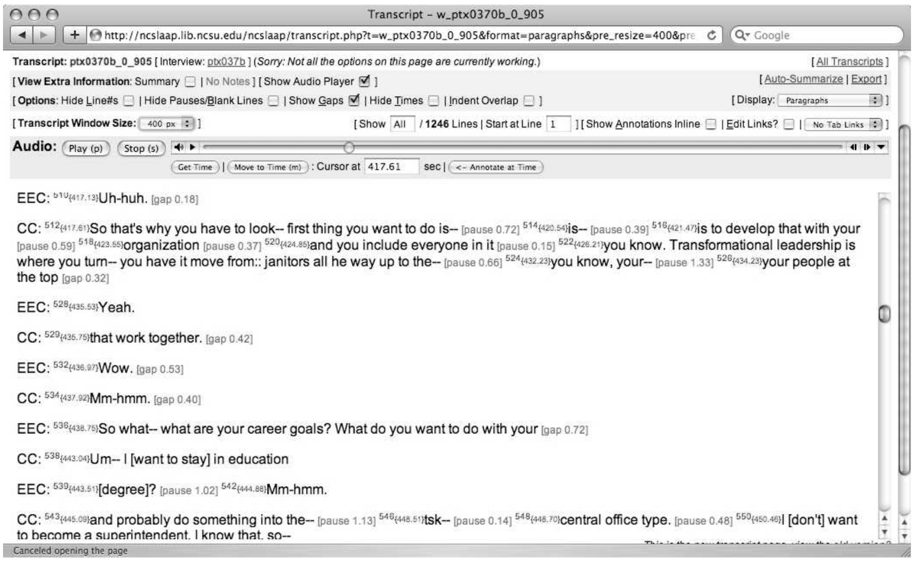

TRADITIONAL ABSTRACTION IN SOCIOLINGUISTIC DATA
quantified data
e.g. variable tabulations
transcripts
annotations
and other qualitative data
acoustic measurements
and other instrumental data
METADATA
audio/video recording
"DATA"
real-world speech event
DATA
RE-CONCEIVED ABSTRACTION IN SOCIOLINGUISTIC DATA
annotations
and other qualitative data
transcripts
quantified data
e.g. variable tabulations
acoustic measurements
and other instrumental data
audio/video recording
real-world speech event
more abstract
less abstract

FIGURE 1 – Layers of abstraction in sociolinguistic data (from KENDALL, 2008, p. 346, Figure 5)

The "re-conceived" model of Figure 1 focuses primarily on maintaining linkages between levels and types of annotation. As in the hypothetical example discussed for the traditional model, I may wish to transcribe the recording and then to extract quantitative data from that. However, here the emphasis would be on maintaining links between each of these layers of data with the other layers. This is achieved through a focus on accurate time-stamping and the development and use of software built for time-aligned linguistic (or at least audio) annotation. Returning to Hymes – who wrote "the most common, the most serious, defect in most reports of speaking probably is that the message form, and, hence, the rules governing it, cannot be recaptured" (1974, p. 54) – we can observe that, while an accurate transcript may capture the lexical and syntactic form of an utterance, no transcript or text-based annotation can be expected to accurately encapsulate its full form, such as its prosody, the nuanced particulars of the speaker's voice, and so on.

The Sociolinguistic Archive and Analysis Project (SLAAP; <http://ncslaap.lib.ncsu.edu/>; cf. KENDALL, 2007, 2008) and the Online Speech/Corpora Archive and Analysis Resource (OSCAAR; <http://oscaar.ling.northwestern.edu>; cf. KENDALL, 2010) are two examples of ways that one might approach implementing this sort of model. Both of these projects feature a time-aligned transcription model which is dynamically linked to the underlying audio recordings and to any additional researcher notes or quantitative data. For example, Figure 2 displays one view of SLAAP's transcript feature for a stored recording. In addition to the transcript text, the user has direct access to the recording audio, as well as to fine-grained information about where silences occur and their lengths. Users can also get "close up" views of individual transcript lines, as in Figure 3, which displays the text of a line along with the audio itself, as well as a spectrogram and pitch track for the utterance (created dynamically from the audio). Users can extract phonetic information directly from this view (only pitch data is illustrated in Figure 3).



Transcript: ptx0370b_0_905 [ Interview: ptx037b ] (Sorry: Not all the options on this page are currently working.) [ All Transcripts ]

[ View Extra Information: Summary □ | No Notes ] [ Show Audio Player ☑ ] [ Auto-Summarize | Export ]

[ Options: Hide Line#s □ | Hide Pauses/Blank Lines □ | Show Gaps ☑ | Hide Times □ | Indent Overlap □ ] [ Display: Paragraphs ]

[ Transcript Window Size: 400 px ] [ Show All / 1246 Lines | Start at Line 1 ] [ Show Annotations Inline □ | Edit Links? □ | No Tab Links ]

Audio: Cursor at 417.61 sec

EEC: 510{417.13}Uh-huh. [gap 0.18]

CC: 512{417.61}So that's why you have to look-- first thing you want to do is-- [pause 0.72] 514{420.54}is-- [pause 0.39] 516{421.47}is to develop that with your [pause 0.59] 518{423.55}organization [pause 0.37] 520{424.85}and you include everyone in it [pause 0.15] 522{426.21}you know. Transformational leadership is where you turn-- you have it move from:: janitors all he way up to the-- [pause 0.66] 524{432.23}you know, your-- [pause 1.33] 526{434.23}your people at the top [gap 0.32]

EEC: 528{435.53}Yeah.

CC: 529{435.75}that work together. [gap 0.42]

EEC: 532{436.97}Wow. [gap 0.53]

CC: 534{437.92}Mm-hmm. [gap 0.40]

EEC: 536{438.75}So what-- what are your career goals? What do you want to do with your [gap 0.72]

CC: 538{443.04}Um-- I [want to stay] in education

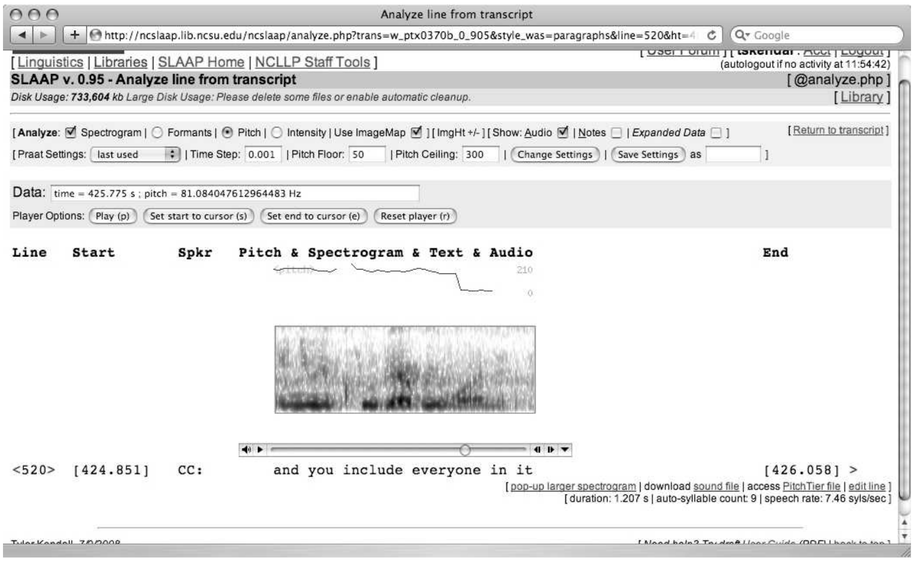
EEC: 539{443.51}[degree]? [pause 1.02] 542{444.88}Mm-hmm.

CC: 543{445.09}and probably do something into the-- [pause 1.13] 546{448.51}tsk-- [pause 0.14] 548{448.70}central office type. [pause 0.48] 550{450.46}I [don't] want to become a superintendent. I know that, so--

Canceled opening the page

Transcript - w_ptx0370b_0_905
http://ncslaap.lib.ncsu.edu/ncslaap/transcript.php?t=w_ptx0370b_0_905&format=paragraphs&pre_resize=400&pre
Google
Play (p)
Stop (s)
Get Time
Move to Time (m)
<- Annotate at Time

FIGURE 2 – A transcript view in SLAAP (a sociolinguistic interview with "CC", a Mexican American female in Southern Texas; "EEC" is the interviewer)



[ Linguistics | Libraries | SLAAP Home | NCLLP Staff Tools ] (autologout if no activity at 11:54:42)

SLAAP v. 0.95 - Analyze line from transcript [ @analyze.php ]

Disk Usage: 733,604 kb Large Disk Usage: Please delete some files or enable automatic cleanup. [ Library ]

[ Analyze: ☑ Spectrogram | ○ Formants | ● Pitch | ○ Intensity | Use ImageMap ☑ ] [ ImgHt +/- ] [ Show: Audio ☑ | Notes □ | Expanded Data □ ] [ Return to transcript ]

[ Praat Settings: last used | Time Step: 0.001 | Pitch Floor: 50 | Pitch Ceiling: 300 | Change Settings | Save Settings as ]

Data: time = 425.775 s ; pitch = 81.084047612964483 Hz

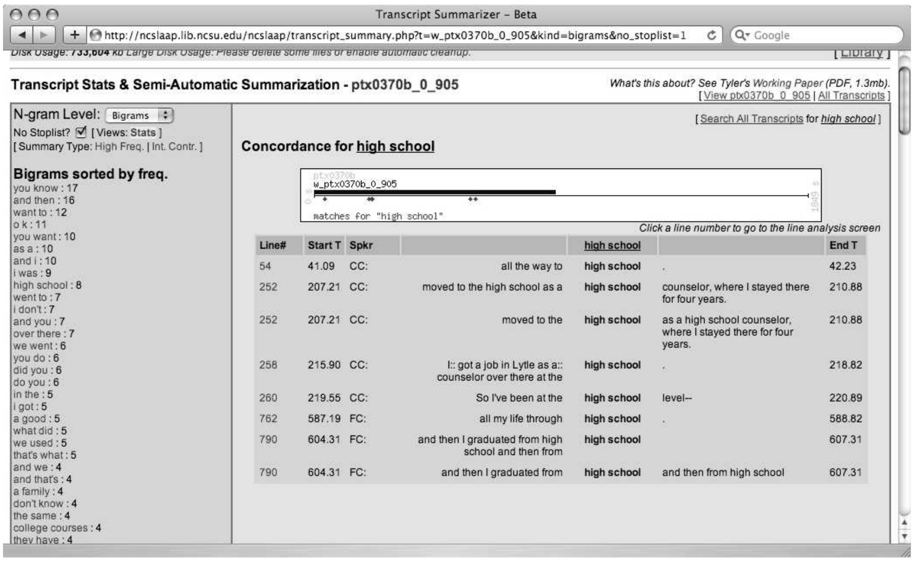
```
Line    Start        Spkr   Pitch & Spectrogram & Text & Audio          End

<520>   [424.851]    CC:       and you include everyone in it           [426.058] >
```

[ pop-up larger spectrogram | download sound file | access PitchTier file | edit line ]
[ duration: 1.207 s | auto-syllable count: 9 | speech rate: 7.46 syls/sec ]

Analyze line from transcript
http://ncslaap.lib.ncsu.edu/ncslaap/analyze.php?trans=w_ptx0370b_0_905&style_was=paragraphs&line=520&ht=4
Google
Player Options: Play (p)
Set start to cursor (s)
Set end to cursor (e)
Reset player (r)
<pitch>
210
0

FIGURE 3 – SLAAP's "close up" of the line "and you include everyone in it" from Figure 2

Transcripts in SLAAP are dynamic entities and can be reformatted in numerous ways, from textual representations, like the columnar format suggested by Ochs (1979), to various graphical formats (screenshots of these other transcript views and a fuller discussion of SLAAP's transcript model are available in KENDALL, 2007, 2008). Traditional corpus analysis features are available, such as in Figure 4, which displays the highest frequency bigrams (on the left) and a sample concordance (on the right; for the phrase "high school") from the same transcript shown in Figures 2 and 3. Since all of the utterances are time-stamped, SLAAP is able to show a graphical timeline (at the top-right) indicating where each of the concordance lines occurs in the recordings (the single line that extends the length of the timeline image represents the temporal duration of the full recording; the filled bar that extends roughly across the left-half of the line represents the transcribed portion of the recording; the dots below the lines show when the concordance lines occur in time). While SLAAP's maintenance of the linkage between audio and text is a centerpiece of the archive, transcripts can also be exported as plain text (or as Praat TextGrids; BOERSMA; WEENINK, 2010) and then manipulated via standard corpus or text analysis tools.


Transcript Summarizer – Beta
http://ncslaap.lib.ncsu.edu/ncslaap/transcript_summary.php?t=w_ptx0370b_0_905&kind=bigrams&no_stoplist=1
Google
ptx0370b
w_ptx0370b_0_905
0
1849
matches for "high school"

Disk Usage: 733,604 kb Large Disk Usage: Please delete some files or enable automatic cleanup. [ Library ]

Transcript Stats & Semi-Automatic Summarization - ptx0370b_0_905

What's this about? See Tyler's Working Paper (PDF, 1.3mb).
[ View ptx0370b_0_905 | All Transcripts ]

N-gram Level: Bigrams
No Stoplist? ☑ [ Views: Stats ]
[ Summary Type: High Freq. | Int. Contr. ]

Bigrams sorted by freq.

| | |
|---|---|
| you know | 17 |
| and then | 16 |
| want to | 12 |
| o k | 11 |
| you want | 10 |
| as a | 10 |
| and i | 10 |
| i was | 9 |
| high school | 8 |
| went to | 7 |
| i don't | 7 |
| and you | 7 |
| over there | 7 |
| we went | 6 |
| you do | 6 |
| did you | 6 |
| do you | 6 |
| in the | 5 |
| i got | 5 |
| a good | 5 |
| what did | 5 |
| we used | 5 |
| that's what | 5 |
| and we | 4 |
| and that's | 4 |
| a family | 4 |
| don't know | 4 |
| the same | 4 |
| college courses | 4 |
| they have | 4 |

[ Search All Transcripts for *high school* ]

Concordance for high school

Click a line number to go to the line analysis screen

| Line# | Start T | Spkr | | high school | | End T |
|---|---|---|---|---|---|---|
| 54 | 41.09 | CC: | all the way to | **high school** | . | 42.23 |
| 252 | 207.21 | CC: | moved to the high school as a | **high school** | counselor, where I stayed there for four years. | 210.88 |
| 252 | 207.21 | CC: | moved to the | **high school** | as a high school counselor, where I stayed there for four years. | 210.88 |
| 258 | 215.90 | CC: | I:: got a job in Lytle as a:: counselor over there at the | **high school** | . | 218.82 |
| 260 | 219.55 | CC: | So I've been at the | **high school** | level-- | 220.89 |
| 762 | 587.19 | FC: | all my life through | **high school** | . | 588.82 |
| 790 | 604.31 | FC: | and then I graduated from high school and then from | **high school** | | 607.31 |
| 790 | 604.31 | FC: | and then I graduated from | **high school** | and then from high school | 607.31 |

FIGURE 4 – A concordance view from SLAAP for "high school"
in the transcript shown in Figures 2-3

The connection between the audio recording and the transcript (and other annotation layers) is not the only step available towards spoken language corpora that fit the "re-conceived" model of Figure 1, but it is, I believe, a large step towards improved spoken language data. Further, SLAAP's transcript implementation has been shown here only as *one* demonstration of a way that this can be accomplished[12] and SLAAP, itself, is meant only as one possible example. The TalkBank website (<http://www.talkbank.org/>; MacWHINNEY, 2007) provides another excellent example of the advantages of time-aligned annotation linked to audio (and video) via specialized software, as does the Origins of New Zealand English (ONZE) project (<http://www.lacl.canterbury.ac.nz/onze/>; GORDON; MACLAGAN; HAY, 2007) and the growing list of projects using the ONZE Miner software (recently renamed as LaBB-CAT; <http://onzeminer.sourceforge.net/>; FROMONT; HAY, 2008). Finally, the Annotation Graph Toolkit (<http://agtk.sourceforge.net/>; BIRD; LIBERMAN, 2001) provides a formal framework for the development of these kinds of interfaces to data. Such systems, by basing the annotation on the temporal record of the recording, allow for multiple versions of annotation (and multiple versions even of transcription) and give the end-users, the analysts, the ability to customize their interfaces with the data.

The "re-conceived" model of abstraction for (socio)linguistic data in Figure 1 is perhaps less a proposal for the future than it is a way to think about and steer the changes that are occurring in the ways that audio-based spoken language recordings are manageable and increasingly managed. By focusing on building flexible annotation systems that maintain links through various levels of annotation and, most importantly, to the source recording, we can build corpora, which, instead of needing to be "tamed", can be utilized in a richer variety of ways than currently possible. I believe these sorts of models present the best opportunities for fruitful future work at the interface of corpus linguistics and sociolinguistics. They also would yield more flexible spoken language corpora for a range of applications beyond sociolinguistics.

[12] Other features in SLAAP also seek to minimize the separation of annotation from the source recording. For example, in addition to its transcript features, SLAAP has tools developed specifically for variationist sociolinguistic analysis that also follow a similar time-stamped and linked model. Analysts can extract and code variables (cf. WOLFRAM, 1993; TAGLIAMONTE, 2006) directly from the audio player or from the transcript views. These variable codes are stored along with their time-stamps and users can later return directly to the moment in the audio associated to each extracted variable at the click of the mouse. See Kendall (2007, 2008) for more on SLAAP's variable analysis features.

## 6. Conclusion

In this paper, I have outlined some areas where corpus linguistics and sociolinguistics have strong existing connections and some areas where these connections are less strong.[13] I have also discussed some wished for items for the future – namely, a broader range of "unconventional" corpora, which document a diverse range of language varieties, and an orientation to corpus-based data that maintains its connection to its context (and audio or video recording) and minimizes the amount of abstraction away from the actual source speech (or writing). These are advancements that I believe would greatly aid sociolinguistic research, as well as non-sociolinguistically oriented corpus-based research, and would build stronger bridges between sociolinguists and corpus linguists. The bulk of this paper has approached the relationship between sociolinguistics and corpus linguistics primarily from the perspective of sociolinguistics and, as such, has largely framed its discussion in terms of

[13] Further, I have focused in this paper on corpora, i.e. data, rather than other areas of intersection among corpus linguistic and sociolinguistic methods and practice. However, much could also be said about these other areas of overlap. For instance, both approaches involve extensive use of quantitative methods, although these exact methods differ in significant (but sometimes subtle) ways. Traditional corpus linguistic quantitative methods, in their focus on (normalized) frequencies of occurrence, can fail to account for what is *not* in a corpus. As D'Arcy (2005, in preparation) indicates through an analysis of discourse particle "like", corpus linguists might benefit from greater attention to variationist sociolinguistic quantitative methods (e.g., variable analysis and its principal of accountability; cf. LABOV, 1972b; TAGLIAMONTE, 2006), which attend not only to how many times the form of interest was realized by language users, but also to what *else* was realized in the places where that form was a relevant option. Meanwhile, much can also be said about Variable Rule Analysis (Varbrul), which for over three decades has been the dominant statistical technique in the sociolinguistic literature. Varbrul, a specialized form of logistic regression developed specifically for sociolinguistic variable analysis (CEDERGREN; SANKOFF, 1974) was a huge advancement over other available techniques for multivariate analysis when it was first developed, but in recent years, an array of powerful statistical techniques have been developed in corpus linguistics and other areas of language research (cf. BAAYEN, 2008; JOHNSON, K., 2008; GRIES, 2009; JOHNSON, D. E., 2009) that are, oftentimes, relevant and more appropriate for sociolinguistic analysis than Varbrul. This has been a point of contention among some language researchers in recent years, but, I believe, sociolinguists are rapidly incorporating these available techniques (see in particular JOHNSON, K., 2008, p. 174-180 and JOHNSON, D. E., 2009) and that this is becoming an area of fruitful, cooperative methodological advancement.

what corpus linguistics "can do" for sociolinguistic research. Yet, these suggestions have important ramifications on corpus linguistics more generally and I hope these ramifications are clear to readers: The development of more spoken language corpora, from a range of varieties and with more flexible annotation, will benefit corpus linguistic research widely.

As my discussion of Eckert's "three waves" account of the development of (variationist) sociolinguistic research indicates, it will likely be the case that much important sociolinguistic work remains heavily engaged in and devoted to a kind of analysis that is likely impossible through the use of corpora. Although, at the same time, as Baker (2010) points out, tools from corpus linguistics can still be used for examining transcribed data, regardless of the overall direction the research or data takes (provided it is transcribed, of course). Software-based archives, like that demonstrated by SLAAP above, can help bring corpus-based methods and a more explicit focus on data to sociolinguistic research, even that which is not interested in large-scale analysis.

I would like to end by posing the question: What can corpus linguists do *now* to best advance sociolinguistic research and to best promote the use of corpora and corpus methodologies in sociolinguistics? There are clearly several answers to this question and while others may respond differently, my own wish would be that corpus linguists (especially those who have extensive experience in corpus development) work directly with sociolinguists (especially those who focus on field-based research and ethnography) to develop sociolinguistically rich, "unconventional" corpora, to make those corpora publically available to researchers, and to work towards developing best-practices for the corpus-like treatment of *sociolinguistic* (spoken language) data. As I have argued elsewhere (KENDALL, 2008), sociolinguistic data and data management practices could greatly benefit from the knowledge and expertise of corpus linguists and language documentarians. Luckily, with the growth of projects like ONZE and LANCHART, and corpora like COLT and the LIC, I believe that we are on our way towards achieving this needed collaboration.

## Acknowledgments

I thank the editors and the anonymous reviewers for excellent and helpful comments on an earlier draft of this paper. I also thank Gerard Van Herk, for many conversations relating to these topics, and Charlotte Vaughn, who helped with the original conception and production of Figure 1. Any errors, of course, remain my own.

## References

ALLEN, W.; BEAL, J.; CORRIGAN, K.; MAGUIRE, W.; MOISL, H. A linguistic "time-capsule": The Newcastle Electronic Corpus of Tyneside English. In: BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007b. V. 2: Diachronic Databases.

ANDERSON, W. Corpus linguistics in the UK: Resources for sociolinguistic research. *Language and Linguistics Compass*, v. 2, n. 2, p. 352-371, 2008.

BAAYEN, R. H. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press, 2008.

BAKER, P. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press, 2010.

BAUER, L. Inferring variation and change from public corpora. In: CHAMBERS, J. K.; TRUDGILL, P.; SCHILLING-ESTES, N. (Ed.). *The Handbook of Language Variation and Change*. Malden, MA / Oxford: Blackwell, 2004.

BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007a. V. 1: Synchronic Databases.

BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007b. V. 2: Diachronic Databases.

BENDER, E. Corpus methods for sociolinguistics. Workshop at New Ways of Analyzing Variation (NWAV) 31. Palo Alto, CA: Stanford University, 2002. Available at: <http://faculty.washington.edu/ebender/corpora_sociolx.html>. Retrieved: April 1, 2011.

BIBER, D.; FINEGAN, E. Drift and the evolution of English style: A history of three genres. *Language*, v. 65, n. 3, p. 487-517, 1989.

BIRD, S.; LIBERMAN, M. A formal framework for linguistic annotation. *Speech Communication*, v. 33, n. 1-2, p. 23-60, 2001.

BOERSMA, P.; WEENINK, D. Praat: Doing phonetics by computer. 2010. Software. Available at: <http://www.fon.hum.uva.nl/praat/>. Retrieved: September 18, 2010.

BUCHOLTZ, M.; HALL, K. Gender, sexuality and language. In: BROWN, K. (Ed.). *Encyclopedia of Language and Linguistics*. 2. ed. Oxford: Elsevier, 2006. V. 4.

BURNARD, L. (Ed.). *Reference guide for the British National Corpus (XML edition)*. Published for the British National Corpus Consortium by Oxford University Computing Services, 2007. Available at: <http://www.natcorp.ox.ac.uk/docs/URG/>. Retrieved: April 1, 2011.

CEDERGREN, H.; SANKOFF, D. Variable rules: Performance as a statistical reflection of competence. *Language,* v. 50, n. 2, p. 333-355, 1974.

CHESHIRE, J. Sex and gender in variationist research. In: CHAMBERS, J. K.; TRUDGILL, P.; SCHILLING-ESTES, N. (Ed.). *The Handbook of Language Variation and Change*. Malden, MA / Oxford: Blackwell, 2004.

CHILDS, B.; VAN HERK, G.; THORBURN, J. Safe harbour: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory*, v. 7, n. 1, p. 163-180, 2011.

COUPLAND, N. *Style*: Language variation and Identity. Cambridge: Cambridge University Press, 2007.

CRAIG, H.; WASHINGTON, J. *Malik Goes to School:* Examining the Language Skills of African American Students from Preschool to 5th Grade. Mahwah, NJ: Lawrence Erlbaum Associates, 2006.

D'ARCY, A. Tracking the development of discourse 'like' in contemporary (Canadian) English. Doctoral Thesis Proposal. University of Toronto, Toronto, Canada, March 16, 2005.

D'ARCY, A. Counting matters: Normalization and accountability. In preparation.

D'ARCY, A.; TAGLIAMONTE, S. Prestige, accommodation, and the legacy of relative *who*. *Language in Society*, v. 39, n. 3, p. 383-410, 2010.

DILLARD, J. L. *Black English:* Its History and Usage in the United States. New York: Random House, 1972.

ECKERT, P. Variation, convention, and social meaning. Paper presented at the 2005 Annual Meeting of the Linguistic Society of America, Oakland, CA, 2005.

ECKERT, P. Three waves of variation study: The emergence of meaning in the study variation. Under review. Available at: <http://www.stanford.edu/~eckert/PDF/ThreeWavesofVariation.pdf>. Retrieved: September 7, 2010.

FASOLD, R. *Tense Marking in Black English*: A Linguistic and Social Analysis. Washington, DC: Center for Applied Linguistics, 1972.

FRANCIS, W. N.; KUČERA, H. Brown Corpus manual. Revised and amplified, 1979. Available at: <http://icame.uib.no/brown/bcm.html>. Retrieved: September 7, 2010.

FROMONT, R.; HAY, J. ONZE Miner: The development of a browser-based research tool. *Corpora,* v. 3, p. 173-193, 2008.

GORDON, E.; MACLAGAN, M.; HAY, J. The ONZE Corpus. In: BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora.* New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007b.

GREGERSEN, F. The data and design of the LANCHART study. *Acta Linguistica Hafniensia*, v. 41, p. 3-29, 2009.

GRIES, St. Th. Exploring variability within and between corpora: Some methodological considerations. *Corpora,* v. 1, n. 2, p. 109-151, 2006.

GRIES, St. Th. *Statistics for Linguistics with R:* A Practical Introduction. Berlin: De Gruyter Mouton, 2009.

HEATH, S. B. *Ways with Words*: Language, Life, and Work in Communities and Classrooms. New York: Cambridge University Press, 1983.

HYMES, D. *Foundations in Sociolinguistics*: An Ethnographic Approach. Philadelphia: University of Pennsylvania Press, 1974.

JOHNSON, D. E. Getting off the Goldvarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass,* v. 3, n. 1, p. 359-383, 2009.

JOHNSON, K. *Quantitative Methods in Linguistics*. Malden, MA / Oxford: Blackwell, 2008.

KENDALL, T. Enhancing sociolinguistic data collections: The North Carolina Sociolinguistic Archive and Analysis Project. *University of Pennsylvania Working Papers in Linguistics*, v. 13, n. 2, p. 15-26, 2007.

KENDALL, T. On the history and future of sociolinguistic data. *Language and Linguistics Compass*, v. 2, n. 2, p. 332-351, 2008.

KENDALL, T. Developing web interfaces to spoken language data collections. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, v. 1, n. 2, Chicago: University of Chicago, 2010.

KENDALL, T.; BRESNAN, J.; VAN HERK, G. The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory*, forthcoming.

KENDALL, T.; VAN HERK, G. (Ed.). Corpus linguistics and sociolinguistic inquiry. *Corpus Linguistics and Linguistic Theory*, Special issue, v. 7, n. 1, 2011.

KENDALL, T.; WOLFRAM, W. Local and external standards in African American English. *Journal of English Linguistics*, v. 37, n. 4, p. 305-330, 2009.

KRETZSCHMAR, W. Jr. *The Linguistics of Speech*. Cambridge: Cambridge University Press, 2009.

KRETZSCHMAR, W. JR.; ANDERSON, J.; BEAL, J.; CORRIGAN, K.; OPAS-HÄNNINEN, L. L.; PLICHTA, B. Collaboration on Corpora for Regional and Social Analysis. *Journal of English Linguistics*, v. 34, n. 3, p. 172-205, 2006.

LABOV, W. The social motivation of a sound change. *Word*, v. 19, p. 273-309, 1963.

LABOV, W. *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics, 1966.

LABOV, W. *Language in the Inner City:* Studies in the Black English Vernacular. Philadelphia: University of Pennsylvania Press, 1972a.

LABOV, W. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press, 1972b.

LIDDICOAT, A. *An Introduction to Conversation Analysis*. London / New York: Continuum, 2007.

MACWHINNEY, B. The TalkBank Project. In: BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007a.

MALLINSON, C.; CHILDS, B. Communities of practice in sociolinguistic description: Analyzing language and identity practices among Black women in Appalachia. *Gender and Language*, v. 1, n. 2, p. 173-206, 2007.

MCDAVID, R.; MCDAVID, V. The relationship of the speech of American Negroes to the speech of whites. *American Speech*, v. 26, n. 1, p. 3-17, 1951.

MCENERY, T.; WILSON, A. *Corpus Linguistics*. 2. ed. Edinburgh: Edinburgh University Press, 2001.

MCENERY, T.; XIAO, R.; TONO, Y. *Corpus-based Language Studies*: An Advanced Resource Book. New York / London: Routledge, 2006.

MILROY, L.; GORDON, M. *Sociolinguistics*: Methods and Interpretation. Malden, MA / Oxford: Blackwell, 2003.

NEWMAN, J. Spoken corpora: Rationale and application. *Taiwan Journal of Linguistics*, v. 6, n. 2, p. 27-58, 2008.

OCHS, E. Transcription as theory. In: OCHS, E.; SCHIEFFELIN, B. (Ed.). *Developmental Pragmatics*. New York: Academic Press, 1979.

POPLACK, S. The care and handling of a mega-corpus: The Ottawa-Hull French Project. In: FASOLD, R.; SCHIFFRIN, D. (Ed.). *Language Change and Variation*. Amsterdam / Philadelphia: John Benjamins, 1989.

POPLACK, S. Foreward. In: BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007a.

POPLACK, S.; TAGLIAMONTE, S. *African American English in the Diaspora*. Malden, MA / Oxford: Blackwell, 2001.

RAYSON, P.; LEECH, G.; HODGES, M. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, v. 2, n. 1, p. 133-152, 1997.

RICKFORD, J. R. *African American Vernacular English*: Features, Evolution, Educational Implications. Malden, MA / Oxford: Blackwell, 1999.

ROMAINE, S. Corpus linguistics and sociolinguistics. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus Linguistics*: An International Handbook. Berlin / New York: Mouton de Gruyter, 2008.

SÄILY, T. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, v. 7, n. 1, p. 119-141, 2011.

SÄILY, T.; SUOMELA, J. Comparing type counts: The case of women, men and *-ity* in early English letters. In: RENOUF, A.; KEHOE, A. (Ed.). *Corpus linguistics*: Refinements and reassessments. Amsterdam: Rodopi, 2009.

SANKOFF, D.; LABERGE, S. The linguistic market and the statistical explanation of variability. In: SANKOFF, D. (Ed.). *Linguistic Variation*: Models and Methods. New York: Academic Press, 1978.

SANKOFF, D.; SANKOFF, G. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In: DARNELL, R. (Ed.). *Canadian Languages in their Social Context*. Edmonton, Alberta: Linguistic Research, 1973.

SCHMID, H-J. Do men and women really live in different cultures? Evidence from the BNC. In: WILSON, A.; RAYSON, P.; MCENERY, T. (Ed.). *Corpus Linguistics by the Lune*: A Festschrift for Geoffrey Leech. Frankfurt: Peter Lang, 2003.

SCHNEIDER, E. *Focus on the USA*. Amsterdam / Philadelphia: John Benjamins, 1996.

SHUY, R.; WOLFRAM, W.; RILEY, W. *Field Techniques in an Urban Language Study*. Washington, D.C.: Center for Applied Linguistics, 1968.

SIMONS, G.; BIRD, S.; SPANNE, J. (Ed.). Best practice recommendations for language resource description. Open Language Archives Community document, 2008. Available at: <http://www.language-archives.org/REC/bpr-20080711.html>. Retrieved: September 18, 2010.

SMITHERMAN, G. *Talkin and Testifying*: The Language of Black America. Boston: Houghton Mifflin, 1977.

SMITHERMAN, G. (Ed.). *Black English and the Education of Black Children and Youth*: Proceedings of the National Symposium on the King Decision. Detroit, MI: Center for Black Studies, Wayne State University, 1981.

STENSTRÖM, A-B.; ANDERSEN, G.; HASUND, I. K. *Trends in Teenage Talk*: Corpus Compilation, Analysis, and Findings. Amsterdam / Philadelphia: John Benjamins, 2002.

TAGLIAMONTE, S. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press, 2006.

TORGERSEN, E. N.; GABRIELATOS, C.; HOFFMANN, S.; FOX, S. A corpus-based study of pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory*, v. 7, n. 1, p. 93-118, 2011.

WEINREICH, U.; LABOV, W. HERZOG, M. Empirical foundations for a theory of language change. In: LEHMANN, W. P.; MALKIEL, Y. (Ed.). *Directions for Historical Linguistics*. Austin, TX: University of Texas Press, 1968.

WENGER, E. *Communities of Practice*: Learning, Meaning, and Identity. Cambridge: Cambridge University Press, 1998.

WOLFRAM, W. *A Sociolinguistic Description of Detroit Negro Speech*. Washington, DC: Center for Applied Linguistics, 1969.

WOLFRAM, W. Identifying and interpreting variables. In: PRESTON, D. (Ed.). *American Dialect Research*. Amsterdam / Philadelphia: John Benjamins, 1993.

WOLFRAM, W.; THOMAS, E. R. *The Development of African American English*. Malden, MA / Oxford: Blackwell, 2002.

Recebido em 09/09/2010. Aprovado em 08/05/2011.

# The future of multimodal corpora

## *O futuro dos corpora modais*

Dawn Knight*
The University of Nottingham
Nottingham / UK

ABSTRACT: This paper takes stock of the current state-of-the-art in multimodal corpus linguistics, and proposes some projections of future developments in this field. It provides a critical overview of key multimodal corpora that have been constructed over the past decade and presents a wish-list of future technological and methodological advancements that may help to increase the availability, utility and functionality of such corpora for linguistic research.

KEYWORDS: Multimodal corpus linguistics; resources; software; availability; usability.

RESUMO: Este artigo apresenta um balanço do estado da arte da linguística de corpus multimodal e propõe a projeção de desenvolvimentos futuros nessa área. Um resumo crítico dos corpora multimodais-chave que foram construídos na última década é apresentado, assim como uma lista de desenvolvimentos tecnológicos e metodológicos futuros que podem auxiliar na disponibilização e utilização, bem como na funcionalidade, de tais corpora para a pesquisa linguística.

PALAVRAS-CHAVE: Linguística de corpus multimodal; recursos; programas computacionais; disponibilidade; usabilidade.

* dawn.knight@nottingham.ac.uk

## 1. Introduction

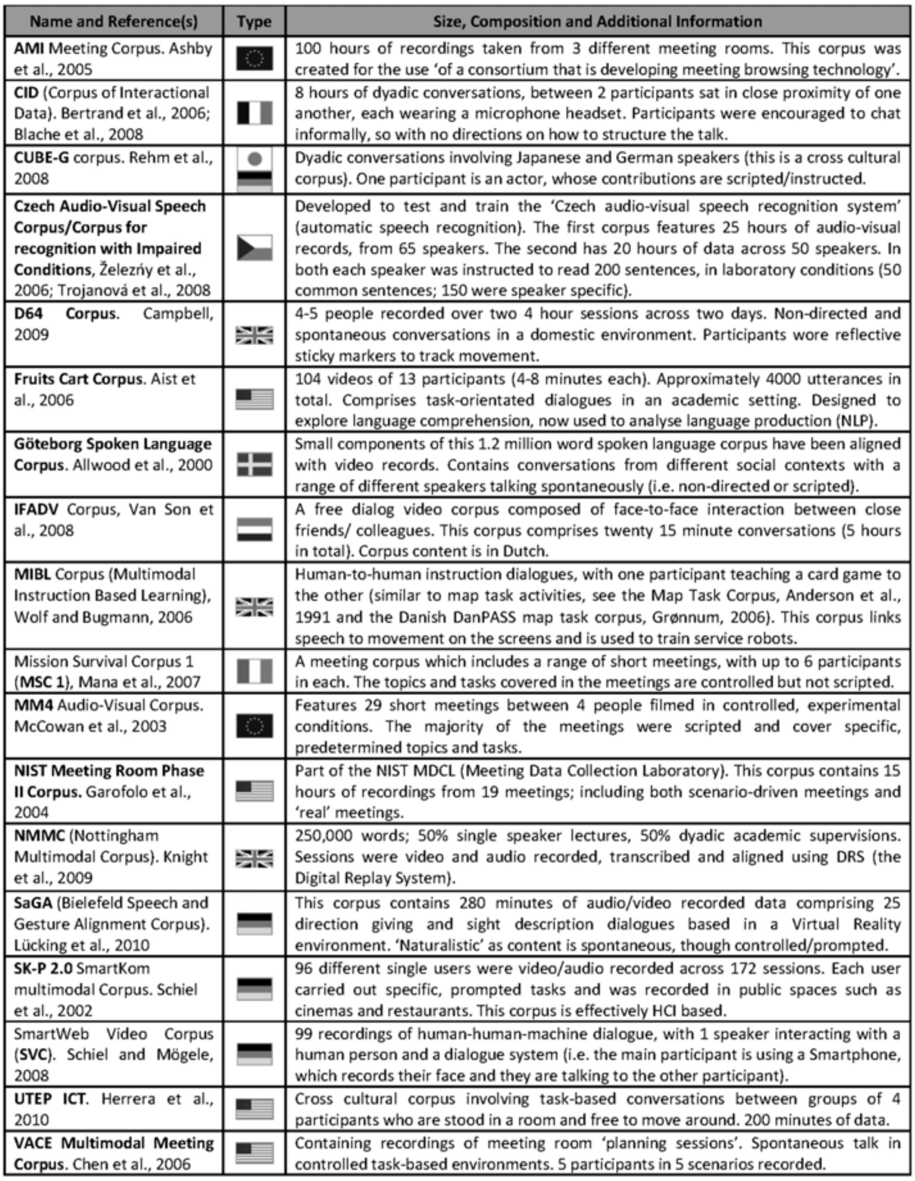
The surge in technological advancements witnessed since the latter part of the last century has provided the linguist with better tools for recording, storing and querying multiple forms of digital records. This has provided the foundations for the recent surge in interest in multimodal corpus linguistics.

A multimodal corpus, for the purpose of the current paper, is best defined as 'an annotated collection of coordinated content on communication channels including speech, gaze, hand gesture and body language, and is generally based on recorded human behaviour' (FOSTER; OBERLANDER, 2007, p. 307-308). The integration of textual, audio and video records of communicative events in multimodal corpora provides a platform for the exploration of a range of lexical, prosodic and gestural features and for investigations of the ways in which these features interact in real-life discourse.

Unlike monomodal corpora, which have a long history of use in linguistics, the construction and use of multimodal corpora is still in its relative infancy, with the majority of research associated with this field spanning back only a decade. Despite this, work using multimodal corpora has already proven invaluable for answering a variety of linguistic research questions, questions that are otherwise difficult to consider (see ALLWOOD, 2008 for further details).

The utility of corpus-based research and methods is in fact becoming popular in a range of different academic disciplines and fields of research, far beyond linguistics. For example, the processes of construction itself is of interest to computer scientists, while the tools developed can be utilised to answer questions posed by behaviourists, psychologists, social scientists and ethnographers. This means that multimodal corpora and corpus-based methods and related projects, which are often necessarily interdisciplinary and collaborative, receive ever-increasing support from academic researchers, funding councils and commercial third parties, something which is likely to be sustained well in to the future.

As a review of the current landscape, however, this paper primarily aims to provide an overview of selected multimodal corpora that have either already been built, or are currently under construction. An index of these corpora is provided in Figure 1, overleaf. The paper examines the types of data they contain, the applications of these datasets and ways in which they are limited. This is followed by a projection of ways such corpora can be further developed, improved or expanded in the future.



| Name and Reference(s) | Type | Size, Composition and Additional Information |
|---|---|---|
| **AMI** Meeting Corpus. Ashby et al., 2005 | 🇪🇺 | 100 hours of recordings taken from 3 different meeting rooms. This corpus was created for the use 'of a consortium that is developing meeting browsing technology'. |
| **CID** (Corpus of Interactional Data). Bertrand et al., 2006; Blache et al., 2008 | 🇫🇷 | 8 hours of dyadic conversations, between 2 participants sat in close proximity of one another, each wearing a microphone headset. Participants were encouraged to chat informally, so with no directions on how to structure the talk. |
| **CUBE-G** corpus. Rehm et al., 2008 | 🇯🇵<br>🇩🇪 | Dyadic conversations involving Japanese and German speakers (this is a cross cultural corpus). One participant is an actor, whose contributions are scripted/instructed. |
| **Czech Audio-Visual Speech Corpus/Corpus for recognition with Impaired Conditions**, Železńy et al., 2006; Trojanová et al., 2008 | 🇨🇿 | Developed to test and train the 'Czech audio-visual speech recognition system' (automatic speech recognition). The first corpus features 25 hours of audio-visual records, from 65 speakers. The second has 20 hours of data across 50 speakers. In both each speaker was instructed to read 200 sentences, in laboratory conditions (50 common sentences; 150 were speaker specific). |
| **D64 Corpus**. Campbell, 2009 | 🇬🇧 | 4-5 people recorded over two 4 hour sessions across two days. Non-directed and spontaneous conversations in a domestic environment. Participants wore reflective sticky markers to track movement. |
| **Fruits Cart Corpus**. Aist et al., 2006 | 🇺🇸 | 104 videos of 13 participants (4-8 minutes each). Approximately 4000 utterances in total. Comprises task-orientated dialogues in an academic setting. Designed to explore language comprehension, now used to analyse language production (NLP). |
| **Göteborg Spoken Language Corpus**. Allwood et al., 2000 | 🇸🇪 | Small components of this 1.2 million word spoken language corpus have been aligned with video records. Contains conversations from different social contexts with a range of different speakers talking spontaneously (i.e. non-directed or scripted). |
| **IFADV** Corpus, Van Son et al., 2008 | 🇳🇱 | A free dialog video corpus composed of face-to-face interaction between close friends/ colleagues. This corpus comprises twenty 15 minute conversations (5 hours in total). Corpus content is in Dutch. |
| **MIBL** Corpus (Multimodal Instruction Based Learning), Wolf and Bugmann, 2006 | 🇬🇧 | Human-to-human instruction dialogues, with one participant teaching a card game to the other (similar to map task activities, see the Map Task Corpus, Anderson et al., 1991 and the Danish DanPASS map task corpus, Grønnum, 2006). This corpus links speech to movement on the screens and is used to train service robots. |
| Mission Survival Corpus 1 (**MSC 1**), Mana et al., 2007 | 🇮🇹 | A meeting corpus which includes a range of short meetings, with up to 6 participants in each. The topics and tasks covered in the meetings are controlled but not scripted. |
| **MM4** Audio-Visual Corpus. McCowan et al., 2003 | 🇪🇺 | Features 29 short meetings between 4 people filmed in controlled, experimental conditions. The majority of the meetings were scripted and cover specific, predetermined topics and tasks. |
| **NIST Meeting Room Phase II Corpus.** Garofolo et al., 2004 | 🇺🇸 | Part of the NIST MDCL (Meeting Data Collection Laboratory). This corpus contains 15 hours of recordings from 19 meetings; including both scenario-driven meetings and 'real' meetings. |
| **NMMC** (Nottingham Multimodal Corpus). Knight et al., 2009 | 🇬🇧 | 250,000 words; 50% single speaker lectures, 50% dyadic academic supervisions. Sessions were video and audio recorded, transcribed and aligned using DRS (the Digital Replay System). |
| **SaGA** (Bielefeld Speech and Gesture Alignment Corpus). Lücking et al., 2010 | 🇩🇪 | This corpus contains 280 minutes of audio/video recorded data comprising 25 direction giving and sight description dialogues based in a Virtual Reality environment. 'Naturalistic' as content is spontaneous, though controlled/prompted. |
| **SK-P 2.0** SmartKom multimodal Corpus. Schiel et al., 2002 | 🇩🇪 | 96 different single users were video/audio recorded across 172 sessions. Each user carried out specific, prompted tasks and was recorded in public spaces such as cinemas and restaurants. This corpus is effectively HCI based. |
| SmartWeb Video Corpus (**SVC**). Schiel and Mögele, 2008 | 🇩🇪 | 99 recordings of human-human-machine dialogue, with 1 speaker interacting with a human person and a dialogue system (i.e. the main participant is using a Smartphone, which records their face and they are talking to the other participant). |
| **UTEP ICT**. Herrera et al., 2010 | 🇺🇸 | Cross cultural corpus involving task-based conversations between groups of 4 participants who are stood in a room and free to move around. 200 minutes of data. |
| **VACE Multimodal Meeting Corpus**. Chen et al., 2006 | 🇺🇸 | Containing recordings of meeting room 'planning sessions'. Spontaneous talk in controlled task-based environments. 5 participants in 5 scenarios recorded. |

FIGURE 1: An index of multimodal corpora

## 2. Multimodal Corpora: analysing discourse 'beyond the text'

### 2.1. Current multimodal corpora

There are two broad 'types' of researchers who are interested in multimodal corpus linguistics, as identified by Gu (2006). Firstly, there are those who are interested in undertaking 'multimodal and multimedia studies of discourse', addressing more social science based issues, with a concern on 'human beings' (GU, 2006, p. 132).

Secondly, there are those interested in the construction of multimodal corpora as an explorative exercise, tackling specific technological challenges of assembling and (re)using these datasets, and evaluating how this is best achieved; that is, which software and hardware tools to use etc. Many of these researchers are more interested in 'how to improve human-computer interaction' (GU, 2006, p. 132, also see KNIGHT *et al.*, 2009 for further discussion and associated examples).

Similar to current monomodal corpora, the contents of multimodal corpora, the ways in which they are recorded, their size, and so on, are highly dependent on the aims and objectives that they are intended to fulfil; the specific research questions that want to be explored or the specific technological or methodological questions that require answering by those developing and/or using the corpus. Given this, there are a variety of different forms of multimodal corpora and related research projects, all with, to some degree, bespoke characteristics regarding:

- **Design and infrastructure**: Concerning what the data in the corpus looks like; what sorts of recordings are included and the basic design methodology used to collect, compile, annotate and represent this data.
- **Size and scope**: Amount of data (in terms of hours and/or word count) and the variation in the types included (in terms of the range of speakers or different contexts included and so on).
- **Naturalness**: How 'natural' or 'real' (authentic) the data is perceived to be; whether it is scripted and/or structured or more spontaneous.
- **Availability and (re)usability**: Access rights to data, whether corpora are published and can be utilised and analysed by other researchers.

Each of these will be discussed at length in the subsequent sections of this paper.

## 2.2. Design and infrastructure

While research using audio recordings of conversation has had a long history in corpus-based linguistics, the use of digital video records as 'data' is still fairly innovative. The specific strategies and conventions used to compile (record), annotate and represent/replay video records for a multimodal corpus therefore generally differ from one to the next (for further discussions on each of these processes, see KNIGHT *et al.*, 2009).

No formally agreed, standardised approach exists for recording data for multimodal corpora and although each current corpus, as seen in figure 1, tends to utilise a range of highly specialised equipment in a fixed, predefined, thus *replicable* set-up, the exact nature of this setting is not necessarily consistent from one to the next. Specific forms of equipment, where they are located and even the file formats that they record in are subject to variation.

Further to this, as discussed extensively in Knight *et al.* (2009), various different schemes exist to mark up, code and annotate multimodal data, and as yet no standard approach is used across all multimodal corpora (although the International Standards for Language Engineering, ISLE project acknowledges the need for such, DYBKJÆR; OLE BERNSEN, 2004, p. 1). As Baldry and Thibault note (2006, p. 148):

> In spite of the important advances made in the past 30 or so years in the development of linguistic corpora and related techniques of analysis, a central and unexamined theoretical problem remains, namely that the methods adapted for collecting and coding texts isolate the linguistic semiotic from the other semiotic modalities with which language interacts.... [In] other words, linguistic corpora as so far conceived remain intra-semiotic in orientation.... [By] contrast multimodal corpora are, by definition, inter-semiotic in their analytical procedures and theoretical orientations.

Extensive deliberation also exists about what aspects should actually be marked up and how; so which specific non-verbal behaviours (patterns of gesticulation) or prosodic features should be annotated and so on. This problem is also true for the software used in order to undertake the processes of coding, annotation, synchronisation and representation (for a more in depth discussion on each of these processes please refer to KNIGHT, 2011).

While an increasing number of multimodal projects, particularly those linked to the multimodal corpora workshop series,[1] are using the software tool Anvil[2] (KIPP, 2001; KIPP *et al.*, 2007), others favour ELAN[3], DRS[4] (FRENCH *et al.*, 2006; GREENHALGH *et al.*, 2007) or EXMARaLDA[5]. Given this, standardised procedures for carrying out these processes would thus be welcomed and are perhaps a priority for the future of research in this field.

## 2.3. Size, scope and range

Figure 1 indicates that few multimodal corpora extend beyond a few thousand words in size. While the AMI corpus (see ASHBY *et al.*, 2005) comprises an impressive 100 hours of video, the majority of this data exists solely as video records. In other words many videos have yet to be transcribed, so the actual *size* of this corpus as a functional multimodal (i.e. text and video based) tool is not especially large. Other multimodal corpora contain only a few hours of video and/or a limited number of words.

This issue of size is especially noteworthy because current monomodal corpora pride themselves on the fact that they extend into multi-million word datasets, such as the British National Corpus (BNC), the Bank of English (BoE) and the Cambridge International Corpus (CIC). The BNC contains 100 million words of British English (90% written, 10% spoken); the BoE stands at over 650 million words (75% written, 25% spoken) and the CIC corpus has recently hit the 1 billion word mark.

Obviously, the advantage of using text-based discourse in the compilation of corpora is that large quantities of data are readily available,

[1] Details of the multimodal corpora workshop series on multimodal corpora, tools and resources can be found at: <http://www.multimodal-corpora.org>.

[2] ANVIL is a frame accurate multimodal annotation and visualisation tool, available for free from: <http://www.dfki.de/~kipp/anvil/>.

[3] ELAN is a 'professional tool for the creation of complex annotations on video and audio resources' which is available to download for free at: <http://www.lat-mpi.eu/>.

[4] DRS, The Digital Replay System, is a multimodal corpus construction and replay tool which is available to download for free at: <http://sourceforge.net/projects/thedrs/>.

[5] Exmeralda, Extensible Markup Language for Discourse Annotation, 'is a system of concepts, data formats and tools for the computer assisted transcription and annotation of spoken language, and for the construction and analysis of spoken language corpora' which is available to download for free at: <http://www.exmaralda.org/en_index.html>.

already machine-readable and/or relatively easy to get hold of, so the process of assembling such databases is relatively straightforward. The process of compiling spoken components or indeed purely spoken corpora is renowned as being a more lengthy process. This is because spoken data needs to be recorded before it is transcribed, annotated and coded before it is integrated into the corpus. As, it is estimated, the process of transcription alone takes a trained researcher up to ten hours to tackle one hour of audio, compiling spoken corpora is often a long and arduous process. For this reason spoken corpora tend to be of a smaller size, such as the five million word CANCODE[6] corpus.

Adding further 'multimodal' levels and perspectives to corpora compounds this problem as recording, aligning and transcribing (if at all) different streams of data is naturally more time consuming and technically difficult than when dealing with a single stream. Furthermore, if specific gestures are to be annotated, the processes of defining, marking up and coding these add further complexity to the construction of these datasets as, it is generally considered, 'the most labour-intensive part for acquiring a multimodal corpus is the annotation of the data, in particular for the visual modality' (FANELLI *et al.*, 2010, p. 70). However, over time we have witnessed an increase in the availability of technical resources for not only recording but also processing, aligning and archiving multimodal corpora, so it is likely that these limitations will become less inhibiting in the future.

Further to size, current multimodal corpora are somewhat limited in terms of *scope*. The majority of the corpora seen in figure 1 tend to be domain specific, mono-lingual (aside from CUBE-G) and/or of a specialist nature, so built of one form of data recorded in a given discourse context. AMI, the MM4 Audio-Visual Corpus, MSC1, the VACE Multimodal Meeting Corpus and the NIST Meeting Room Phase II Corpus all feature records of interaction from a professional meeting room. In these meeting-based corpora, the primary motivation behind the associated research (and corpus construction) is to enable the development and integration of technologies for displaying and researching

[6] CANCODE stands for Cambridge and Nottingham Corpus of Discourse in English. This corpus has been built as part of a collaborative project between The University of Nottingham and Cambridge University Press with whom sole copyright resides. CANCODE comprises five million words of (mainly casual) conversation recorded in different contexts across the British Isles.

meeting room activity specifically. In some of these corpora, the content is scripted or pre-planned to a certain extent and/or the conditions in which the recordings take place are controlled and experimental, with participants being told specifically where to sit and so on.

So, despite the commendable size of AMI, the utility of this corpus for general corpus linguistic research is perhaps limited. As with specialised monomodal corpora such the MICASE corpus[7] of academic discourse and the Wolverhampton Business English Corpus,[8] the contextual and compositional specificity of the data included means it is not necessarily appropriate for addressing research questions that focus on the more interpersonal aspects of communication (for example), beyond this formal, professional contextual domain. This is because the meeting room environment is generally understood as not being particularly conducive to the frequent occurrence of more informal, interpersonal language and/or behaviours. The specialised nature of these corpora potentially affects the spontaneity of the content included (a facet discussed in more detail below), as the constrained nature of the discourse context influences the content and structure of the discourse.

A similar criticism is valid for the NMMC which includes only lecture and supervision data (i.e. academic), and can also be extended to the map or task-based corpora, which prompt highly structured and sometimes scripted content (examples include CUBE-G, the Czech Audio-Visual Speech Corpus, Fruits Carts Corpus, MIBL, SaGA and UTEP ICT).

Further to this, the NMMC was initially designed to allow the application of a 2D digital tracker onto the derived images (see Knight et al., 2006 for further details), as a means of defining patterns of gesticulation. Therefore, recordings are all close up, focusing mainly on the head and torso of participants in order to produce high *quality* images to support the use of the tracking software. Thus while patterns of hand, arm and head movements can be analysed in this data, other bodily actions and spatial positions (i.e. proxemics), for example, cannot. Therefore researchers interested in

[7] MICASE, the Michigan Corpus of Academic English, is a 1.7 million word corpus of transcribed interactions recorded at the University of Michigan. For more information, see: <http://lw.lsa.umich.edu/eli/micase/index.htm>.

[8] The Wolverhampton Business English Corpus is comprises 10 million words of written English from the business domain. These texts were collected between 1999 and 2000. For more information, see: <http://www.elda.org/catalogue/en/text/W0028.html>.

researching a range of different behaviours would perhaps find the NMMC dataset limited (see BRÔNE *et al.*, 2010 for further discussion). This is true for other examples of corpora using more laboratory based and/or situated, static, recording methodologies, as detailed in Figure 1.

If the NMMC utilised recordings of participants at a greater distance away, thus capturing more aspects of the bodily movement, it is unlikely that the tracking system, which required a face-on and in-focus image, could be utilised. This would make the data recorded unfit for its original intended purpose. Overall, it is difficult to maintain a balance between the quality of corpus data and its potential usability, a balance which is somewhat constrained by the limitations of recordings equipment used to collect it. This makes the criticisms of the balance between the relative quality and reusability of multimodal corpus data particularly difficult to resolve/overcome.

The only corpora featured in figure 1 (above) that are exempt from this criticism of 'scope' are D64, components of the Goteborg Spoken Language Corpus, IFADV, SK-P and the SmartWeb Video Corpus. These corpora are either mobile based, so are not fixed to specific geographical or social contexts (SK-P and the SmartWeb Video Corpus) or include data which is seen to be 'spontaneous' and 'naturalistic'; featuring speakers who are static but who are discussing a range of self-selected topics (elements of the Goteborg Spoken Language Corpus and IFADV) and are perhaps, as is the case of D64, recorded in relaxed and familiar domestic settings.

## 2.4. Naturalness

Support for using corpora in linguistic research was traditionally founded on the notion that while 'introspective data is artificial…..corpora are natural, unmonitored sources of data' (McENERY; WILSON, 1996, p. 8, also see McCARTHY, 2001, p. 125 and MEYER, 2002, p. 5). Corpora therefore provide records of discourse as it is used in real-life contexts, that is, language as it is performed; rather than relying on more rationalistic, intuitive accounts (as previously advocated by CHOMSKY, 1965).

Constructing and utilising authentic, *naturalistic* language records is also a real aim for those working with multimodal data; an aim which has proven to be difficult to fully achieve. By definition alone, this notion of naturalness is abstract and interpretive. As an idealised concept, it is best described as that language which is used in our daily lives; unconstrained and fluid, changeable from one context to the next.

Following this definition, and given the matters discussed in section 2.4, the proposed naturalness of the data contained in those corpora listed in figure 1 can be brought under scrutiny. As the recording set-ups used are generally fixed, laboratory based and/or feature specialist environments with participants; they are thus far from 'unconstrained' and 'context-free'. Oertel et al. suggest that current set-ups effectively exist on a cline, a 'spectrum', as seen in Figure 2 (2010, p. 28).

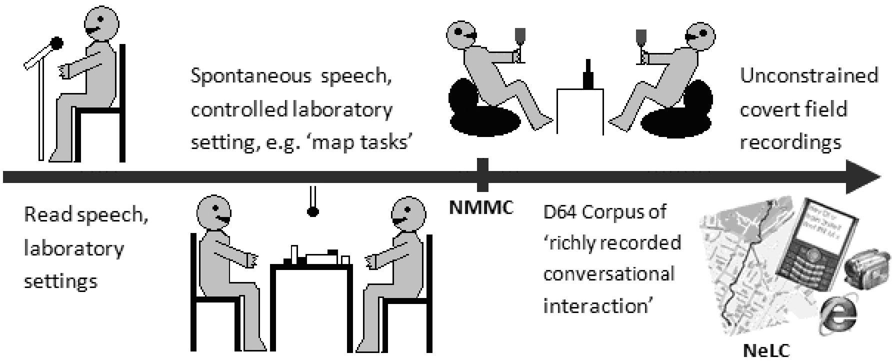


FIGURE 2: 'Spectrum of observation scenarios ranging from highly controlled to truly ethological' (based on OERTEL *et al.*, 2010, p. 28)

At the extreme left of the spectrum exists the highly conditioned and scripted forms of corpora such as CUBE-G and Czech Audio-Visual Speech corpus. This progresses to dyadic, but situated, records of speakers in controlled scenarios (such as the Fruit Carts Corpus, MIBL Corpus and SaGA Corpus) through to more spontaneous forms of 'richly recorded' datasets taken from more informal contexts, such as domestic settings (the D64 corpus for example). At the right side of the spectrum we see unconstrained covert field recordings.

To develop corpora which are as *naturalistic* as possible then, it is suggested that the form of recording set-up positioned to the far right of this figure would be most effective. This would thus include data recorded in dynamic environments; on the move and in a variety of different contexts, away from the standardised, fixed and situated setting. While no corpus of this nature has been fully developed as yet, plans to do so are currently underway at the University of Nottingham (see section 3.1 for more details).

Not only does the recording context, that is the physical setting, potentially compromise this notion of naturalness in corpus development, but so too does the equipment used in this context. Audio and video recorders can impact on the data due to the 'observer's paradox' (LABOV, 1972), whereby participants may (sub)consciously adjust their behaviours because they are aware that they are being filmed. Given that video cameras, in particular, are quite obtrusive, and technically it is not ethical to 'hide' these or other recording devices (without the participant's consent), it is difficult to minimise the potential effect this will have on how *naturalistic* behaviours are.

In addition to cameras and microphones, in order to track gestures the D64 corpus, for example, also required participants to wear reflective sticky markers during the recording phase. Again these markers are somewhat invasive and detrimental to the perceived naturalness of the recorded data as they are 'not only time-consuming and often uncomfortable' to wear but 'can also significantly change the pattern of motion' (FANELLI *et al.*, 2010, p. 70, also see FISHER *et al.*, 2003). However, as a means for capturing bodily movements and sequences of gestures accurately, the use of these markers is unavoidable, as they provide the best method for accurately capturing patterns of discrete body movements. So, as a means of fitting the future research needs of this particular corpus, the use of these devices cannot be legitimately criticised (although in terms of multimodal corpora as 'generic' tools, the reverse is the case).

Fanelli et al. suggest the utility of 3D capture techniques for gesture tracking as an alternative, more unobtrusive alternative to sticky markers. This is something that is still under development by a range of different researchers (i.e. a proven accurate version of such a utility has yet to be released).

Arguably the most naturalistic of the those multimodal corpora listed in figure 1 are the CID, UTEP ICT, SVC and the D64 corpus (despite its' use of sticky markers). The CID contains recordings of German interaction between dyads of people sitting next to each other. The participants are encouraged to discuss any topic or issue they wish, in a bid to provide accounts of conversational data which is as true to 'real-life' as possible. However, again, the conditions in which these recordings took place are to a certain extent experimental, with participants sitting in a closed laboratory and wearing headset microphones.

Participants in the UTEP ICT corpus were also required to wear microphones, although these were wireless and pin-on. For this corpus, cameras are placed around the room as unobtrusively as possible, with participants standing in the middle of the room, able to move freely around the room as desired. Although the content is described as spontaneous, a key limitation of this corpus is that discussions are task based and specifically 'designed to elicit a range of dialog behaviours' (HERRERA *et al.*, 2010, p. 50).

The SVC adopts a recording approach which is even less context-specific and more 'mobile'. It uses portable Smartphone devices to record a range of different public spaces, both inside and outside, with varying light conditions and acoustic properties (SCHIEL; MÖGELE, 2008, p. 2). However, the Smartphone devices are only used to record single participants in these corpora, despite the fact the SVC is based on dyadic conversations. This limits the potential for exploring patterns in dyadic or group behaviour in the data. Furthermore the quality of these recordings is not particularly good and only specific sequences of behaviour, facial expressions and head movements are captured at a high resolution. So for potential reuse in studies which look at other forms of gesticulation, proxemics or other features, this dataset is limited. Though, in truth, this is perhaps more a limitation of the equipment specifications than the recording design methodology. An additional, more general limitation of these corpora is that they are both task-orientated, so although discourse is occurring in real-life contexts, the prescribed nature of the tasks again affects the spontaneity and perceived *naturalness* of the data.

Finally, the D64 corpus is an English based corpus which has been recorded in arguably the most naturalistic setting; that is a domestic living room (see CAMPBELL, 2009), aiming to record language in a truly social situation, so 'as close to an ethological observation of conversational behavior as technological demands permit' (OERTEL *et al.*, 2010, p. 27). Conversations were recorded over long periods of time, the topics of which were not scripted or prompted. As with the UTEP ICT, participants were able to move around the room as they so wished, although they notably did remain seated for the majority of the time. Interestingly 'to add liveliness to the conversation, several bottles of wine were consumed during the final two hours of recording' (OERTEL *et al.*, 2010, p. 27). While the raw data for this corpus is now available, the edited version, complete with transcriptions, codes, tags and so on has yet to be released.

### 2.5. Availability and (re)usability

As Brône et al. note even now 'truly multimodal corpora including visual as well as auditory data are notoriously scarce' (2010, p. 157), as few have been published and/or are publicly available and no ready-to-use large corpus of this nature is currently commercially *available*.

This is due to a variety of factors, but is most strongly linked to 'privacy and copyright restrictions' (van SON *et al.*, 2008, p. 1). Corpus project sponsors or associated funding bodies enforce restrictions on the distribution of materials, and prescriptions related to privacy and anonymity in multimodal datasets reinforce such constraints. Although, notably, plans to publish/release data contained within the D64 (CAMPBELL, 2009) and NOMCO corpora (an 'in-progress' cooperative corpus development project between Sweden, Denmark and Finland focusing on human-human interaction, see BOHOLM; ALLWOOD, 2010) have been confirmed for the near future, these have yet to come to fruition.

### 2.6. Section overview

In brief, shortcomings of current multimodal corpora and related research approaches and methodologies can be summarised as follows:

- **Design:** Multimodal corpora tend to include synchronised video, audio and textual records designed and constructed primarily to meet a specific research need and/or to answer particular questions.
- **Infrastructure**: Strategies and conventions used to record, mark-up, code, annotate and interrogate multimodal corpora vary dramatically from one corpus to the next. Standardised procedures for each of these processes have yet to be developed and/or agreed.
- **Size**: They are all fairly limited in size, compared to their monomodal equivalents. Multi-million word multimodal corpora do not exist as yet.
- **Scope**: The majority of these corpora tend to be domain specific, mono-lingual and/or are of a specialist nature (i.e. recorded in one discourse context). In some of these, the content is also pre-planned or scripted, and the conditions under which they are recorded are experimental and controlled.

- **Naturalness**: The controlled recording conditions, settings and obtrusive equipment used may compromise the extent to which the data contained within the majority of multimodal corpora is spontaneous and 'naturalistic'.
- **Availability and (re)usability**: No widely available, large scale corpus has been published to date.

The next section outlines ways in which these may be overcome in the future of research in this field.

## 3. Future developments for multimodal corpora

### 3.1. Making multimodal corpora 'bigger' and 'better'

While section 2 focused on outlining some limitations related to current multimodal corpus linguistics, the following section seeks to propose some solutions which may help to change the landscape of this area of research for the future.

Firstly, perhaps the obvious solution to criticisms related to the size, scope and availability of multimodal corpora is to strive for the development of bigger, more diverse datasets. Paradoxically, 'what is meant by large corpora is however quite a relative notion' in conventional linguistic research (BLACHE *et al.*, 2008, p. 110). 'In some linguistic fields such as syntax, for instance, corpora of several million words are used, whereas in prosody where most of the annotations are made manually, a few hours of speech are considered as a large corpus' (BLACHE *et al.*, 2008, p. 110). So the appropriate size of a corpus, whether it be mono or multimodal, can only really be determined in the light of what it is to be used for. This means it is perhaps ill informed to qualify size as a strength or shortcoming of those corpora in figure 1 (as addressed in section 2.3) given that, as with the monomodal counterparts, the data in multimodal corpora tends to be research specific, specialist and/or domain specific.

Further to this, 'since language text is a population without limits, and a corpus is necessarily finite at any one point; a corpus, no matter how big, is not guaranteed to exemplify all the patterns of the language in roughly their normal proportions' (SINCLAIR, 2008, p. 30). Corpora are necessarily 'partial', as it is impossible to include *everything* in a corpus as the methodological and practical processes of recording and documenting natural language are selective; ergo 'incomplete' (THOMPSON, 2005, also see OCHS, 1979; KENDON, 1982, p. 478-479 and CAMERON, 2001, p. 71). This is true irrespective of whether a corpus is specialist or more general in nature.

Yet, in an ideal scenario, current multimodal corpora would be larger and more extensive in order to allow them to be more representative of a wider range of language samples/types, to enable the linguist to make better informed observations of language-in-use from a multitude of different perspectives. Further to this, multimodal corpora should accommodate a range of other forms of media, beyond the standard of video, audio and textual data and associated metadata. This projected strand of corpus research and compilation thus works on the understanding that 'communication is not only a linguistic process, but also a multimodal exchange of meaningful information' (BOYD; HEER, 2006). Communication in the digital age is performed via a multitude of multimedia platforms with real-life, everyday discourse witnessing an ever increasing use of digital devices in a variety of different contexts. It is thus vital that we attempt to embrace this evolution in the next phase of multimodal corpus development.

As already noted, early efforts to capture the fluidity and complexity of context (see GOODWIN, 2000, 2007) in real-life discourse have been made by researchers who developed the SVC corpus. The DReSS II project,[9] based at the University of Nottingham, builds on this further. The project is focusing on assembling a corpus of everyday (inter)actions from various different resources, incorporating not only text-based data, such as SMS messages, interaction in virtual environments (for example instant messaging logs and entries on personal notice boards), but also audio and video records from face-to-face conversation, as well GPS logs and a range of other media types. This project is still in progress.

The compilation of such heterogeneous data may enable us to extrapolate further information about communication across a range of different speakers, mediums and environments. In theory, this could assist in the questioning of the extent to which language choices are determined by different spatial, temporal and social contexts in communication.

In reality, there are obviously a whole host of ethical, practical and methodological problems that need to be faced when constructing such

[9] For more information, results and publications from DReSS, please refer to the main project website: http://web.mac.com/andy.crabtree/NCeSS_Digital_Records_Node/Welcome.html

corpora. The realisation of these aims and the successful development of heterogeneous multi-context corpora is heavily reliant on: technological advancements; on the constant refinement of systems that will enable the capture and structuring of natural language-in-use; as well as software that will promote the interrogation of different multimodal datasets. Constraints attributed to questions of scalability are also obviously inherent to the practical implementation of this 'next-step', since, as already identified, the processes of recording, transcribing, time-stamping and coding data remain very time-consuming despite the availability of software for this (for detailed discussions and specific examples of these, see KNIGHT *et al.*, 2009).

Such problems may deter linguists from attempting to create multimodal corpora of this nature because, to date, simple solutions to these problems have yet to emerge. This includes matters of what and how behaviours are quantified, queried and represented to the linguist, and how patterns are statistically assessed and/or analysed.

### 3.2. Software and hardware requirements

Given that NELC (Nottingham eLanguage Corpus, developed as part of the DReSS II project) is to include multiple forms of varying media types, there are lots of issues to be addressed regarding the optimum ways in which these are recorded, processed, stored and accessed/interrogated by the linguist. The methods employed at each of these stages naturally differ from each media type because they are stored in a variety of file formats, and are typically visualised and represented in different ways. Therefore better devices for recording multiple forms of data, in synchronicity and at a high quality, need to be developed. This will help to enhance the speed at which corpora are composed, giving researchers the chance to extend the size of their corpora at speed.

While cameras and Dictaphones and other recording hardware of an ever increasingly higher specification are constantly being developed, the mobility and functionality of these still recommend that the situated forms of laboratory type recordings will yield the best results. Numerous cameras can be positioned in various locations around the room in order to capture participants from multiple perspectives, from close up and head on (which would support the use of tracking software on resultant images when analysing the data) to birds eye views or more panoramic shots. Similarly eye or movement tracking equipment (such as the sticky markers discussed earlier) can be worn, as required, by participants, in static environments.

More mobile toolkits, as called for here, are becoming increasingly available, although they are still somewhat primitive as the quality of recordings, or the length allowed for recordings, for example, is limited (for an example of such a toolkit under development, see the DReSS II website for more information, also see CRABTREE; RODDEN, 2009).

It would also prove beneficial to look to develop more enhanced tools for the automatic transcription of data. While such tools are currently in existence (such as Speechware[10]), it is widely acknowledged that these are far from accurate, especially when recording spontaneous dyadic or group conversation. Given this, these tools are rarely used in monomodal or multimodal corpus construction.

Thirdly, semi-automated processes of annotating data would also ease the speed at which multimodal corpora are developed and analysed. This may take the form of those digital tracking devices discussed above, designed to allow users to automatically define and subsequently encode specific features of interest in video data (according to specific parameters set by the analyst), to allow for larger scale explorations of language and gesture-in-use to be undertaken with ease (see KNIGHT *et al.*, 2006; BRÔNE *et al.*, 2010 and JONGEJAN, 2010). Although in practice, since such technologies are still 'developing', these tracking techniques are far from perfect, so at present they remain a speculative *potential* rather than *functional* part of the multimodal Corpus Linguistic approach.

Finally, software to support the representation and meaningful interrogation of these datasets needs to be developed as again no standard procedures exist for this in current multimodal corpus methodology. Knight *et al.* identify the following features as being essential to interrogate heterogeneous corpus toolkits, although utilities are likely to need to extend beyond these (2010, p. 17):

- The ability to search data **and** metadata in a principled and specific way, within and/or across the three global domains of data:
    - Devices/ data type(s)
    - Time and/or 'location
    - Participants' given contributions

[10] Speechware is an automatic transcription and speech recognition tool. For more information, visit the following website: <www.speechware.be/en/company.php>.

- Tools that allow for the frequency profiling of events/ elements within and across domains (providing raw counts, basic statistical analysis tools, and methods of graphing such).
- Variability in the provisions for transcription and the ability for, for example, representing simultaneous speech and speaker overlaps.
- Graphing tools for mapping the incidence of words or events, for example, over time and for comparing sub-corpora and domain specific characteristics.

These will seek to build on, combine and extend the functionalities of common monomodal corpus analytical tools, such as those provided by WordSmith Tools (SCOTT, 1999), Sketch Engine (KILGARRIFF *et al.*, 2004) and WMatrix (RAYSON, 2003), as well other forms of social science and qualitative data research software (as mentioned in section 3.1 above). Ideally, such tools should also be free/open source since, to date, much of the field has been monopolised by pay-for-prescription tools and datasets as monies are perhaps necessary to fund the development, maintenance and sustainability of corpus infrastructure (as although funding is often available, commercialisation is often a by-product of this). This somewhat inhibits the accessibility of tools to certain users. Open source software and uiltiities will, in comparison, enhance accessibility for all and will promote the cross fertilization of corpus based methods into other linguistic fields and beyond.

Thankfully, a range of sophisticated corpus tools are being developed in this research 'space', aiming to support some or all of the utilities listed above, within an open-source corpus workbench, including ELAN, DRS, Exmeralda and Anvil. While these tools mainly support corpus construction, maintenance and analysis without providing any corpus 'data' of their own, they set a great example of the potential for the availability of corpus tools for the future.

## 4. Summary

Multimodal corpora are an important resource for studying and analysing the principles of human communication' (FANELLI *et al.*, 2010). Multimodal datasets function to provide a more lifelike representation of the individual and social identity of participants, allowing for an examination of prosodic, gestural and proxemic features of the talk in a specific time and place. They thus reinstate partial elements of the reality of discourse, giving each speaker and each conversational episode a specific distinguishable identity. It is only when these extra-linguistic and/or paralinguistic elements are represented in records of interaction that a greater understanding of discourse can be generated, following linguistic analyses.

This paper has outlined various strengths shortcomings of current (early) multimodal linguistic corpora. It has focused on outlining characteristics of the basic design and infrastructure of (early) multimodal corpora; their size and scope; the quality and authenticity/naturalness of data contained in them and their availability and (re)usability. The paper has offered some reflections on the strengths of current multimodal corpora alongside some recommendations and a projective 'wish-list' for key areas of development that are likely to be addressed in the future of this area.

The successful implementation of these prospective advancements is heavily reliant on institutional, national and international collaborative interdisciplinary and multidisciplinary research strategies and funding. This is because 'modern research is increasingly complex and demands an ever widening range of skills…..often, no single individual will possess all the knowledge, skills and techniques required' (for discussion on the advantages of cross and multi-disciplinary research see NEWELL, 1984; KATZ; MARTIN, 1997 and GOLDE; GALLAGHER, 1999, p. 281). It is difficult to gauge whether all or any of these projections will ever be fully met, or how the multimodal landscape will look in the next decade or so, although it can be asserted with a fair amount of confidence, that interest in these corpora and associated methodologies will attract an ever increasingly amount of interest as time goes on and our digital worlds continue to expand.

## Acknowledgments

The research on which this paper is based is funded by the UK Economic and Social Research Council (ESRC), grant numbers RES-149-25-0035 & RES-149-25-1067.

## References

AIST, G.; ALLEN, J.; CAMPANA, E.; GALESCU, L.; GÓMEZ GALLO, C.; STONESS, S.; SWIFT, M.; TANENHAUS, M. Software architectures for incremental understanding of human speech. In: Interspeech 2006. *Proceedings...* Pittsburgh PA, USA: Interspeech, 2006.

ALLWOOD, J. Multimodal corpora. In: LÜDELING, A.; KYTÖ, M. (Ed.). Corpus Linguistics: An International Handbook. *HSK - Handbücher zur Sprach und Kommunikationswissenschaft*, v. 29, n. 1-2, p. 207-225, 2008.

ALLWOOD, J.; BJÖRNBERG, M.; GRÖNQVIST, L.; AHLSEN, E.; OTTESJÖ, C. The Spoken Language Corpus at the Department of Linguistics, Göteborg University. *Forum: Qualitative Social Research*, v. 1, n. 3, 2000. Available at: < http://www.qualitative-research.net/index.php/fqs/article/view/1026>. Retrieved: 12 Jul. 2010.

ANDERSON, A.; BADER, M.; BARD, E.; BOYLE, E.; DOHERTY, G. M.; GARROD, S.; ISARD, S.; KOWTKO, J.; McALLISTER, J.; MILLER, J.; SOTILLO, C.; THOMPSON, H. S; WEINERT, R. The HCRC Map Task Corpus. *Language and Speech*, v. 34, p. 351-366, 1991.

ASHBY, S.; BOURBAN, S.; CARLETTA, J.; FLYNN, M.; GUILLEMOT, M.; HAIN, T.; KADLEC, J.; KARAISKOS, V.; KRAAIJ, W.; KRONENTHAL, M.; LATHOUD, G.; LINCOLN, M.; LISOWSKA, A.; MCCOWAN, I.; POST, W.; REIDSMA, D.; WELLNER, P. The AMI Meeting Corpus. In: Measure Behaviour 2005. *Proceedings...*Wageningen, NL: Measuring Behavior, 2005.

BALDRY, A.; THIBAULT, P.J. *Multimodal Transcription and Text Analysis*: A multimedia toolkit and course book. London: Equinox, 2006.

BERTRAND, R.; BLACHE, P.; ESPESSER, R.; FERRE, G.; MEUNIER, C.; PRIEGO-VALVERDE, B.; RAUZY, S. Le CID: Corpus of Interactional Data -protocoles, conventions, annotations. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence* (TIPA) v. 25, p. 25-55, 2006.

BLACHE, P.; BERTRAND, R.; FERRÉ, G. Creating and exploiting multimodal annotated corpora. In: LREC 2008. *Proceedings...*Marrakech, Morocco: Sixth International Conference on Language Resources and Evaluation (LREC), 2008. p. 110-115. Available at: < http://www.lrec-conf.org/proceedings/lrec2008/>. Retrieved: July 12, 2010.

BOHOLM, M.; ALLWOOD, J. Repeated head movements, their function and relation to speech. In: LREC 2010. *Proceedings...*Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

BOYD, D.; HEER, J. Profiles as conversation: Networked identity performance on Friendster. In: HICSS 2006. *Proceedings*... Hawaii: Hawaii International Conference of System Sciences (HICSS-39), 2006.

BRÔNE, G., OBEN, B.; FEYAERTS, K. InSight Interaction- A multimodal and multifocal dialogue corpus. In: LREC 2010. *Proceedings*... Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

CAMERON, D. *Working with spoken discourse*. London: Sage, 2001.

CAMPBELL, N. Tools and Resources for Visualising Conversational-Speech Interaction. In: KIPP, M.; MARTIN, J.-C.; PAGGIO, P.; HEYLEN, D. (Ed.). *Multimodal Corpora*: From Models of Natural Interaction to Systems and Applications. Springer: Heidelberg, 2009.

CHEN, L.; TRAVIS-ROSE, R.; PARRILL, F.; HAN, X.; TU, J.; HUANG, Z.; HARPER, M.; QUEK, F.; MCNEILL, D.; TUTTLE, R.; HUANG, T. VACE *Multimodal Meeting Corpus*. Lecture Notes in Computer Science, v. 3869, p. 40-51, 2006.

CHOMSKY, N. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press, 1965.

CRABTREE, A.; RODDEN, T. Understanding interaction in hybrid ubiquitous computing environments. In: ACM 2009. *Proceedings*... Cambridge, ACM: 8th International Conference on Mobile and Ubiquitous Multimedia. Available at: <http://portal.acm.org/toc.cfm?id=1658550&type=proceeding&coll= GUIDE&dl =GUIDE&CFID=96741701&CFTOKEN= 20154123>. Retrieved: July 12, 2010.

DYBKJÆR, L.; OLE BERNSEN, N. Recommendations for natural interactivity and multimodal annotation schemes. In: LREC 2004. *Proceedings*...Lisbon: Language Resources and Evaluation Conference (LREC) Workshop on Multimodal Corpora, 2004.

FANELLI, G.; GALL, J.; ROMSDORFER, H.; WEISE, T.; VAN GOOL, L. 3D Vision Technology for Capturing Multimodal Corpora: Chances and Challenges. In: LREC 2010. *Proceedings*...Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

FISHER, D.; WILLIAMS, M.; ANDRIACCHI, T. The therapeutic potential for changing patterns of locomotion: An application to the acl deficient knee. In: ASME 2003. *Proceedings*... Miami, Florida: ASME Bioengineering Conference, 2003.

FOSTER, M.E.; OBERLANDER, J. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, v. 41, n. 3/4, p. 305–323, 2007.

FRENCH, A.; GREENHALGH, C.; CRABTREE, A.; WRIGHT, W.; BRUNDELL, B.; HAMPSHIRE, A.; RODDEN, T. Software Replay Tools for Time-based Social Science Data. In: ICeSS 2006. *Proceedings*... Manchester, UK: 2nd annual international e-Social Science Conference, 2006. Available at: <http://www.ncess.ac.uk/events/conference/2006/papers/>. Retrieved: July 12, 2010.

GARFOLO, J.; LAPRUN, C.; MICHEL, M.; STANFORD, V.; TABASSI, E. The NIST Meeting Room Pilot Corpus. In: LREC 2004. *Proceedings*...Lisbon, Portugal: 4th Language Resources and Evaluation Conference (LREC), 2004.

GOLDE, C.M; GALLAGHER, H.A. The challenges of conducting interdisciplinary research in traditional Doctoral programs. *Ecosystems*, v. 2, p. 281-285, 1999.

GOODWIN, C. Action and embodiment within situated human Interaction. *Journal of Pragmatics*, v. 32, n. 10, p. 1489-522, 2000.

GOODWIN, C. Participation, stance and affect in the organisation of activities. *Discourse and Society*, v. 18, n. 1, p. 53-73, 2007.

GREENHALGH, C.; FRENCH, A.; TENNANT, P.; HUMBLE, J.; CRABTREE, A. From ReplayTool to Digital Replay System. In: ICeSS 2007. *Proceedings*... Ann Arbor, Michigan, USA: 3rd International Conference on e-Social Science, 2007. Available at: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.100.755>. Retrieved: July 12, 2010.

GRØNNUM, N. DanPASS - a Danish phonetically annotated spontaneous speech corpus. In: LREC 2006. *Proceedings*...Genoa, Italy: 5th LREC conference, 2006.

GU, Y. Multimodal text analysis: A corpus linguistic approach to situated discourse. *Text and Talk*, v. 26, n. 2, p. 127-167, 2006.

HERRERA, D.; NOVICK, D.; JAN, D.; TRAUM, D. The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus. In: LREC 2010. *Proceedings*... Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

JONGEJAN, B. Automatic face tracking in Anvil. In: LREC 2010. *Proceedings*... Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

KATZ, J.S.; MARTIN, B.R. What is research collaboration? *Research Policy*, v. 26, p. 1-18, 1997.

KENDON, A. The organisation of behaviour in face-to-face interaction: observations on the development of a methodology. In: SCHERER, K.R.; EKMAN, P. (Ed.). *Handbook of Methods in Nonverbal Behaviour Research*. Cambridge: Cambridge University Press, 1982.

KILGARRIFF, A.; RYCHLÝ, P.; SMR•, P.; TUGWELL, D. The sketch engine. In: EU-RALEX 2004. *Proceedings*... International Congress, Lorient, France: In Proceedings of EU-RALEX, 2004.

KIPP, M. Anvil – A generic annotation tool for multimodal dialogue. In: INTERSPEECH 2001. *Proceedings*... Aalborg, Denmark: 7th European Conference on Speech Communication and Technology 2nd INTERSPEECH Event, 2001.

KIPP, M.; NEFF, M.; ALBRECHT, I. An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources and Evaluation*, v. 41, n. 3/4, p. 325-339, 2007.

KNIGHT, D. *Multimodality and active listenership*: A corpus approach. London, UK: Continuum Books, 2011.

KNIGHT, D.; BAYOUMI, S.; MILLS, S.; CRABTREE, A.; ADOLPHS, S.; PRIDMORE, T.; CARTER, R. Beyond the Text: Construction and Analysis of Multimodal Linguistic Corpora. In: ICeSS 2006. *Proceedings*... Manchester, UK: 2nd International Conference on e-Social Science, 2006. Available at: <http://www.ncess.ac.uk/events/conference/2006/papers/>. Retrieved: July 12, 2010.

KNIGHT, D.; EVANS, D.; CARTER, R.; ADOLPHS, S. Redrafting corpus development methodologies: Blueprints for 3rd generation "multimodal, multimedia" corpora. *Corpora*, v. 4, n. 1, p. 1-32, 2009.

KNIGHT, D.; TENNENT, P.; ADOLPHS, S.; CARTER, R. Developing heterogeneous corpora using the Digital Replay System (DRS). In: LREC 2010. *Proceedings*... Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

LABOV, W. *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press, 1972.

LÜCKING, A.; BERGMAN, K.; HAHN, F.; KOPP, S; RIESER, H. The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In: LREC 2010. *Proceedings*... Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

MANA, N.; LEPRI, B.; CHIPPENDALE, P.; CAPPELLETTI, A.; PIANESI, F.; SVAIZER, P.; ZANCANARO, M. Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection. In: ICMI 2007. *Proceedings*... Nagoya, Japan: Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information, ICMI'07.

McCARTHY, M.J. *Issues in Applied Linguistics*. Cambridge: Cambridge University Press, 2001.

MCCOWAN, S.; BENGIO, D.; GATICA-PEREZ, G.; LATHOUD, F.; MONAY, D.; MOORE, P.; WELLNER; BOURLAND, H. Modelling Human Interaction in Meetings. In: IEEE ICASSP 2003. *Proceedings*... Hong Kong: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2003.

McENERY, T.; WILSON, A. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

MEYER, C.F. *English corpus linguistics*: An introduction. Cambridge: Cambridge University Press, 2002.

NEWELL, W.H. Interdisciplinary curriculum development in the 1970's: the paracollege at St. Olaf and the Western College Program at Miami University. In: JONES, R.M.; SMITH, B.L (Ed.). *Against the current*: reform and experimentation in higher education. Cambridge: Schenkman, 1984.

OCHS, E. Transcription as theory. In: OCHS, E.; SCHIEFFELIN, B.B. (Ed.). *Developmental Pragmatics*. New York: Academic Press, 1979.

OERTEL, C.; CUMMINS, F.; CAMPBELL, N.; EDLUND, J.; WAGNER, P. D64: A Corpus of Richly Recorded Conversational Interaction. In: LREC 2010. *Proceedings*... Mediterranean Conference Centre, Malta: LREC Workshop on Multimodal Corpora, 2010.

RAYSON, P. Matrix: *A statistical method and software tool for linguistic analysis through corpus comparison*. (Doctoral thesis) – Department of Linguistics and English Language/Lancaster University, Lancaster, 2003.

REHM, M.; NAKANO, Y.; HUANG, H-H.; LIPI, A-A.; YAMAOKA, Y.; GRÜNEBERG, F. Creating a standardized corpus of multimodal interactions for enculturating conversational interfaces. In: IUI ECI 2008. *Proceedings*... Gran Canaria: IUI-Workshop on Enculturating Interfaces (ECI), 2008.

SCHIEL, F.; MÖGELE, H. Talking and Looking: the SmartWeb Multimodal Interaction Corpus. In: LREC 2008. *Proceedings*... Sixth International Conference on Language Resources and Evaluation (LREC), 2008. Available at: <http://www.lrec-conf.org/proceedings/lrec2008/>. Retrieved: July 12, 2010.

SCHIEL, F.; STEININGER, S.; TÜRK, U. The SmartKom Multimodal Corpus at BAS. In: LREC 2002. *Proceedinngs*... Las Palmas, Gran Canaria, Spain: 3rd Language Resources and Evaluation Conference (LREC), 2002.

SCOTT, M. *Wordsmith Tools* [Computer program]. Oxford: Oxford University Press, 1999.

SINCLAIR, J. Borrowed ideas. In: GERBIG, A.; MASON, O. (Ed.). *Language, people, numbers* - Corpus Linguistics and society. Amsterdam: Rodopi BV, 2008.

THOMPSON, P. Spoken Language Corpora. In: WYNNE, M. (Ed.). *Developing Linguistic Corpora*: a Guide to Good Practice. Oxford: Oxbow Books, 2005.

TROJANOVÁ, J.; HRÚZ, M.; CAMPR, P.; žELEZNÝ, M. Design and Recording of Czech Audio-Visual Database with Impaired Conditions for Continuous Speech Recognition. In: LREC 2008. *Proceedings*... Marrakech, Morocco: Sixth International Conference on Language Resources and Evaluation (LREC) 2008. Available at: < http://www.lrec-conf.org/proceedings/lrec2008/>. Retrieved: July 12, 2010.

VAN SON, R. J. J. H.; WESSELING, W.; SANDERS, E.; VAN DER HEUVEL, H. The IFADV corpus: A free dialog video corpus In: LREC 2008. *Proceedings*... Marrakech, Morocco: Sixth International Conference on Language Resources and Evaluation (LREC), 2008. Available at: <http://www.lrec-conf.org/proceedings/lrec2008/>. Retrieved: July 12, 2010.

WOLF, J.C.; BUGMANN, G. Linking Speech and Gesture in Multimodal Instruction Systems. In: IEEE RO-MAN 2006. *Proceedings*... Plymouth, UK: 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06), 2006.

žELEZNY, M.; KRNOUL, Z.; CÍSAR, P.; MATOUšEK, J. Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. *Signal Processing*, v. 83, n. 12, p. 3657-3673, 2006.

Recebido em 13/07/2010. Aprovado em 08/05/2011.

# Corpora and historical linguistics

## *Corpora e linguística histórica*

Merja Kytö*
Uppsala University
Uppsala / Sweden

ABSTRACT: The present article aims to survey and assess the current state of electronic historical corpora and corpus methodology, and attempts to look into possible future developments. It highlights the fact that within the wide spectrum of corpus linguistic methodology, historical corpus linguistics has emerged as a vibrant field that has significantly added to the appeal felt for the study of language history and change. In fact, according to a historical linguist with more than fifty years of experience, "[w]e could even go as far as to say that without the support and new impetus provided by corpora, evidence-based historical linguistics would have been close to the end of its life-span in these days of rapid-changing life and research, increasing competition on the academic career track and the methodological attractions offered to young scholars" (RISSANEN, forthcoming). Historical corpora and other electronic resources have also made the study of language history attractive: working on them engages students in an individual and interactive way that they find appealing (CURZAN 2000, p. 81).

KEYWORDS: Electronic historical corpora; corpus methodology; electronic resources; interdisciplinary collaboration.

RESUMO: Este artigo objetiva fazer um levantamento e avaliar o estado da arte dos corpora históricos eletrônicos e da metodologia de estudos de corpora, assim como sugerir possíveis desenvolvimentos futuros na área. Destaca-se que dentro do espectro metodológico da linguística de corpus, a linguística de corpus histórica emergiu como um campo de investigação vibrante que tem adicionado interesse ao estudo da história e da mudança linguística. De acordo com um pesquisador da área com mais de cinqüenta anos de experiência, "pode-se dizer que sem o apoio e o novo ímpeto trazidos pelos corpora, a linguística histórica baseada em evidências teria estado próxima ao fim de sua vida nesses tempos de rápidas mudanças de vida e de pesquisa, aumentando a competição na carreira acadêmica e nas atrações metodológicas oferecidas aos jovens pesquisadores (RISSANEN, no prelo). Corpora históricos e outros recursos eletrônicos têm também tornado o estudo da história da língua atraente: eles engajam a atenção dos estudantes tanto de forma individual quanto interativa (CURZAN 2000, p. 81).

PALAVRAS-CHAVE: Corpora históricos eletrônicos; metodologia de estudos de corpora; recursos eletrônicos; colaboração interdisciplinar.

* merja.kyto@engelska.uu.se

## 1. Introduction

The title of this article, "Corpora and historical linguistics", is likely to have meant something different to linguists some thirty to forty years ago than what it is taken to mean today. Similarly, "historical corpus linguistics" might well have been considered an instance of tautology, given that, apart from re-construction, all historical linguistics is in a wide sense corpus-based. If 'a corpus' is taken to be, as most would agree, "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (MEYER, 2002, p. xi), 'a historical corpus' is "intentionally created to represent and investigate past stages of a language and/or to study language change" Claridge (2008, p. 242). These definitions apply to two types of historical corpora, pre-electronic ones that antedate the advent of the computer, and electronic ones that exploit computer technology, the difference accounting for the above change in the use of terminology.

The present article aims to survey and assess the current state of electronic historical corpora and corpus methodology, and attempts to look into possible future developments. To begin with, it is important to keep in mind that within the wide spectrum of corpus linguistic methodology, historical corpus linguistics has emerged as a vibrant field that has significantly added to the appeal felt for the study of language history and change. In fact, according to a historical linguist with more than fifty years of experience, "[w]e could even go as far as to say that without the support and new impetus provided by corpora, evidence-based historical linguistics would have been close to the end of its life-span in these days of rapid-changing life and research, increasing competition on the academic career track and the methodological attractions offered to young scholars" (RISSANEN forthcoming). Historical corpora and other electronic resources have also made the study of language history attractive: working on them engages students in an individual and interactive way that they find appealing (CURZAN, 2000, p. 81).

Such corpus-based projects as biblical concordances, early grammars and early dictionaries bear witness to the painstaking nature of manual work involved in the use of pre-electronic corpora comprising one text or several texts (MEYER, 2008, p. 1). In the 1970s and 1980s, when it became possible to compile and analyse large-scale electronic corpora far more rapidly than had been the case with pre-electronic corpora (JOHANSSON, 2008, p. 33), historical linguists found themselves at the threshold of a new era. When describing this transitional stage in his introduction to the panel discussion devoted to "Issues in historical linguistics" at the 30th ICAME (International

Computer Archive of Modern and Medieval English) conference in May 2008, the convenor, Christian Mair (University of Freiburg), pointed out that "long before the advent of computers, monumental corpus projects were conceived which in some instances were later digitised and have continued into the present".[1] An example of such projects is the Corpus Inscriptionum Latinarum, which was started in 1853 and which "includes the Latin inscriptions from the entire area of the former Roman empire, arranged by region and by inscription-type" and which since its foundation has been "the standard edition of the epigraphic legacy of ancient Rome" (http://cil.bbaw.de/ ). On the other hand, the Thesaurus Linguae Graecae, a research centre at the University of California, Irvine, founded in 1972, set out to represent "the first effort in the Humanities to produce a large digital corpus of literary texts". It has so far "collected and digitized most literary texts written in Greek from Homer to the fall of Byzantium in AD 1453", with the goal "to create a comprehensive digital library of Greek literature from antiquity to the present era" (<http://www.tlg.uci.edu/>). Similarly, it was not until 1970's that we could also trace the first large-scale historical electronic corpus project aimed at documenting a period of the English language in toto (ca. 450-1100), that is, the Dictionary of Old English Corpus in Electronic Form, "a complete record of surviving Old English except for some variant manuscripts of individual texts" (<http://ota.ahds.ac.uk/headers/2488.xml>).

The ensuing tradition of English historical corpus linguistics has been particularly rich and has presented, a constantly growing family of historical corpora which documents periods extending from thirty years (or shorter spans of time) to a millennium. There is an increasing interest in historical corpora for many other modern languages, among them German and Mittelhochdeutsche Begriffsdatenbank, the Bonner Frühneuhochdeutsches Korpus and DeutschDiachronDigital, French and Textes de Français Ancien, Spanish and Corpus del Español, and Portuguese and Corpus do Português, to name just a few (for further examples and references, see CLARIDGE, 2008 and XIAO, 2008). There have also been signs in cross-linguistic historical corpus compilation projects as will be shown in the present article later on. Even though English historical corpora will serve as the basis for the discussion in the present article, it is hoped that the methodological issues raised, or most of them, can largely be taken to pertain to historical corpora in general.

[1] I am indebted to Christian Mair for permission to cite his script.

This article is organised as follows. After some preliminary remarks (section 1), resources and methodology in historical corpus linguistics will be discussed (section 2). The rationale of the approach will be examined (2.1), and the types of historical corpora available or underway (2.2.) will be surveyed along with the tools enabling historical corpus analysis (2.3). Section 3 will be devoted to an assessment of the developments in the field, with a discussion of recent advances and remaining bottleneck areas. The main themes will be resources and their potential for enhancement and new projects (3.1), prospects of searchability and corpus annotation (3.2), the need to enhance access to and information on historical electronic resources (3.3), and the need to promote interdisciplinary collaboration. A summary of future directions and desiderata will conclude the article (section 4).

## 2. Resources and methodology

### 2.1 The rationale of the historical corpus linguistic approach

There are a number of reasons why it makes sense to study the history of a language and language change using corpus linguistic methodology. These will be touched upon in the present section as they also tend to lie behind corpus compilation methodology and guide the developments in the field (see section 3).

A useful discussion of the benefits brought by the corpus linguistic approach to the study of language change can be found in Curzan (2008). In the study of language change, the aim is often to detect and substantiate general trends in language development. For this, one needs easy access to large amounts of data representative of different registers and levels of language use. Computerised corpora allow the study of stages of linguistic development from a contrastive or comparative perspective. They also facilitate the statistical analysis of relationships between linguistic phenomena and linguistic or extralinguistic factors at work in language change. By drawing attention to the influence of language use on language structure, and by offering access to often less well-known texts outside the literary canon, historical corpora and other related electronic resources have become of great interest to those working with functional linguistic approaches. They have also contributed to bringing the study of the past and present of a language together by serving as a testing ground for, for instance, modern sociolinguistic theory and by making those interested in present-day grammar look at recent and on-going change in systematic and empirical terms to avoid the pitfalls of anecdotal observation (MAIR, 2008, p. 1111-1112). Access to computerised data has also meant an increase in the awareness of the importance of language-theoretical considerations in linguistic research: it has become much less acceptable to simply collect examples and present them without paying attention to language theory or generalisation than it was in the days of pre-electronic historical language study (RISSANEN forthcoming). Finally, the fact that historical linguists seldom have access to stratified, balanced corpora that would cover the full range of diachrony and/or genres investigated has meant that more open-ended and unbalanced electronic data sources need to be resorted to in search for further materials. Indeed, the work done in the field has made many question the notion of all too restrictive a definition for a 'corpus' that may not serve the broad spectrum of linguistic research as well as a more generous definition often seems to do. Accordingly, in addition to traditional stratified corpora, the present article will consider further electronic resources such as large-scale electronic text collections, electronic text editions, linguistic atlases and dictionaries.

The increasing popularity of corpus linguistic methodology in the study of language change also obviously has to do with the kind of research questions that we can reasonably ask when using historical corpora. Attempts to answer these questions have also contributed to advances in the area. The use of extensive textual evidence was already a landmark of the research carried out on pre-electronic corpora, and changes in "the different ways of saying more or less one and the same thing" had been addressed by scholars back in time, with attention paid to factors taken to explain the loss or emergence of linguistic forms. However, with the advent of electronic corpora, it has been the process of change itself, and the transmission or implementation stages in it that have emerged as perhaps of major interest. To demonstrate how the rivalry of variant forms in, for instance, the development of second-person address pronouns proceeded across time, genre and different groups of language users requires a carefully selected dataset that enables generalisations (cf. WALKER, 2007). This line of research had already been fuelled by the interest felt in the 1970's and 1980's in the question of how language theory could best explain or account for change. Among the influential works in this respect can be mentioned, for instance, Weinreich, Labov and Herzog (1968), Samuels (1972), Lass (1980) and Romaine (1982), all of which paid attention to the importance of the empirical study of language variation and change. Examples of recent work in historical sociolinguistics include the study of the macro-level spread of language change in the Early Modern English period (e.g. NEVALAINEN; RAUMOLIN-BRUNBERG, 2003) and micro-level change with individual language users in focus (e.g. NURMI, NEVALA; PALANDER-COLLIN, 2009). This research has helped trace changes originating 'from below', an area of special interest in terms of actuation and spread of change. In register and genre studies, the development of genres has attracted attention, and the history of written English, for instance, has been approached as the history of registers showing shifting relationships to the more oral style that characterises at least less formal registers of spoken language (BIBER; FINEGAN, 1989; 1992; 1997).

Another boosting factor contributing to the interest felt for historical corpora was the emergence and consolidation of the historical pragmatics approach starting in the 1990s and onward. Since Jucker (1995), historical pragmaticians have found computerised data useful for systematic analysis of historical dialogue features and dialogues (JUCKER; FRITZ; LEBSANFT, 1999b, p. 17; FITZMAURICE; TAAVITSAINEN, 2007; cf. KYTÖ, 2010, p. 33-34). In this approach, pragmatic meanings and the changes in their realisations over time are of interest, as in the study of, for instance, speech acts (e.g. JUCKER; TAAVITSAINEN, 2000; 2008a; 2008b; TAAVITSAINEN; JUCKER, 2007; 2008a; 2008b), and grammaticalisation, pragmaticalisation, and lexicalisation phenomena in the history of English (e.g. Brinton, 1996, 2006). In historical socio-pragmatics, the focus is on pragmatic uses and their developments over time across male and female language users representative of various social ranks (e.g. LUTZKY, 2009; CULPEPER; KYTÖ, 2010). Yet another approach that has encouraged the use of historical corpora includes cognitive semantics and prototype semantics that study the emergence of meanings and their expressions in human cognition, central vs. more peripheral meanings, and changes in these relations over time (e.g. RISSANEN *et al.*, 2007). These are all examples of analytical frameworks where the use of historical corpora and corpus linguistic techniques enables large-scale and sophisticated analyses and adds to the coverage and reliability of results. The criteria adopted for the compilation of corpora also offer a convenient short-cut for investigating the possible influence of extralinguistic factors on developments. Among the texts, of special interest are those reflecting informal, everyday language, or offering access to 'non-standard' language use (CLARIDGE; KYTÖ, 2010). Corpus linguistic methodology also enables statistical analyses that are beyond the traditional manual approach (e.g.

collostructional and keyword analyses, n-grams; for problems in practical applications with historical data, see 3.2).

## 2.2 Types of historical corpora and other electronic resources

According to McEnery and Wilson ([1996] 2001, p. 123), computerised resources and tools used to analyse them have become part of most research on historical linguistics today. Regarding English, there are currently thirty to forty English historical corpora available or underway, amounting to more than 130 million words, excluding the 400-million-word Corpus of Historical American English and the 100-million-word Time Corpus; if we deduct from this figure the 52-million-word Old Bailey Corpus (see below), the materials amount to some 78 million words. In the literature, the available corpora have been deemed to give a fair picture of the development of English vocabulary and grammar from the earliest times to our own days (CLARIDGE, 2008; RISSANEN forthcoming). However, there are gaps in coverage, to be discussed in section 3.1 below. In addition to historical corpora, resources containing historical material come to us in other forms that enable us to use them as corpora. It is often necessary for historical linguists to use various types of electronic (and non-electronic) resources in their hunt for information. This section surveys some of the main resource types by way of a background to the discussion of future desiderata in the field. In addition to stratified multigenre and specialised corpora, attention will be paid to large-scale text collections, electronic text editions, linguistic atlases and dictionaries (for further discussion, see KYTÖ, 2010 and forthcoming).

Multigenre corpora aim at representing a wide variety of registers and language use across several centuries in order to allow investigations of long-term developments in usage. The first stratified electronic historical corpus of English was The Helsinki Corpus of English Texts. Extending from 700's to 1710, this corpus of 1.5 million words spans from the Old English through the Middle English to the Early Modern English period and contains samples of genres such as law, philosophy, history writing, science, handbooks, travelogues, (auto)biographies, fiction, drama, private and official correspondence, and the Bible. A good number of these are represented across the corpus (e.g. law, philosophy, science, handbooks) while others only appear for a certain period or periods (e.g. homilies for the Old and Middle English periods, romances for the Middle English period, and trial proceedings for the Early Modern English period). ARCHER (A Representative Corpus of Historical English

Registers) (1.7 million words) is another multigenre corpus, extending from 1650 to 1990 and containing partly the same genres as the Helsinki Corpus, for instance, science, fiction, drama and correspondence. While the Helsinki Corpus only contains British English texts, ARCHER contains both British and American English texts. Historical corpora are mostly associated with the written medium, and texts that have been taken to reflect past 'spoken' interaction, phonological spellings or orthoepists' comments have been used as a way of obtaining indirect evidence of past spoken language. However, there is an increasing interest in historical corpora containing spoken texts that could provide direct evidence of the spoken medium. The Diachronic Corpus of Present-Day Spoken English (800,000 words) is such a corpus: it contains samples of recent English, drawing from the ICE-GB (the British component of the International Corpus of English (ICE), collected in the early 1990s) and the London-Lund Corpus of Spoken English (late 1960s-early 1980s). This multigenre corpus contains genres such as face-to-face and telephone conversations, broadcast discussions and interviews, spontaneous commentary, parliamentary language, legal cross-examination, and prepared speech.

As the data yielded by multigenre corpora tend to break down across the genres and periods distinguished, multigenre corpora are typically suitable for diagnostic purposes, pointing to trends that can be verified with the help of further data found in specialised corpora, for instance. Specialised corpora tend to focus on a genre (or related genres), a period, a certain aspect of language use, or even a single text or author. Examples of the last-mentioned are the Electronic Beowulf and the Shakespeare Corpus. Other types of specialised corpora have often been compiled to facilitate observing language change from a specific analytical framework (or a number of them). Thus the Corpora of Early English Correspondence (5.1 million words, letters from the early 1400s to 1800) were compiled to allow historical sociolinguistic study; Corpus of Early English Medical Writing 1375-1800 (estimated 3.8 million words, medical texts of various types) for observing stylistic change in early medical English; A Corpus of English Dialogues 1560-1760 (1.2 million words, dialogic texts) to allow the study of early speech-related language; Zurich English Newspaper Corpus (1661-1791) (1.6 million words, newspapers), and the Lampeter Corpus of Early Modern English Tracts (1640-1740) (1.2 million words, pamphlets and other tracts) for studies of language use in the public domain. Examples of period-specific and/or genre-specific corpora are the above-mentioned Dictionary of Old English Corpus in Electronic Form; A Corpus of Nineteenth-Century English (1800-1900, 1 million words, seven genres, British English only); the Time Corpus (or Time Magazine Corpus of American English, 1923-2006, 100 million words); and A Corpus of Historical American English (400+ million words, 1810's-2000's, popular magazines, newspapers, and academic writing). The last-mentioned is also an example of specialised historical corpora that focus on transplanted regional varieties. Among other such corpora can be mentioned A Corpus of Irish English (14th-20th centuries, 550,000 words) and the (Corpus of Oz Early English (1788-1900, 2 million words).

Like present-day corpora, historical corpora can also contain parts-of-speech or other grammatical or textual annotation. Examples of such corpora are the Parsed Corpus of Early English Correspondence (2.2 million words), which is available in plain text files, part-of-speech tagged files, and syntactically parsed files, with metadata about the letters (date, authenticity, recipient classification) and correspondents (name, date of birth, gender, etc.). The annotation scheme used for this corpus had earlier been applied to Penn-Helsinki Parsed Corpus of Middle English (second edition) and the Penn-Helsinki Parsed Corpus of Early Modern English. A remarkably richly annotated and manually checked resource is the above-mentioned Diachronic Corpus of Present-Day Spoken English, which comes with the ICECUP search suite and allows one "to perform a variety of different queries, including using the parse analysis *in* the corpus to construct Fuzzy Tree Fragments *to search* the corpus" (http://www.ucl.ac.uk/english-usage/projects/dcpse/).

In addition to stratified historical corpora proper, electronic versions of early texts have been made available in the form of facsimile or plain text files in huge computerisation projects such as the Literature Online collection (Lion), the Early English Books Online (EEBO), and its chronological sequel the Eighteenth Century Collections Online (ECCO). The Lion collection "offers the full text of more than 350,000 works of poetry, drama and prose in English from the eighth century to the present day", and "more than 800 classic literary essays, from the sixteenth century to the early twentieth". Further, Lion also provides links to more than 8,000 additional electronic texts from third-party internet sites. Importantly, "[a]ll texts are reproduced faithfully from the original printed sources without silent emendation" (http://lion.chadwyck.co.uk/marketing/editpolicy2.jsp). EEBO comprises over 22 million digital page images from "virtually every work printed in England, Ireland, Scotland, Wales and British North America and works in English printed elsewhere from 1473–1700" (http://eebo.chadwyck.com/home). Similarly, ECCO is a large-scale collection, comprising more than 136,000 titles in 26 million digital facsimile pages. ECCO covers a wide range of subject areas, among them literature and language, law, history and geography, social sciences and fine arts, medicine, science and technology, and religion and philosophy (<http://mlr.com/DigitalCollections/products/ecco/>). (For limitations set to searchability, see 3.2.)

The above text collections provide useful material for the study of language change even though they were not compiled for primarily linguistic research. Other such very large-scale collections, although more specialised, include newspaper texts. Among these are the ProQuest Historical Newspapers collection (www.proquest.com) and the Times Digital Archive (www.gale.cengage.com). The former is a massive collection that offers "full-text and full-image articles for [36] significant newspapers dating back to the 18th Century [1764-2008]" and mostly comprises sources representing American English. The latter represents British English and contains over 7.6 million articles published in The Times starting in 1785 over a period of more than 200 years. There are also smaller collections such as North American Review (Library of Congress), Blackwood's Edinburgh Magazine (Bodleian Library online), The Collected Works of Abraham Lincoln (Humanities Text Initiative online, University of Michigan) and American Whig Review (Library of Congress) (for references and further information, see MacQUEEN, 2010). Another specialised large-scale collection is The Proceedings of the Old Bailey, London's Central Criminal Court, 1674 to 1913 (Old Bailey Corpus). The Old Bailey Corpus provides "[a] fully searchable edition of the largest body of texts detailing the lives of non-elite people ever published, containing 197,745 criminal trials held at London's central criminal court" (http://www.oldbaileyonline.org/). The web site provides access to 190,000 images of the original pages of the Proceedings and 4,000 pages of Ordinar's Accounts, in addition to historical, social and other support material. This resource was originally intended for the use of historians, but a project aiming at converting the digitised transcripts into a linguistic corpus is underway at the University of Giessen, Germany (HUBER, 2007): mark-up will be provided to distinguish direct speech from the rest of the text in a 134-million-word section of the full corpus; this section will also be tagged for parts of speech. Sociolinguistic mark-up will be entered for about half of the material qualifying as direct speech (i.e. for ca. 57 million words out of the 113 million words comprising direct speech) (<http://www.uni-giessen.de/oldbaileycorpus/index.php>).

In addition to ready-made large-scale text collections, it is also possible to look for electronic texts on internet sites, for instance at the Project Gutenberg site (<http://www.gutenberg.org/wiki/Main_Page>) or from distribution houses such as the Oxford Text Archive (note that such material may be of uneven reliability in terms of editions used, the accuracy of the text, etc.). The Corpus of Late Modern English Texts, Extended Version (1710-1920) (15 million words) was compiled using texts available in these sources (see De Smet, 2005).

The possibility of combining digital manuscript images with searchable transcriptions and textual annotation has increased the interest in electronic text editions, especially such as are intended to render the original manuscript text as faithfully as possible (for recent work, see e.g HONKAPOHJA; KAISLANIEMI; MARTTILA, 2009, and KYTÖ; GRUND; WALKER forthcoming, and references therein). These editions can be used as electronic corpora and they also lend themselves to further digital applications such as hypertext databases. Compared with most historical corpora based on imprint material, the time-consuming nature of transcription work generally limits the text length of electronic editions. Examples of electronic text editions include collections such as the Corpus of Scottish Correspondence (1500-1730, 256,000 words), An Electronic Text Edition of Depositions 1560-1760 (267,000 words) and The Middle English Grammar Corpus (1100-1500, 450,000 words), and single texts such as Electronic Beowulf and A London Provisioner's Chronicle, 1550-1563, by Henry Machyn. Manuscript-based digitised transcriptions of early texts are also available in linguistic atlases such as A Linguistic Atlas of Early Middle English 1.1 (1150-1325) (c. 650,000 words) and A Linguistic Atlas of Older Scots, Phase 1 (1380-1500), both follow-up projects to the hard-copy Linguistic Atlas of Late Modern English (LALME) (1350-1450), which is being revised and digitised into an e-LALME version.

Electronic dictionaries are powerful tools that facilitate looking up information on words and phraseology. They do not of course generally provide such contexts as full-text corpora do for individual search items, but the information extracted can be used for follow-up searches in historical corpora proper. Large-scale dictionaries, which aim at covering the history of a language's vocabulary, are long-term projects going far back in time. Among such projects are the Oxford English Dictionary Online (OED Online) for English, Der digitale Grimm for German, and Svenska Akademiens ordbok for Swedish.

More specialised electronic dictionaries focus on a certain period as, for instance, the Dictionary of Old English and the Middle English Dictionary, or are digitised versions of early dictionaries such as Samuel Johnson's Dictionary of the English Language (1773 [1755]) (McDERMOTT, 1996). A collection of digitised early dictionaries is available in the Lexicons of Early Modern English (1480-1702) database, a multilingual resource that currently comprises close to 580,000 word entries drawn from 168 searchable lexicons (e.g. monolingual, bilingual, and polyglot dictionaries, hard-word glossaries and spelling lists) digitised from early imprints or manuscripts (LANCASHIRE, 2006).

## 2.3 Tools for historical corpus analysis

The basic tools used by historical corpus linguists do not differ essentially from those used for searching present-day material. Among these tools are word lists and concordances, combined with more sophisticated methods such as collocate, keyword, or n-gram (or lexical bundle or multi-word expression) analysis. Search programs currently available on the market are WordSmith Tools, MonoConc Pro, Corpus Presenter and Xaira. The last-mentioned provides advanced graphical support for investigating results. The powerful statistical computing and graphics program R can also be used to process language data (<http://www.r-project.org/>). (For a useful discussion of data retrieval software, see WYNNE, 2008).

Among the resources that provide a search engine of their own are, for instance, the Penn Parsed Corpora of Historical English and the Parsed Corpus of Early English Correspondence, which have been annotated for the purposes of the CorpusSearch 2 program (<http://corpussearch.sourceforge.net/index.html>), or the Corpus of Irish English, the Middle English Medical Texts and the Early Modern English Medical Texts (parts of the above-mentioned Corpus of Early English Medical Writing), and An Electronic Text Edition of Depositions 1560–1760, which each come with a customised Corpus Presenter application. Another solution has been opted for in the Corpus of Historical American English which can be accessed via a search interface allowing one to investigate, for instance, changes in the frequency of words and phrases, parts of words, grammatical constructions and collocates. Large-scale text collections (Lion, the Old Bailey Corpus) most often provide a search engine of their own. As these collections were not primarily designed for linguistic searches, applying the search engines to solve linguistic research questions seldom works adequately. Overall, using search programs on historical data is not altogether unproblematic, especially as regards spelling variation, a feature characteristic of pre-standard varieties (see 3.2).

## 3. Assessing the field: recent advances and bottleneck areas

As shown above, significant progress has been made in the production of historical corpora and other electronic resources over the past few decades. However, there are still problems in various areas that would benefit from further attention. A number of these will be addressed in the following. To begin with, gaps in the present coverage will be discussed, with special reference to the field of English historical linguistics, again with the aim that similar problem areas could be identified for other languages. Attention will then be drawn to recent advances in the corpus compilation "philosophies" that often lie behind corpus projects and the potential they have for further advances. Related to this, the question of comparability between different corpora will be highlighted, and attention also paid to various linguistico-philological issues in corpus compilation (3.1). Issues with searchability, corpus annotation, and spelling variation, referred to above, will be discussed along with the ways in which problems in these areas hamper the full use of, for instance, statistical tools in the study of language change (3.2). The remaining points taken up pertain to corpus linguistics in general but are nevertheless worth considering as regards historical corpus linguistics, in particular. These include copyright questions, and how to inform the community of linguists and other potential users of the availability and properties of historical corpora (3.3). Finally, a call will be made for enhancing awareness among historical corpus linguists of the benefits brought about by the interdisciplinary framework (3.4).

### 3.1 Resources: potential for enhancement and new projects

Regarding gaps in textual coverage in English historical corpora, according to Rissanen (forthcoming), "[t]he chronological coverage of the corpora is uneven, however, and does not give us a sufficient amount of information on all genres or regional varieties, or the language use of different social groups. More corpora are needed and their use should be made easier and more efficient by new software developments, both as concerns search engines and annotation." Claridge (2008, p. 245-246) goes even farther saying that "[w]hile the textual situation becomes better after the Middle Ages with regard to both amount and variation, the historical corpus linguist will always face shortages of some nature before the late 19th century". Compilers and users of historical corpora need to accept the sad fact that a lot of valuable material has been lost in fires, floods, wars, or in other circumstances (for instance, only very little evidence of English is preserved from the Early Middle English period, 1250-1350, as a consequence of political circumstances that led to Anglo-Norman and French being the languages of the ruling ranks). Also, the time distance between the date of the original text and the copy preserved to us can cover several generations of language users, making it difficult to draw conclusions about usage in the time of the original. This can be the case not only with medieval texts but also even in the early modern period (for instance, many sixteenth-century trial proceedings survive in seventeenth-century copies only, see CULPEPER; KYTÖ, 2010, p. 50-51). Nor are early texts easily accessible, especially if available only in manuscript form. There are also socio-historical and cultural constraints such as poor levels of literacy and writing skills, and limited access to formal education, which hampered the production of early texts. The lower and middle segments of society, in particular, were subject to illiteracy, so the language of the social and educational elite, and especially male writers, tends to dominate in historical corpora leaving language of women and representatives of the lower echelons underrepresented (CLARIDGE, 2008, p. 248). Finally, nor do we always know for certain whether it was a scribe or the ascribed author who produced the text. This can be the case with early letters written in the Middle Ages or with even much later letters. For instance, we have valuable 'non-standard' material in the so-called 'pauper letters' from the eighteenth and early nineteenth century, written by ordinary people on the verge of poverty to their overseers (Sokoll, 2001). An electronic corpus of these letters is now underway (by Mikko Laitinen, see RAUMOLIN-BRUNBERG, 2003), but what will limit the use of the material is that it is often unclear whether a letter was written by the ascribed author or by another person hired to do the job.

It is important that compilers of future historical corpora pay attention to the above problems and that they document their compilation decisions in clear terms in user guides, corpus manuals and like material that will accompany the release versions of the corpora. It would be all too time-consuming and virtually impossible for end-users to replicate the research done to find out about the background of texts included in historical corpora. For instance, early imprints of one and the same work may differ in details owing to compositors having made changes to the type in individual copies. For later verification purposes, it is necessary for the respective corpus file or manual to contain bibliographical reference information on the specific copy used for the corpus. Overall, assessing the reliability and validity of source texts as evidence of language use from the past periods is of prime importance to any historical corpus compilation project. For instance, text editions come in varying quality and based on varying editorial policies. Careful attention needs to be paid to the relationship of text editions to the original texts, and to keeping end-users aware of the value of the evidence drawn from them (for further discussion, see KYTÖ; WALKER, 2003; KYTÖ; PAHTA, forthcoming).

Despite the above considerations, there is a lot of potential in the various corpus compilation "philosophies" to enhance extant historical corpora and to develop new ones. As mentioned above, the first structured historical corpora containing early English were multigenre corpora intended for the study of language variation and change across the centuries. The underlying hypothesis was that comparative analysis of written texts which stand at different distances from speech may help us in our attempts to envisage what past 'spoken' language might have been like and that it is also possible to extrapolate from informal writing about everyday language use (KYTÖ; RISSANEN, 1983; RISSANEN, 1986, 1999). Commendably, such corpora are still being compiled as, for instance, the Leuven English Old to New (LEON) corpus, which is intended to span from the 900's to the twenty-first century (PETRÉ, 2009). The earlier corpora are also being enhanced in view of more sophisticated use, as is the case with for instance ARCHER (YÁÑEZ-BOUZA, 2011).

At the same time projects focusing on specialised corpora have produced a growing body of innovative research in areas such as historical sociolinguistics, genre and register studies, and the study of 'spoken' interaction in the past. All these directions are to be encouraged as the research carried out within these frameworks has significantly added to our knowledge of language history and processes of change. The results obtained in historical sociolinguistics have helped evaluate and re-assess some of the findings presented in modern sociolinguistic research. Similarly, systematic evidence-based genre and register studies have helped map and account for stylistic and grammatical shifts in language use from medieval to modern times in a way that would hardly have been possible without the support of historical corpora. The study of 'spoken' interaction in the past is also of special interest: while dialogic face-to-face interaction has been considered relevant in actuation of change (MILROY, 1992; TRAUGOTT; DASHER, 2002; CULPEPER; KYTÖ, 2010), historical evidence of it has been preserved only in written form. Even though texts containing early speech-related or speech-like language, whether in the form of dialogues (e.g. trial proceedings, drama) or private correspondence, cannot be expected to have preserved speech with the accuracy that modern audio-recording devices do, they are valuable as they can be studied "as communicative manifestations in their own right" (JACOBS; JUCKER, 1995, p. 9). There is also an interest in this approach among those working on the history of other languages than English as can be seen in works such as Collins' 2001 study of speech-reporting strategies in a substantial corpus of medieval Russian trial transcripts, and in articles included in Journal of Historical Pragmatics.

The above-mentioned Diachronic Corpus of Present-Day Spoken English allows the systematic study of change in spoken English in real-time, but only for a relatively brief period of time. More than 130 years have passed since the Chicago Daily Tribune (9 May, 1877) reported on the 'talking-machine' that Thomas Alva Edison was working on and that he later on that year presented as a phonograph, the first device able to record and replay the sound. This leaves us with oceans of material for historical corpus compilers to explore. A fascinating example of a study based on extensive audio-recordings provided by New Zealand's 'mobile disk unit' gives us information on how the earliest New Zealand-born settlers spoke and how this new variety of English first spoken in the 1850s developed (GORDON *et al.*, 2009). Having access to structured sets of early audio-recorded materials would enable real-time and apparent-time research on language change based on direct spoken language evidence. Such corpus compilation projects would contribute to current resources in most valuable ways.

As has been shown above, historical corpora have widened the spectrum of texts beyond those, mainly literary, that have traditionally been considered by language historians. It is desirable that historical corpus compilers continue to explore such materials further. More resources containing women's language, and language of untutored writers, or writers with little formal education are on end-users' wish list. This also holds for resources containing evidence of early 'spoken' interaction, and dialectal, regional or other 'non-standard' usage.

Considering the spread of English as an international world language, there is plenty of room for corpus projects aimed at recording the historical stages of the emergence and subsequent development of various transplanted varieties. It would also be fascinating to have access to materials representative of the development of individual genres or genre families across time periods.

An example of such a project underway is the Corpus of English Religious Prose (KOHNEN, 2007), which aims at documenting the history of English religious writing. On the whole, genres of chronological continuity would merit better attention, among them legal language, history writing, handbooks, science, philosophy, travelogues, (auto)biography, fiction, drama, and verse. As a genre may also change across time as regards stylistic and other conventions, attention should be paid to genre definitions across the diachrony; it may be difficult to see whether what we have at hand is language change or only change in genre conventions (cf., e.g., BIBER; FINEGAN, 1989).

But there is also room for new areas of interest. One so far rather neglected an area is the historical cross-linguistic perspective. Only very little has been done to compile historical parallel corpora that would combine different languages. A step in that direction has been the GerManC project launched at the University of Manchester to compile a representative historical corpus of written German for the years 1650-1800. The project aims at providing "a basis for comparative studies of the development of the grammar and vocabulary of English and German and the way in which they were standardized". For this end, the GerManC corpus has been structured and designed "to parallel that of similar historical linguistic corpora of English, notably the ARCHER corpus". The compilation team are collaborating with representatives of the ARCHER team to maximise the degree of comparability between the corpora. Once complete, the GerManC corpus "will contain 2000-word samples from nine genres: drama, newspapers, sermons, personal letters and journals (to represent orally oriented registers) and narrative prose (fiction and biographies), academic, medical and legal texts (to represent more print-oriented registers)" (http://www.llc.manchester.ac.uk/research/projects/germanc/). Another example is the "Three centuries of drama dialogue: A cross-linguistic perspective" project underway at Uppsala University. In its current pilot stages, this project aims at an English-Swedish Drama Dialogue corpus containing drama texts in English and Swedish from the three periods, 1725-1750, 1825-1850 and 1925-1950. The North Sea area offers ample opportunities for the compilation of interesting cross-linguistic historical corpora that could provide material for comparisons with Germanic and Romance languages. There are also counterparts for comparisons in the form of parallel corpora containing present-day language.

A further neglected area in historical corpus compilation is language teaching. There has been an increasing interest among historical pragmaticians in dialogues found in language teaching books (e.g. HÜLLEN, 1995; WATTS, 1999; for these and further references, see CULPEPER; KYTÖ, 2010, p. 45). A Corpus of English Dialogues 1560-1760 contains didactic works, a subsection of which is devoted to language teaching manuals. Language teaching texts have been separated from the other didactic works in this corpus owing to their special characteristics and socio-historical background. On the one hand, these texts are realistic in their display of language use they aim to teach. On the other hand, they also contain features uncharacteristic of authentic language use situations such as long vocabulary lists (CULPEPER; KYTÖ, 2010, p. 46-48). The target language may also have influenced to varying degrees the dialogues in which the teaching materials are couched (KYTÖ; WALKER, 2006, p. 23; CULPEPER; KYTÖ, 2010, p. 48). The texts included in this corpus were intended to teach English to the French and French to the English, with one text aimed at teaching German to the English. However, the material remains scanty in view of in-depth studies, and given the interest in present-day language teaching materials, more historical texts in searchable form would be welcome. Related to this, one new avenue would be the compilation of corpora containing early grammarians' and orthoepists' works. These have always been of major interest to historical linguists as, among other things, they provide glimpses of contemporaneous views of language use.

Regarding other forms of electronic resources than structured corpora, electronic text editions are an area that would deserve much more attention than is the case today. Libraries, archives and record offices contain great amounts of valuable manuscript material which, if scanned or transcribed, provided with metadata annotation, and, ideally, accompanied by manuscript images or samples of them, would be of the utmost interest to the research community. Transcriptions aiming at rendering the language and other features of the original manuscripts as faithfully as possible within the limitations set by modern typography and electronic processing facilities are to be encouraged (for linguistic annotation, see 3.2). Electronic editions of early imprints would also be welcome, especially in areas such as science and handbooks, where images play an important role and multimodal applications would enhance the value of the material. As for linguistic atlases that contain the texts they are based on, such as A Linguistic Atlas of Early Middle English, the work is only in its infancy. As for the history of English, dialect maps of regions or localities from the Old English and the early modern period would be of great value, to complement the current Middle English atlas projects.

Gaps in coverage often necessitate looking for data from a number of corpora. The question is to what extent corpora compiled on varying principles are comparable. There are examples of corpora that represent as perfect a match as is possible considering that genres may also change in time and that sources such as newspapers may be discontinued. The family of 'Brown corpora' presents a case of a number of matching corpora designed to enable one-to-one comparisons. These corpora follow the one-million-word Brown Corpus (or A Standard Corpus of Present-day Edited American English, for Use with Digital Computers) released in 1964, and include the LOB corpus (or Lancaster-Oslo/Bergen Corpus) of British English (1978), and their counterparts Frown Corpus (Freiburg-Brown Corpus of American English) and F-LOB (Freiburg-LOB Corpus of British English) (1999 original versions, 2007 POS-tagged versions). These match in size and composition, with the only difference that while the Brown and LOB corpora were compiled to represent language from 1961, Frown and F-LOB include sources 30 years after, from 1991. Two further family members are underway, the BLOB-1931 corpus sampled from the period 1928-1934, with a focus on 1931, and another from 1901, to provide further sources for comparison on the British English axis. These corpora allow systematic study of for instance recent and on-going change in English grammar, and the linguistic and social factors that are influencing processes of change (see, e.g., LEECH *et al.*, 2009).

However, gaps in textual representation, differences in period divisions and classification of social strata, and other such features usually entail that comparisons across corpora can seldom be straightforward; instead, further consideration and adjustments are needed on the part of end-users. It is of course desirable that future corpus compilers pay attention to previous compilation plans when launching their projects in order to facilitate research across historical corpora. This is also of prime importance for future annotation projects.

## 3.2 Issues of searchability and corpus annotation

In addition to enhancing extant resources and creating new ones, compilers and end-users of historical corpora would need to collaborate with computational linguists to a greater extent than has been the case so far. There is a general lack of consensus on platforms, and searching historical corpora, large-scale text collections and electronic dictionaries is not always as unproblematic as one could wish.

As mentioned above, many of the search engines that come with large-scale collections are not primarily intended for linguistic study but rather for identifying quotations in literary works (e.g. Lion) or for extracting historical information (e.g. the Old Bailey Corpus). Similarly, the EEBO and ECCO images are searchable only in the sense that one can look for a word or phrase and get a list of the full-text contexts of all instances, with the possibility of clicking over to the facsimile of the page (the same goes for ECCO). On the other hand, the results cannot be concordanced, and one has to find ways to determine the approximate number of words in the corpus in order to approximate an incidence figure for the expression at hand (for such techniques applied to very large-scale historical newspaper collections, see MacQueen, 2010, chapter 5). However, the bibliographical information on the EEBO texts can be searched. In addition, the Text Creation Partnership (TCP) at the University of Michigan has so far stored some 25,000 books in the collection in the form of searchable plain texts. Further, the search engine accompanying a central source such as the Corpus of Middle English Prose and Verse ("at present, sixty-two texts are available; about eighty others will be added soon, with another 150 smaller texts in preparation", see http://quod.lib.umich.edu/m/mec/about/) lists occurrences text by text separately, as they are not given conveniently in one and the same file. This invaluable resource and many others such as the Dictionary of Old English Corpus would benefit from a retrieval program that would make it easier to sort the texts by date, dialect, and genre, and to create subcorpora according to these parameters (Rissanen forthcoming). As implied above, it is also often surprisingly difficult, if not altogether impossible, to obtain word counts for each text (needed for counting the incidence figures for a linguistic feature per a certain text length, for instance) or download them for further *in situ* annotation or other processing.

The search programs available can be used for many basic and even advanced search tasks, but depending on the research questions and the type of material one is working on, professional computer programming skills are often needed to extract the kind of data one is after. Interesting results can also be achieved by exploring methodologies applied in other fields. For instance, as there is generally no coding for pragmatic phenomena such as speech acts in historical corpora, historical pragmaticians will need to develop methodologies to locate their data. Accordingly, for their study of compliments and gender in the history of English, Taavitsainen and Jucker  developed an "ethnographic" method: to pin down "what was considered proper and polite, particularly in association with gender", they collected speech-act labels such as 'compliment', 'compliments', 'complement', 'complements' and their spelling variants (TAAVITSAINEN; JUCKER, 2008b, p. 207, with reference to ROMAINE, 2003, p. 104-105). The aim of the searches was "to locate relevant passages for qualitative assessment"; TAAVITSAINEN; JUCKER, 2008b, p. 208; for methodology, see also JUCKER; SCHNEIDER; TAAVITSAINEN; BREUSTEDT, 2008). The method has also been applied successfully to the study of apologies (JUCKER; TAAVITSAINEN, 2008b).

The searchability of a corpus is crucially dependent on how the corpus has been annotated. Again, there is a lack of consensus on this point, and compilers of historical corpora have been slow or even reluctant to apply standards such as the Text Encoding Initiative (TEI) Guidelines (P5) (<http://www.tei-c.org/index.xml>). Many of the better known corpora are annotated for the main textual features but not all, and not as exhaustively as could have been the case. The features that an end-user would need to be able to learn about with little effort include, for instance, the title of the text, date(s) (if composition and copy diverge), text-type/genre, content description, level of formality, medium (written/spoken), language use (prose/verse; dialect; foreign languages etc.), authenticity of the document (autograph/copy etc.), references to established citation systems, the original/edition used for the corpus, and other bibliographical information. Certain author properties would also be useful information: age, gender, social rank/class, parentage, education, profession(s), residence, dialect, type of possible author-recipient relationship (if interactive) etc. Coding plans paying attention to both the writer/speaker and the addressee/interlocutors are to be encouraged. For instance, the Sociopragmatic Corpus, part of A Corpus of English Dialogues 1560-1760, has been annotated for both speaker and addressee properties, turn by turn. Interrogating this corpus for advanced searches requires a customised search engine; a similar approach was adopted when coding the speaker turns for the above-mentioned English-Swedish drama corpus.

Enhancing the searchability of historical electronic resources is not a straightforward task. There are a number of factors complicating annotation efforts, and it is no surprise that the amount of grammatically annotated historical material is still relatively scant in comparison to corpora containing annotated present-day material. There are historical corpora that have been tagged completely by manual means, for instance, the German Bonner Frühneuhochdeutsch Korpus (CLARIDGE, 2008, p. 254-255), but resorting to automatic tagging and manual checking to correct tagging errors has also been attempted. As tagging systems and software have mostly been developed for present-day standard varieties, they run into problems when trying to deal with historical varieties that tend to vary internally and present unanticipated language structure and spelling variation. Compared with modern texts that can be tagged automatically at the rate of about 96-97%, Early Modern English material presents lower rates, from 80% to 95%, depending on the date of the text (CLARIDGE, 2008, p. 254). Manual checking and correction is usually required to produce more reliable results; for instance, a considerable amount of manual labour was needed to annotate the York-Helsinki Parsed Corpus of Old English Poetry, the York-Helsinki Parsed Corpus of Old English Prose, the Penn-Helsinki Parsed Corpus of Middle English, the Penn-Helsinki Parsed Corpus of Early Modern English and the Penn Parsed Corpus of Modern British English (1700-1914, close to 1 million words). Syntactic annotation (parsing) in the three Penn Parsed Corpora of Historical English "permits searching not only for words and word sequences, but also for syntactic structure" (<http://www.ling.upenn.edu/hist-corpora/>). In addition to syntactic annotation, the Parsed Corpus of Early English Correspondence contains parts-of-speech tagging.

Examples of semantic tagging of historical data are few. A notable exception is the Mitterhochdeutsche Begriffsdatenbank (Middle-High German Conceptual Database), which "provides very powerful **search functions** for a large number of the most important works of Middle-high German literature, with linguistic and semantic search criteria" and "a **Wordindex with Concepts** for the lemmas and words in the database" (http://mhdbdb.sbg.ac.at:8000/index.en.html). There has also been pilot work on Early Modern English newsbooks (613,000 words) by (re)training the UCREL Semantic Analysis System (USAS) to cope with this historical variety with the help of the web-based corpus tool Wmatrix (ARCHER; MCENERY; RAYSON; HARDIE, 2003). This tool, and the subsequent Wmatrix2, was originally developed for modern varieties, so the mismatch between the tags adopted for modern texts and those required by the historical material caused some problems. Similarly, the tool had difficulties in dealing with automated grammatical annotation and variant spellings. By way of remedy, the historical validity of the semantic tag set will be improved in future work with the help of the Historical Thesaurus of English (<http://libra.englang.arts.gla.ac.uk/historicalthesaurus/aboutproject.html>) and by pre-processing the texts to be tagged with a variant spelling detector (VARD, see below) (ARCHER, forthcoming). Semantic tagging of historical texts is clearly a field full of promise and in need of further work.

As seen above, spelling variation presents a problem for automatic annotation and searching of historical texts, and there has been some tension between the respect felt by historical linguists for the source text and the demands set by searchability. Only a little over a decade ago, we could read that "[i]n English studies, normalization and/or regularization have never been popular. As to their role in machine-readable corpus compilation, the common opinion seems to be that compilers ought to reproduce the specific features of their source text and not smooth them away. In line with this common understanding, hardly any studies concerning normalization or regularization can be found" (MARKUS, 1997, p. 211). To normalise or not to normalise, that was the hotly debated question for quite some time, with those remaining in the minority who advocated the need for normalised versions of the text. Over the past few years, interest in techniques such as keyword and n-gram analyses has certainly promoted the awareness of the value of texts displaying regularised spelling. One way out of the faithfulness *vs.* ease of retrievability dilemma is to represent both original and regularised spelling versions of the corpus, through an annotation system (as in the Lancaster Newsbook Corpus), or through a multi-level architecture, or through a link to a normalised index.

Also, over the past few years, significant advances have been made in variant spelling research with the help of the Variant Detector (VARD) computer program (<http://ucrel.lancs.ac.uk/VariantSpelling/>; see, also, RAYSON *et al.*, 2007). The current version, VARD2, "is intended to be a pre-processor to other corpus linguistic tools such as keyword analysis, collocations and annotation (e.g. POS and semantic tagging), the aim being to improve the accuracy of these tools" (<http://www.comp.lancs.ac.uk/~barona/vard2/>) (see BARON; RAYSON, 2008). The approach is to produce a list of variant spellings, which are manually matched to normalised forms. The variant detector computer program inserts modern equivalents of these forms when they appear in a given text, while preserving the original variant. This approach proved to be very effective. So far over 50,000 variants have been identified from analysis of different historical texts, and empirical studies of spelling variation across the sixteenth to the nineteenth centuries have been carried out. Even though the tool was designed specifically to deal with Early Modern English spelling variation, it has the potential to work on any form of spelling variation and in any language after training the program with a relevant dictionary and spelling rules. The program has already been applied to for instance A Corpus of English Dialogues 1560-1760, the Corpora of Medical Writing, ARCHER, the Innsbruck Computer Archive of Machine-Readable English Texts, the Lampeter Corpus, the Shakespeare Corpus, and EEBO texts to quantify the level and development of spelling variation in the history of English, and to identify spelling patterns across periods and genres (BARON; RAYSON; ARCHER, 2009a, 2009b; BARON; RAYSON, 2009). Clearly, tools such as VARD2 show the way to future development of software and have great potential to enhance the searchability of historical texts.

Having access to normalised spelling versions of historical corpora would thus facilitate the use of sophisticated statistical analyses. For instance, keyword analyses can be used to study the various ways in which texts function, their related semantic spaces and collocational patterns (WYNNE, 2008, p. 730-734; ARCHER, 2009). Similarly, n-gram analyses based on multi-word sequences located by the computer can be used to study recurrent phraseology across the history of a language (for the principle, see WYNNE, 2008, p. 734-735; on lexical bundles in Early Modern vs. Present-day English trials and play texts, see CULPEPER; KYTÖ, 2010, chapter 5). Further, by using a data-driven bottom-up clustering method Gries and Hilpert (2008) identified historical stages in the data based on differing quantitative distributions. The data, originally collected and exploited for Hilpert (2006), had been drawn from the Penn-Helsinki Parsed Corpus of Early Modern English and the Corpus of Late Modern English Texts, with the different spelling variants harmonised to their present-day counterparts (Gries and Hilpert, 2008: 65). The study showed that, for instance in the case of the verbal complementation of 'shall', the three consecutive 140-year periods that had been distinguished as a result of pooling together the original six successive 70-year periods in the corpora did not tally with the way in which the data actually distributed, falling instead into two 180-year groups in quantitative terms. Discoveries such as these are important in that they enable language historians to gain fresh insights and approach language change from a novel perspective. Clearly, developing such techniques, and providing versions of historical corpus texts that enable their use, are among the top priorities in historical corpus linguistics.

### 3.3 Access to and information on historical electronic resources

Copyright restrictions are an unquestionable bottleneck in the corpus compilation effort, and historical corpora are no exception in this respect. Applying for permission to use and distribute texts in electronic form can be a time-consuming and costly enterprise. Libraries and archives may sometimes be much more forthcoming than publishing houses. Some improvement has been shown recently by, for instance, the Wellcome Library in London, where a generous approach has been adopted for granting permission to use text and images; the British Library and local archives also tend to be generous, apart from requests concerning images, whose use and distribution usually cost considerable sums. Historical corpus compilers are fortunate in that a lot of material has fallen out of copyright. One solution might be to work with editions that are out of copyright, but a potential drawback is that such sources may reflect out-dated linguistic evidence. Also, even though early imprints have fallen out of copyright, libraries usually stipulate that no material from them be distributed to a third party without due application for permission. Compilers of historical corpora have adopted various solutions to the copyright problem, and some of them are worth discussing in the present context.

One way has been, if perhaps only for a transitional period, to publish those parts of the corpus for which copyright is available, as has been done with the Corpus of Early English Correspondence Sampler, which contains half a million of the overall 2.6 million words included in the original Corpus of Early English Correspondence; the rest of the materials could be consulted on an in-house basis. This was also the method applied to the sampler versions of the Innsbruck Computer Archive of Machine-Readable English Texts corpora. A further solution has been to aim at international collaboration within which resources can be shared on a collaborative basis; an example of this is the ARCHER consortium, which pools a number of scholars in many countries in Europe and in the U.S. and, even though no material can be distributed, the consortium is able to offer access to the materials on an in-house basis (YÁÑEZ-BOUZA, 2011). Yet another way is the one chosen for the Time Corpus and the Corpus of Historical American English: the corpus texts are made searchable via a web-based interface that enables a wide range of queries with KWIC displays showing the hit word(s) surrounded by 40 to 60 words or 180 to 200 words in expanded view. This solution is allowed by U.S. copyright law when no more than a certain percentage of each text is displayed to the end-user and when the original text cannot be cut and pasted together from the concordance lines. Even though the raw texts have not been made available, there is great search potential in the solution adopted (DAVIES, 2010, p. 414). Efforts to solve copyright problems will continue to be an important part of the historical corpus compilation initiative.

It is not always easy to obtain accurate and up-to-date information on electronic resources regarding whether the work on them has been completed or is still underway, for example. A recent tool designed to distribute information on English language corpora is the Corpus Resource Database (CoRD) web site at the VARIENG research unit at the University of Helsinki (<http://www.helsinki.fi/varieng/CoRD/index.html>). All descriptions have been submitted or approved by the compilers of each corpus. Each entry contains a set of core information, including a brief description of the corpus, its contents and structure, the names of the compilers, recommended reference line, copyright details, and availability. Other useful information is also offered, including the principles followed in the compilation of the corpus, its annotation conventions, and a bibliography of research conducted using a particular corpus. Compilers of English language corpora can be encouraged to send descriptions of their corpora to the site, and one would welcome similar initiatives for other languages.

### 3.4 Interdisciplinary considerations

There has been an increasing interest in corpus linguistic techniques among, for instance, literary scholars, discourse analysts, historians and ethnographers. This interdisciplinarity is natural in view of the present trend in historical linguistic research which emphasizes the influence of extralinguistic factors on variation and change in the history of a language. Large-scale text collections have proved useful especially for literary scholars to work on but smaller corpora can also be useful objects of study. Corpora containing full texts, such as the Corpus of Middle English Prose and Verse and the Innsbruck Computer Archive of Machine-Readable English Texts offer valuable material for literary and socio-historical research. Electronic editions such as the An Electronic Text Edition of Depositions 1560-1760 (ETED) which make available transcriptions of early official documents are of interest not only to historical linguists but also to legal and social historians.

Overall, the use made of electronic historical texts is diversifying, and it would benefit the research community if collaboration were increased and efforts pooled across disciplinary borders (WYNNE, 2010, p. 425). For instance, historical linguists have a lot to learn from the methodologies applied by social, political, legal, cultural, and other historians, and from the results they have obtained in their research. In terms of software and other developments, historical corpus linguists should perhaps be more active about reaching out and making their voices heard (CURZAN forthcoming). This would make it easier to make innovative use of even resources that have not necessarily been developed for linguistic research in the first place.

## 4. Outlook: prospects of historical corpus linguistics

The future prospects of historical corpus linguistics look favourable. As for the English language, there are already vast amounts of digitised material enabling the study of not only the history of the language and literature but also of various aspects of social, political and cultural history in the English-speaking parts of the world. There is also a growing interest in corpus compilation and exploitation and there are also many other areas for further work are many. These aspirations are becoming increasingly felt for many other languages as well. Such positive developments in the field are very much the result of a large body of inspiring research carried out on the extant resources so far. But there is nevertheless plenty of room for further work. Historical corpus linguistics is still very much in the stage where new and exciting discoveries are made but less attention is being paid to the synergetic effects that will become manifest only when resources and research agendas are pooled, and collaboration is extended across interdisciplinary borders.

By way of summary, the proposed list of desiderata for future developments in historical corpus linguistics is here divided into three overarching categories: i) enhancing and adding to the resources and methodologies for studying long-term and recent change, ii) ensuring comparability and links across corpora, other electronic resources, and software, and iii) increasing our knowledge of the sociohistorical and cultural context of corpus texts, with special reference to interdisciplinary considerations. We would benefit from creating further resources that contain everyday, colloquial, utilitarian or non-standard language, spoken and speech-related language, language of women and lower social ranks, language representative of early transplanted varieties and their pidgin and creole-based off-shoots, cross-linguistic material, and early manuscript material in transcriptions faithful to their respective source texts. Further, the present wish list also includes developing linguistically and historically responsible corpus compilation strategies and new corpus compilation "philosophies" aiming at novel explanatory models. This means paying special attention to extralinguistic and linguistic annotation, handling spelling variation, and developing search tools and statistical approaches well suited for interrogating and analysing early texts.

## References

### Corpora and other electronic resources

ARCHER (A Representative Corpus of Historical English Registers), version 3.1. 2006. See http://www.llc.manchester.ac.uk/research/projects/archer/.

BLOB-1931 corpus. In progress. Compiled by Geoffrey Leech, Paul Rayson and Nick Smith. See http://www.helsinki.fi/varieng/CoRD/corpora/BLOB-1931/index.html.

Bonner Frühneuhochdeutsch Korpus. See http://korpora.zim.uni-duisburg-essen.de/Fnhd/.

Brown Corpus (A Standard Corpus of Present-day Edited American English, for Use with Digital Computers). 1964 (original version). Compiled by W. Nelson Francis and Henry Kučera (Brown University, Providence, Rhode Island). See http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/index.html.

Corpus Inscriptionum Latinarum. See http://cil.bbaw.de/.

Corpus of Early English Correspondence. 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin (Department of English, University of Helsinki). For the corpus family, see http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html.

Corpus of Early English Correspondence Sampler. 1998. Compiled by Jukka Keränen, Minna Nevala, Terttu Nevalainen, Arja Nurmi, Minna Palander-Collin and Helena Raumolin-Brunberg (Department of English, University of Helsinki).

Corpus of Early English Medical Writing 1375–1800. In progress. Compiled by Irma Taavitsainen, Päivi Pahta et al. (University of Helsinki). See Middle English Medical Texts and Early Modern English Medical Texts.

A Corpus of English Dialogues 1560–1760. 2006. Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University). See http://www.helsinki.fi/varieng/CoRD/corpora/CED/index.html.

Corpus of Historical American English. 2010. Compiled by Mark Davies (Brigham Young University). See http://corpus.byu.edu/coha/.

A Corpus of Irish English. 2003. Compiled by Raymond Hickey (University of Duisburg-Essen). See http://www.uni-due.de/CP/CIE.htm.

Corpus of Late Modern English Texts (Extended Version). 2006. Compiled by Hendrik De Smet (Department of Linguistics, University of Leuven). See http://www.helsinki.fi/varieng/CoRD/corpora/CLMETEV/.

Corpus of Middle English Prose and Verse. See http://quod.lib.umich.edu/m/mec/about/.

A Corpus of Nineteenth-century English. Compiled by Merja Kytö (Uppsala University) and Juhani Rudanko (University of Tampere). See Kytö, Merja, Juhani Rudanko and Erik Smitterberg (eds.), Nineteenth-century English: Stability and Change. Cambridge: Cambridge University Press, 2006.

Corpus of Oz Early English. 1998–2004. Compiled by Clemens Fritz (Free University of Berlin). See http://www.helsinki.fi/varieng/CoRD/corpora/COOEE/.

Corpus of Scottish Correspondence, 1500–1715. Compiled by Anneli Meurman-Solin (University of Helsinki). See http://www.helsinki.fi/varieng/CoRD/corpora/CSC/index.html.

The Diachronic Corpus of Present-day Spoken English. 2010. See http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm.

Dictionary of Old English. See http://www.doe.utoronto.ca/.

Dictionary of Old English Corpus in Electronic Form. 2004. Compiled by Antonette diPaolo Healey, Dorothy Haines, Joan Holland, David McDougall, Ian McDougall and Xin Xiang (University of Toronto). See http://www.doe.utoronto.ca/pub/corpus.html; for earlier versions, see http://www.doe.utoronto.ca/pub/pub.html; see, also, http://www.doe.utoronto.ca/.

Der digitale Grimm. See http://www.lehrer-online.de/digitaler-grimm.php.

Early English Books Online (EEBO). See http://eebo.chadwyck.com/home.

Early Modern English Medical Texts. 2010. Compiled by Irma Taavitsainen, Päivi Pahta, Turo Hiltunen, Ville Marttila, Martti Mäkinen, Maura Ratia, Carla Suhr and Jukka Tyrkkö. CD-ROM with software by Raymond Hickey included in Irma Taavitsainen and Päivi Pahta (eds.), Early Modern English Medical Texts: Corpus Description and Studies. Amsterdam: John Benjamins. See http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/EMEMTindex.html. See, also, Corpus of Early English Medical Writing.

ECCO, see Eighteenth Century Collections Online.

EEBO, see Early English Books Online.

Eighteenth Century Collections Online (ECCO). See http://www.gale.cengage.com/pdf/facts/ECCO.pdf.

e-LALME. See http://www.ling.ed.ac.uk/research/ihd/projectsX.shtml.

Electronic Beowulf. 2003. Edited by Kevin Kiernan. See http://www.uky.edu/~kiernan/eBeowulf/guide.htm.

An Electronic Text Edition of Depositions 1560–1760. Forthcoming (2011). See Kytö, Merja, Peter J. Grund and Terry Walker, Testifying to Language and Life in Early Modern England. Including a CD containing An Electronic Text Edition of Depositions 1560–1760 (ETED). Amsterdam/Philadelphia: John Benjamins.

English Books Online (EEBO). See http://lion.chadwyck.co.uk.

English-Swedish Drama Dialogue. In progress. Compiled by Linnéa Anglemark, Merja Kytö, Ulla Melander Marttala and Mats Thelander (Department of English and Department of Scandinavian Languages, Uppsala University).

F-LOB (Freiburg-LOB Corpus of British English). 1999 (original version), 2007 (POS-tagged version). Original version compiled by Christian Mair (Albert-Ludwigs-Universität Freiburg); POS-tagged version compiled by Christian Mair (Albert Ludwigs-Universität Freiburg) and Geoffrey Leech (Lancaster University). See http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/index.html.

Frown Corpus (Freiburg-Brown Corpus of American English). 1999 (original version), 2007 (POS-tagged version). Original version compiled by Christian Mair (Albert-Ludwigs-Universität Freiburg); POS-tagged version compiled by Christian Mair (Albert Ludwigs-Universität Freiburg) and Geoffrey Leech (Lancaster University). See http://www.helsinki.fi/varieng/CoRD/corpora/FROWN/index.html.

GerManC. In progress. Compiled by Martin Durrell, Paul Bennett, Silke Scheible and Richard J. Witt. See http://www.llc.manchester.ac.uk/research/projects/germanc/.

The Helsinki Corpus of English Texts. 1991. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English) (Department of English, University of Helsinki). See http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html.

Historical Thesaurus of English. See http://libra.englang.arts.gla.ac.uk/historicalthesaurus/aboutproject.html.

Innsbruck Computer Archive of Machine-Readable English Texts. Compiled by Manfred Markus (University of Innsbruck). See, e.g., http://www.anglistikguide.de/cgi-bin/ssgfi/anzeige.pl?db=lit&ew=SSGFI&nr=000753.

International Corpus of English (ICE). See http://ice-corpora.net/ice/.

Johnson, Samuel, see McDermott 1996.

Lampeter Corpus of Early Modern English Tracts. See http://www.helsinki.fi/varieng/CoRD/corpora/LC/index.html and http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM.

Lancaster Newsbook Corpus. See http://www.lancs.ac.uk/fass/projects/newsbooks/.

Lexicons of Early Modern English. See http://leme.library.utoronto.ca/public/.

LOB (Lancaster-Oslo/Bergen Corpus). 1976 (original version), 1986 (POS-tagged version). Compiled by Geoffrey Leech, Stig Johansson, Knut Hofland and Roger Garside. See http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html.

A Linguistic Atlas of Early Middle English (1150–1325), version 2.1. 2008. Compiled by Margaret Laing and Roger Lass (Edinburgh: The University of Edinburgh). See http://www.lel.ed.ac.uk/ihd/laeme1/laeme1.html.

A Linguistic Atlas of Older Scots, Phase 1 (1380–1500). 2008 (current version). © 2007 Edinburgh: The University of Edinburgh. See http://www.helsinki.fi/varieng/CoRD/corpora/LAEME/index.html and http://www.lel.ed.ac.uk/ihd/laos1/laos1.html.

Lion, see Literature Online.

Literature Online (Lion). See http://lion.chadwyck.co.uk/.

A London Provisioner's Chronicle, 1550–1563, by Henry Machyn: Manuscript, Transcription, and Modernization. 2006. Edited by Richard W. Bailey, Marilyn Miller and Colette Moore. Ann Arbor, Michigan: University of Michigan Press; Scholarly Publishing Office of the University of Michigan University Library. See http://quod.lib.umich.edu/m/machyn/.

London-Lund Corpus of Spoken English. 1980 (original version). Compiled by Jan Svartvik. See http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM.

McDermott, Ann (ed.). 1996. Samuel Johnson: Dictionary of the English Language on CD-ROM. Cambridge: Cambridge University Press. See http://xml.coverpages.org/cup-johnson.html.

Middle English Dictionary. 2001. See http://quod.lib.umich.edu/m/med/.

The Middle English Grammar Corpus. See http://www.arts.gla.ac.uk/SESLL/EngLang/ihsl/projects/MEG/meg.htm.

Middle English Medical Texts. 2005. Compiled by Irma Taavitsainen, Päivi Pahta and Martti Mäkinen. CD-ROM with software by Raymond Hickey. Amsterdam: John Benjamins. See http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/MEMTindex.html. See, also, Corpus of Early English Medical Writing.

Mitterhochdeutsche Begriffsdatenbank. See http://mhdbdb.sbg.ac.at:8000/index.en.html.

Old Bailey Corpus. See The Proceedings of the Old Bailey, London's Central Criminal Court, 1674 to 1913 available at http://www.oldbaileyonline.org/. See, also, http://www.uni-giessen.de/oldbaileycorpus/index.php.

Oxford English Dictionary Online (OED Online). See http://www.oed.com/.

Oxford Text Archive. See http://ota.ahds.ac.uk/.

Parsed Corpus of Early English Correspondence, parsed version. 2006. Annotated by Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki.

Parsed Corpus of Early English Correspondence, tagged version. 2006. Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki.

Parsed Corpus of Early English Correspondence, text version. 2006. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin, with additional annotation by Ann Taylor. Helsinki: University of Helsinki and York: University of York.

Penn Parsed Corpus of Modern British English. See http://www.ling.upenn.edu/hist-corpora/.

Penn-Helsinki Parsed Corpus of Early Modern English. See http://www.ling.upenn.edu/hist-corpora/.

Penn-Helsinki Parsed Corpus of Middle English (second edition). See http://www.ling.upenn.edu/hist-corpora/.

Project Gutenberg. See http://www.gutenberg.org/wiki/Main_Page.

ProQuest Historical Newspapers. See www.proquest.com.

Shakespeare Corpus. See, e.g., http://www.lexically.net/downloads/corpus_linguistics/shakespeare_corpus_readme.txt.

Sociopragmatic Corpus, a Specialised Sub-Section of A Corpus of English Dialogues 1560–1760. 2007. Compiled by Jonathan Culpeper and Dawn Archer (Lancaster University).

Svenska Akademiens ordbok. See http://g3.spraakdata.gu.se/saob/.

Thesaurus Linguae Graecae. See http://www.tlg.uci.edu/.

Time Corpus (Time Magazine Corpus of American English). 2007. Compiled by Mark Davies (Brigham Young University). See http://corpus.byu.edu/time/.

Times Digital Archive. See www.gale.cengage.com.

York-Helsinki Parsed Corpus of Old English Poetry. 2001. Compiled by Susan Pintzuk and Leendert Plug (University of York). See http://www-users.york.ac.uk/~lang18/pcorpus.html.

York-Helsinki Parsed Corpus of Old English Prose. 2003. Compiled by Ann Taylor, Anthony Warner, Susan Pintzuk, Frank Beths (University of York). See http://www-users.york.ac.uk/~lang22/YcoeHome1.htm.

Zurich English Newspaper Corpus. Version 1.0. 2004. Compiled by Udo Fries, Hans Martin Lehmann, Beni Ruef, Peter Schnieder, Patrick Studer, Caren auf dem Keller, Beat Nietlispach, Sandra Engler, Sabine Hensel and Franziska Zeller (University of Zurich). See http://www.helsinki.fi/varieng/CoRD/corpora/ZEN/index.html.

## Other references

ARCHER, D. (Ed.). *What's in a word-list?* Investigating word frequency and keyword extraction. Farnham: Ashgate, 2009.

ARCHER, D. Data retrieval in a diachronic context: The case of the historical English courtroom. In: NEVALAINEN, T.; TRAUGOTT, E. (Ed.). *A handbook to the history of English*. Oxford: Oxford University Press, forthcoming.

ARCHER, D.; McENERY, T.; RAYSON, P.; HARDIE, A. Developing an automated semantic analysis system for Early Modern English. In: ARCHER, D.; RAYSON, P.; WILSON, A.; McENERY, T. (Ed.). Corpus Linguistics 2003 Conference. *Proceedings*... (UCREL technical paper number 16). Lancaster: UCREL, Lancaster University, 2003.

BARON, A.; RAYSON, P. VARD 2: A tool for dealing with spelling variation in historical corpora. In: Postgraduate Conference in Corpus Linguistics. *Proceedings*... Aston University, Birmingham, 22 May 2008. Retrieved May 5, 2011 from http://acorn.aston.ac.uk/conf_proceedings.html.

BARON, A.; RAYSON, P. Automatic standardization of texts containing spelling variation, how much training data do you need? In: MAHLBERG, M.; GONZÁLEZ-DIAZ, V.; SMITH, C. (Ed.). Corpus Linguistics Conference, CL2009. *Proceedings*... University of Liverpool, UK, 20-23 July 2009. Available at: <http://ucrel.lancs.ac.uk/publications/cl2009/>. Retrieved: May 5, 2011.

BARON, A.; RAYSON, P.; ARCHER, D. Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies*, v. 20, n. 1, p. 41-67, 2009a.

BARON, A.; RAYSON, P.; ARCHER, D. Automatic standardization of spelling for historical text mining. In: 'Digital Humanities 2009'. *Proceedings*... University of Maryland, USA, 22-25 June 2009. 2009b.

BIBER, D.; FINEGAN, E. Drift and the evolution of English style: A history of three genres. *Language*, v. 65, n. 3, p. 487-517, 1989.

BIBER, D.; FINEGAN, E. 1992. The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries. In: RISSANEN, M.; IHALAINEN, O.; NEVALAINEN, T.; TAAVITSAINEN, I. (Ed.). *History of Englishes*: new methods and interpretations in historical linguistics. Berlin/New York: Mouton de Gruyter, 1992. (Topics in English Linguistics 10)

BIBER, D.; FINEGAN, E. Diachronic relations among speech-based and written registers in English. In: NEVALAINEN, T.; KAHLAS-TARKKA, L. (Ed.). *To explain the present*: studies in the changing English language in honour of Matti Rissanen. Helsinki: Société Néophilologique, 1997. (Mémoires de la Société Néophilologique 52)

BRINTON, L. J. *Pragmatic markers in English*: grammaticalization and discourse functions. Berlin/New York: Mouton de Gruyter, 1996. (Topics in English Linguistics 19)

BRINTON, L. J. Pathways in the development of pragmatic markers in English. In: van KEMENADE, A.; LOS, B. (Ed.). *The handbook of the history of English*. London: Blackwell, 2006.

CLARIDGE, C.. Historical corpora. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

CLARIDGE, C.; KYTÖ, M. 2010. Non-standard language in earlier English. In: HICKEY, R. (Ed.). *Varieties of English in writing*. The written word as linguistic evidence. Amsterdam/Philadelphia: John Benjamins, 2010. (Varieties of English Around the World G41)

COLLINS, D. E. *Reanimated voices*. Speech reporting in a historical-pragmatic perspective. Amsterdam/Philadelphia: John Benjamins, 2001. (Pragmatics & Beyond New Series 85)

CORPUS PRESENTER. Available at: <http://www.uni-due.de/CP/>.

CORPUS RESOURCE DATABASE (CoRD). Available at: <http://www.helsinki.fi/varieng/CoRD/index.html>.

CULPEPER, J.; KYTÖ, M. *Early Modern English dialogues*: spoken interaction as writing. Cambridge: Cambridge University Press, 2010.

CURZAN, A. English historical corpora in the classroom: the intersection of teaching and research. *Journal of English Linguistics*, v. 28, n. 1, p. 77-89, 2000.

CURZAN, A. Historical corpus linguistics and evidence of language change. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

CURZAN, A. The electronic life of texts: Insights from corpus linguistics for all fields of English. In: KYTÖ, M. (Ed.). *English corpus linguistics*: crossing paths. Amsterdam: Rodopi, forthcoming.

DAVIES, M. More than a peephole: using large and diverse online corpora. In: POPE, C. W. (Ed.). Special issue on the bootcamp discourse and beyond. *International Journal of Corpus Linguistics*, v. 15, n. 3, p. 412-418, 2010.

DE SMET, H. A corpus of Late Modern English texts. *ICAME Journal*, v. 29, p. 69-82, 2005.

FITZMAURICE, S. M.; TAAVITSAINEN, I. (Ed.). *Methods in historical pragmatics*. Berlin/New York: Mouton de Gruyter, 2007. (Topics in English Linguistics 52)

GORDON, E.; CAMPBELL, L.; HAY, J.; MACLAGAN, M.; SUDBURY, A.; TRUDGILL, P. *New Zealand English*. Its origins and evolution. Cambridge: Cambridge University Press, 2009.

GRIES, St. Th.; HILPERT, M. The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora*, v. 3, n. 1, p. 59-81, 2008.

HILPERT, M. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory*, v. 2, n. 2, p. 243-256, 2006.

HONKAPOHJA, A.; KAISLANIEMI, S.; MARTTILA, V. Digital editions for corpus linguistics: representing manuscript reality in electronic corpora. In: JUCKER, A. H.; SCHREIER, D.; HUNDT, M. (Ed.). Corpora: pragmatics and discourse. 29th International Conference on English Language Research on Computerized Corpora (ICAME 29). *Papers*... Ascona, Switzerland, 14-18 May 2008. Amsterdam: Rodopi, 2009.

HUBER, M. The Old Bailey Proceedings, 1674-1834: evaluating and annotating a corpus of 18th- and 19th-century spoken English. In: MEURMAN-SOLIN, A.; NURMI, A. (Ed.). *Annotating variation and change*. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki, 2007. (Studies in Variation, Contacts and Change in English 1). Available at: <http://www.helsinki.fi/varieng/journal/volumes/01/huber/>. Retrieved: May 5, 2011.

HÜLLEN, W. A close reading of William Caxton's Dialogues: "… to lerne Shortly frenssh and englyssh". In: JUCKER, A. H. (Ed.). *Historical pragmatics*: pragmatic developments in the history of English. Amsterdam/Philadelphia: John Benjamins, 1995. (Pragmatics & Beyond New Series 35)

JACOBS, A.; JUCKER, A. H. The historical perspective in pragmatics. In: JUCKER, A. H. (Ed.). *Historical pragmatics*. Pragmatic developments in the history of English. Amsterdam/Philadelphia: John Benjamins, 1995. (Pragmatics & Beyond New Series 35)

JOHANSSON, S. Some aspects of the development of corpus linguistics in the 1970s and 1980s. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

JUCKER, A. H. *Historical pragmatics*: pragmatic developments in the history of English. Amsterdam/Philadelphia: John Benjamins, 1995. (Pragmatics & Beyond New Series 35)

JUCKER, A. H.; FRITZ, G.; LEBSANFT, F. *Historical dialogue analysis*. Amsterdam/Philadelphia: John Benjamins, 1999a. (Pragmatics & Beyond New Series 66)

JUCKER, A. H.; FRITZ, G.; LEBSANFT, F. Historical dialogue analysis: roots and traditions in the study of the Romance languages, German and English. In: JUCKER, A. H.; FRITZ, G.; LEBSANFT, F. (Ed.). *Historical dialogue analysis*. Amsterdam/Philadelphia: John Benjamins, 1999b. (Pragmatics & Beyond New Series 66)

JUCKER, A. H.; SCHNEIDER, G.; TAAVITSAINEN, I.; BREUSTEDT, B. Fishing for compliments: precision and recall in corpus-linguistic compliment research. In: JUCKER, A. H.; TAAVITSAINEN, I. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008. (Pragmatics and Beyond New Series 176)

JUCKER, A. H.; TAAVITSAINEN, I. Diachronic speech act analysis: the case of insults. *Journal of Historical Pragmatics*, v. 1, n. 1, p. 67-95, 2000.

JUCKER, A. H; TAAVITSAINEN, I. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008a. (Pragmatics and Beyond New Series 176)

JUCKER, A. H.; TAAVITSAINEN, I. Apologies in the history of English: routinized and lexicalized expressions of responsibility and regret. In: JUCKER, A. H.; TAAVITSAINEN, I. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008b. (Pragmatics and Beyond New Series 176)

KOHNEN, T. Text types and the methodology of diachronic speech act analysis. In: FITZMAURICE, S. M.; TAAVITSAINEN, I. (Ed.). *Methods in historical pragmatics*. Berlin/New York: Mouton de Gruyter, 2007. (Topics in English Linguistics 52)

KYTÖ, M. Data in historical pragmatics. In: JUCKER, A. H.; TAAVITSAINEN, I. (Ed.). *Historical pragmatics*. Berlin/New York: Walter de Gruyter, 2010. (Handbooks of Pragmatics 8)

KYTÖ, M. Corpus linguistics. In: BERGS, A.; BRINTON, L. (Ed.). *Historical linguistics of English*: an international handbook. Berlin/New York: Walter de Gruyter, forthcoming. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft).

KYTÖ, M.; GRUND, P. J.; WALKER, T. *Testifying to language and life in Early Modern England*. Amsterdam/Philadelphia: John Benjamins, forthcoming (2011). Including a CD containing An Electronic Text Edition of Depositions 1560-1760 (ETED)

KYTÖ, M.; PAHTA, P. Evidence from historical corpora up to the twentieth century. In: NEVALAINEN, T.; TRAUGOTT, E. (Ed.). *A handbook to the history of English*. Oxford: Oxford University Press, forthcoming.

KYTÖ, M.; RISSANEN, M. The syntactic study of Early American English: the variationist at the mercy of his corpus? *Neuphilologische Mitteilungen*, v. 84, n. 4, p. 470-490, 1983.

KYTÖ, M.; WALKER, T. The linguistic study of Early Modern English speech-related texts: how "bad" can "bad" data be? *Journal of English Linguistics*, v. 31, n. 3, p. 221-248, 2003.

KYTÖ, M.; WALKER, T. *Guide to A Corpus of English Dialogues 1560-1760*. Uppsala: Acta Universitatis Upsaliensis, 2006. (Studia Anglistica Upsaliensia 130)

LANCASHIRE, I. *Introduction*. 2006. Available at: <http://leme.library.utoronto.ca/public/intro.cfm>. Retrieved: May 5, 2011.

LASS, R. *On explaining language change*. Cambridge: Cambridge University Press, 1980.

LEECH, G.; HUNDT, M.; MAIR, C.; SMITH, N. *Change in contemporary English*: a grammatical study. Cambridge: Cambridge University Press, 2009.

A LINGUISTIC ATLAS OF LATE MEDIAEVAL ENGLISH. Compiled by McINTOSH, A.; SAMUELS, M. L.; BENSKIN, M.; with the assistance of Margaret Laing and Keith Williamson. 4 v. Aberdeen: Aberdeen University Press, 1986.

LUTZKY, U. *Discourse markers in Early Modern English*: The case of 'marry', 'well' and 'why'. 325 p. 2009. (PhD thesis) – Vienna University.

MacQUEEN, D. S. *The integration of MILLION into the English system of number words*: a diachronic study. Frankfurt am Main/Berlin, etc.: Peter Lang, 2010. (English Corpus Linguistics 11)

MAIR, C. Corpora and the study of recent change in language. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

MARKUS, M. Normalization of Middle English prose: possibilities and limits. In: LJUNG, M. (Ed.). Corpus-based studies in English. Seventeenth International Conference on English Language Research on Computerized Corpora. *Papers...* Stockholm, May 15-19, 1996. Amsterdam/Atlanta, GA: Rodopi, 1997.

McENERY, T.; WILSON, A. *Corpus linguistics*: an introduction. Second edition. Edinburgh: Edinburgh University Press, 2001 [1996].

MEYER, C. F. *English corpus linguistics*: an introduction. Cambridge: Cambridge University Press, 2002.

MEYER, C. F. Pre-electronic corpora. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

MILROY, J. A social model for the interpretation of language change. In: RISSANEN, M.; IHALAINEN, O.; NEVALAINEN, T.; TAAVITSAINEN, I. (Ed.). *History of Englishes*: new methods and interpretations in historical linguistics. Berlin and New York: Mouton de Gruyter, 1992. (Topics in English Linguistics 10)

MONOCONC PRO. Available at: <http://www.athel.com/mono.html>.

NEVALAINEN, T.; RAUMOLIN-BRUNBERG, H. *Historical sociolinguistics*: language change in Tudor and Stuart England. London/New York/Toronto: Pearson Education, 2003.

NURMI, A.; NEVALA, M.; PALANDER-COLLIN, M. (Ed.). *The language of daily life in England* (1400-1800). Amsterdam/Philadelphia: John Benjamins, 2009. (Pragmatics & Beyond New Series 183)

PETRÉ, P. *Leuven English Old to New (LEON)*: some ideas on a new corpus for longitudinal diachronic studies. Paper given at the Middle and Modern English Corpus Linguistics conference, University of Innsbruck, 5-9 July 2009.

THE R PROJECT FOR STATISTICAL COMPUTING. Available at: <http://www.r-project.org/>.

RAUMOLIN-BRUNBERG, H. Review of SOKOLL, T. (Ed.). Essex pauper letters 1731-1837. Records of Social and Economic History, New Series 30. Published for The British Academy by Oxford University Press, 2001. *Historical Sociolinguistics and Sociohistorical Linguistics*, v. 3, 2003. Available at: <http://www.let.leidenuniv.nl/hsl_shl/sokoll.htm> Retrieved: May 5, 2011.

RAYSON, P,; ARCHER, D.; BARON, A.; SMITH, N. Tagging historical corpora – the problem of spelling variation. In: Digital Historical Corpora. Dagstuhl-Seminar 06491, International Conference and Research Center for Computer Science. *Proceedings*... Schloss Dagstuhl, Wadern, Germany, December 3rd-8th 2006. 2007. Available at: <http://drops. dagstuhl.de/opus/volltexte/2007/1055/. Retrieved: May 5, 2011.

RISSANEN, M. Variation and the study of English historical syntax. In: SANKOFF, D. (Ed.). *Diversity and diachrony*. Amsterdam/Philadelphia: John Benjamins, 1986. (Current Issues in Linguistic Theory 53)

RISSANEN, M. Syntax. In: LASS, R. (Ed.). *The Cambridge history of the English language*, v. III, 1476-1776. Cambridge: Cambridge University Press, 97-109. p. 187-331.

RISSANEN, M. Corpora and the study of the history of English. In: KYTÖ, M. (Ed.). *English corpus linguistics*: crossing paths. Amsterdam: Rodopi, forthcoming.

RISSANEN, M.; KAHLAS-TARKKA, L.; HINTIKKA, M.; McCONCHIE, R. (Ed.). *Change in meaning and the meaning of change*: studies in semantics and grammar from Old to Present-day English. Helsinki: Société Néophilologique, 2007. (Mémoires de la Société Néophilologique de Helsinki 72)

ROMAINE, S. *Socio-historical linguistics*: its status and methodology. Cambridge: Cambridge University Press, 1982. (Cambridge Studies in Linguistics 34)

ROMAINE, S. Variation in language and gender. In: HOLMES, J.; MEYERHOFF, M. (Ed.). *The handbook of language and gender*. Malden, MA: Blackwell, 2003. (Blackwell Handbooks in Linguistics 13)

SAMUELS, M. L. *Linguistic evolution, with special reference to English*. Cambridge: Cambridge University Press, 1972. (Cambridge Studies in Linguistics 5)

SOKOLL, T. (Ed.). *Essex pauper letters*, 1731-1837. Published for The British Academy by Oxford University Press, 2001. (Records of Social and Economic History, New Series 30)

TAAVITSAINEN, I.; JUCKER, A. H. Speech act verbs and speech acts in the history of English. In: FITZMAURICE, S. M.; TAAVITSAINEN, I. (Ed.). *Methods in historical pragmatics*. Berlin/New York: Mouton de Gruyter, 2007. (Topics in English Linguistics 52)

TAAVITSAINEN, I.; JUCKER, A. H. Speech acts now and then: towards a pragmatic history of English. In: JUCKER, A. H.; TAAVITSAINEN, I. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008a. (Pragmatics and Beyond New Series 176)

TAAVITSAINEN, I.; JUCKER, A. H. "Methinks you seem more beautiful than ever": compliments and gender in the history of English. In: JUCKER, A. H.; TAAVITSAINEN, A. H. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008b. (Pragmatics and Beyond New Series 176)

TEXT ENCODING INITIATIVE (TEI). Available at: <http://www.tei-c.org/index.xml>.

TRAUGOTT, E. C.; DASHER, R. B. *Regularity in semantic change*. Cambridge: Cambridge University Press, 2002. (Cambridge Studies in Linguistics 96)

VARD2. Available at: <http://www.comp.lancs.ac.uk/~barona/vard2/>.

WALKER, T. *Thou and You in Early Modern English dialogues*: trials, depositions, and drama comedy. Amsterdam/Philadelphia: John Benjamins, 2007. (Pragmatics & Beyond New Series 158)

WATTS, R. J. Refugiate in a strange countrey: learning English through dialogues in the 16th century. In: JUCKER, A. H.; FRITZ, G.; LEBSANFT, F. (Ed.). *Historical dialogue analysis*. Amsterdam/Philadelphia: John Benjamins, 1999. (Pragmatics & Beyond New Series 66)

WEINREICH, U.; LABOV, W.; HERZOG, M. I. Empirical foundations for a theory of language change. In: LEHMANN, W. P.; MALKIEL, Y. (Ed.). *Directions for historical linguistics*: a symposium. Austin/London: University of Texas Press, 1968.

WORDSMITH TOOLS. Available at: <http://www.lexically.net/wordsmith/>.

WYNNE, M. Searching and concordancing. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

WYNNE, M. Interdisciplinary relationships. In: POPE, Caty Worlock (Ed.). Special issue on the bootcamp discourse and beyond. *International Journal of Corpus Linguistics*, v. 15, n. 3, p. 425-427, 2010.

XAIRA. Available at: <http://www.oucs.ox.ac.uk/rts/xaira/>.

XIAO, R. Well-known and influential corpora. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

YÁÑEZ-BOUZA, N. ARCHER past and present (1990-2010). *ICAME Journal*, v. 35, p. 205-236, 2011.

Recebido em 21/01/2011. Aprovado em 08/05/2011.

# Corpus linguistics and second/foreign language learning: exploring multiple paths

## *Linguística de corpus e aprendizagem de segunda língua/língua estrangeira: explorando caminhos múltiplos*

Fanny Meunier*
Catholic University of Louvain
Louvain-la-Neuve / Belgium

ABSTRACT: The aim of this article is twofold: first, to briefly assess the influence that corpus linguistic research has had on second/foreign language learning so far, and second, to suggest future directions for a more coherent and well thought out integration of corpora in instructed settings. In section 1, the influence of *native* and *learner* corpus research on second/foreign language learning will be assessed in turn, and some reasons for the overall lack of uptake of corpora in educational contexts will be put forward. In section 2, I will argue that multiple paths will have to be explored for a better integration of corpora in instructed settings. The fact that various – and sometimes even radically opposite – directions will be proposed might appear conflicting at first sight, but it will be demonstrated that opting for a multiplicity of perspectives is the only way to lay the foundations of a healthy cross-fertilization between corpus linguistics and the current multi-faceted language learning and teaching cultures.

KEYWORDS: corpus linguistics; second/foreign language corpora; applications; second/foreign language teaching and learning.

RESUMO: O objetivo deste artigo é duplo: primeiramente, avaliar de forma sucinta a influência que a pesquisa da linguística de corpus tem tido sobre a área de aprendizagem de segunda língua/língua estrangeira; e, em segundo lugar, sugerir direções futuras para uma integração mais coerente e bem refletida sobre a integração de corpora aos ambientes de ensino. Na seção 1, a influência da pesquisa de corpora de língua nativa e corpora de aprendizes na aprendizagem de segunda língua/língua estrangeira será avaliada e algumas razões para a não-adoção dessas metodologias nos ambientes de ensino serão apontadas. Na seção 2, argumentarei que diferentes caminhos deverão ser adotados para que haja uma melhor integração entre corpora e ambientes de ensino. O fato de vários – e às vezes, até mesmo – caminhos opostos, serem propostos, pode parecer conflitante à primeira vista. Mas, será mostrado que, a opção por perspectivas múltiplas é o único caminho para que se estabeleçam bases saudáveis para a interação entre as culturas das áreas de estudos de corpora e aprendizagem e ensino de línguas.

PALAVRAS-CHAVE: linguística de corpus; corpora de segunda língua; aplicações; ensino/aprendizagem de segunda língua/língua estrangeira.

* fanny.meunier@uclouvain.be

## 1. Corpus linguistic research and second/foreign language learning: a brief state-of-the art

When trying to assess the influence of corpus linguistic research on second/foreign language learning (henceforth abbreviated as L2[1] language learning or L2L), it seems reasonable to address the influence of native and learner corpus research separately given that they have had rather different implications on instructional settings.

### 1.1 Native corpus research and L2 language learning

Widdowson (2004, p. 357), in an article addressing the recent trends in English language teaching, acknowledges the impact of technology on the current modes of language use and communication but also on "ways in which the language so used is recorded and analysed". He adds (2004, p. 357) that "the most striking development in linguistic description over the past twenty years has been the use of the computer to collect and analyse vast corpora of actually occurring language data" and speaks of an "abundance of dictionaries and grammatical **descriptions** which are corpus-based and which chart the patterns of the contemporary usage of English". In Meunier & Gouverneur (2009) we also argued that corpora have found their way to the offices of major ELT publishers and are increasingly used as a source of authentic data to inform new series of **reference and pedagogical materials** such as dictionaries, grammar or vocabulary books. Cambridge University Press offers a 'real English guarantee'[2] to the buyers and users of their material, Longman assures its readership that [they] 'only see real English, as it is really used'.[3] As for MacMillan, the use of their World English Corpus is described as 'a unique modern database of over 200 million words revealing fresh information on how words are used and natural examples of English as it is written and spoken now!'[4]

Römer (2006, p. 121) states however that "despite the progress that has been made in the field of corpus linguistics and language teaching, the **practice** of ELT has so far been largely unaffected by the advances of corpus research";

[1] In this article, no distinction will be made between second and foreign language learning, hence the general L2 abbreviation (which also encompasses the learning of possible third, fourth, etc. languages).

[2] See <http://www.cambridge.org/elt/corpus/corpus_based_books.htm>.

[3] See <http://www.longman.com/dictionaries/corpus/index.html>.

[4] See <http://www.macmillandictionary.com/aboutcorpus.htm>.

Breyer (2009) concurs by highlighting the opposition between the enthusiasm of the research community and the dearth of applications of corpus tools and resources in the classroom. There is thus a clear divide between the exponentially growing number of publications in applied native corpus research and the introduction of corpus data in reference books and teaching materials on the one hand and everyday teaching practices on the other. Furthermore, as English is the language that has been described most fully thanks to corpus methods, the gap or opposition mentioned by Römer and Breyer is likely to be even much wider for other languages.

Several reasons account for the lack of uptake of corpus-oriented tools and methods in the classroom, and I will expand on four of them. One rather fundamental reason is that the enthusiasm for corpus methods is mostly expressed by linguists and that the authority of the linguists does not always find an echo among teachers. As Widdowson (2004, p. 359) puts it whilst "the case for 'real' English" is in itself very appealing to teachers, it is often proposed "**on the authority of the linguists**". As a result, the so-called 'corpus revolution' (RUNDELL; STOCK, 1992) may either not have reached the teachers yet, or may be intentionally rejected by them. In the unintentional case, teachers are often not aware of the possibilities offered by corpora or not aware of the changes that corpus methods have brought to materials that they are sometimes actually using. This may be due to a lack of information at the pre- or in-service teacher training levels and/or to the sometimes too vague statements found in the introductions to teaching materials, which might refer to the corpus-informed nature of the materials but not explicitly list the pedagogical implications of this corpus-based nature. As for the intentional negation of the benefits of corpus research in L2L, it may be caused by the oft-cited 'ivory tower effect', i.e. the perception by teachers that linguists work in their offices at university and have no idea of what teaching is about - and this despite the fact that some of those linguists are also teachers. This feeling of distance is usually reinforced by the fact that the types of examples or applications provided in the literature are often meant for EAP/ESP audiences. The collection of corpora (be they native or learner corpora) is usually coordinated by university teams and a vast majority of applied uses of corpora are found at university level (see for instance FEAK; SWALES, 2010; JONES; SCHMITT, 2010). This focus on advanced and specialized levels of proficiency does not always facilitate a transposition to learners with less advanced, non-academic needs.

Another reason explaining the lack of uptake of corpus methods in instructional settings (and one that is especially worth taking into account by corpus linguists advocating the applied relevance of their research whilst not being directly involved in teaching) is that the importance of using **authentic, corpus-based descriptions of the target language** is **only one line of thinking among many other influential ones** in L2 language learning. Socio-politically oriented considerations have led some to reject the promotion of standard native speaker usage as the norm for L2L as a manifestation of linguistic imperialism (PHILLIPSON, 1992) that must be abolished. This corresponds to what Davies (1996) and Widdowson (2004) call the 'conspiracy' view.

More practical and methodologically oriented views have however also been put forward against the use of a native speaker model in L2L: it might set goals that are unachievable, unrealistic and unnecessary for the actual needs of the learners, and might not be appropriate to the socio-cultural conventions of the groups acquiring the L2.

Another key issue that has attracted controversy in L2 language learning and teaching circles is the **issue of frequency** which is unmistakably entwined with corpus linguistic research. As Gries (forthcoming) explains, many linguistic fields have witnessed "a development towards more rigorous data analysis: statistical analysis of various levels of complexity have become a mainstream component of linguistic analysis". This quantitative/statistical development can be considered as most welcome as it definitely promotes objectivity in research. Two main uses of frequency can be distinguished in L2L: a) the role of input frequency on L2 acquisition and b) the use of frequency information to help select which aspects of language (be they words, expressions, grammatical structures, errors, etc.) deserve more attention form the part of the learners and/or teachers. Whilst the role of frequency effects in SLA has clearly been demonstrated (see for instance ELLIS, N., 2002a; 2002b; SIYANOVA; SCHMITT, 2008; COLLINS; ELLIS N., 2009; ELLIS, N.; LARSEN-FREEMAN, 2009), the picture is perhaps less clear when it comes to potential applications to the teaching of foreign languages. Leech (forthcoming, 2011) argues that when applied to teaching, the frequency principle is often interpreted as 'more frequent = more useful to teach'. Frequency lists are certainly not unknown to teachers and many are familiar with the notion of threshold levels (see van EK; ALEXANDER, 1980), that of vocabulary size needed to read and understand unsimplified texts (see HIRSH; NATION, 1992), or also the existence of an academic word list (COXHEAD, 2000).

Some teachers also use existing web tools that include vocabulary frequency lists for teaching and learning purposes, together with a variety of corpus tools (see for instance the Lexical Tutor, <http://www.lextutor.ca/>, maintained by Cobb at the Université du Québec à Montréal). The existence of frequency lists and corpus tools that can help access frequency information should however not be considered as an end in itself, and, whilst stressing the important role of frequency in L2L, Leech (forthcoming, 2011) nonetheless warns against a naïve interpretation of the frequency principle when it comes to teaching (see section 2.3 for further comments).

A last factor accounting for the lukewarm reception of corpora in the classrooms is the **lack of empirical studies exploring the actual impact of corpus methods on the learning outcomes**. The results of the few studies available (see for instance YOON, 2005, VANNESTAL; LINDQUIST, 2007; BELZ; VYATKINA, 2008; BOULTON, 2009; BREYER, 2009) present a contrasted picture and show that using corpora with students may require substantial support in some cases, that it takes time and practice to help students become independent users, that it does not appeal to all the students, and that it may prove beneficial for some skills and tasks but not for others. Yoon's study (2005) shows for instance that the use of corpora in writing classes provides students with common usage and collocation patterns that can be recycled immediately in their own writing, helps them develop longer-term cognitive skills (such as a greater awareness for lexico-grammatical aspects), and promotes independent learning. Vannestal and Lindquist (2007) find similar results but add that weak students find corpus consultation difficult or boring, and that some students do not find corpora very useful to help them improve their grammatical knowledge of the target language. All the studies listed above also stress the fact that teachers who want to use corpora with their students need to have a good understanding of the multi-faceted aspects of corpus literacy if they want the experiment to be successful. More research on the impact and learning outcomes of corpus methods is definitely in order to provide clearer evidence on the types of tasks and skills that would benefit most from a corpus approach.

Whilst far from being exhaustive,[5] the list of arguments provided in the preceding paragraphs nevertheless sheds some light on the reasons that have

[5] The access to well-equipped computer rooms with up-to-date software and hardware has for instance not been mentioned.

put a break on the expansion, integration, acceptance and understanding of native corpus research in educational settings.

## 1.2 Learner corpus research and L2 language learning

The fact that L2L has remained largely unaffected by the advances of native corpus research is even truer when one looks at the L2 teaching/learning applications of learner corpus research. Learner corpora are sometimes described as the 'missing link' in (EAP) pedagogy (see GILQUIN *et al.*, 2007) and they provide one ideal type of data to help linguists and teachers attest actual learners' needs on the basis of a careful analysis of their productions. Yet, despite the potential of learner corpora, Granger (2009, p. 24) provides a critical evaluation of their actual contribution to SLA and foreign language teaching and writes that "there is undeniably very little evidence of fully-fledged up-and-running applications".

It would be wrong to state that learner corpora have not been used by ELT publishers, but it must be acknowledged that they have been used to a much smaller extent than native corpora. When publishers refer to learner corpora, they seem to privilege in-house learner corpora.[6] Another problem is that, as was the case for the use of native corpora (see section 1.1), the exact use that is being made of the learner corpus is not always clearly documented.[7] This leads to the sometimes surprising inclusions of what Granger (2010, p. 32) describes as error notes "apparently based on learner corpora". She gives the examples of the use of *attend* instead of *wait for* (*Her mother was attending her outside the car) or of *piece* instead of *room* (*There are en suite bathrooms in every piece), examples not even attested once in the International Corpus of Learner English (GRANGER *et al.*, 2009), a learner corpus containing 3.5 million words produced by over 6,000 learners from 16 different mother tongue backgrounds.

[6] See for instance CUP and the Cambridge Learner Corpus at <http://ww.cambridge.org/elt/corpus/learner_corpus2.htm> or Longman and the Longman Learner Corpus at <http://www.longman.com/dictionaries/corpus/learners.html>.

[7] De Cock et al. (2007)'s section *Improving your writing skills*, which is included in the CD-ROM version of the Macmillan English Dictionary for Advanced Learners, is well-documented and constitutes a welcome exception to the publishers' use of in-house learner corpora.

Another reason accounting for the limited impact of learner corpora in instructional settings is put forward by Flowerdew, who states that in most studies of learner corpora "the implications for pedagogy are not developed in any great detail with the consequences that the findings have had little influence on [...] syllabus and materials design" (1998, p. 550). There is an urgent need to go beyond the usual last paragraph of articles (or last slide of conference presentations) stating that 'foreign language instruction could profit from this kind of investigation' and efforts should be made towards providing teachers with ready-to-use teaching materials, or at least free access, user-friendly and ready-to-use platforms which they could use to collect, analyse and exploit learner corpora on a more regular basis.

I have also recently argued (MEUNIER, 2010) that an additional reason accounting for the lack of direct influence of learner corpus studies on L2 syllabuses and materials[8] is that the topics covered in most existing learner corpora are often miles away from the everyday needs of a vast majority of L2 school teachers who target the L2 for general purposes, often for a teenage audience. Finding a learner corpus that meets their needs comes close to looking for a needle in a haystack. The Common European Framework of Reference for Languages (CEFR, Council of Europe 2001, p. 52) suggests that the following thematic categories should be addressed in L2 for General Purposes (EGP): personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather. Few, if any, native or learner corpus studies provide easily transferrable research results which could be integrated in a syllabus addressing these themes. Corpus compilers will urgently have to address the learner's needs for what Braun (2005) calls 'pedagogical relevance' or what Belz and Viyatkina (2008) call 'authentication'.

## 2. Corpus linguistic research meets second/foreign language learning: exploring multiple paths for a balanced integration

A multiplicity of paths will have to be explored if a fuller integration between the two domains is to be achieved. Various – and sometimes even

[8] This does not apply to language for academic/specific purposes – domains where teachers do actually use native and advanced learner corpora (see for instance Flowerdew, 2003; Gilquin *et al.*, 2007, or PAQUOT, 2008).

radically opposite – directions are proposed in the coming sections. I believe that this multi-directionality is necessary to promote a healthy cross-fertilization between corpus linguistics and the current multi-faceted language learning cultures. Multi-directionality is also the only way to help teachers and material designers cater for the particular needs of specific learning populations in no less specific socio-cultural contexts.

### 2.1 Go global: expand your horizons

The call for more **cross-disciplinarity** between the various research paradigms involved in L2L has repeatedly been made. Recent calls include Sorace (2010), who states that a lot of relevant SLA research is done in other fields and ignored by SLA researchers. Similarly, Norris (2010), a famous proponent of task-based language learning,[9] stresses the importance of understanding instructed SLA (ISLA) and of taking into account the needs of the teachers and learners in classrooms. In Granger and Meunier (2010), we plead for a closer integration between SLA and learner corpus research (LCR) and show that SLA studies can greatly benefit from the solid empirical base provided by learner corpora research tools and methods, whilst LCR needs a more solid grounding in SLA theory.

Whilst it is obviously impossible to acquire unlimited expertise (foreign/second language teacher, SLA researcher, corpus linguist, sociolinguist, computer programmer, statistician, etc.) it is nevertheless essential to make a conscious effort to be open to other academic cultures and working environments and hence, to leave one's comfort zone. An implicit corollary of cross-fertilization is that each field should make yet another conscious effort to highlight the convergences between its own domain and other related domains. This, in turn, implies a certain degree of elaboration, specification and sometimes even simplification. Some researchers set the example and provide clear introductions to their research area. Römer and Wulff (2010), in a paper entitled *Applying corpus methods to written academic texts: Explorations of MICUSP*, provide a most welcome step-by-step introduction to the central techniques in corpus analysis intended for students and/or corpus

[9] Task-based learning involves goal-oriented communicative activities, with a specific outcome, where the emphasis is on exchanging meanings and not on producing specific language forms (see WILLIS, 1996).

novices. This type of publication can only be encouraged as it corresponds to what Römer calls the missionary work of corpus linguists, i.e. the need to convince teachers, students, materials writers, and syllabus designers that corpora can be of great use in their everyday work (2006: 128). Similar missionary work from other fields must also be encouraged.

One welcome development illustrating the benefits of cross-fertilization is the increasing use of **triangulation methods** in L2L studies. Triangulation can for instance be obtained by combining corpus and experimental data (SIYANOVA; SCHMITT, 2008), or by revisiting/replicating earlier SLA studies on (more) learner corpus data (as exemplified by HOUSEN, 2002, who revisited the verb morphemes study initially carried by DULAY *et al.*, 1982, or by WULFF *et al.*, 2009, who reinterpreted tense-aspect studies conducted by BARDOVI-HARLIG, 1998; 2002). Also important among triangulation methods is what Ellis, Simpson-Vlach & Maynard (2008) call the validation of the instructional value, i.e. the assessment by experienced language instructors and testers of the teaching-worthiness of the linguistic output obtained thanks to corpus metrics.

The challenge of going global and expanding the horizons can also be taken up within the corpus linguistics field. **More (learner) corpora** should be collected to represent:

- **more languages**, to counterbalance the predominance of anglo-saxon native and learner corpora and to foster the computer-aided analysis of different languages and language families,
- **more communicative modes**: spoken corpora, interactional corpora (classroom interactions, authentic interactions representing what Wagner (2010) calls 'language learning in the wild', multimodal corpora, corpora of textbook materials, etc.,
- **more text types and genres**, to cover text types which are less represented in corpora to date (letters, emails, twits, leaflets, TV programmes, book synopses, recipes, short notes, chat room logs, etc.),
- **more longitudinal language data**, from beginners to advanced levels, from children to adults, from L1 to L2s, but also attrited language and language impaired data,
- **more variables**: more language learning variables should be collected and encoded at the time of corpus collection (proficiency, language aptitude, motivation, more precise description of the task, of temporal, social or situational settings, etc.).

Adopting a cross-disciplinary perspective, and making a disciplinary effort to expand data types and quality within corpus linguistics, will lead to a better understanding of the processes at play in L2 learning and acquisition. It will also make it possible for more social, cultural and situational variables to be taken into account in instructed environments.

## 2.2 Go local: create a sense of community

The call for going global is mainly meant for researchers and teachers. If one adopts a more **learner-centred** view, a reverse call - viz. going local - is probably more appropriate.

One way of encouraging learners to use corpora is to enhance the **pedagogical relevance** (BRAUN, 2005) and **authentication** (BELZ; VIYATKINA, 2008) of corpus use (see section 1.2). In *From Corpus to Classroom*, O'Keeffe *et al.* (2007, p. xi) mention the "frequent mismatch between corpus linguistic research and what goes on into materials and resources, and what goes on in the language classroom". It is actually fair to argue that corpora will only be used by language learners if they can interpret, analyse and understand them in a personally meaningful way (BELZ; VIYATKINA, 2008). A direct involvement of learners in corpus collection and use corresponds to what Granger (2009, p. 25) calls corpora for immediate pedagogical use (IPU), i.e. data "collected by teachers as part of their normal classroom activities [...] [and where] the learners are at the same time producers and users of the corpus data". Examples of such IPU include telecollaborative interactions (oral or written) during which learners build personal relationships with other speakers. Those other speakers can be native speakers (as in BELZ; VIYATKINA's, 2008 study) but they can also be other non-native speakers. Once the oral and written interactions are archived they can be accessed and explored to serve as a basis for pedagogical interventions. Teachers can focus on specific linguistic forms produced by the learners themselves, in the context of **meaningful interactions** (see KASPER; ROSE, 2002) in **communicative tasks**. Learners are then more likely to feel a sense of authentication and pedagogical relevance.

Braun (2006)'s project provides another good illustration of pedagogical relevance; she uses a small English Interview corpus (ELISA) containing 26 interviews of approximately 10 minutes each, for a total of about 60,000 words. Despite its limited size in words, the corpus covers a variety of communicatively relevant topics from the broad area of professional, social and cultural life. Braun (2006) also argues that homogeneity and **topical**

**relevance** are more important than representativeness in the traditional sense, and that learners and teachers are more likely to adopt a more **qualitative approach** to corpus analysis as it is more appropriate and manageable for them.

Teachers who decide to join the corpus bandwagon will definitely have to go '**beyond the pen and paper**' (WIBLE, 2008) but this does not necessarily imply a technological big bang. Some teachers will no doubt be better served than others (well-equipped computer labs, school technicians, projects carried out on a large scale) but it has been shown that corpus collection and use can be started on a smaller scale. Illustrations of larger-scale projects, which nonetheless promote a sense of community, are presented in Wible *et al.* (2001) and Simpson *et al.* (2002). Wible *et al.* explain how learner corpora can be collected and annotated by teachers, and subsequently be used for further pedagogical exploitation. They have set up an interactive online environment in which essays written by learners, together with the comments provided by teachers are archived in a searchable online database. Corpus collection and error annotation are **integrated in the normal teaching activities**: the student composes an essay offline and hands it in to the teacher over the Internet; the teacher marks the essay and sends it back to the student who revises his/her essay. During the correction phase, teachers insert comments in the student's essay by keying in a comment or by choosing a comment from an already existing 'Comment Bank' (e.g. '*wrong verb*', '*wrong tense*', etc.). Learners can get a cumulative comment list which will give them an idea of their most error-prone patterns; teachers can use these lists to design exercises that target learners' most frequent errors. In addition, the online corpus can be searched to get more instances of error-prone patterns. The human and computing resources required in that project are impressive (large number of teachers and learners taking part in the project all over Taiwan). As for the MICASE project (Michigan Corpus of Spoken Academic English), presented in Simpson *et al.* (2002), it consists of a collection of nearly 1.8 million words of spoken academic English recorded on the University of Michigan campus, and transcribed into searchable documents. MICASE can freely be browsed and searched online, notably to find recurrent grammatical and phraseological patterns or track generalized changes in speech patterns as people gain experience of university culture and academic speech.

Regardless of the size of the corpus collected, such projects subscribe to the learning-driven data methodology advocated by Seidlhofer (2002) by promoting a **learner-centred**, **context-dependent and culture-bound approach**. The fact that learners analyse their own productions favours the individualization of learning (and teaching), and helps learners monitor their own production and the effects of their production on others (MEUNIER, 2010).

To achieve the sense of community mentioned in the heading of the present section it is also essential to promote **care and attentiveness** in the use of corpora, both from a teacher- and learner- perspective. This care and attentiveness can be offered to teachers thanks to projects such as the Web 2.0 ERC (see <http://www.web20erc.eu/>), a new European Union funded education project to help educators who find ICT confusing and difficult to access. As for learners, Vannestal and Lindquist (2007) insist that using corpora with students requires time and a large amount of introduction and support, an issue that should certainly not be neglected when teachers opt for a corpus approach.

Going local will undoubtedly help promote corpus literacy as a useful tool for the empowerment of learning and teaching communities alike.

### 2.3. Let computer technology, frequencies and figures help and inform you; do not let them dictate

It would be most unreasonable to minimize the impact of technology without which frequency lists, specifications of textual features across languages, text types and genres, pattern grammar (the corpus-driven approach to the lexical grammar of English), collostructions (degree of attraction or repulsion between words and constructions), word sketches[10] *(*summaries of a word's lexical and grammatical collocational behaviour), or data-driven learning activities would still be unknown to date. The corpus revolution would have been impossible without the exponential increase in computing power, storage capacities, and programming and analysis skills of competent language experts. Jarvis (forthcoming) states that an important characteristic of (learner) corpus analysis is its heavy reliance on computer automation for purposes of discovering patterns in the data. Because of the size and complexity of most language corpora, it would be infeasible to perform comprehensive analyses of the data without computer automation. Gries *et al.* (2010, p. 4) also convincingly argue that "maybe the most dramatic changes that the field of corpus linguistics is witnessing these days concerns its methodologies… [and that] the field of corpus linguistics is rapidly being enriched with methodological

[10] For illustrations of word sketches, see <http://www.webdante.net/the_project.html>

expertise borrowed from other fields such as statistics, computational linguistics, and even artificial intelligence."

This said, the fact that sophisticated multi-factorial analyses are needed to refine the results of corpus studies should not divert our attention away from other key issues in language learning. The validation of the instructional value of the results of corpus studies, together with the pedagogical relevance of tools and methods used, have already been expanded on in sections 2.1 and 2.2 respectively. In the present section, I would like to come back to the frequency issue mentioned in section 1.1., and particularly to the 'more frequent = more useful to teach' approach.

A first warning must be made against an exclusive instructional focus on the more frequent lexical items in vocabulary lists, be they single words or multiword units. Vocabulary acquisition studies have demonstrated that higher **proficiency levels** correlate with the knowledge of less frequent words together with the knowledge of phraseological (and less common) uses of frequent words. It is therefore essential to gradually cover the whole frequency spectrum and even to come back to very frequent items in more advanced stages of acquisition in order to cover their phraseological and less common uses.

A slightly different perspective could probably be adopted for grammatical and syntactic patterns. While presenting highly infrequent structures to learners for **receptive** purposes makes sense, it is much less sensible to prompt learners to use these structures in **productive** tasks.

Whilst access to frequencies (in all its guises) is per se a very good thing, Leech (forthcoming, 2011) also warns that frequency counts are *least* useful when they are based on a general corpus covering the range of the language and are *more* useful if they are more specific, i.e. differentiated for mode, register, text type or region. A desirable evolution in corpus linguistics would then be to provide teachers and learners with more specific lists in line with teachers' and learners' communicative needs. On a more negative note, however, teachers and learners alike might wonder where to draw the line. The tensions between the precision and accuracy of the descriptions provided by corpus specialists can be perceived by teachers (and learners) as 'too much of a good thing', as shown by Coxhead (2008) who analysed the learners' negative perceptions of the importance of learning multi-word units when one word does the 'communication' trick.

In sum, I would recommend a flexible, teacher-validated and informed use of frequency lists. Teachers and learners would be wrong to do without or to ignore the information contained in frequency lists, but they would be equally wrong to abide by them dogmatically.

## 3. Concluding remarks

The influence of native and learner corpus research on second/foreign language learning has been discussed in section 1. There is no denying that a perceptible divide exists between the numerous publications in applied corpus research and the actual use of corpus data in instructional settings. Rather than sticking to that rather pessimistic conclusion, I have expanded on four possible reasons which may explain why instructional settings have tended to shy away from corpus use. Acknowledging these issues and actually addressing them is a vital step in promoting corpus uptake.

In section 2, I have put forward some suggestions to foster a healthy cross-fertilization between corpus linguistics and the current multi-faceted language learning and teaching cultures. I mentioned the importance of going global and expanding our horizons by encouraging cross-disciplinarity, promoting the use of triangulation methods in L2 studies, and further refining the learner, task and situational variables in the compilation of new types of corpora. I also suggested an opposite trend, which consists in going local and creating a sense of community. Taking a successful digital turn requires pedagogical relevance and authentication. If corpus methods are to be integrated in normal teaching activities they must be learner-centred, context-dependent and culture-bound. Time, care and attentiveness are also essential when promoting corpus literacy as empowerment tools for learners and teachers. The third line of discussion was devoted to frequencies. I have highlighted their overall importance in corpus studies but have nevertheless suggested the need for a flexible, teacher-validated and informed use of frequencies for pedagogical purposes.

## References

BARDOVI-HARLIG, K. Narrative structure and lexical aspect: Conspiring factors in second language acquisition of tense-aspect morphology. *Studies in Second Language Acquisition*, 20, p. 471-508, 1998.

BARDOVI-HARLIG, K. Analyzing aspect. In: SALABERRY, R.; SHIRAI, Y. (Ed.). *Tense-aspect morphology in L2 acquisition.* Amsterdam: John Benjamins, 2002.

BELZ, J.; VYATKINA, N. The Pedagogical Mediation of a Developmental Learner Corpus for Classroom-Based Language Instruction. *Language Learning & Technology*, v. 12, n. 3, p. 33-52, 2008.

BOULTON, A. Testing the Limits of Data-Driven Learning: Language Proficiency and Training. *ReCALL*, v. 21, n. 1, p. 37-54, 2009.

BRAUN, S. From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, v. 17, n. 1, p. 47-64, 2005.

BRAUN, S. ELISA - a Pedagogically Enriched Corpus for Language Learning Purposes. In: BRAUN, S.; KOHN, K.; MUKHERJEE, J. (Ed.). *Corpus Technology and Language Pedagogy*: New Resources, New Tools, New Methods. Frankfurt: Peter Lang, 2006.

BREYER, Y. Learning and Teaching with Corpora: Reflections by Student Teachers. *Computer Assisted Language Learning*, v. 22, n. 2, p. 153-172, 2009.

COLLINS, L. ; ELLIS, N. C. (Ed.). Input and second language construction learning: frequency, form, and function. Special issue. *Modern Language Journal*, v. 93, n. 3, 2009.

COXHEAD, A. A New Academic Word List. *TESOL Quarterly*, 34, v. 2, p. 213-238, 2000.

COXHEAD, A. Phraseology and English for academic purposes: Challenges and opportunities. In: MEUNIER, F.; GRANGER, S. (Ed.). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: Benjamins, 2008.

DAVIES, A. Ironising the Myth of Linguicism. Review of Linguistic Imperialism, by Robert L.H. Phillipson. Oxford: Oxford University Press, 1992. *Journal of Multilingual and Multicultural Development*, v. 17, n. 6, p. 485-496, 1996.

DULAY, H.C.; BURT, M.; KRASHEN, S. *Language Two*. Rowley: Newbury House, 1982.

ELLIS, N. C. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, v. 24, n. 2, p. 249-60, 2002a.

ELLIS, N. C. Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*, v. 24, p. 297-339, 2002b.

ELLIS, N. C.; LARSEN-FREEMAN, D. Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning*, v. 59, 2009. Supplement 1, p. 93-128.

ELLIS, N.C.; SIMPSON-VLACH, R.; MAYNARD, C. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, v. 42, p. 375-396, 2008.

ELLIS, R. *The Study of Second Language Acquisition*. Oxford: Oxford University Press, 1994.

ELLIS, R. *The Study of Second Language Acquisition.* 2. ed. Oxford: Oxford University Press, 2008.

FEAK, C.B.; SWALES, J. M. Writing for publication: Corpus-informed materials for post-doctoral fellows in perinatology. In: HARWOOD, N. (Ed.). *English Language Teaching Materials*. Theory and Practice. Cambridge: Cambridge University Press, 2010.

FLOWERDEW, L.J. Corpus Linguistic Techniques Applied to Textinguistic. *System*, v. 26, n. 1, p. 545-556, 1998.

GILQUIN, G.; GRANGER, G.; PAQUOT, M. Learner corpora: the missing link in EAP pedagogy. THOMPSON, P. (Ed.). *Corpus-based EAP Pedagogy. Special issue of the Journal of English for Academic Purposes*, v. 6, n. 4, p. 319-335. 2007.

GRANGER, S.; MEUNIER, F. SLA research and learner corpus research: Friend or foe? Paper presented at the *20th* A*nnual Conference of the European Second Language Association*. Reggio Emilia, Italy, 1-4 September, 2010.

GRANGER, S. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In: AIJMER, K. (Ed.). *Corpora and Language Teaching.* Studies in Corpus Linguistics 33. Amsterdam: John Benjamins, 2009.

GRANGER, S. Vingt ans d'analyse de corpus d'apprenants : leçons apprises et perspectives. In : CAPPEAU, P. ; CHUQUET, H. ; VALETOPOULOS, F. (Ed.). *L'exemple et le corpus*. Quel statut? Travaux linguistiques du CerLiCo. Numéro 23. Presses Universitaires de Rennes, 2010.

GRANGER, S.; DAGNEAUX, E.; MEUNIER, F.; PAQUOT, M. *The International Corpus of Learner English*. Version 2. Handbook and CD-Rom, Louvain-la-Neuve: Presses Universitaires de Louvain, 2009.

GRIES, St. Th. Statistical tests for the analysis of learner corpus data. In : DIAZ-NEGRILLO, A.; THOMPSON, P.; BALLIER, N. (Ed.). *Multidisciplinary perspectives to learner corpora*. Amsterdam & Philadelphia: John Benjamins. Forthcoming.

GRIES, St. Th.; WULFF, S.; DAVIES, M. (Ed.) *Corpus-linguistic applications*. Current studies, new directions. Amsterdam/New York: Rodopi, 2010.

HIRSH, D.; NATION, P. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language,* v. 8, n. 2, p. 689-696, 1992.

HOUSEN, A. A corpus-based study of the L2-acquisition of the English verb system. In: GRANGER, S.; HUNG, J.; PETCH-TYSON, S. (Ed.). *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Amsterdam: John Benjamins, 2002.

JARVIS, S. Data Mining with Learner Corpora: Choosing Classifiers for L1 Detection. In: MEUNIER, F.; DE COCK, S.; GILQUIN, G.; PAQUOT, M. (Ed.). *A Taste for Corpora*. In honour of Sylviane Granger. Amsterdam/Philadelphia: Benjamins. Forthcoming 2011.

JONES, M.; SCHMITT, N. Developing materials for discipline-specific vocabulary and phrases in academic seminars. In: HARWOOD, N. (Ed.). *English Language Teaching Materials*. Theory and Practice. Cambridge: Cambridge University Press, 2010.

LEECH, G. Frequency, corpora and language learning. In: MEUNIER, F.; DE COCK, S.; GILQUIN, G.; PAQUOT, M. (Ed.). *A Taste for Corpora*. In honour of Sylviane Granger. Amsterdam/Philadelphia: Benjamins, forthcoming 2011.

MEUNIER, F.; GOUVERNEUR, C. New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material. In: AIJMER, K. (Ed.). *Corpora and Language Teaching*. Amsterdam & Philadelphia: Benjamins, 2009.

MEUNIER, F. Learner Corpora and English Language Teaching: Checkup Time. *Anglistik: International Journal of English Studies*, v. 21, n. 1, p. 209-220, 2010.

NORRIS, J. Understanding instructed SLA: Constructs, contexts, and consequences. Plenary address at the *20th A*nnual Conference of the European Second Language Association. Reggio Emilia, Italy, 1-4 September, 2010.

O'SULLIVAN, I. Enhancing a Process-Oriented Approach to Literacy and Language Learning: The Role of Corpus Consultation Literacy. *ReCALL*, v. 19, n. 3, p. 269-286, 2007.

O'KEEFFE, A.; MCCARTHY, M.; CARTER, R. *From Corpus to Classroom*. Language use and language teaching. Cambridge: Cambridge University Press, 2007.

PHILLIPSON, R. *Linguistic Imperialism*. Oxford: Oxford University Press, 1992.

RÖMER, U.; WULFF, S. Applying corpus methods to writing research: Explorations of MICUSP. *Journal of Writing Research*, v. 2, n. 2, p. 99-127, 2010.

RÖMER, U. Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments. GAST, V. (Ed.). The Scope and Limits of Corpus Linguistics – Empiricism in the Description and Analysis of English. Special Issue: *Zeitschrift für Anglistik und Amerikanistik*, v. 54, n. 2, p. 121-134, 2006.

RUNDELL, M.; STOCK, P. The Corpus Revolution. *English Today,* v. 8, p. 9-14, 1992.

SEIDLHOFER, B. Pedagogy and local learner corpora: Working with learning-driven data. In: GRANGER, S.; HUNG, J.; PETCH-TYSON, S. (Ed.). *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Amsterdam: John Benjamins, 2002.

SIMPSON, R. C.; BRIGGS, S. L.; OVENS, J.; SWALES, J. M. *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan, 2002.

SIYANOVA, A.; SCHMITT, N. L2 Learner Production and Processing of Collocation: A Multi-study Perspective. *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, v. 64, n. 3, p. 429-458, 2008.

SORACE, A. SLA as bilingualism: Or, it's time to see the forest for the trees. Plenary address at the *20th* A*nnual Conference of the European Second Language Association.* Reggio Emilia, Italy, 1-4 September, 2010.

VAN EK, J.A.; ALEXANDER, L.G. *Threshold Level English*. Oxford: Pergamon, 1980.

VANNESTAL, M.E.; LINDQUIST, H. *Learning English Grammar with a Corpus*: Experimenting with Concordancing in a University Grammar Course. ReCALL, v. 19, n. 3, p. 329-350, 2007.

WIBLE, D. Multiword expressions and the digital turn. In: MEUNIER, F.; GRANGER, S. (Ed.). *Phraseology in Foreign Language Learning and Teaching.* Amsterdam & Philadelphia: Benjamins, 2008.

WIBLE, D.; KUO, C. W.; CHIEN, F., LIU, A.; TSAO, N. L. A webbased EFL writing environment: integrating information for learners, teachers, and researchers. *Computers and Education*, v. 37, p. 297-315, 2001.

WIDDOWSON, H. A perspective on recent trends. In: HOWATT, A. P. R.; WIDDOWSON, H. (Ed.). *A History of English Language Teaching.* Second edition. Oxford: Oxford University Press, 2004.

WILLIS, J. *A Framework for Task-Based Learning*. Longman. 1996.

WULFF, S.; ELLIS, N.C.; RÖMER, U.; BARDOVI-HARLIG, K.; LEBLANC, C.J. The Acquisition of Tense–Aspect: Converging Evidence From Corpora and Telicity Ratings. *The Modern Language Journal*, v. 93, n. 3, p. 354-369, 2009.

YOON, H. *An investigation of students' experiences with corpus technology in second language academic writing dissertation*. 2005. PhD (Dissertation, Degree Doctor of Philosophy) – Graduate School of The Ohio State University, 2005. Available at: <http://etd.ohiolink.edu/send-pdf.cgi/Yoon%20Hyunsook.pdf? osu1109806353>

Recebido em 22/09/2010. Aprovado em 08/05/2011.

# Spoken corpora and pragmatics

# *Corpora orais e pragmática*

Massimo Moneglia*
University of Florence
Firenze / Italy

ABSTRACT: The goal of this paper is to present arguments in favour of two points related to the study of oral corpora and pragmatics: a) at the level of annotation, corpora must ensure the parsing of the speech flow into utterances on the basis of prosodic cues and provide an easy access to the acoustic source; b) at the level of sampling, corpora must ensure the maximum representation of context variation, rather than speaker variation. We will present the reasons which support the very basic prosodic annotation of speech (prosodic boundaries) as a means to obtain relevant data from the speech flow. Starting from our present knowledge about the distribution of speech acts types in spoken corpora, we will present the reasons why building corpora in accordance to a context variation strategy should expand our knowledge of pragmatics. Additionally, we will claim that prosody is the necessary interface between locutive and illocutive acts and we will show that a deeper prosodic analysis is necessary to grasp unknown speech act types from language usage. Finally, we will briefly sketch the main assumptions of the Language into Act Theory (CRESTI, 2000) which is dedicated to the link between prosody and pragmatics and helps make explicit core aspects of pragmatic knowledge.

KEYWORDS: oral corpora; pragmatics; annotation; sampling; speech act types; prosody; Language into Act Theory.

RESUMO: O objetivo deste artigo é apresentar argumentos favoráveis a dois pontos relacionados ao estudo de corpora orais e pragmática: a) no nível da anotação, os corpora devem garantir o processamento do fluxo discursivo em enunciados, baseando-se em chaves prosódicas, e oferecer fácil acesso aos arquivos de som; b) no nível da amostragem, os corpora devem garantir a representatividade máxima de variação contextual, ao invés de variação de falantes. Apresentaremos os motivos que sustentam a escolha das fronteiras prosódicas como o referencial básico para a anotação prosódica da fala, como uma forma relevante de se obterem dados importantes do fluxo discursivo. Partindo do nosso conhecimento atual sobre a distribuição tipológica de atos de fala em corpora orais, apresentaremos as razões pelas quais a construção de corpora de acordo com a estratégia da variação contextual deve expandir o nosso conhecimento sobre pragmática. Adicionalmente, defenderemos que a prosódia é a interface necessária entre atos locutórios e ilocutórios e mostraremos que uma análise prosódica mais profunda é necessária para que se obtenham atos de fala desconhecidos a partir do uso da língua. Por fim, esboçaremos rapidamente os principais pressupostos da Teoria da Língua em Ato (CRESTI, 2000), a qual se debruça sobre a ligação entre a prosódia e a pragmática e auxilia na explicitação dos principais aspectos do conhecimento pragmático.

PALAVRAS-CHAVE: corpora orais; pragmática; anotação; amostragem; tipologia dos atos de fala; prosódia; Teoria da Língua em Ato.

* moneglia@unifi.it

## 1. Introduction

In my view, in order for spoken corpora to be exploited in a way that will enhance our knowledge in the domain of pragmatics to be built, two main basic strategies should be followed: a) at the level of annotation, corpora must ensure the parsing of the speech flow into utterances on the basis of prosodic cues and provide an easy access to the acoustic source; b) at the level of sampling, corpora must ensure the maximum representation of context variation, rather than speaker variation. These criteria, which have been applied in the construction of the C-ORAL-ROM corpus (CRESTI; MONEGLIA, 2005) and have been in practice at the LABLITA lab at the University of Florence, have ensured a good basis for grounding pragmatic concepts on actual speech data (CRESTI, 2000; CRESTI; FIRENZUOLI, 2001; FIRENZUOLI, 2003; SCARANO, 2003; FROSALI, 2006, CRESTI; MONEGLIA, 2010; CRESTI; MONEGLIA; TUCCI, in press).

The goal of this paper is to present arguments in favour of these two choices. In 2 we will present the reasons which support the very basic prosodic annotation of speech (prosodic boundaries) as a means to obtain relevant data from the speech flow. In 3, starting from our present knowledge about the distribution of speech acts types in spoken corpora, we will present the reasons why building corpora in accordance to a context variation strategy should expend our knowledge of pragmatics. In 4, we will claim that prosody is the necessary interface between locutive and illocutive acts and we will show that a deeper prosodic analysis is necessary to grasp unknown speech act types from language usage. In 5 we will briefly sketch the main assumptions of the Language into Act Theory (CRESTI, 2000) which is dedicated to the link between prosody and pragmatics and helps make explicit core aspects of pragmatic knowledge. According to this theory it is possible to identify the components of the utterance responsible for the illocutionary activity (Comment Unit) and to get clear distinctions between the main pragmatic functions allowed by the language structure, i.e., illocutionary activity and dialogue regulation activity.

More generally, in this paper, we will argue that the possibility to get robust knowledge about language structures that govern speech act performance in the ordinary use of language depends on a better understanding of the link between prosody and pragmatics. This relation and the need for a corpus-based strategy in pragmatic studies are both fundamental steps for grounding pragmatics on strong empirical evidence.

## 2 . Basic prosodic annotation for the exploitation of spoken corpora

### 2.1. Pragmatic units of reference for spoken language and prosody

If pragmatics is to profit from the huge amount of evidence which can be derived from contemporary corpus linguistics, these corpora must provide language data which are proper objects for pragmatic analysis; i.e., units of reference within the corpus which show pragmatic qualities. The series of lexical entries which constitute the speech flow (wording) do not provide this minimal linguistic entity directly.

In the case of written language, the nature of the linguistic units ranking above word level is clear. Although it may be chosen at different levels, i.e. argument structures, sentences or clauses, or head dependent structures (ABEILLÉ, 2003), written language can be properly parsed according to syntactic and semantic principles. Conversely, the identification of the units of reference in a spoken corpus can hardly be identified through the same syntactic and semantic devices (BLANCHE-BENVENISTE, 1997; BIBER *et al.*, 1999; CRESTI, 2000; MILLER; WEINERT, 1998; IZRE'EL, 2005).

Reference units for spontaneous speech are commonly identified with the term "utterance". The utterance might be anchored to syntactic and/or semantic properties as well. For instance, it can be identified with a syntactic *clause* (MILLER; WEINERT, 1998), or, as in The Longman Grammar, with a *C-Unit* with or without a clause structure (BIBER, *et al.*, 1999). Clair Benveniste proposed to identify the nucleus of an utterance in a macro-syntactic domain based on a *noyau* bearing a modal value (BLANCHE-BENVENISTE, 1997; BENVENISTE *et al.*, 1990). The definition of such an entity is a complex matter when its annotation in the speech flow is required. The main problem is that in spoken language a lot of configurations that are not clauses may turn out to be utterances in the speech flow. Almost 1/3 of speech events, according to the C-ORAL-ROM for the Romance languages and the Longman Grammar for English, do not have a verb and therefore do not show a clear syntactic structure. (BIBER *et al.*, 1999; MONEGLIA, 2005; MONEGLIA, 2006).

The following example taken from the LABLITA corpus of spoken Italian corresponds to one dialogic turn in which one speaker performs a word sequence. Considering the mere linear word sequence, no configuration pattern can be clearly identified and, from a pragmatic point of view, it is not possible to decide what the pragmatic value of any group of words is.

*SUS: *lei gliene serve una anch'a lei una in più o no no lei ha questa*
[you need one more too or not no you have this one]

According to pragmatic tradition (AUSTIN, 1962), the utterance is *the minimal linguistic entity such that can be pragmatically interpreted*; i.e. the linguistic entity that is 'concluded' and 'autonomous' from a pragmatic point of view (QUIRK *et al.*, 1985; CRESTI, 2000), but pragmatics can hardly benefit from corpus data if the object that carries pragmatic qualities, i.e. the utterance, is not identified in a corpus. For instance the above sequence cannot be interpreted even if one knows the context of the utterance (the speaker has been asked by a professor to make photocopies of a paper).

In this frame a speech event may also be identified as a dialogue act, and recorded in a dialogue representation scheme. This solution has been clear ever since the origin of corpus linguistic studies (see SINCLAIR; COULTHARD, 1975, and the literature cited below), but the task is hard to be undertaken and the identification of dialogue acts are difficult to be agreed upon, given that speech acts are also quite underdetermined (FAVA, 1995; KEMPSON, 1977).

In any event, however, this task necessarily requires considering the acoustic information, since the evaluation of the prosodic performance is crucial to determine the value of a speech act. Therefore, the access to acoustic information is the basic requirement for whatever exploitation of spoken corpora in the domain of pragmatics.

In the above example, the solution could be: "listen to a speech extract and provide your parsing of the speech flow into utterances". But what determines the parsing of the speech flow once the acoustic information is provided? The operative criteria which lead to the annotation of utterance boundaries in the speech flow must be explicit in order for the obtained data to be reliable and consistent for pragmatic and linguistic studies.

Approaches may diverge on this. My point is that the reference unit for spoken language is not underdetermined if pragmatic and prosodic features of speech are taken into account. Classic studies on prosody have always highlighted the fact that utterances end with a terminal profile (CRYSTAL, 1975; KARCEVSKY, 1931) and this quality is clearly perceived by speakers in conjunction with the assignment of an illocutionary value to a stretch of speech. From this point of view, this simple property can be considered a property equivalent to speech acts, to be used as a heuristic to determine the utterance boundaries in the speech flow: each string ending with a perceptively relevant terminal break is an utterance, in principle matching with a speech act (MONEGLIA, 2005).

According to this method, the speech flow and its transcription can be easily parsed. In the above example, when the grouping of words through intonation is considered , the identification of speech act boundaries is "naturally" guaranteed and it turns out that the above mysterious dialogic turn is made up of four utterances (marked by the terminal signs "?" or "//").

*SUS: *lei /gliene serve una anch'a lei ? una in più / o no ? no // lei ha questa //*
[you / (do) you need one also for you ? one more / or not ? no // you have this one //]

The criterion for the identification of utterance boundaries is intonation-based. This criterion does not imply the evaluation of the different intonation features and their categorization (evaluation of the movement types, tones, levels, focal points), which is very complex, but is only based on perception: detection of terminal and non-terminal prosodic breaks.[1] These cues are so prominent that they require little training to be recognized. Moreover, the experience of corpus annotation has shown that the perception of terminal breaks is consistent at a cross-linguistic level; English, Dutch, Italian, French, Spanish, European Portuguese, Brazilian Portuguese, Hebrew – all have been the object of this annotation with successful results (IZRE'EL *et al.*, 2005; AMIR *et al.*, 2004; MONEGLIA *et al.*, 2005; MONEGLIA *et al.*, 2010; BUHMANN *et al.,* 2002).

This practice allows for the possibility to get low cost information on speech acts from huge amounts of corpus data. It is reliable from the point of view of the detection of utterances in the speech flow. In this approach, the parsing of the speech flow into discrete speech events is not a function of the recognition of a specific speech act type by the labeler in any annotation schema, since the assignment of utterance boundaries is independently motivated. This property is in some sense quite widely recognized. Some spoken dialogue annotation tasks, for instance, the DRI/DAMSL and HCRC system dialog act codings, work under the same assumption (see CARLETTA *et al.,* 1996; JURAFSKY *et al.,* 1997) i.e. the dialogue act labeling and segmentation of 'utterances' is understood to proceed in tandem.

[1] Prosodic breaks must not be mixed up with pauses when looking at utterance boundaries. In around 60% of cases, pauses act as a re-enforcement of terminal prosodic breaks; however also around 40% of non terminal breaks are accompanied by a pause. See Moneglia (2005).

The correspondences between labeled Break Indices (i.e. intonational and intermediate phrases) and the majority of dialogue act boundaries are compatible with results from earlier studies about the relationship between intonational features and discourse (e.g. LEHISTE, 1975; NAKATANI *et al.,* 1995; SWERTS, 1997, SHRIBERG *et al.*, 1998). Dialogue act boundaries usually coincided with intonation boundaries in the MAP TASK corpus with matches of 88% for HCRC moves, and 84% for DAMSL dialogue acts (see below).

However, working within this frame, the prosodic boundaries strategy has not been really exploited having as an end the annotation of dialogue acts. Although the coding scheme for dialogue acts provides a closed list of possible moves, a competent speaker may find it difficult to identify and define the performed act. The replicability of the coding scheme is, as a matter of fact, one of the main problems for the annotation of dialogue acts, even in quite restricted domains.

For this reason, once the utterance limits are identified, the language string corresponding to an utterance is the linguistic entity which is suitable for receiving a certain tag. In other words, the definition of utterance limits is a matter of direct perception, while the assignment of a specific value to a dialogue act is a categorization issue, involving our knowledge of linguistic values. One can count speech acts without a clear agreement on their illocutionary value.

The annotation of utterance boundaries, according to the annotation of prosodic breaks perceived with a terminal value, does not go hand in hand with the ability to assign a specific value to an utterance (categorization task), but rather with the judgment that the utterance is an object of interpretation in the world.

In other words, a competent speaker can agree with the fact that the utterance being regarded can be interpreted, but may diverge, for many reasons, as to the specific value to be assigned to the utterance itself. The capacity to assign the quality of "being interpretable in the world" to a stretch of speech follows from this "illocutionary principle" and is a function of perception that is based on unconscious features.

This idea is not foreseen in the Searlian paradigm (SEARLE, 1983), in which intentional activities, such as language understanding in this case, are up to consciousness. Understanding that a stretch of speech is an object of interpretation in the world is not a function of the conscious assignment of a specific interpretation.

## 2.2. Speech act performance and syntactic relations

The traditional point of view that the reference unit of spontaneous speech can be detected when the relation among words generates autonomous compositional elements is strongly challenged by the prosodic strategy just described. Most scholars partake the view that prosodic criteria must be considered but always in conjunction with syntactic and semantic evidence (BENVENISTE, 1997, Rhapsodie Project). This is reasonable, but may lead to thorny problems if pragmatics is to be taken seriously.

The following argument shows, in particular, that the parsing of the speech flow into discrete units according to the detection of terminal prosodic breaks is not only a necessary condition, but also a sufficient condition. Speech event boundaries can be identified through prosodic boundaries apart from any other semantic or syntactic consideration, since prosodic boundaries are a function of speech act performance.

The underdeterminacy of syntactic structures in spontaneous speech is not only linked to the absence of verbs, but also to pragmatic activity itself. For instance, when a verbal proposition may, in principle, be figured out from the speech data, this not always provides the actual structure of a speech act. The dialogic turns reported below from the French and Portuguese C-ORAL-ROM collection are presented in a bare transcription without prosodic tagging. The two strings show the same superficial structure, which is a verbal nucleus followed by an adverbial expression:

| | |
|---|---|
| *EMA: ça c' est clair de plus en plus | [this is clear more and more] |
| *NOE: estive no Chiado há pouco tempo | [I have been in the Chiado recently] |

In both cases, on the basis of semantic and syntactic considerations, it is possible to figure out one proposition; i.e. one verbal nucleus with an adverbial extension. However, if the prosodic information provided by terminal breaks is considered, the two strings show different pragmatic properties since only the second one is accomplished within the boundary of one utterance, while in the first case the adverbial expression performs an independent monorematic utterance, and it is an independent 'adverbial clause', as in the following notation.

| | |
|---|---|
| * EMA: [1] ça c' est clair // [2] de plus en plus // | [this is clear // more and more //] |
| *NOE: [1] estive no Chiado / há pouco tempo // | [I have been in the Chiado / recently //] |

This has consequences on pragmatic grounds: given that two illocutionary activities are accomplished by the speaker, we cannot figure out only one syntactic structure. The idea (quite *a prioristic* indeed) that speech acts are just a matter of performance and that one single syntactic program is "executed in two utterances" is not consistent with the pragmatic interpretation. If pragmatics regards units of reference corresponding to speech acts and their structure, it cannot be admitted that "one speech act performance" is the performance of two speech acts. What is actually performed by the speaker must be taken seriously by the theory. Therefore, in no circumstance will the adverb modify the verb, since it gives rise to an independent reference unit. The syntax of pragmatic units is not independent.

Of course the reverse is also true. Although in principle an adverbial clause can accomplish one utterance, in no circumstance, in the second turn, does the adverb perform an independent act, since it does not follow a terminal break and does not bear any illocutionary value alone. In summary, the access to the prosodic information determines the structure of the speech flow; it does not read the structure of independently motivated semantic/syntactic entities.

## 3. Speech act variation and the representation of the language usage

### 3.1 Corpus-based detection of speech act variability

A corpus-based pragmatics must provide the means to specify what the speech acts actually performed in ordinary conversation are, and what differentiated linguistic properties they show. Searle's taxonomy (SEARLE, 1969) is still probably the most influential speech act classification. The taxonomy was set up at the end of the sixties within a logical paradigm and is based on lexical properties. In his conception, the linguistic counterpart of a speech act, i.e. the utterance, is equivalent to a *performative* predicate applied to a proposition [$F$ (p) = u]. On the basis of a "Principle of Expressibility" a correspondence between *speech act* types and performative verbs is established. Speech acts are defined in accordance, as a set of performative verbs belonging to five classes sharing a set of necessary and sufficient conditions of application (SEARLE, 1969).

However, when carrying out corpus-based experimental research, this point of view turns out to be not adequate to capture real data. The richness of the actions carried out in ordinary conversations is not recorded by the list of performative verbs, and the *Expressibility Principle* does not provide a valid heuristic to detect actual speech acts, which have, almost always, a *primitive* form in spontaneous speech. More specifically, while importance is given to linguistic actions that never occur in corpora or are rare, several – even very common – *speech acts* are not identified, since they have no equivalent performative (*refusal, deixis, call, instructions*). Even more intriguing: in spontaneous speech, although a performative sentence may, in principle, provide possible paraphrases for an utterance in its primitive form, there is no guarantee that the act actually performed belongs to that type.

The general point is that classical speech act theory lacks giving the appropriate value to prosody, which is the real means used in speech to express speech act types. For this reason, our knowledge about the set of speech act types that are possible in language (and their definition) is still very far from a satisfactory state. We will discuss this in 4. Nevertheless, corpus-based studies solely provide the most promising data which can increase our understanding in this domain.

Let us take a look at some findings in corpus-based speech act detection and classification.

A lot of work was being done at the end of the 90's in the domain of Dialogue Act Modeling for Automatic Recognition of Conversational Speech. In the map tasking coding scheme (ANDERSON *et al.*, 1991), the set of possible dialogue acts (moves in the map task) were investigated and the relevant link between prosodic and discourse structures underlined. Stirling *et al.* (2001) studied these moves and their relation to prosody in accordance with the HCRC map task coding scheme (CARLETTA *et al.*, 1996, 1997) and the 'Switchboard' version of the DRI/DAMSL scheme (JURAFSKY *et al.,* 1997) in detail. The following is the set of acts identified in the richer DAMSL coding scheme.

TABLE 1
DAMSL coding scheme

| SWBD-DAMSL codes (based on JURAFSKY *et al.*, 1997) | |
|---|---|
| **Forward-Communicative-Functions:**<br>*Statement*<br>statement-non-opinion<br>statement-opinion<br>*Influencing-addressee-future-action*<br>open option<br>yes-no-question (info-request)<br>wh-question (info-request)<br>open question (info request)<br>or-question (info request)<br>or-clause after y/n question (info request)<br>rhetorical question (info request)<br>declarative question (info request)<br>tag question<br>action-directive<br>*Committing-speaker-future-action*<br>offer<br>commit<br>*Other-forward-function*<br>conventional opening<br>conventional closing<br>explicit performative<br>exclamation<br>other forward function<br>thanking<br>you're welcome<br>apology | **Backward-communicative-functions:**<br>*Agreement*<br>accept<br>accept-part<br>maybe<br>reject-part<br>reject<br>hold<br>*Understanding*<br>signal non-understanding<br>acknowledge<br>backchannel in question form<br>acknowledge-answer<br>repeat phrase<br>collaborative completion<br>summarize/reformulate<br>appreciation<br>sympathy<br>downplayer<br>correct-misspeaking<br>*Answer*<br>yes answers<br>no answers<br>affirmative non-yes answers<br>negative non-no answers<br>other answers<br>no plus expansion<br>yes plus expansion<br>statement expanding y/n answer<br>expansions of y/n answers<br>dispreffered answers |

However, map task is spontaneous speech recorded in one quite peculiar situation only. Current trends in corpora which document a huge variety of sociolinguistic and pragmatic domains show that the set of possible speech acts may vary in accordance with the variety of contexts that are sampled in the corpus.

Yuki, Abe and Lin (2005) show that 50 types of functions have been identified by the Usage Based Linguistic Informatics Group (UBLI) in Japanese corpora, and more types have been extracted from the other foreign languages. A reduced table of the more frequent acts, derived from the combination and comparison among the previous ones, has been proposed below.

TABLE 2
40 Functions (YUKI; ABE; LIN, 2005)

| | | | | |
|---|---|---|---|---|
| Greetings | Thanking | Attracting attention | Introducing oneself | Apologizing |
| Giving | Saying goodbye | Asking Information (price) | Asking Information (experience) | Telling one's plan |
| Asking Information (degree) | Asking Information (time) | Asking Information (number) | Saying how and why | Asking skill and ability |
| Asking Information (existence and place) | Asking Information (attribute) | Saying one's opinion | Saying one's taste (thing) | Saying one's taste (behaviour) |
| Stating procedure and order | Asking what one is | Saying how one acts under certain circumstance | Comparing (comparative and superlative degree) | Suggesting |
| Confirming duty/negating | Prohibiting | Instructing | Asking for unacceptable thing | Confirming duty / affirming |
| Inviting | Advising | Demanding | Stating one's hope | Introducing someone |

The analysis carried out during the last decade based on our Italian corpora has led to the identification of a larger set of about 90 *speech act types* in speech (CRESTI; FIRENZUOLI, 2001; FIRENZUOLI, 2003).

TABLE 3
LABLITA Corpus Based Reference Table of Speech Acts Classes and Types

| **REPRESENTATIVES** | Objection | **DIRECTIVES** | Announcing | **EXPRESSIVES** | Complaint | **RITES** |
|---|---|---|---|---|---|---|
| Concluding | Confirmation | Distal recall - not visible object | Advising | Exclamation | Imprecation | Thanks |
| Weak assertion | Approval | Distal recall - visible object | Warning | Expression of contrast | Insinuation | Greetings |
| Answering | Disapproval | Proximal recall | Suggesting | Expression of obviousness | Derision | Apologies |
| Commentary | Agreement | Distal deixis | Proposal | Softening | Provocation | Welcome |
| Strong assertion | Disagreement | Proximal deixis | Recommend | Expression of surprise | Reproaching | Congratulation |
| Identification | | Presenting (object/event) | Invite | Expression of fear | Hint | Wishes |
| Verification | | Introducing (person) | Prompt | Expression of relief | Encouragement | Compliments |
| Claim | | Request of information | Permit | Expression of uncertainly | Assuring | Declarations of legal value |
| Hypothesis / Supposition | | Request of action | Authorize | Expression of doubt | Threatening | *condemnation* |
| Explanation | | Order | Prohibition | Expression of Certainty | Giving up | *condolences* |
| Inference | | Total question | Instruction | Expression of wish | | *baptism* |
| Definition | | Partial question | | Expression of Disbelief | | *promise* |
| Narration | | Alternative question | | Expression of Pity | | *bet* |
| Describing | | Request of confirmation | | Irony | | |
| Quotation | | Reported speech | | Regret | | **REFUSALS** |

The LABLITA research, based on a corpus of 10h; 9300 utterances, ranging over a large variety of informal situations,[2] also showed that 90% of utterances perform a set of roughly 30 speech act types, which are the more common in everyday conversation. The relative frequency of speech act classes in this corpus is the following:[3]

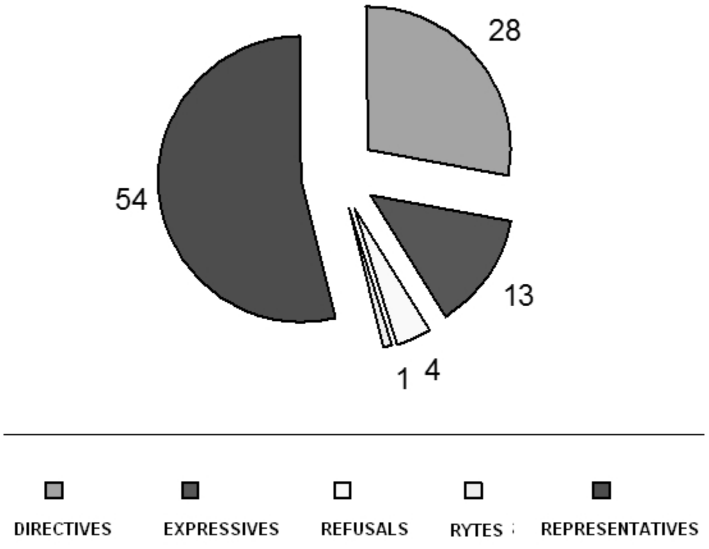


FIGURE 1 - Percentage of utterances of 5 Illocutionary classes in a sampling of the LABLITA Corpus (FIRENZUOLI, 2003)

Despite the difference in annotation strategy and label choice, these tagsets record at least some similar items. Comparing the three tables we can observe that the LABLITA tagset is larger, but the intersection of recorded speech act types is quite reduced. One easy conclusion: should one wish to represent the spontaneous speech universe in order to capture the variety of possible speech act types, the constitution criteria must ensure the widest

[2] Ratio of utterance sampling is 15%.

[3] Despite huge theoretical differences in the definition, the LABLITA illocutionary classes turn out very close to Searle's ones, with minimal adjustment (Commissives are not considered, Declarations are named Rites, an idiosyncratic class "Refusal" which cover a frequent act "NO!" in speech).

possible variation in speech contexts, and the lowest control in the speech event, which is exactly the opposite of what collections that are restricted to a specific task do.

A second requirement follows from the fact that, in the previous LABLITA pie, rites are unduly underrepresented. This is contradicted by the objective frequency of *salutation*, *thanks* and other everyday conventional declarations in daily life. This shortcoming depends on the practical choices made in transcribing samples. The beginning of interactions is almost always avoided since interactions start being more natural (ignoring the recording apparatus) after a while; the end of interactions are almost never sampled, since samples are always shorter than the whole of the interaction. Therefore, if this kind of act should be investigated, the corpus sampling must provide data regarding full pragmatic interactions. This means, more general criteria of corpus design should be integrated with criteria regarding how sessions are sampled. In this case, the map task strategy prevails.

The variety of types of conventional activities allowed by the linguistic system is obviously to be found within the main classes of Representatives, Directives and Expressives which record the highest number of speech act instances and, therefore, contain the relevant variation. This is the main area for future corpus-based research that is, however, strongly dependent on annotation schemas and identification criteria.

## 3.2 Corpus Sampling

The setting up of spontaneous speech language resources must ensure a huge corpus variety to allow speech act type detection. This requirement is similar to what happens with lexical variation. The representation of a sufficient number of contexts, covering relevant types of speech events in the universe, is the only possible strategy to get data for a frequency lexicon. A high-frequency lexicon may be underrepresented in specific pragmatic domains which, on the other hand, may maximise the probability of occurrence of low-frequency lexical items. The linguistic properties of the speech events vary in connection with non-linguistic variations. The connection between non-linguistic variation and linguistic variation goes beyond the frequency of lemmas: while lexical variation depends on topics, pragmatic variation depends on the needs of the interactive context and on the speaker's personal attitude and habits in that context. The goal to represent the variety of speech acts performed in everyday life from language usage data poses a problem of representation that is common in corpus linguistics, but is particularly sensitive in the spoken domain. There are relevant technical constraints to speech recording that are not present in written resources. Moreover, speech performance consistently varies from context to context and from individual to individual depending on many parameters.

According to sociolinguistic studies (LABOV, 1966; BERRUTO, 1987; BIBER, 1988; DE MAURO *et al.*, 1993; GADET, 1996) and also to recent initiatives for the annotation of corpus metadata (IMDI), the spontaneous speech universe foresees variation parameters that can be divided into three main groups: a) channel parameters; b) contextual parameters; d) demographic parameters.

*Channel variation*

a. Face-to-face interactions in natural contexts
b. Telephone recordings
c. New media audiovisual interactions
d. Human / machine interactions
e. Media productions
f. Written to be spoken

*Contextual variations parameters*

a. *Structure of the linguistic event*: speech events having a dialogue or a multi-dialogue structure vs. monologues
b. *Social context*: interactions belonging to *family* and *private life* vs. *interactions taking place in public*
c. *Domain of use:* domains of social environments, activities and professional domains such as law, business, research, education, politics church, etc.
d. *Genre:* lesson, debate, chat, row, storytelling, professional explanation, interview *etc*
e. *Register*: context requirements regarding formal register vs. informal language uses

*Demographic parameters:* the main sociologic qualities of speakers

a. Age
b. Sex
c. Education

d. Occupation
e. Geographical origin
f. Social class
g. City vs. Country

The impact of such variation parameters on the spontaneous speech universe cannot be pre-theoretically foreseen as for instance in the written part of the BNC. To provide a significant sampling of the population according to demographic parameters and then record them across their lifespan is, in principle, the best strategy. If the socio-demographical sampling of the population is valid, the linguistic sampling will also be valid as far as this population will be recorded through all relevant contexts of the day. All contexts occurring in society will have probability of occurrence according to their frequency in the life of the population and at the same time all language styles and personal variations due to sociologic qualities will be captured.

CoSIH (IZRE'EL *et al.*, 2001) was designed to ensure this. Day-long recordings of 950 informants representing all social and ethnic groups of the Israeli population have been planned over a one-year period. In this procedure informants are captured in recordings while they go through all the contextual and interpersonal situations that occur in the day, so ensuring speech data that are balanced at the same time both at sociological and contextual variation levels.

However, the CoSIH approach is not easily pursued. Indeed, to my knowledge, no corpus has been accomplished at present with this approach. From a practical point of view the recording of most contexts of use requires setting up a recording apparatus beforehand, and those situations remain excluded if not planned. The strategy is also difficult to be applied for legal or moral reasons in countries where the signed agreement of each intervenient in a recording is required beforehand, and the recording of many professional situations like business transactions are constrained by strict rules that go beyond the expressed agreement of the speakers.

If the strategy is not followed coherently the result will have exactly those variations that are significant for speech act variation reduced. For instance the BNC tries to integrate demographic and contextual criteria and dedicates almost half of its spoken part to recordings provided by a significant sampling of the British population. Subjects were asked to record their conversations during a certain period of time, so testifying the actual use of spoken language in accordance with the variation caused by speaker's parameters. However, in practice, the results are limited to the sole context of chat at home, which is the easiest situation for recording, but provides a reduced variation of speech activities.

It should be clear that providing data through a statistically significant sampling of the population does not imply that all linguistic variations in the corpora are due to the sociolinguistic qualities of the speakers, i.e. age, education, geographical origin, role of the speaker in society (MORENO-FERNÁNDEZ, 2005). For instance, a story told to a child and a row between husband and wife, which are my favorite examples, vary a lot depending on topics, language register, lexical choice and syntactic complexity according to the socio-cultural level of the speakers. However, crucially for pragmatics, the illocutionary quality of the utterances recorded therein can be better foreseen on the basis of context requirements. *Veiled threats*, *protests* and *refusals* will have high probability in a row regardless of the demographic sample. On the other hand, reported speech, narration, explanations have high probability of occurrences in storytelling.

In short, a sociological sampling of the population is valid in so far as it also captures relevant context variations, which is highly predictive of illocutionary variation. Therefore the sampling strategy must capture a huge amount of context variation in order to be a valid source of data for pragmatics. Assuming this conclusion, the comparison with lexical frequency corpus sampling needs can help us understand what the guidelines for setting up a valid corpus for pragmatics study should be. To the ends of lexical frequency the variation in topics and the wording actually used can be derived from a higher or lower probability of occurrence of those arguments in the world. The sampling is adequate when all parameters have probability of occurrence in their relative frequency. This need is much less relevant for speech act types, for which the goal is not to retrieve the relative frequency of a type, but rather, at present, to identify most possible types and their qualities. For this reason the contextual variation testified in the corpus must focus on variation of contexts rather than on their probability of occurrence.

## 4. Corpus data and the definition of *speech act* types

### 4.1 The illocutionary values of intonation

The definition of the value of an utterance as a conventional activity performed in the speech flow is strongly dependent on interpretations that may be highly subjective. It can reach a sufficient degree of inter-rater agreement when the appropriate tagset for a well-defined situation is applied, as is the case with map task (CARLETTA *et al.*, 1997), but remains vague in an unlimited context.

For instance the following dialogic turn has been interpreted within the LABLITA tagset as a sequence of one question, an alternative question; a self-answer; and one act of conclusion (tags in the annotation line %ill).

*SUS: [1] *lei /gliene serve una anch'a lei ? [2] una in più / o no ? [3] no //* [4] *lei ha questa //*
[you / (do) you need one also for you ? one more / or not ? no // you have this one //]
%ill: [1] question; [2] alternative question; [3] self-answer; [4] conclusion

This annotation has been done mainly on the basis of the interpretation of the value conveyed by the prosody of each utterance and the value of the semantic content in that context. As a matter of fact, in corpus-based research relevant *speech acts* are not identified either on the basis of the occurrence of performative verbs or on the basis of the logic of conversation, as in the Searle/ Grice paradigm (SEARLE, 1975, GRICE, 1975).

Let's concentrate on the fourth utterance that has been tagged as a "conclusion". It must be highlighted that, on the basis of possible performative paraphrases and contextual adequacy, the value *verification* could also have been assigned, or alternatively the simple value *assertion*. This question could be considered underdetermined in principle. We will show in the following that it is not undeterminate, if the differential value of prosody is considered. We will see that the exploitation of prosodic cues is crucial if spoken corpora must contribute to pragmatics.

In ordinary speech, prosody is essential to pragmatic interpretation. In natural speech a language string cannot receive an interpretation at all without prosody, which is the necessary interface between the illocutionary and the locutionary act. This is obvious when speech acts like *assertions, orders* or *questions* are concerned. It is well known that every language has melodic shapes conventionally codified to express sentence modalities, and this is one of the main functions of prosody. For instance, the following Italian phrasing (*gira a destra* [turn right]) can perform in a given context either one *order* or one *assertion* according to its prosodic form.

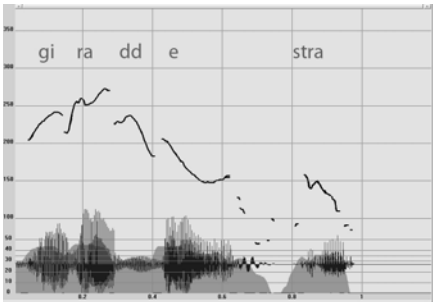
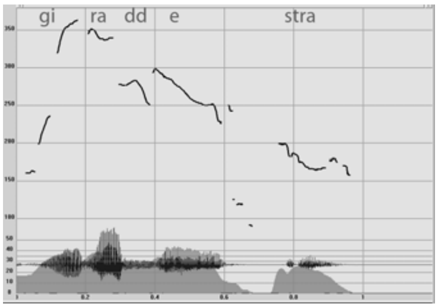
| 1) Answer / Assertion<br>*gira a destra* [(It) turns right] | 2) Order<br>*gira a destra* [turn right!] |
|---|---|
| - Does Viale Canova still continue after the Square?<br>- It turns to the right | - While driving<br>– Turn to the right ! |
|  |  |

FIGURE 2 - Answer vs. Order[4]

Although the theoretical framework for the description of prosodic features may vary, it can be verified that the prosodic form of the two acts have differential features. The following graph shows the two previous F0 curves overlapped in transparency.

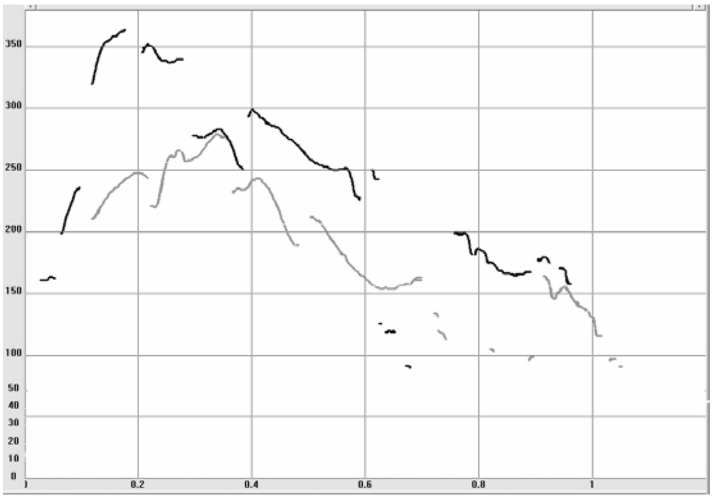


FIGURE 3 - Answer (gray) vs. Order (black) overlapped in transparency

[4] These and the following graphs have been generate by the speech software WINPITCH-PRO and correspond to the same female voice.

Very roughly speaking the prosodic nucleus of an assertive utterance (in gray) is characterized by:

Rising-falling movement. Rising at a mid F0 value followed by a gradual falling

- the post tonic syllable is longer
- mid intensity

The nucleus of an order (in black) is characterized by:

- rising-falling movement. Short optional rising preparation followed by a rapid falling (tail) on the tonic syllable, starting at high F0 values
- the post tonic syllable is short
- high intensity

Most scholars will also agree that the above profiles have a differential pragmatic value; i.e. they are a necessary feature in order for the utterance to be interpreted as an order or an assertion. This can be easily verified in an experimental setting. Given a pragmatic context requiring an order, the replacement of the appropriate order with the same stretch of speech intonated with an assertive profile is meaningless. We have carried this experiment out and the result is impressive (see below).

However, it is much less obvious to what extent the relation between prosody and speech acts characterizes ordinary speech and to what extent the study of prosodic profiles retrieved in spoken corpora can really help to characterize the system governing speech act performance. As we have just observed in the previous example, an *assertive* act can, in principle, be interpreted as a *conclusion or alternatively as a verification* and the potential adequacy in the context of an equivalent performative sentence ("I conclude that ..." "I verify that ...") does not really select the actual interpretation. As a matter of fact, we do not have explicit criteria to distinguish an "assertion" from a "conclusion" in the set of utterances which commit the speaker with the truth.

Moreover, sometimes the label derived from the interpretation of corpus data is not a performative verb. For instance, *to instruct* is not a performative verb, but the language activity to give *instruction* has been retrieved in all previous corpus studies as such. What ensures that this activity corresponds to an illocutionary act? Can we set up the conditions determining that an *instruction* is performed rather than just an *order* or a generic directive act? A tagset of 90 labels for speech act types needs very detailed specifications in order to be applied; otherwise, the definition of types within each illocutionary class remains underdetermined.

### 4.2 Prosody and "Empirical" Pragmatics

In this section the methodology for studying how prosody contributes to the illocutionary interpretation of spoken utterances will be sketched. To this end the interpretations derived from corpus analysis must be challenged on empirical grounds. The prosodic profile of the utterance to which the tag has been assigned must be repeated with different locutive content by different speakers. The appropriate context eliciting the act can be defined. The following is a summary of the standard empirical methodology for the exploitation of corpus data to the ends of empirical pragmatics:

*Corpus driven induction*

- collection of speech acts occurrences that have been judged to be of the same illocutionary type during corpus annotation

*Positive repetition of the profile in controlled elicitation context*

- operative description of the pragmatic characters of the elicitation context
- production and validation of a fictional elicitation context for one comment with the appropriate profile
- repetition and validation by different locutors of the profile as a function of the elicitation context
- adjustment and definition of the pragmatic features of the elicitation context that better allow the production of the profile

*Substitution test*

- the stretch of speech performed with the appropriate profile is substituted by the stretch of speech with other profiles
- competent speakers evaluate whether or not the profile fits the circumstances

*Differential prosodic properties*

- repetition of the profile on different accentual structures in the elicitation context
- description of the differential prosodic properties of the profiles
- synthetic modification of necessary features and validation of the range of accepted modification in the eliciting situation (not discussed in this paper).

Below, the overall problem of what determines the interpretation of a speech act in ordinary speech will be grounded, through questioning whether or not we can find differential prosodic features between the two acts studied, respectively *assertion* vs. *conclusion* and *order* vs. *instruction*.

Figure 4 presents the profiles found in our corpus in utterances respectively tagged as *answer*, *conclusion*, *instruction* and *order* that have been repeated by the same female locutor in experimental setting with respect to the same Italian locutive content "Gira a destra" [turn on the right (one)] which has been chosen for its semantic ambiguity . Figure 5 presents the overlapping transparencies of the two pairs of curves:

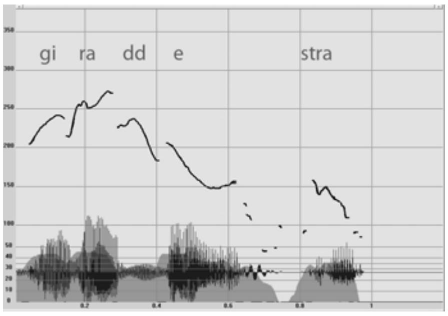
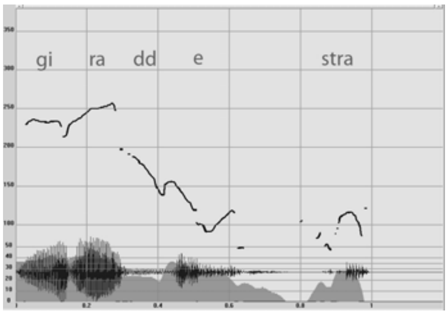
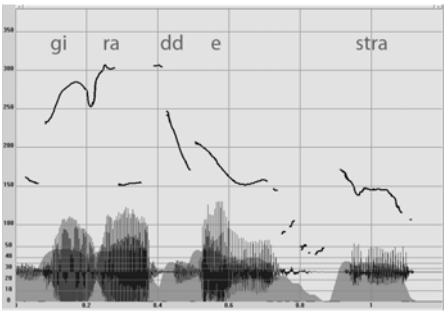
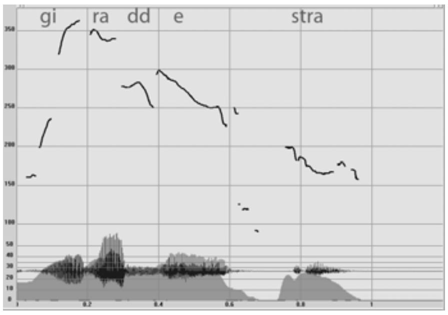
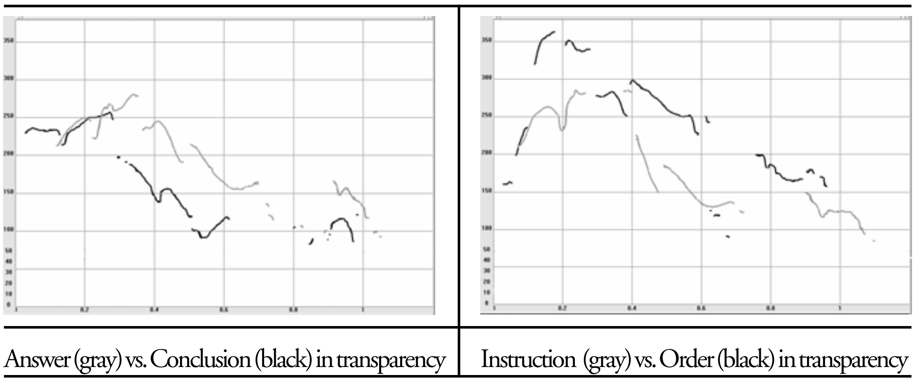
| 1) Answer / Assertion<br>*gira a destra* [(It) turns right] | 2) Conclusion<br>*gira a destra* [(It) turns right] |
|---|---|
| - Does Viale Canova still continue after the Square ?<br>- It turns right ! | [trying to understand the direction of a river from an old map]<br>- It turns right |
|  |  |
| 3) Instruction<br>*gira a destra* [(turn right] | 4) Order<br>*gira a destra* [turn right!] |
| - Sorry, do you know in which room is the Latin exam?<br>- Do you see the corridor? Turn right | [While directing somebody who cannot see]<br>– Turn right ! |
|  |  |

FIGURE 4 - "*gira a destra*" [(0 turns right] with the prosodic profiles of Anwer (1); Conclusion (2); Instruction (3); Order (4)

As can be seen from the overlapping of the two pairs of acts, the $F_0$ profiles of each speech act type belonging to the same illocutionary class are quite different.



Answer (gray) vs. Conclusion (black) in transparency | Instruction (gray) vs. Order (black) in transparency

FIGURE 5 - $F_0$ Illocutionary acts of the same illocutionary class in overlapping transparencies

The table below reports the main prosodic characteristics (with regard to $F_0$ and Duration properties) that have been highlighted in the study of the profiles in accordance with the IPO system.[5]

[5] The movements of the nucleus of the tone unit are described according to the IPO terminology in ‘t Hart, Collier, & Cohen 1990. The following is the matrix of possible movements.

| | /1/ | /2/ | /3/ | /4/ | /5/ | /A/ | /B/ | /C/ | /D/ | /E/ |
|---|---|---|---|---|---|---|---|---|---|---|
| rise | + | + | + | + | + | - | - | - | - | - |
| early | + | - | - | - | + | - | + | - | - | + |
| late | - | + | - | + | - | - | - | + | + | - |
| spread | - | - | - | + | - | - | - | - | + | - |
| full | + | + | + | + | - | + | + | + | + | - |

The system works according to the following parameters: Direction of Movement (rise-fall) / Position of Movement in the syllable (early-late) / Duration on syllables (spread or not) / If the movement cover or not the maximum-minimum excursion (full or not). See Firenzuoli (2003) for a more comprehensive framework.

TABLE 3
$F_o$ Features

| *Illocution* | Onset | Level | Range | Nucleus | Structure | Alignment |
|---|---|---|---|---|---|---|
| Answer | Mid/Low | Mid | Mid<br>100/200 Hz male<br>200/300 Hz female | [1A][D] | (Pre-nucleus)<br>(Tail) | Right side of the tone unit<br>[1A]<br>[D] on the focus |
| Conclusion | Low | Low | 100/150 Hz Male<br>150/250 Hz Female | [D] | (Pre-nucleus) | Right side of the tone unit |
| Instruction | Mid | Mid/High | Strong<br>80/250 Hz male<br>150/300 Hz female | [1A] [D]<br>(Final Plateau) | (Pre-nucleus) | Left side of the tone unit<br>[1A]<br>Right side<br>[D]<br>(Final Plateau) |
| Order | Mid/High | High | Strong<br>150/250 Hz male<br>100/400 Hz female | [1A]<br>sudden | (Pre-nucleus)<br>(Tail) | Left side of the tone unit<br>[1A] |

TABLE 4
Duration Features

| Illocution | Syllable Length of the nucleus | Speed of the utterance |
|---|---|---|
| Answer | Around 200ms. | Mid |
| Conclusion | Around 350 ms | Mid |
| Instruction | Around 250ms. | Slow |
| Order | Around 150ms. | High |

The speech acts with the prosodic profiles recorded in the above tables have been challenged. If a specific prosodic profile constitutes a differential feature in order to attribute the requested illocutionary value in the appropriate pragmatic context, then the illocutionary value is conventionally codified within the language system (reglardless of its lexical performative encoding). This is what is required by the system. The following tables summarise the pragmatic features that are needed to characterize the elicitation contexts for *answer / conclusion / order / instruction*.

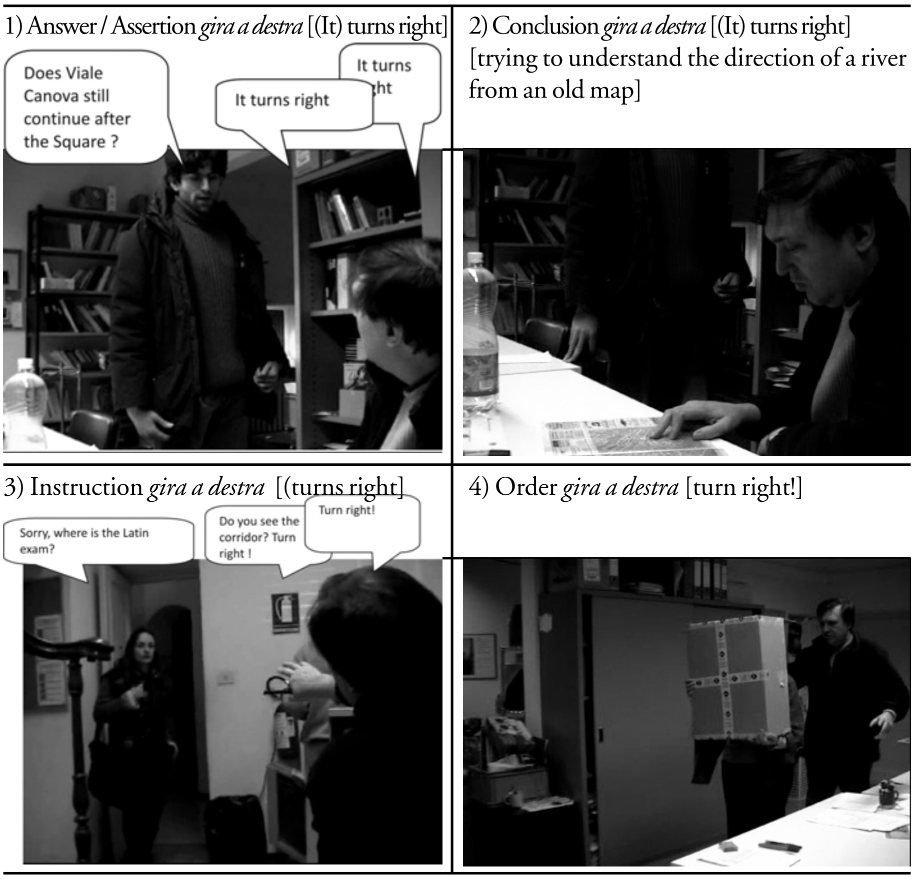
TABLE 5
Pragmatic features of the elicitation context for Assertion vs. Conclusion

| | **Answer** | **Conclusion** |
|---|---|---|
| communication channel | open | open |
| attention | shared | shared |
| prossemic between the speakers | direct interaction | no interaction |
| intentional features of the process | cognitive | cognitive |
| effect | shared information focus | indirect information focus |
| modifications in the partner | cognitive | not implied |
| perceptual characters of the referred objects in the pragmatic/cognitive context | no restriction | proximal |
| preparatory condition in the speaker | question by the hearer | problem in the context |
| preparatory condition in the hearer | expectation | no restriction |

TABLE 6
Pragmatic features of the elicitation context for Order vs. Instruction

| | **Order** | **Instruction** |
|---|---|---|
| communication channel | open | open |
| attention | shared | shared |
| prossemic between the speakers | direct interaction | direct interaction |
| intentional features of the process | behavioral | cognitive |
| effects | modification of the world | modification of knowledge and abilities |
| modifications in the partner | operative | cognitive |
| perceptual characters of the referred ontological entity in the pragmatic/cognitive context | presence of the referred ontological entity in the context | possibility to explore the context |
| preparatory condition in the speaker | social role and/or pragmatic skill | knowledge |
| preparatory condition in the hearer | possibility of intervention in the pragmatic situation | need of know-how |

These features are instantiated in scenes performed by actors and represented in Figure 6. Scenes are eliciting context for the appropriate prosodic profile.


Does Viale Canova still continue after the Square ?
It turns right
It turns ght
Sorry, where is the Latin exam?
Do you see the corridor? Turn right !
Turn right!

1) Answer / Assertion *gira a destra* [(It) turns right]

2) Conclusion *gira a destra* [(It) turns right] [trying to understand the direction of a river from an old map]

3) Instruction *gira a destra* [(turns right]

4) Order *gira a destra* [turn right!]

FIGURE 6 - Elicitation contexts for Answer, Conlusion, Instruction, Order

In the elicitation contexts the profiles work fine with the appropriate illocutionary values and are easily replicated by the speakers. Elicitation features are quite compulsory. For instance, in context 6.2, we discovered that as soon as the actor addresses the utterance to the hearer looking at him, despite the intention to replicate the *conclusion profile*, the outcome bears the *answer profile*, while the utterance can be naturally performed with the *conclusion profile* when the speaker does not look at him, but rather concentrates on the object he is evaluating. The feature "no interaction" is therefore a necessary trait for acts of *conclusion*. If the *conclusion profile* is forced in the context eliciting the answer, the speech act is judged as "depressed utterance". If, on the other hand, the *answer profile* is forced in the eliciting context of a conclusion, than the utterance is not judged as a conclusion any more, but rather as a simple assertion.

In the case of Figure 6.4 we discovered that the *order profile* is hard to be replicated in all contexts in which the hearer understands what to do on the basis of his evaluation of the information provided in the order. The *instruction profile* is performed instead of the *order profile*. In context 6.4, this is not the case and the order profile is easily elicited. The differential feature "behavioural" vs. "cognitive" is crucial to foresee if the prosodic profile of "order" or the prosodic profile "instruction" will be performed. When the instruction profile and the order profile are forced in a context that is adequate to the other, the result is totally unacceptable.

We must underscore that features that are responsible for the above systematic prosodic variation bearing illocutionary value; i.e. the underling linguistic form of the speech act, cannot be identified on the basis of any deductive process. In particular, "shared attention without eye contact" for *conclusion* and "Cognitive vs. Operational process in the interlocutor" for *order* and *instruction*, must be considered idiosyncratic constraints (naturalistic) that can display their pragmatic relevance on the basis of an empirical investigation.

The idea that context determines the illocutionary interpretation of utterances, which originated from Austin, does not fit in with the above experimental data. In order to get an utterance appropriate interpretation, a specific prosodic profile is required as a necessary condition. From a logical point of view, we would have expected, for instance, the *instruction*-intonated utterance and the *order* to be both acceptable in the above contexts, since they belong to the *directive* class; but the intonation requirement is compulsory. While the context supports and elicits the prosodic performance of the utterance, it does not determine its value. Therefore we can infer that the distinctions between order vs. instruction and answer vs. conclusion respectively follow from conventional features borne by prosody, and are therefore genuine illocutionary distinctions codified within the language system.

The identification of these speech act types, as many others, is strictly dependent on the availability of large corpus data in which those acts have probability of occurrence. The definition of the pragmatic constraints to the performance of those acts (i.e. the semantic forms underlying them) is not associated to a lexical item but rather to prosodic forms. Therefore, in summary, corpus data enhance pragmatic theory in two main respects, both crucial:

a) The possibility to have a clear picture of the natural speech act types actually performed in ordinary language usage requires corpora covering a huge variety of contexts in which those acts have probability of occurrence and a long work of annotation and experimental verification; b) the inner semantic form of speech act types can be derived from pragmatic investigations which are grounded on experimental data rather than on sole logical inference.

## 5. Information patterning and pragmatic functions in the Language into Act Theory

### 5.1 The *comment* unit

Prosody carries out various functions: a) segmentation of the speech continuum into groups (structural function); b) expression of differential modal acts (statement, order, question etc.) regardless of their segmental content; c) expression of emotions and attitudes (not considered in this paper) (BOLINGER, 1972; 1989; CRYSTAL, 1969; CRYSTAL; QUIRK, 1964; DANEš, 1960; LADD, 1980; ROSSI, 1985; 1999).

The structural function is in principle separated from modal and expressive functions and for what regards corpus data it is linked to the need for annotation of prosodic parsing inside speech transcriptions (BRAZIL, 1995). For instance the annotation of prosodic parsing in the Santa Barbara Corpus of American English (DUBOIS *et al.*, 2000) foresees the marking of both terminal and non terminal breaks. In that coding scheme (CHAFE, 1993) Intonation Units are considered basic organizational units of speech, but according to the overall conception of spoken language (CHAFE, 1987; 1994) they represents *ideas* activated at the consciousness level and bring about the flow of thought rather than speech acts. The previous arguments regarding the importance of terminal breaks for marking speech acts boundaries go in a different direction. In this section we will present other arguments to demonstrates that, besides the segmentation of the speech flow into speech acts, the internal prosodic parsing of the utterance is also relevant to pragmatics if we want to exploit spoken corpora for its ends.

Although the number of utterances without internal prosodic segmentation is consistent in interactive speech, in the majority of cases utterances are not composed of a single word-grouping, but correspond to a complex pattern.[6] From this point of view the utterance has often been

[6] In the Italian C-ORAL-ROM subcorpus, the percentage of utterances made up of groupings of more than one word is over 57%, but in the formal domain it is generally much higher (CRESTI; MONEGLIA, 2005).

indicated by a dual functional opposition (WEILL, 1844; MATHESIUS, 1929;) in terms of theme/propos (BALLY, 1950), theme/rheme (Prague functionalism, SORNICOLA; SVOBODA, 1989), topic/comment (HOCKETT, 1958), topic/focus (CHOMSKY, 1971; JACKENDOFF, 1972; LAMBRECHT, 1994), given/new (HALLIDAY, 1976), prefix/noyau (BLANCHE-BENVENISTE, 1987; 2003). More recently a lot of scholars interested in dialogue structure and pragmatics have also focused on other components of the utterance that have a clear pragmatic value, that is, discourse-makers and the functions they carry out to regulate the dialogue (SCHIFFRIN, 1987; BAZZANELLA, 1995; BAZZANELLA *et al.*, 2008).

Although almost all researchers noticed that word grouping is marked by prosody, only few of them have exploited this property to study the pragmatic organization of speech. The research carried out at LABLITA in the frame of the Language Into Act Theory (CRESTI, 1994; 2000; CRESTI; MONEGLIA, 2010) points out that the annotation of prosodic parsing is strictly necessary to specify functional structures of the utterance and specifically those components that have pragmatic values.[7]

In this frame every utterance corresponds to an information pattern which is systematically signalled by an intonation pattern whose units are marked by non-terminal prosodic breaks. The intonation pattern is therefore isomorphic to an information pattern. The most important innovation brought by the Language into act Theory is that information is ruled within actual spoken language use according to pragmatic principles (CRESTI, 1987). The core of the utterance does not correspond to a predication or to a focus, but rather to the expression of the illocutionary value. Crucially the expression of the illocutionary force is up to one and only one word grouping within a prosodic envelope. The information unit so defined (the Comment) accomplishes the illocutionary force and for this reason it is the only unit which is necessary and sufficient to give rise to an utterance.

In other words, in spontaneous interactive speech, if the utterance is simple, i.e. it is made up of one prosodic envelope only, then it does not show an information structure and necessarily bear the prosodic cues for its illocutionary interpretation. If on the contrary the utterance is made up of more then one envelope, than only one of them bears illocutionary cues (the

[7] The Informational Patterning Theory has been introduced in Cresti (1987) and Cresti (1994) and developed in many publications after the reference book (CRESTI, 2000). See also the debate on *Macro-syntax* in Scarano (2003).

Comment). For this reason the Comment unit constitutes the core information unit of the utterance, i.e. the utterance cannot be interpreted at all if this unit is erased. On the other hand, in a complex utterance all units other than the Comment can be erased without compromising the possibility of an interpretation to be assigned.

This is a severe constraint regarding the way the illocutionary force of the utterance is expressed in spontaneous speech. Whatever the length of the utterance parsed by prosody might be, there is always one and only one prosodic unit which bears the illocutionary cues allowing its interpretation.

For instance, let us consider the first utterance of the previous dialogic turn which is made up of two prosodic units:

*SUS: [1] *lei /gliene serve una anch'a lei ?COM* [you / (do) you need one also for you ?]

The second unit is the Comment and it is necessary and sufficient for the interpretation. This is not because of its sentential form. For instance, in the following examples, taken from C-ORAL-ROM and C-ORAL-BRASIL, a clause structure appears in the second unit, but this is not the Comment. The first unit can be interpreted in isolation, while the second, although it contains a verb, can be erased without the loss of illocutionary value.

*LIA: Baratti /COM mi pare fosse <stato> // PAR [ifamcv01] [the Baratti Goulf, I think It was that place]
*LUZ: duas vagas /COM eu acho //PAR [bfamdl03] [two positions, I think]

The second unit has a propositional form, but it cannot be interpreted in isolation since it lacks the prosodic information which specifies how to relate it to the world. If the first unit is erased, it is perceived as "suspended". In the following verbless utterance (again, an answer), the illocutionary cues are in the second unit, which, again, cannot be erased.

*LID: il mi' bisnonno /TOP Pietro //COM [ifamdl02] [My grandfather, Peter]
*LAU: departamento /TOP Artes Plásticas //COM [bfamdl03] [Department, Fine Arts]

It may be interesting to note that we cannot provide any distributional evidence for the above assertions without taking into account the prosodic form and the ability by competent speakers to assign or not an interpretation. In other words, the distributional evidence requires both an acoustic source and speaker's judgments.

In light of the above considerations we can underscore the following requirements for pragmatic analysis of spontaneous speech corpora. The internal organization of the utterance into prosodic units is an essential feature of spoken corpora annotation. In order to identify the illocutionary values expressed by prosody, the speech flow must be parsed into terminal and non terminal prosodic units and for each utterance the autonomous unit must be identified by competent speakers. This task highlights the core of the utterance and distinguishes the illocutionary information unit (the Comment) from all other units. The pragmatic definition of the Comment unit within the utterance structure is a crucial finding to bootstrap the illocutionary values conveyed by prosody from corpora. The relevant prosodic cues are foreseen in one unit of the utterance only, whatever the length of the utterance might be, so making this task affordable.

## 5.2 Speech acts and Dialogue acts

Within the utterance, various types of information units, all optional from an informational point of view, can surround the Comment.[8] They also correspond to prosodic units and are divided into two classes dedicated to different types of functions: a) the textual construction of the utterance (Topic, Appendix, Parenthesis, Locutive Introducer – not considered here); b) its communicative support (Incipit, Phatic, Allocutive, Conative, Connector.[9] These last units are of special interest to pragmatics, and the prosodic annotation of internal prosodic boundaries of the utterance is again essential to their corpus based identification.

In the last twenty years, new data derived from a better consideration of spoken language have made it clear that language expressions proper have not been clearly identified in the grammatical tradition. These expressions, usually referred to with the term "Discourse markers" (SCHIFFRIN, 1987), are dedicated to perform the peculiar pragmatic functions that are required to manage the dialogic interaction. More specifically, they are signals directed to the interlocutor, carrying some specific functions. For instance:

[8] The main aspects of the informational patterning of the utterance are described in Cresti (2000).

[9] "Substantial" vs. "Regulatory" Intonation units, in Chafe's terms.

- turn taking
- attention request
- opening of the communication Channel
- phatic function
- keeping the communication channel open
- reception control

Discourse markers are found in all languages and can be roughly identified on the lexical level. For instance, in English expressions like *listen*, *guys, I mean*; in Spanish *o sea* , *pues nada*; in French *hein alors donc écoute*; in Italian *senti*, *guarda*, *allora*, *eh*; Brazilian Portuguese *né*, *cara, oi*̉ , may play this role.

Although most of these expressions might be, in many cases, assigned to their traditional Part of Speech, in certain positions of the speech flow, they lose their usual meaning and morphosyntactic value and play a dialogue regulation function instead. Moreover, in conjunction with this peculiar value, these expression are not any more compositional, i.e. they do not contribute to the propositional meaning of the utterance, and can, in principle, be eliminated without effect on the propositional meaning itself (BAZZANELLA, 2006; SCHOURUP, 1999; FRASER, 2006). There is no agreement as to the number of Discourse Markers, their functions, nor criteria for their definition (FISCHER, 2006).

This important chapter of present day's pragmatic theory can strongly benefit from corpus data only if the prosodic and informational properties of discourse markers are taken into account. Indeed, and also in this case most authors have noted the strong correlation with prosody; in particular, discourse markers tend to occur in the dedicated tone unit in which they are isolated.

Indeed, this property fits in with the general framework of the Language into Act Theory. Discourse markers are just information units and, in accordance with a general principle, they show in one-to-one correspondence with prosodic units. Discourse markers fit in with units playing a set of functional roles in the frame of dialogue regulation.

The prosodic character can allow the individuation of discourse markers in speech, i.e. it helps to specify when the above lexical items are compositional elements which play their usual PoS role and conversely when they work as dialogue regulators. For instance in the following two examples the Italian verb *guarda* [look] is used as a discourse marker in a *conative* function (to push the interlocutor to a shared point of view (CRESTI, 2000). It is isolated respectively in first and final position, it is not compositional and for this reason it is hardly thought of as a verb.

*LIA: guarda / io ti dico così // [look / I will say this way //
*SRE: ti stavamo aspettando / guarda // [we were waiting for you / look ]

The prosodic break is a necessary feature in order to get a Conative dialogue act. It would be unacceptable to group the same stretch of speech within one sole prosodic envelope:

* guarda io ti dico così // [look I will say this way //
* ti stavamo aspettando guarda // [we were waiting for you look ]

The opposite occurs when the same expression (*guarda*) is a verb in a functional relation with its propositional object, as in the following example:

*MAX: guarda come tu stavi // [look in what a shape you were]

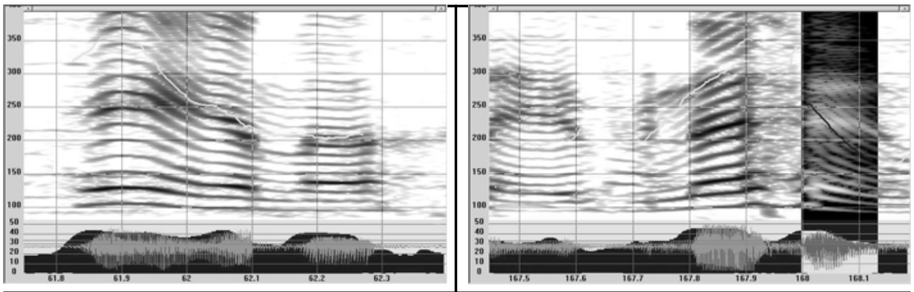
The expression is not isolated in a distinct prosodic unit and since it is a verb, in no circumstances could it be interpreted as a means to accomplish a dialogue act. The prosodic parsing is the main heuristic to foresee whether the expressions commonly used as discourse markers play a dialogic function or, on the contrary, participate in the construction of the propositional content of the utterance. The frequencies of dialogue-type functions are very high and, more than 50% of utterances performed in spontaneous conversations present these kinds of devices (FROSALI, 2006).

Dialogic units are information units with pragmatic value which on one side must be distinguished from the locutive expressions which contribute to the propositional meaning and on the other from the linguistic units which perform illocutive acts. In other words, from a pragmatic point of view, a clear distinction between speech acts and discourse particles must be made.

The occurrence of a discourse marker in speech may be recorded as a dialogue act in an annotation schema and listed within the series of natural speech acts. This is conceivable since these expressions do not have a propositional value but rather perform an interactive function. The DAMSL coding scheme, for instance, takes this view. However, this is misleading, since these entities are not autonomous from a linguistic point of view, as a speech act should be. In the Language into Act theory these expressions are considered information units within the utterance with a dialogue regulation function, and are clearly distinguished form units which specify their relation to the world (Comment).

For instance Raso & Mello (forthcoming) have distinguished *allocutives* from the *recall illocution* (either proximal or distal). This is an interesting case, since specifically the proper name of the interlocutor can be used in both functions. Indeed grammars consider both under the category of vocative. *Allocutives*, however, are dialogic units of the utterance which play a cohesive and empathic function. They specify to whom the message is directed keeping his attention. *Recall* is a speech act whose function is to get the attention of somebody toward the speaker "opening a close communication channel".

The following examples by the same speaker clearly show their prosodic and functional distinction.



| | |
|---|---|
| *FLA: *Rena* // =COM= [Proximal call] | *FLA: é uma neve / =COM= né / =PHA= *Rena* // =ALC= |

FIGURE 7 - Proximal call vs. Allocutive Dialogic unit in Brazilian Portuguese

Recall illocutions correspond to a Comment information unit of a one word utterance, showing a higher intensity, a much higher duration and a functional focus that allows for their interpretability in isolation, besides presenting a high F0 variation. On the other hand, Allocutives have a flat or falling profile, without focus, low intensity and duration (roughly 1/5 of the recall) and are one prosodic unit of a complex utterance. The Allocutive cannot be interpreted in isolation. For instance, cutting off the rest of the utterance and keeping "Rena" in the second utterance, the stretch of speech cannot be interpreted as a speech act. On the other hand, if the call is inserted within a stretch of speech both in starting or final position, it gives rise to a terminal break and the stretch of speech is considered a sequence of utterances, i.e. it cannot be integrated within the utterance with another Comment unit.

## Conclusions

Our knowledge about the set of possible speech act types is very far from a satisfactory state. The availability of large collections of spoken language corpora is an opportunity for present day pragmatics to bootstrap speech act variation from the actual use of language, so grounding pragmatic knowledge on strong empirical evidence. However corpora compilation and annotation must follows a set of requirements to this end. At the level of compilation, corpora should ensure the maximum context variation to give, in principle, probability of occurrences to all speech act types, whose variation does not depend no either speakers or topics, but rather on what activities are relevant in a given interaction.

The identification of the linguistic counterparts of speech acts in the speech flow (utterances) is the main requirement for what regards corpus annotation. We have shown that the parsing of spontaneous speech into utterances should be a function of prosodic annotation, since prosodic breaks (and terminal breaks in special) are a necessary correlation of the speech act performance. This prosodic parsing has a crucial advantage: it is easily recovered by competent speakers, while both syntactic structure and pragmatic categorization of the speech flow are strongly underdetermined. The access to acoustic information which provides prosodic evidence is, therefore, the basic requirement for whatever exploitation of spoken corpora in the domain of pragmatics.

More generally, empirical pragmatics relies heavily on the study of prosody as far as the exploitation of spontaneous speech data is concerned. The relation prosody/speech acts is crucial for speech act categorization, since in the ordinary use of language, speech acts types are necessarily performed through conventional prosodic forms. More specifically, we have shown that prosody is a differential feature of speech act types, and that those features are strictly required to accomplish the appropriate acts in their elicitation contexts. In short, prosody is the necessary interface between locutive and illocutive acts.

Finally, under the Language into Act theory, we have proposed that the link between prosody and pragmatics also influences the internal information structure of the utterance. Only one prosodic unit within the utterance is devoted to the expression of the illocutionary cues (Comment Unit) and, for this reason, it constitutes the core of the utterance itself. This step allows a clear distinction between the main pragmatic functions performed within the utterance: illocutionary activity performed by the Comment Unit and dialogue regulation activities performed by other units and referred back to the Comment.

## References

ABEILLE, A. *Treebanks Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic, 2003.

AMIR, N.; SILBERT-VARODZ, V.; IZRE'EL, S. (2004), Characteristics of intonation unit boundaries in spontaneous spoken Hebrew: Perception and acoustic correlates. *SProSIG*, p. 677-680, 2003.

ANDERSON, A.; BADER, M.; BARD, E.; BOYLE E.; DOHERTY, G.; GARROD, S.; ISARD, S.; KOWTKO, J.; MCALLISTER, J.; MILLER, J.; SOTILLO, C.; THOMPSON, H.; WEINERT, R. The HCRC map task corpus. *Language and Speech,* v. 34, p. 351-366, 1991.

AUSTIN, L. J. *How to Do Things with Words*. Oxford: Oxford University Press, 1962.

BALLY, C. *Linguistique Générale et Linguistique Française*. Berne: Francke Verlag, 1950.

BAZZANELLA, C. I segnali discorsivi. In: RENZI, L.; SALVI, G.; CARDINALETTI, A. (Ed.). *Grande Grammatica di Consultazione*. Bologna: Il Mulino, 1995.

BAZZANELLA, C.; BOSCO, C.; GILI FIVELA, B.; MIECZNIKOWSKI, J.; TINI BRUNOZZI, F. Polifunzionalità dei segnali discorsivi, sviluppo conversazionale e ruolo dei tratti fonetici e fonologici. In: PETTORINO, M.; GIANNINI, A.; VALLONE, M.; SAVY, R. (Ed.). *La comunicazione parlata*. Napoli: Liguori, 2008. v. II.

BERRUTO, G. *Sociolinguistica dell'Italiano Contemporaneo*. Roma: La Nuova Italia Scientifica, 1987.

BIBER, D. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press, 1988.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. *The Longman Grammar of Spoken and Written English*. London / New York: Longman, 1999.

BLANCHE-BENVENISTE, C. *Approches de la Langue Parlée en Français.* Paris: Ophrys, 1997.

BLANCHE-BENVENISTE, C.; BILGER, M.; ROUGET, Ch.; VAN DEN EYNDE, K.; MERTENS, P. *Le Français Parlé:* Études Grammaticales. Paris: Éditions du C.N.R.S., 1990.

BLANCHE-BENVENISTE, C. Le recouvreman de la syntaxe et de la macro-syntaxe. In: SCARANO, A. (Ed.). *Macro-syntaxe et Pragmatique. L*'analyse Linguistique de l' Oral. Roma: Bulzoni, 2003.

BNC http://www.natcorp.ox.ac.uk/

BRAZIL, D. *A Grammar of Speech.* Oxford: Oxford University Press, 1995.

BOLINGER, D. L. (Ed.). *Intonation*: Selected readings. Harmondsworth: Penguin, 1972.

BUHMANN, J.; CASPERS, J.; VAN HEUVEN, V.; HOEKSTRA, H.; MARTENS, J-P.; SWERTS, M. Annotation of prominent words, prosodic boundaries and segmental lengthening by no-expert transcribers in the spoken Dutch corpus. *In Proceedings of LREC 2002* Paris: ELRA. p 779-785, 2002.

CARLETTA, J.; ISARD, A.; ISARD, S.; KOWTKO, J.; DOHERTY-SNEDDON, G.; ANDERSON, A. HCRC dialogue structure coding manual. HCRC/TR-82. Human Communication Research Centre, University of Edinburgh, 1996.

CARLETTA, J.; ISARD, A.; ISARD, S.; KOWTKO, J.; DOHERTY-SNEDDON, G.; ANDERSON, A. The reliability of a dialogue structure coding scheme. *Computational Linguistics,* v. 23, n. 1, p. 13-31, 1997.

CHAFE, W. Cognitive constraints on information flow. In: TOMLIN, R. (Ed.). *Coherence and grounding in discourse*. Amsterdam: John Benjamins, 1987.

CHAFE, W. (1993). Prosodic and functional units of language. In: EDWARDS, Jane A.; LAMPERT, Martin D. (Ed.). *Talking data:* Transcription and coding methods for language research. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.

CHAFE, W. *Discourse, consciousness, and time:* The flow and displacement ofconscious experience in speaking and writing. Chicago / London: The University of Chicago Press, 1994.

CHOMSKY, N. Deep Structure, Surface Structure and Semantic Interpretation. STEIMBERG, D.; JACOBOVITS, L. (Ed.). *Semantics*: an Interdisciplinary Reader. Cambridge: Cambridge University Press, 1971.

CRESTI, E. L'articolazione dell'informazione nel parlato. In: AA.VV. *Gli Italiani Parlati:* Sondaggi sopra la Lingua d'oggi. Firenze: Accademia della Crusca, 1987.

CRESTI, E. Information and intonational patterning in Italian. In: FERGUSON, B.; GEZUNDHAJT, H.; MARTIN, P. (Ed.). *Accent, Intonation, et Modéles Phonologiques.* Toronto: Editions Mélodie, 1994.

CRESTI, E. *Corpus di Italiano Parlato*. Firenze: Accademia della Crusca, 2000. v. I-II, CD-ROM.

CRESTI, E.; FIRENZUOLI, V. Illocution and intonational contours in Italian. *Revue Française de Linguistique Appliquée*, v. IV, n. 2, p. 77-98, 2001.

CRESTI, E.; MONEGLIA, M. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: Benjamins, 2005.

CRESTI, E.; MONEGLIA, M. Informational Patterning Theory and the Corpus based description of Spoken language. The composiotionality issue in the Topic Comment pattern. In: MONEGLIA, M.; PANUNZI, A. (Ed.). *Bootsrapping Information from Corpora in a Cross Linguistic Perspective*. Firenze: FUP, 2010.

CRESTI, E.; MONEGLIA, M.; TUCCI, I. Annotation de l'entretien avec Anita Musso selon la Théorie de la langue en acte. In: LEFEUVRE, F.; MOLINE, E. (Ed.). *Unités syntaxiques et Unités prosodiques*, *Langue Française*, 2011.

CRYSTAL, D.; QUIRK, R. *Systems of Prosodic and Paralinguistic Features in English.* The Hague: Mouton, 1964.

CRYSTAL, D. *The English Tone of Voice*. London: Edward Arnold, 1975.

DANEŠ, F. Sentence intonation from a functional point of view. *Word*, v. 16, p. 34-55, 1960.

DE MAURO, T.; MANCINI, F.; VEDOVELLI, M.; VOGHERA, M. *Lessico di Frequenza dell'Italiano Parlato.* Milano: ETAS, 1993.

DUBOIS, J. W.; CHAFE, W.; MEYER, C.; THOMPSON, S. A. *Santa Barbara Corpus of Spoken American English – Part 1*. Linguistic Data Consortium, 2000.

FAVA, E. Tipi di atti e tipi di frase. In: RENZI, L.; SALVI, G.; CARDINALETTI, A. (Ed.). *Grande Grammatica Italiana di Consultazione*. Bologna: Il Mulino, 1995.

FIRENZUOLI, V. *Le Forme Intonative di Valore Illocutivo dell'Italiano Parlato:* Analisi Sperimentale di un Corpus di Parlato Spontaneo (LABLITA). 2003. PhD (Thesis) – Università di Firenze, Firenze.

FISCHER, K. (Ed.). *Approaches to discourse particles*. Studies in Pragmatics 1. Bingley, UK: Emerald Group Publishing, 2006.

FRASER, B. Towards a Theory of Discourse Markers. In: FISCHER, K. (Ed.). *Approaches to discourse particles*. Studies in Pragmatics 1. Bingley, UK: Emerald Group Publishing, 2006.

FROSALI, F. Il lessico degli ausili dialogici. In: CRESTI, E. (Ed.). *Prospettive nello studio del lessico italiano* (Atti del IX Congresso SILFI), Firenze: FUP, 2006.

GADET, F. Variabilité, variation, variété. *Journal of French Language Studies*, v. 1, p. 75-98, 1996.

GRICE, H. Logic and Conversation. In: COLE, P.; MORGAN, G. *Speech Acts.* Syntax and semantics. New-York: Academic Press, 1975. v. 3.

HALLIDAY, M.A.K. *System and Function in Language*: Selected Papers. London: Oxford University Press, 1976.

‘t HART, J.; COLLIER, R.; COHEN, A. *A Perceptual Study on Intonation.* An Experimental Approach to Speech Melody. Cambridge: Cambridge University Press, 1990.

HOCKETT, C. F. *A Course in Modern Linguistics*. New York: The Macmillan Company, 1958.

IMDI http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf.

IZRE’EL, S. Intonation Units and the Structure of Spontaneous Spoken Language: A View from Hebrew. In: *Proceedings of the IDP05 on Discourse-Prosody Interfaces*, 2005.

IZRE’EL, S.; HARY, B.; RAHAV, G. Designing *CoSIH*: The corpus of spoken Israeli Hebrew. *International Journal of Corpus Linguistics*, v. 6, p. 171-197, 2001.

JACKENDOFF, R. *Semantic Interpretation in Generative Grammar*. Cambridge Mass: MIT Press, 1972.

JURAFSKY, D.; SCHRIBERG, L.; BIASCA, D. Switchboard SWBD-DAMSL Shallow-Discourse-Function-Annotation Coder’s Manual, Draft 13. Technical Report TR 97-02. Institute for Cognitive Science, University of Colorado at Boulder, 1997.

KARCEVSKY, S. Sur la phonologie de la phrase. In: *Travaux du Cercle linguistique de Prague* IV, p. 188-228, 1931.

KEMPSON, R. *Semantic Theory*. Cambridge:  Cambridge University Press, 1977.

LABOV, W. *The Social Stratification of English in New York City*. Washington D.C.: Center for Applied Linguistics, 1966.

LADD, D. R. *The structure of the Intonational Meaning*. London: Bloomington, 1980.

LAMBRECHT, K. *Information structure and sentence form*. Cambridge: Cambridge University Press, 1994.

LEHISTE, I. The phonetic structure of paragraphs. In: COHEN, A.; NOOTEBOOM, S. (Ed.). *Structure and Process in Speech Perception*. Berlin: Springer-Verlag, 1975.

MATHESIUS, V. La linguistica funzionale. In: SORNICOLA, R.; SVOBODA, A. (Ed.). (1991). *Il campo di tensione. La sintassi della scuola di Praga*. Napoli: Liguori, 1929.

MILLER, J.; WEINERT, R. *Spontaneous Spoken Language*. Oxford: Clarendon Press, 1998.

MONEGLIA, M. The C-ORAL-ROM resource. In: CRESTI, E.; MONEGLIA, M. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: Benjamins, 2005.

MONEGLIA, M. Units of Analysis of Spontaneous Speech and Speech Variation in a Cross-linguistic Perspective. In: KAWAGUCHI, Y.; ZAIMA, S.; TAKAGAKI, T. (Ed.). S*poken Language Corpus and Linguistics Informatics.* Amsterdam: John Benjamins, 2006.

MONEGLIA, M.; FABBRI, M.; QUAZZA, S.; PANIZZA, A.; DANIELI, M.; GARRIDO, J. M.; SWERTS, M. Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM corpus. In: E. CRESTI; MONEGLIA, M. (Ed.). *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins, 2005.

MONEGLIA, M.; RASO, T.; MALVESSI-MITTMANN, M.; MELLO, H. Challenging the perceptual relevance of prosodic breaks in multilingual spontaneous speech corpora: C-ORAL-BRASIL / C-ORAL-ROM in Speech Prosody 2010, W1.09, Satellite workshop on Prosodic Prominence: Perceptual, Automatic Identification Chicago. Available at: <http://aune.lpl.univ-aix.fr/~sprosig/sp2010/papers/102010.pdf>.

MORENO FERNANDEZ, F. Corpus of spoken Spanish language – The representativeness Issue. KAWAGUCHI *et al.* (Ed.). *Usage-Based Linguistics Informatics*. Amsterdam: John Benjamins, 2005.

NAKATANI, C.; GROSZ, B.; HIRSCHBERG, J. Discourse structure in spoken language: studies on speech corpora. In: Proc. AAAI-95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, 1995.

QUIRK, R.; GREENBAUM, S.; LEECH, G.; SVARTVIK, J. *A Comprehensive Grammar of the English Language*. London / New York: Longman, 1985.

RAPSODIE Project http://rhapsodie.ilpga.fr/wiki/Chaine_de_traitement

RASO, T.; MELLO, H. Allocutives as discourse markers: a comparative corpus-based study for Italian, Spanish, European Portuguese and Brazilian Portuguese. *Proceedings of the 2th International Pragmatics Conference.* Manchester, 3-8 July 2011. Forthcoming.

ROSSI, M. L'intonation et l'organisation de l'énoncé. *Phonetica*, v. 42, p. 135-153, 1985.

ROSSI, M. *L'intonation, le Système du Français*: Description et Modélisation. Paris: Ophrys, 1999.

SCARANO, A. (Ed.). *Macro-syntaxe et Pragmatique.* L'analyse Linguistique de l'Oral. Roma: Bulzoni, 2003.

SCHIFFRIN, D. *Discourse Markers.* Cambridge: Cambridge University Press, 1987.

SCHOURUPS, L. Discourse markers. *Lingua*, v. 107, p. 227-265, 1999.

SEARLE, J. *Speech Acts:* An Essay in the Philosophy of Language. Cambridge: Cambridge University Press, 1969.

SEARLE, J. *Intentionality.* An essay in the Philosophy of the Mind. Cambridge: CUP, 1983.

SEARLE, J. Indirect speech acts. In: COLE, P.; MORGAN, J. L. (Ed.). *Syntax and Semantics, 3:* Speech Acts. New York: Academic Press, 1975.

SHRIBERG, E.; BATES, R.; STOLCKE, A.; TAYLOR, P.; JURAFSKY, D.; RIES, K.; COCCARO, N.; MARTIN, R.; METEER, M.; VAN ESS-DYKEMA, C. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, v. 3-4, p. 443-492, 1998. Special issue on Prosody and Conversation, 41.

SINCLAIR, J. M.; COULTHARD, R. M. *Towards of Analysis of Discourse*: The English Used by Teachers and Pupils. London: Oxford UP, 1975.

SORNICOLA, R.; SVOBODA, A. *Il campo di tensione*. Napoli: Liguori, 1989.

STIRLING, J.; FLETCHER, I.; MUSHIN, R.; WALES, L. Representational issues in annotation: Using the Australian map task corpus to relate prosody and discourse structure. *Speech Communication,* v. 33, p. 113-134, 2001.

SWERTS, M. Prosodic features at discourse boundaries of different strength. *J. Acoust. Soc. Amer.* v. 101, p. 514-521, 1997.

SWERTS, M.; GELUYKENS, R. The prosody of information units in spontaneous monologues. *Phonetica*, v. 50, p. 189-196, 1993.

WEIL, H. 1844. De l'ordre des mots dans les langues anciennes comparées aux langues modernes. In: *The order of words in the ancient languages compared with that of the modern languages translation*, by C.W. SUPER. Amsterdam: Benjamins, 1978.

WINPITCH-PRO http://www.winpitch.com/

YUKI, K.; ABE, K.; LIN, C. Development and assessment of TUFS Dialogue Module-Multilingual and Functional Syllabus. In: KAWAGUCHI *et al. Usage-Based Linguistics Informatics*. Amsterdam: John Benjamins, 2005.

Recebido em 13/04/2011. Aprovado em 26/05/2011.

# Corpora and Cognitive Linguistics

## *Corpora e linguística cognitiva*

John Newman*
University of Alberta
Edmonton / Canada

ABSTRACT: Corpora are a natural source of data for cognitive linguists, since corpora, more than any other source of data, reflect "usage" – a notion which is often claimed to be of critical importance to the field of cognitive linguistics. Corpora are relevant to all the main topics of interest in cognitive linguistics: metaphor, polysemy, synonymy, prototypes, and constructional analysis. I consider each of these topics in turn and offer suggestions about which methods of analysis can be profitably used with available corpora to explore these topics further. In addition, I consider how the design and content of currently used corpora need to be rethought if corpora are to provide all the types of usage data that cognitive linguists require.

KEYWORDS: corpora, cognitive linguistics, metaphor, polysemy, synonymy, prototypes, constructional analysis, statistics, R statistical program

RESUMO: Corpora são uma fonte natural de dados para a linguística cognitiva, uma vez que, estes, mais que qualquer outra fonte de dados, refletem "o uso" – a noção que é frequentemente apontada como tendo importância crítica para o campo da linguística cognitiva. Corpora são relevantes para todos os principais tópicos de interesse da linguística cognitiva: metáfora, polissemia, sinonímia, protótipos e análise construcional. Neste artigo, considerarei cada um desses tópicos e oferecerei sugestões sobre quais os métodos de análise podem ser utilizados com os corpora disponíveis para melhor se explorarem esses tópicos. Adicionalmente, discuto como a arquitetura e o conteúdo dos corpora atualmente disponíveis necessitam ser repensados se pretenderem oferecer todos os tipos de dados de uso necessários às análises da linguística cognitiva.

PALAVRAS-CHAVE: Corpora, linguística cognitiva, metáfora, polissemia, sinonímia, protótipos, análise construcional, estatística, programa estatístico R

* john.newman@ualberta.ca

## 1. Cognitive Linguistics

In its broadest sense, *cognitive linguistics* is concerned with general principles that provide some explanation for all aspects of language, including principles drawn from disciplines other than linguistics (cf. EVANS; GREEN, 2006, p. 27-28). Any intellectual movement which attempts to be so all-encompassing in its scope and so multi-disciplinary in its approach must inevitably lead to a proliferation of data types, methodologies, and strategies of persuasion. One may not be convinced that the field of cognitive linguistics will achieve all the goals it has set itself–which intellectual movement has?–but one can be grateful to the rise of cognitive linguistics for the balance it brings to the study of language, offering linguists a more rounded and more complete agenda for research than the relatively circumscribed, self-absorbed, self-referential, inward-looking kind of theorizing which constituted mainstream linguistic research, at least in syntax, in the latter half of the twentieth century.

In a reflective critique of the current state of cognitive linguistics, especially its methodologies, Geeraerts (2006, p. 29) refers to the "growing tendency of Cognitive Linguistics to stress its essential nature as a usage-based linguistics", a tendency which points unequivocally in the direction of corpus-based research. Certainly, there are now important collections of papers exemplifying corpus-based methods for a cognitive-linguistic audience, e.g., Gries and Stefanowitsch (2006), Stefanowitsch and Gries (2006), and Lewandowska-Tomaszczyk and Dziwirek (2009). Geeraerts choice of words–"growing tendency"–hints, however, at the unsettled role of usage-based methods in the field. On the one hand, it is sometimes claimed that such methods are a crucial component of a cognitive linguistic approach (cf. EVANS; GREEN, 2006, p. 108). Clearly, however, not everyone who purports to be a cognitive linguist has seen usage-based methods in quite the same way. If they did, there would not be a "growing tendency" to rely on such; instead, the use of these methods would be firmly entrenched in the practice of cognitive linguistics.

In the following sections, I comment on a variety of ways in which corpora can be exploited in the study of topics which have been central in the field of cognitive linguistics: metaphor, synonymy, polysemy, prototypes, and constructional analysis. Throughout, the emphasis will be on methods which I consider most promising in terms of their potential to yield interesting results.[1] The Appendix contains the R scripts (see The R Project for Statistical Computing) I relied on to calculate the statistical measures and visualizations of the data. I will also comment briefly on the kinds of corpora which cognitive linguists should give more attention to.

## 2. Metaphor

A major advance in research on metaphor, and a prominent part of the cognitive linguistic movement, has been the adoption of a conceptually based approach to understanding metaphor. If ARGUMENTS ARE STRUCTURES and LOVE IS WAR, it is the concepts of argumentation, war, love etc. which are at the heart of these metaphors, not the words *argument*, *war*, *love* etc. This reconceptualization of metaphor research opened up fascinating new avenues of research. While acknowledging the conceptual breakthrough behind this development, I suggest that it is now appropriate to broaden the scope of inquiry into metaphor by appealing to more usage-based methods, in particular corpus-based methods, than has been customary. Taking metaphor research in this direction is, in fact, perfectly in keeping with the increasing empiricization of the field. My views on this seem to be simply echoing views already expressed by some researchers within the field. Ray Gibbs, for example, explicitly endorses a greater role for corpus-based research on metaphor in the following remarks:

> But I am most impressed these days with the development of work on corpus linguistics and conceptual metaphor. The research using corpora is important because it forces scholars to be more explicit on the procedures used for identifying metaphor in both language and thought, which is a necessary complement to more traditional introspectionist cognitive linguistic work on conceptual metaphor. (GIBBS, quoted in VALENZUELA, 2009, p. 310)

Later in the same interview, Gibbs calls for more attention to lexical and grammatical behaviors associated with metaphorical usage, which I, too, would advocate.

What might a corpus-based approach to metaphor entail? I see the insights of the cognitive linguistic research as providing a natural starting point for a corpus-based approach, in so far as these insights have identified a host of

[1] I am grateful to Ewa Dąbrowska for sharing the data on English motion verbs in Table 3 and Dagmara Dowbor for sharing her complete data on *over*. I am also grateful for comments by anonymous reviewers on an earlier draft of this article.

metaphorical mappings from source to target domains which in turn should stimulate research into the usage of these metaphors. ARGUMENTS ARE STRUCTURES may be a possible metaphorical mapping, but, compared with other metaphorical mappings, how often does this particular mapping actually occur? Or, how often do argument events show such structure (always, nearly always, hardly ever, etc.)? When discourse participants rely on that metaphor, which lexical items and constructions, as a matter of fact, constitute the vehicle of expression of the metaphor? Which conceptual mappings are more entrenched? We have become accustomed to rethinking syntactic concepts like ditransitive structures, passive structures etc. in terms of actual usage of such structures. In a similar way, it is appropriate to rethink the ARGUMENTS ARE STRUCTURES conceptual mapping, and others like it, in terms of usage. Usage of some metaphors more than others may lead us to insights into processing biases some of which may reflect fundamental aspects of human conceptualization while others may reflect culturally specific preferences. Investigating the degree of entrenchment of a metaphorical structure, established through the study of corpus-based usage, seems the natural next step to take if we are interested in such questions.

The presence of metaphorical usage in a corpus is clearly not something that is easily ascertained, unless the corpus has already been annotated to facilitate such searches.[2] An example of such a corpus would be one which has been annotated for word senses along the lines of WordNet (FELLBAUM, 1998) and related projects such as Euro WordNet (VOSSEN, 1998) and Global WordNet.[3] In WordNet, various senses of a word receive different *sense keys* which identify the sub-senses of a word and a set of semantically similar words, a *synset*, may share the same sense key. So, for example, *war* and *warfare* constitute one synset, sharing a sense key of 1:04:02 representing the shared sense of an active struggle between competing entities. In this sense key, "1" represents the syntactic category of noun, "04" denotes nouns representing acts or actions, and "02" is a unique identifier of the (shared) sense of the lemma

[2] See Philip (in press) for a helpful and critical review of various semi-automated approaches to identifying metaphors in a corpus. In Philip's own approach (2010, in press) *keywords* are first identified (where *keywords* are understood as words which are more common in a statistically significant sense in one corpus than in a reference corpus). Further procedures are followed, focusing on the *low-frequency* content words among the key words. See also Fass (1991) for an interesting computer-based approach to discriminating between metonymy and metaphor.

[3] See <http://www.globalwordnet.org/>.

*war* and the lemma *warfare*. In Table 1 below, we see three different sense keys associated with the pair *war* and *warfare*, the senses they represent, and an example of use of each.

TABLE 1
Selected WordNet sense keys and uses of *war* and *warfare*

| | WORD | SENSE KEY | SENSE | EXAMPLE |
|---|---|---|---|---|
| 1 | *war* | 1:04:00 | the waging of armed conflict against an enemy | *thousands of people were killed in the* ***war*** |
| 2 | *war* | 1:04:01 | a concerted campaign to end something that is injurious | *the* ***war*** *on poverty* |
| 3 | *war* | 1:04:02 | an active struggle between competing entities | *a* ***war*** *of wits* |
| 4 | *warfare* | 1:04:02 | an active struggle between competing entities | *Diplomatic* ***warfare*** |

It is clear how useful this kind of annotation would be if searching for metaphorical uses of both *war* and *warfare* – one would search these words when they occur with sense keys 1:04:01 or 1:04:02. It should be noted, though, that simply searching on these number sequences alone will not uniquely identify metaphorical uses. Other synsets of nouns relating to actions can utilize these same numeric sequences for other kinds of senses. For example, *battle* has senses utilizing these sequences where they indicate the senses shown in Table 2. In this case, the 1:04:01 sense is arguably the metaphorical use and 1:04:02 the more literal use. The sense key establishes a unique sense in relation to lemmas within a synset, but not between synsets. Thus, even with a WordNet annotated corpus, there is no simple search that will identify all metaphorical meanings relating to WAR.[4]

[4] Examples of corpora annotated for English WordNet are SemCor 3.0, based on the BROWN corpus (<http://www.cs.unt.edu/~rada/downloads.html#omwe>) and the Princeton WordNet Gloss Corpus of 1.6 million words (<http://wordnet.princeton.edu/glosstag.shtml>). There is substantial work involved in annotating a corpus for WordNet senses (as there is for most kinds of semantic annotation) which helps explain the lack of wide-spread availability of such corpora. A promising approach to a practical solution for adding WordNet annotations is found in Stamou, Andrikopoulos, and Christodoulakis (2003) where the authors describe a module WnetTag which retrieves and displays all

TABLE 2
Selected WordNet sense keys for *battle*

| | WORD | SENSE KEY | SENSE | EXAMPLE |
|---|---|---|---|---|
| 1 | *battle* | 1:04:01 | an energetic attempt to achieve something | *he fought a* ***battle*** *for recognition* |
| 2 | *battle* | 1:04:02 | an open clash between two opposing groups (or individuals) | *police tried to control the* ***battle*** *between the pro- and anti-abortion mobs* |

Another way of assigning semantic categories to words, though less complete than WordNet in the range of semantic relations to be indicated, is the "UCREL semantic analysis system" (USAS).[5] USAS relies on a classification into broad categories represented by the letters of the alphabet and narrower sub-categories indicated by additional delimiting numbers. For example, the category of government and the public domain is the G category, G2 is crime, law and order, G2.2 is general ethics. There is also a category of names and grammatical words that is assigned to words traditionally considered to be empty of content (i.e., closed class words) and proper nouns. USAS has been developed with automatic semantic tagging in mind and a detailed description of the tagging process and the array of sub-routines required to effectively disambiguate senses can be found in Rayson, Archer, Piao, and McEnery (2004). In principle, a corpus annotated by means of USAS would be of great advantage when it comes to identifying metaphorical usage. As Hardie, Koller, Rayson and Semino (2007) point out, the semantic fields assumed by USAS correspond, approximately, to the "domains" we are familiar with from metaphor theory (WAR, TIME, ACTIONS, STATES etc.). A word such as *campaign* would have multiple semantic tags associated with it, reflecting its varied uses. Hardie *et al.* (2007) discuss the need to implement a "broad-sweep" approach to identifying the relevant semantic tags associated with words like *campaign* where the metaphor researcher might well wish to retrieve both the G3

Wordnet senses of a given term to enable a more efficient annotation by the (human) annotator. For a sophisticated exploitation of WordNet to identify metaphorical senses of words (in the WordNet database, as opposed to searching in a corpus), see Peters and Wilks (2003).

[5] UCREL stands for University Centre for Computer Corpus Research on Language, Lancaster University. More details on USAS can be found at <http://ucrel.lancs.ac.uk/usas/>.

‘warfare’ semantic category and the X7 ‘wanting, planning, choosing’ category since these two categories represent the source and target domains in a usage like *advertising campaign*. USAS has been implemented in the software suite Wmatrix (RAYSON, 2003; 2007), but to date we still lack publicly available corpora annotated in this way.[6]

Typically, then, exploring metaphorical usage in a corpus will require a good deal of inspection and decision-making by a researcher (see the extensive discussion of issues associated with this task in STEEN, 2007). Boers (1999) stands out, still, as a simply designed but very revealing corpus-based study of metaphor, which involved the manual inspection of all instances of HEALTH metaphors in the editorials ofThe Economist magazine over a ten-year period. A corpus was constructed, guided by the occurrence of HEALTH metaphors in the editorials over a ten-year period, of about 1,137,000 words from articles accompanying those editorials. Although the resulting corpus may be small by current standards, the task of identifying all HEALTH metaphors in such a corpus (i.e., HEALTH as the target domain rather than the source domain) is not something one would take on lightly. The metaphors identified by Boers include a wide range of forms, as one might expect: *sickly firms*, *diagnosing a shortage*, *the market cure*, *surgery that costs jobs* etc. Some of the expressions were identified as clear instances of a HEALTH metaphor (e.g., *the market cure*), while others were categorized as vague or ambiguous (e.g., *economic remedy*). Boers relied on the Collins Cobuild English Language Dictionary to help make principled decisions about the two categories: when the first (i.e., more frequent) usage in the dictionary entry actually mentioned the domain of physical health, then its figurative use in the corpus was categorized as clear; otherwise the figurative use was categorized as vague or ambiguous. The quantitative data was then presented in two ways – the clear metaphors only and all the metaphors, though the two sets of data were similar.

A somewhat similar, though methodologically more refined, approach to metaphor identification is MIP, i.e., “metaphor identification procedure”, as developed by the Pragglejaz Group (2007). MIP seeks to make a clear distinction between metaphorical and non-metaphorical usage and does so in a relatively programmatic way. Thus, a researcher is expected to work through

[6] It has been announced that the International Corpus of English (ICE) will be annotated using USAS and Wmatrix. See <http://ice-corpora.net/ice/index.htm> for details of these corpora and updates on progress.

a series of specified steps to arrive at a decision about the metaphorical status of a word, with the word being the relevant unit of interest. The researcher must determine if the word has a more basic contemporary meaning in other contexts and, if so, whether the word in the use being investigated can be understood in comparison with the more basic meaning.[7] The use of STRUGGLE in a context such as *someone struggled to convince X of Y* is claimed to be a clear case of metaphorical use if the basic meaning of STRUGGLE is taken to be 'use ones physical strength against someone or something'. The use of CONVINCE in the same example is taken to be non-metaphorical since there is no other more basic meaning found in other contexts. As in the case of Boers' approach referred to above, MIP relies, in part, on dictionary entries to guide decisions about basic vs. non-basic meaning. The strength of MIP lies in the explicitness of the procedure and the further testing of the reliability of the procedure (the procedure includes a recommendation that researchers report the statistical reliability of their analyses, e.g., measuring reliability across cases and reliability across analysts) in order to arrive at defensible and replicable results. Even so, the authors in the Pragglejaz Group recognize the challenge of applying their procedure, commenting that "it is not a task that can be accomplished easily or quickly" (Pragglejaz Group 2007, p. 36).

There are, however, various ways in which one might try to reduce the amount of manual inspection associated with exploring source and/or target domain behaviors, recognizing that undertaking a full-blown MIP-type analysis of a large corpus is not feasible (see STEFANOWITSCH, 2006, p. 1-6 for a more extensive review of possible methods):

(a) *Use a small corpus to first identify items of interest* (cf. DEIGNAN 2005, p. 93). Cameron and Deignan (2003) begin with a small corpus of 28,285 words of transcribed talk by primary (elementary) school children to identify, exhaustively, forms and patterns relevant to their interest in metaphor. Their method of identifying metaphorical usage appeals to an earlier insightful discussion of issues surrounding metaphoricity in

[7] An extension of MIP is MIPVU, where VU stands for *Vrije Universiteit.* This extension is described in Steen, Dorst, Herrmann, and Kaal (2010). Among other differences with MIP, MIPVU recognizes a three-way distinction when identifying metaphorically used words: clear metaphor-related words, metaphor-related words that are doubtful, and words that are clearly not related to metaphor. The new "doubtful" category is reminiscent of Boers categories of vague/ambiguous cases of metaphorical usage.

Cameron (1999). Metaphorical usage was identified by Cameron, initially, in cases where there was an incongruity between Topic and Vehicle and where a coherent interpretation of that incongruity is possible. However, Cameron (1999) introduced a number of interesting and subtle qualifications to this general approach. For example, she distinguished "insider" and "outsider" perspectives, depending on whether the interpretation is from inside or outside the shared discourse world of speaker and listener. An example such as "This pillow is my spaceship", as spoken by a three-year old, might be seen as metaphorical for a typical adult hearing such an utterance out of context, but the utterance might be better seen as non-metaphorical from the point of view of the child engaged in creating a particular, imaginative scene and assigning a precise role to the pillow. The results from the smaller and fine-tuned exercise were then used by Cameron and Deignan (2003) as the basis for searching a larger 9 million word corpus (transcribed spoken data from the Bank of English). A variant of this approach uses a sample of texts as a way of first identifying metaphorical uses of interest, e.g., Charteris-Black (2004). When the focus of research is the overarching metaphorical structure of a large piece of discourse (e.g., the rhetoric of an election campaign, the metaphors at work in an advertising campaign, the interplay of metaphorical devices in a work of fiction etc.), then inspecting smaller texts is a natural way to sample the larger discourse, before moving on to other methods of analysis.

(b) *Use a large corpus, but create a fixed set of search terms*. This method is suitable when the research can identify key words in a semantic domain, e.g., domestic animals, weather conditions, modes of transport etc. An electronic thesaurus might be a useful tool to create some sets of related terms. Usually, though, it would be almost impossible to anticipate the full range of expressions, which might instantiate a concept. It would seem very unlikely, for example, that Boers findings about HEALTH, referred to above, could be replicated simply by deciding a priori on a set of forms to investigate. Stefanowitsch (2006) adopts a fairly bold approach in his method for studying metaphorical mappings in the British National Corpus. To explore mappings from the source domain of ANGER, for example, Stefanowitsch identified a representative lexical item associated with this domain, choosing the term with highest frequency from within the set of *anger*, *fury*, *rage*, *wrath* etc. The most frequently occurring term is, in fact, *anger*, so the form *anger* is the basis for further exploration of mappings from source domain ANGER. While this method may seem somewhat simplistic in its approach, it proves to be surprisingly revealing for the study of emotion metaphors. Another variant of this method is found in Oster (2010) who explores the concept of FEAR, including metaphoric and metonymic uses, in the very large Corpus of Contemporary American English (COCA, <http://corpus.byu.edu/coca/>). She determines, first, lists of collocates of various forms of *fear* and their contexts (maximum 400 hits for each set of results obtained form various search expressions). The identification of metaphorical usage proceeds by applying an adaptation of the morpheme identification procedure proposed by the Pragglejaz Group (2007), working initially with lists of collocates rather than linear text.

(c) *Use a limited number of concordance lines to inspect results*. Deignan (2005, p. 155-157) relies on a sample of 1,000 concordance lines with *cat(s)* as the search term to investigate metaphorical use of these forms. Ideally, such sample concordance lines would be obtained as a random set faithfully representing the range of genres/texts which are of relevance. Stefanowitsch, in the study referred to above, takes a random sample of 1,000 concordance lines to inspect the use of his key emotion terms.

One can arrive at many insights about metaphor from the application of these methods. Deignan (2005) uncovers quite a variety of results which make a very real contribution to the study of metaphor by relying on relatively simple methods like those above. Her discussion of animal metaphors is a good example of just what can be learned by applying corpus-based methods. Deignan (2005, p. 152-157) investigated metaphorical uses of nouns from the source domain ANIMALS (*pig*, *wolf*, *monkey*, *rat*, *horse* etc.). One finding was that simple equational kinds of expressions, like a much celebrated example in the literature (*Richard is a gorilla*), is exceedingly rare in usage. Instead, the animal noun is typically converted to a verb or adjective (*I was horsing around with Katie*; *the mousy little couple*; *she bitched about Dan* etc.). The conversion from noun to verb/adjective in these cases is presumably related to the fact that we employ animal metaphors to conceptualize human behaviors (prototypically expressed through the verb category in English) and, to a lesser extent, attributes (prototypically expressed through the adjective category in English). One would not be able to appreciate these patterns, at least not with any real supporting evidence, without the aid of a corpus. Boers (1999) found fluctuations in the relative frequency of the health metaphor depending on the month (averaged over the ten-year period), with the highest frequency occurring in the winter months. The explanation that Boers offers is that the winter months are the months when issues of physical health are a relatively salient part of human experience, i.e., the more that health is experienced as an everyday reality, the more likely it is that health will function as a source domain for metaphorical mappings. Boers' study can be seen as further evidence for the experiential grounding of language behavior.

## 3. Polysemy and synonymy

Cognitive linguistics has been especially interested in exploring semantically based word relations, especially polysemy. Elucidating the nature of the relationships between word-senses, as in polysemy, and the basis for such relationships easily leads to broader discussions about the contexts in which certain words or word-senses appear and the nature of extra-linguistic reality. In their Preface to an important collection of papers on polysemy, Ravin and Leacock articulate two key ideas which emerge from the research represented in their volume:

> [...] first, polysemy remains a vexing theoretical problem, leading many researchers to view it as a continuum of words exhibiting more or less polysemy, rather than a strict dichotomy. The second is the increasing realization that context plays a central role in causing polysemy, and therefore should be an integral part of trying to resolve it. Ravin and Leacock (2000, p. v-vi)

There has been some convergence of thinking about how corpora might be enlisted to help integrate context of usage into the analysis of relatedness of word senses (in the case of polysemy) and words (in the case of synonyms).[8] The FrameNet project is one example of this kind of approach in the way it seeks to characterize words and their uses. The FrameNet methodology

[8] WordNet is designed to capture directly facts about polysemy and synonymy and is potentially of great value when integrated with a corpus. See, for example, Davies (2007) who describes an ingenious method of integrating WordNet and the British National Corpus. Davies' proposal facilitates searches based on semantic relationships (e.g., synonyms, hyponyms, hypernyms) and so is potentially useful for tracking down metaphorical usage.

involves, among other things, "examining the kinds of supporting information found in sentences or phrases containing the word in terms of semantic role, phrase type and grammatical function), and building up an understanding of the word and its uses from the results of such inquiry" (FILLMORE; ATKINS, 2000, p. 101). Gries (2006), Gries and Divjak (2009), and Gries and Otani (2010) build upon the same recognition of the role of contextual factors, as evidenced in a corpus, to create their *behavioral profiles* of polysemy and near-synonymy (see below).

A corpus-based approach to analyzing polysemy and near-synonymy would proceed in similar ways, differentiated by the level of analysis: $sense_1$, $sense_2$, $sense_3$ etc. for polysemy, and $word_1$, $word_2$, $word_3$ etc. for near-synonymy. I will illustrate one such approach, guided by both of the key ideas articulated by Ravin and Leacock, as quoted above: the incorporation of a range of contextual factors and the identification of degrees of relatedness between near-synonyms. I will consider the nine slow movement verbs *stagger*, *hobble*, *limp*, *trudge*, *plod*, *amble*, *saunter*, *sidle*, *slink*, already studied by Dąbrowska (2009), but here subjected to a somewhat different analysis. As part of a larger study involving a variety of interesting methods, Dąbrowska had 63 native speakers of English offer definitions of these words and then use the verbs in sentences which illustrate their meanings. While this is not a conventional corpus, which we would normally understand to be a collection of naturally occurring stretches of discourse, the sentences collected in this way can be thought of as an "elicitation corpus" illustrating speakers preferred use of words in context. The sentences were coded for a number of factors: characteristics of the person doing the walking (HUMAN, DRUNK, INJURED etc.), the path (presence of various words/phrases such as *home*, *away*, *in the room*, *into the room*, *from the pub*), the setting (INDOORS, OUTDOORS, COUNTRY) and the manner (CRUTCHES etc.). Dąbrowska's method illustrates perfectly the way in which contextual factors (here, a combination of conceptual/semantic properties and lexical/phrasal forms) can be identified and systematically coded. Twenty sentences for each verb were chosen as the basis for the analysis in Dąbrowska (2009). A subset of the results is given in Table 3. The numbers in this table represent percentages of occurrence of a factor out of the total number of times the verb is used, e.g., 65% of occurrences of *stagger* in the corpus refer to a male walker.

TABLE 3
A subset of six contextual factors relevant to the use of nine verbs, adapted from Dąbrowska (2009: 211, Table 1) and the percentages of occurrence with each verb in a corpus.

| | INJURED | MALE | PLURAL | *in the room* | *from the pub* | *home* |
|---|---|---|---|---|---|---|
| *stagger* | 5 | 65 | 10 | 5 | 40 | 40 |
| *hobble* | 15 | 55 | 0 | 0 | 5 | 0 |
| *limp* | 40 | 60 | 0 | 0 | 0 | 10 |
| *trudge* | 0 | 40 | 45 | 0 | 0 | 20 |
| *plod* | 0 | 45 | 25 | 0 | 0 | 20 |
| *amble* | 0 | 20 | 70 | 0 | 0 | 0 |
| *saunter* | 0 | 60 | 5 | 25 | 0 | 0 |
| *sidle* | 0 | 75 | 5 | 0 | 0 | 0 |
| *slink* | 0 | 25 | 5 | 0 | 0 | 0 |

Dąbrowska judged there to be four clusters of verbs in this group, based on a combination of her own intuitions and an informal similarity judgement study (described in DĄBROWSKA, 2009, p. 210, fn. 6). These are shown in (1).

(1) a. *amble, saunter*
b. *plod, trudge*
c. *sidle, slink*
d. *hobble, limp, stagger*

Relying simply on a visual inspection of the numerical data in Table 3 can quickly leads to quandaries. One can see identical percentages for some sets of verbs: {*saunter*, *sidle*, *slink*}, for example, all show 5% PLURAL in their usage. But what about the pair {*sidle*, *stagger*} where each verb has around 70% MALE factor? Does *sidle* belong more in the {*saunter*, *sidle*, *slink*} group or more in the {*sidle*, *stagger*} group? As we move through more and more data (and remember that Table 3 is only a subset of the complete dataset), "eye-balling" the data to arrive at any satisfying conclusion about the whole dataset becomes impossible. One simply has to turn to statistical methods from the family of multifactorial analysis to make sense of this complexity. One such method is Correspondence Analysis (CA) (BENDIXEN, 2003; GREENACRE, 2007; GLYNN, in press). The overall objective of CA is to represent the maximum possible variance in a plot of few dimensions. The summary statistics given by a CA analysis in the ca package in R shows that two dimensions explain just 63.5% of the variance. Assuming three dimensions explains a full 86.8% of the variance and so we show the results in the three plots Figures 1-3 showing the interaction of dimensions 1x2, 1x3, and 2x3. "Asymmetric" here means that we can inspect how the verbs lie relative to one another and how the contextual factors are spread out relative to the verbs. The verbs are represented by dots and the factors by triangles. The larger the dot/triangle, the more the contextual factor contributes to the correspondence. In Figure 1, an 'inside' (*in the room*) vs. 'outside' (*from the pub*, *home*) orientation is evident, in Figure 2 an 'injured' vs. 'location' orientation, and in Figure 3 we see a three-way contrast between 'injured', 'inside', and 'outside' factors.

From the plots one can appreciate the closeness of some pairs of verbs by their proximity (though finer details are not so easy to see in the reduced size of the plots shown here), e.g., the pairs {*trudge, plod*}, {*slink*, *slide*}, and {*hobble, limp*} in Figure 1. Notice that these pairs correspond closely to some of the clusters that Dąbrowska had arrived at on the basis of intuition and the judgement task. One also sees some associations between the verbs and the contextual factors by their proximities. In Figure 1, for example, *amble* associates closely with PLURAL (70% PLURAL in Table 3); a number of verbs are close to MALE, most of all *sidle* (75% MALE in Table 3). One can also see that *saunter* is the closest to the "indoors"-orientated contextual factor *in the room*, while *stagger* is the closest to the 'outdoors'-oriented contextual factors *home* and *from the pub*, but these are relatively weaker associations (cf. Table 3, where less than 50% of the use of the verbs have these characteristics).

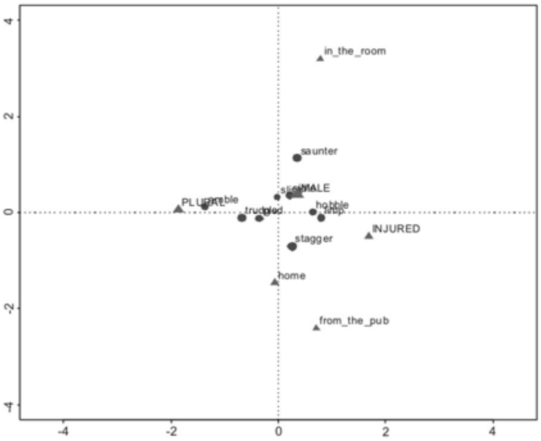


FIGURE 1 - Asymmetric plot of dimensions 1 and 2 from a CA analysis of data in Table 3

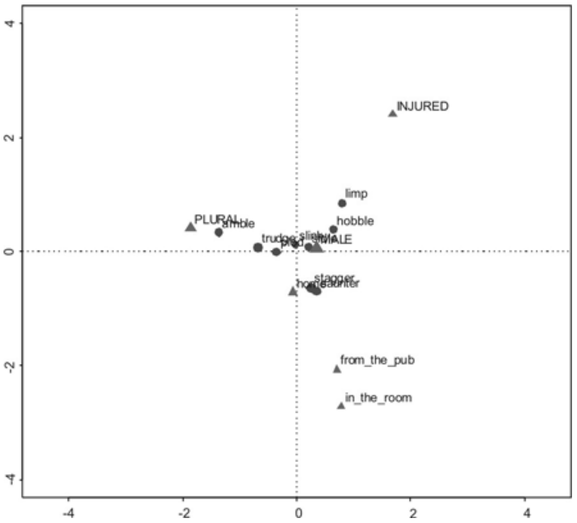


FIGURE 2 - Asymmetric plot of dimensions 1 and 3 from a CA analysis of data in Table 3

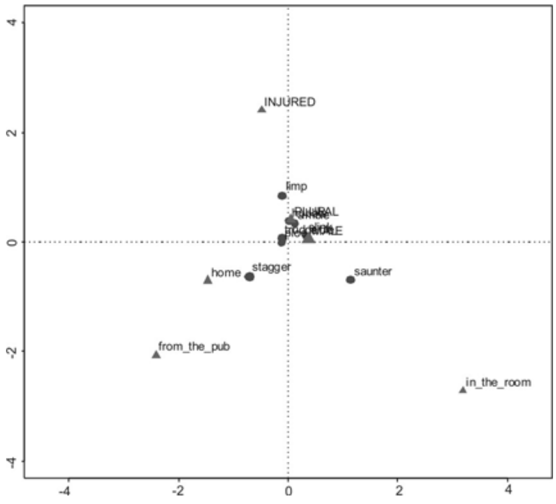


FIGURE 3 - Asymmetric plot of dimensions 2 and 3 from a CA analysis of data in Table 3

Another exploratory method for identifying groupings of verbs is clustering analysis. Clustering analysis subsumes quite a number of techniques, but the basic idea is that numerical data, like that in Table 3, is transformed into a representation by a "distance metric" (a number of options are available), and the transformed data is then the basis for clustering the rows into sub-groups (again, a number of clustering algorithms are available). Further discussion of these techniques can be found in Gries (2009b, p. 306-319) and Baayen (2008, p. 138-148). Continuing with our sample data in Table 3 for the sake of consistency, we choose from a class of clustering methods called "hierarchical agglomerative clustering" which combines or "agglomerates" the most similar cases (rows in Table 3) into groups and then those groups into larger groups to form a tree-like structure called a "dendogram". In the present case, we choose "Canberra" as the distance metric and a widely used clustering algorithm, "Ward" (ROMESBURG, 2004, p. 101-102, 129-135).[9] The result can be seen in Figure 4. Some groupings emerge as "closer" than others in this tree. The same pairs that we were led to using CA appear here, too, as clusters: {*trudge, plod*}, {*slink, sidle*}, and {*hobble, limp*}. The more vertical height there is between sisters, the less close they are. This means that the pairing of *stagger* with {*hobble*, *limp*} turns out to be a relatively close grouping (consistent with Dąbrowska's fourth grouping above). The dendogram itself does not directly say anything about which contextual factors are contributing most to the groupings visible in the hierarchy. However, by astutely manipulating the input to the clustering analysis (by eliminating one or more columns of data), one can identify certain columns of data as being more or less relevant to some clusterings.

In addition to constructing a dendogram, we should follow up with some tests for the reliablility of clustering produced in a dendogram. Figure 5 includes a probability measure added to each partition of the tree below the root, the "AU" value . The AU value is the abbreviation of "approximately unbiased" probability value (Shimodaira 2004; Suzuki and Shimodaira 2006). One can consider that clusters with high AU values are strongly supported by data. In this case the three groupings that we had identified as being relatively strong by the CA analysis do indeed appear with with high AU values: {*trudge, plod*} = 100% AU, {*slink, sidle*} = 98% AU, {*hobble, limp*} = 93% AU.[10] The dendogram also numbers the clusters on the "edges" in the

[9] The combination of the Canberra distance metric and the Ward clustering algorithm follows Gries' (2006) practice.

[10] The percentages can be expected to vary a little when this procedure is repeated, since the algorithm relies on "multiple bootstrapping", i.e., resampling the original data multiple times, so different sets of sampled data are used as the basis for the calculations. In multiple trials on this data, AU percentages did vary, but only in a small range of a few percentage points.

order in which they are formed (1, 2, 3, etc.). The order here is based on the order of combination of the most similar cases: the lower down on the y-axis where the agglomeration takes place, the sooner the combination takes place in building up the dendogram.

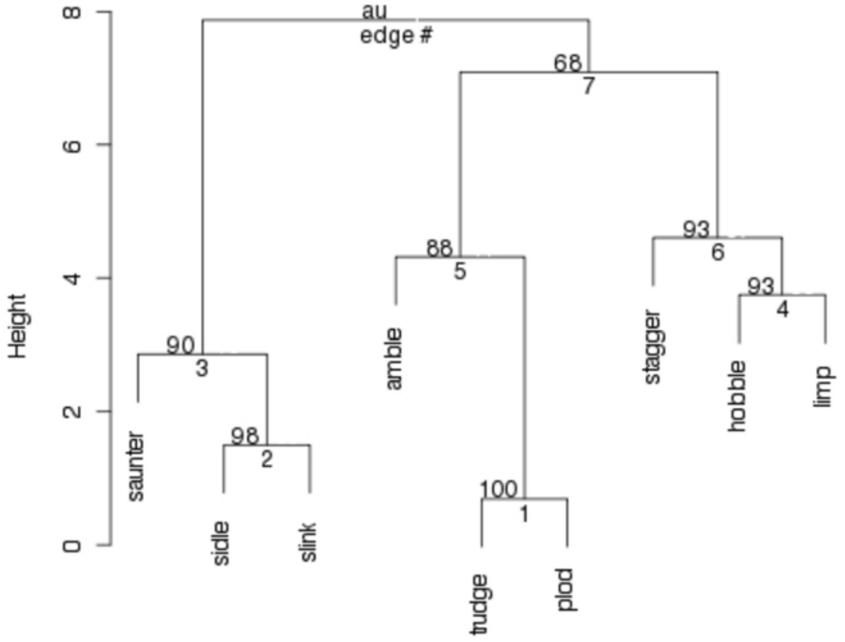


FIGURE 4 - Clustering analysis of data in Table 3

Sometimes, we may have data in a non-numeric format, as in Table 4, taken from Dowbor (ms.). The data in this table are taken from a much larger table in Dowbor (ms.), where Dowbor explored a clustering analysis of the multiple senses of the preposition *over*. Her table was initially constructed in a spreadsheet where she coded each instance of the use of *over* in concordance lines extracted from a corpus – a common way in which such data might be collected. Each sub-sense of over was coded as 'about', 'above', 'across', 'by means of' etc. and these represent the cases to be clustered. Dowbor constructed a number of variables concerning the nature of the verb used with any sub-sense (e.g., dynamicity) and properties of the trajector (TR) and landmark (LM) associated with the use of *over*. Gries and Otani (2010) describe how one might carry out a conversion to numeric-only format, making each feature used in the coding a variable in its own right. The conversion to a numeric table, along with a number of other attractive features (e.g., addition of probability values to the edges of the dendogram), are contained within the "behavioral profiles" R script (GRIES, 2009a).

TABLE 4
Partial data frame of *over* data, adapted from Dowbor (ms.)

| SENSE | Dynamicity | TR | TR_concrete | TR_animate | LM |
|---|---|---|---|---|---|
| about | stative | PSYCH | abstract | non-animate | STIMULUS |
| above | dynamic | THING | concrete | non-animate | PLACE |
| across | stative | PERSON | concrete | animate | PLACE |
| across | stative | THING | abstract | non-animate | PLACE |
| by_means_of | dynamic | COMM | concrete | non-animate | INST |
| during | dynamic | EVENT | abstract | non-animate | TIME |

Polysemy and synonymy can not be properly researched if we insist on only working with discrete yes-no categories. Both types of semantic relationships exhibit gradient properties which must be captured by methods which allow the researcher to appreciate the different degrees to which the usages of a word can be related or the different degrees to which words are similar in their usage. Corpus-based methods which incorporate into their analysis a range of contextual data retrievable from a corpus are particularly suited to revealing this gradience associated with polysemy and synonymy and, moreover, are consistent with the methodological observations made by Ravin and Leacock (2000), cited above. The methods illustrated in this Section presuppose a fair amount of coding of properties of the relevant factors – the data frames which underlie the various statistical analyses above – but the rewards in terms of visualization and conceptualization of the phenomena make this initial outlay of effort well worthwhile.

## 4. Prototypes

The idea of prototypes is pervasive in cognitive linguistics and, indeed, has been one of the hallmarks of the cognitive linguistic movement. Having one central member of a category, a prototype, is just one way that the members of a category may be organized. Other ways in which a category might be organized include the possibility of multiple "local" prototypes, each with a cluster of other members around it or a whole network of relationships which chain together members of a category. The idea of prototypes stands in contrast to the view that membership of a category is matter of strict and necessary conditions on all members, without differentiating degrees of membership. Some of the procedures introduced in other sections (including the following section) could be considered as methods to help identify central members of categories. Bybee (2010, p. 76-104) argues for high token frequency within a construction as playing a key role in the formation of central categories, with type frequency and semantics also contributing in important ways to how the central member behaves (e.g., higher type frequency of a word in a construction contributes to the productivity of the construction more than token frequency does).

Stubbs (2001, p. 84-96) outlines a method for systematically studying the central uses of a word and an associated "lexico-grammatical frame" (= construction). The method requires the analyst to work through collocates to identify patterns which are structurally and informationally salient. His method starts with the top 20 collocates in a span of 4 words to the left and 4 words to the right of the word of interest and then examining 20 random concordance lines for each of these 20 collocates. By systematically working through such data, it is possible to obtain a profile of major tendencies within a construction. Stubbs illustrates the method with the verb UNDERGO and successfully identifies key aspects of the construction shown in Table 5.

TABLE 5
Prototypical usage of UNDERGO, adapted from Stubbs (2001, p. 92, Table 4.1)

| passive/modal | | adjective | abstract noun |
|---|---|---|---|
| forced to | | further | medical procedure |
| required to | UNDERGO | extensive | testing |
| must etc. | | major etc. | change etc. |

When working with multivariate data as in the case of Dowbor's data in Table 4 above, it is clear that there are potentially many possible combinations of features. With such data, one would like to know if certain combinations of features stand out as more significant than others in a statistical sense. Conveniently, there is just such a technique, though it is a technique which does not often appear in the handbooks on statistics. This technique is called Hierarchical Configural Frequency Analysis (HCFA) and explanation and illustrations of the method can be found in von Eye (1990), Lautsch and Weber (1995), von Eye and Lautsch (2003), and Gries (2009b,

p. 248-252). HCFA is a procedure which computes the statistical significance of combinations of features that show up in the variables, i.e., the "levels" of the factors in the analysis. Many calculations can be involved when the procedure works through every possible combination and carries out its calculations. In the present case, I used the hcfa_3-2.R script (GRIES, 2004b). Running this interactive script on the TR and LM variables in all of Dowbor's *over* data, the script produces the results in Table 6 (a subset of the full results produced by the script). In this table, we see the combinations of levels of the factors TR and LM where the observed frequency of a configuration is greater than expected in a statistically significant manner, indicated by the three asterisks in the Decision ("Dec") column. Three other statistical results are shown: contribution to Chi-square, probability value of a Holm adjusted probability value (an adjustment applied in order to obtain a more appropriate measure of the contribution of one test when there are multiple tests producing an overall probability), and a measure of pronouncedness "Q" (an effect size, independent of how large or small the data are). The seven combinations of TR and LM values shown are just those seven combinations of these features which yielded results at a very significant level. The example sentences in (2), taken from the ICE-GB corpus which Dowbor used, illustrate these combinations of features. Note that the identification of what we may call "prototypical uses" of *over* in this manner does not rely upon any prior decision about exactly how many sub-senses the preposition *over* has. Table 3 does include coding into sub-senses such as 'above', 'across' etc. under the SENSE variable, but the results in Table 6 were obtained without any reliance upon this particular variable. This is an attractive way to proceed in light of the suspect nature of claims about the number and nature of polysemous sub-senses of a word (cf. the insightful remarks by TAYLOR, 2006, on the problem of polysemy in general and the polysemy of *over* in particular).

TABLE 6
HCFA results showing statistically significant "types" of TR and LM configurations with *over*

| TR | LM | Freq | Exp | Chisq | P.adj.Holm | Dec | Q |
|---|---|---|---|---|---|---|---|
| [AMOUNT] | AMOUNT | 60 | 9.55 | 266 | 3.059e-27 | *** | 0.137 |
| EVENT | TIME | 56 | 14.66 | 116 | 2.440e-15 | *** | 0.114 |
| THING | PLACE | 50 | 16.67 | 66 | 1.220e-09 | *** | 0.093 |
| PSYCH-STATE | STIMULUS | 35 | 3.25 | 310 | 1.490e-22 | *** | 0.085 |
| STATE | DEPENDENT | 12 | 0.93 | 131 | 5.398e-08 | *** | 0.029 |
| COMMUN | INSTRUMENT | 5 | 0.07 | 367 | 1.635e-06 | *** | 0.013 |
| ATTRIBUTE | STANDARD | 4 | 0.05 | 293 | 5.107e-05 | *** | 0.01 |

(2) a. …*over 700 farms* still cannot sell their meat for human consumption [AMOUNT] X AMOUNT

b. the blood pressure <unclear-words> at such a level after repeat measurements *over a considerable period of time* sometimes as long as six months EVENT X TIME

c. a minute on each side on high and then 5 minutes *over a low flame* will do it THING X PLACE

d. In view of the furore *over the transmission of news from the Falklands* PSYCH-STATE X STIMULUS

e. Abortion is the right of a woman *over her own body* STATE x DEPENDENT ENTITY

f. When digital data are transmitted *over a single parallel interface* there is no crosstalk between the codes COMMUNICATION X INSTRUMENT

g. It offered many advantages *over other systems* including rapid action ATTRIBUTE X STANDARD ITEM

Whether one relies on Stubbs' method of identifying the most common usage of a form or statistically more sophisticated methods like HCFA (and collostructional analysis, to be discussed in Section 5), there are very systematic corpus-based procedures which a cognitive linguist may use to identify prototypical usage of a word or pattern. These methods succeed in leading a researcher to choices for prototypes, including, for some kinds of phenomena, multiple prototypes. The methods have the virtue of being strongly grounded in facts of usage, complementing any other (intuition-based or experimentally based) methods the researcher might employ.

## 5. Constructions

One area of interest in cognitive linguistics relates to a new understanding we have of the relationship between words and the constructions in which they occur. "Construction" here may include the more traditional chunks of text which correspond to the traditional structuralist view of language consisting of constituents like NP, VP, PP etc. But, with a more open attitude to what is of interest in terms of surrounding context, it can be any one of a number of properties of surrounding context in sentences and utterances which might contribute to a corpus-based analysis of a word in context.

To illustrate some methodological possibilities for the analysis of words in constructions, I will consider the collostructional analysis approach pioneered by Stefanowitsch and Gries (STEFANOWITSCH; GRIES, 2003; GRIES, 2004a). I will illustrate the approach through the examples of two related constructions: EXPERIENCE N and EXPERIENCE *of* N, using Mark Davies' COCA corpus as the basis for the calculations.

In the collostructional analysis methodology, one assesses the statistical significance of the association between a construction and an associated word. Consider, for example, the use of *life* in EXPERIENCE *life* in a corpus (where the small caps indicate the lemma, i.e., all the inflected forms of the verb). There are two pairs of contrasting numbers to be considered in evaluating the significance of the frequency of EXPERIENCE *life*: frequencies of the noun *life* in the EXPERIENCE N construction and the total corpus frequency of the noun *life*, frequencies of the EXPERIENCE N construction and the total corpus frequency of all constructions.[11] Since EXPERIENCE N is an instance of a verb phrase, I take the number of (lexical) verbs in the corpus as an approximation of the total number of relevant constructions of the corpus.[12]

[11] There is certainly room for disagreement about what constitutes the "number of constructions" relevant to a problem (cf. SCHMID, 2010 for discussion of this issue).

[12] Bybee (2010, p. 97-101) argues for the prime importance of relative frequency of occurrence within a construction and semantics, rather than Stefanowitsch and Gries' collostructional strength measure. In particular, Bybee objects to the reliance on any assumption that words should be considered as occurring "by chance" in a corpus, an assumption underlying many statistical approaches, including the

Following the procedure described in Stefanowitsch and Gries (2003), we begin by noting the following frequencies in COCA:

(4) From the corpus, we directly obtain:[13]

- the number of EXPERIENCE *life* sequences = 80
- the number of EXPERIENCE + noun sequences = 4,169
- the number of tokens of the noun *life* in the corpus = 293,108
- the number of all lexical verbs in the corpus = 47,560,677

From these we obtain the crucial numbers that constitute the contingency table which is the basis for the statistical calculation. Table 7 presents the 2x2 contingency table (the shaded part in the table) which we will use for the calculation (cf. MANNING; SCHÜTZE, 1999, p. 169-172 for a discussion of the underlying procedure applied to the co-occurrence of words, as opposed to a construction and a word). The numbers in bold in Table 7 are the four numbers from (4), obtained directly from the corpus; other numbers are obtained by subtraction. We then carry out a test of statistical significance, such as the Fisher Exact test, on the numbers in the shaded area and obtain a probability value (6.035451e-18 = 6.035451 with the decimal point moved 18 places to the left). A convenient way to report this value, the "collostructional strength" is to use the negative log to base 10 of this number (=17.22). Intuitively, one can think of this result as follows: there is a total of 47,560,677 datapoints in the total sample; the EXPERIENCE N construction occurs with a probability of 4,169/47,560,677 based on the total number of datapoints; the noun *life* occurs with a probability of 293,108/47,560,677; the joint probability of both *life* and the EXPERIENCE N construction is the product of these two probabilities = 5.402111e-07 which equals 25.69 when applied to

collostructional analysis method. Psycholinguistic evidence can be adduced in support of each of their methods: Bybee cites Bybee and Eddington (2006) in support of her position; Gries, Hampe, and Schönefeld (2005) and Gries, Hampe, and Schönefeld (2010) present evidence for the role of collostructional strength.

[13] One could imagine slightly different ways to obtain the relevant frequencies in COCA. In the present case, I used the following search terms: "[experience].[vv*]" to search for the verb lemma EXPERIENCE; "[nn*]" in the R1 position of "[experience].[vv*]" to search for all tokens of the EXPERIENCE N construction; [vv*] to search for all lexical verb constructions in the corpus.

the total number of datapoints. That is, even if the distributions of *life* and the EXPERIENCE N were completely independent of each other, we would still expect 25.69 occurrences of EXPERIENCE *life*. In fact, *life* occurs 80 times in this construction and this number is more than expected to a statistically significant degree, as shown by the extraordinarily low probability value. So, *life* is "attracted" to the EXPERIENCE N construction.

TABLE 7

Table of frequencies relevant to occurrence of nouns in the EXPERIENCE N construction. Numbers in bold are obtained directly from the corpus. The shaded part is the 2x2 contingency table which is the basis for the statistical calculations

| | life | life nouns | Total |
|---|---|---|---|
| EXPERIENCE N | **80** | 4089 | **4169** |
| EXPERIENCE N | 293028 | 47263480 | 47556508 |
| Total | **293108** | 47267569 | **47560677** |

Fortunately, it is not necessary to perform this sequence of steps manually if we use the coll.analysis R script (GRIES, 2004a). This script will calculate collostructional strength for any number of words relevant to a construction. The results from the script for the words ("collexemes") under consideration are shown in Table 8. As can be seen in this table, the script returns the overall word frequency of a collexeme, its observed and expected frequency within the construction, the reliance measure (here called "faithfulness", abbreviated "faith"), the attraction or repulsion of the collexeme to the construction, and the collostructional strength (as computed by the Fisher Exact test in these tables). A collostructional strength >3 is significant at the p<0.001 level. Our particular calculations for the *life* collexeme above appear here at rank 8, showing the expected frequency of 25.69 and the collostructional strength of 17.2.

TABLE 8
Collostructional profile of nouns occurring in the EXPERIENCE N construction, based on all genres of COCA, ranked by collostructional strength

| RANK | WORDS | WORD. FREQ | OBS. FREQ | EXP. FREQ | FAITH | RELATION | COLL. STRENGTH |
|---|---|---|---|---|---|---|---|
| 1 | difficulty | 13211 | 48 | 1.16 | 0.0036 | attraction | 58.7 |
| 2 | success | 49194 | 52 | 4.31 | 0.0011 | attraction | 36.9 |
| 3 | difficulties | 10123 | 28 | 0.89 | 0.0028 | attraction | 31.4 |
| 4 | depression | 17654 | 32 | 1.55 | 0.0018 | attraction | 30.1 |
| 5 | symptoms | 14539 | 30 | 1.27 | 0.0021 | attraction | 29.9 |
| 6 | feelings | 21766 | 33 | 1.91 | 0.0015 | attraction | 28.5 |
| 7 | pain | 40596 | 40 | 3.56 | 0.0010 | attraction | 27.4 |
| 8 | stress | 24492 | 31 | 2.15 | 0.0013 | attraction | 24.6 |
| 9 | anxiety | 13777 | 24 | 1.21 | 0.0017 | attraction | 22.4 |
| 10 | life | 293108 | 80 | 25.69 | 0.0003 | attraction | 17.2 |
| 11 | problems | 100663 | 43 | 8.82 | 0.0004 | attraction | 15.9 |
| 12 | wash-out | 100 | 6 | 0.01 | 0.0600 | attraction | 15.3 |
| 13 | pleasure | 17729 | 20 | 1.55 | 0.0011 | attraction | 15.2 |
| 14 | discrimination | 10200 | 16 | 0.89 | 0.0016 | attraction | 14.5 |
| 15 | boredom | 2124 | 10 | 0.19 | 0.0047 | attraction | 13.9 |
| 16 | discomfort | 3146 | 11 | 0.28 | 0.0035 | attraction | 13.9 |
| 17 | nausea | 1865 | 9 | 0.16 | 0.0048 | attraction | 12.7 |
| 18 | frustration | 7883 | 13 | 0.69 | 0.0016 | attraction | 12.2 |
| 19 | orgasm | 1380 | 8 | 0.12 | 0.0058 | attraction | 12.0 |
| 20 | burnout | 1513 | 8 | 0.13 | 0.0053 | attraction | 11.7 |

A number of observations may be made based on the results in Table 8. Notice, for a start, that it is not the case that the ordering mirrors relative frequency of occurrence in the pattern. For example, *life* is in the eighth position even though *life* has the highest frequency of all nouns in table. *Life* occurs relatively often in the whole corpus and so, all else being equal, we would expect more occurrences of this noun in the construction under investigation. Instead of *life*, the collexeme showing the strongest collostructional strength is *difficulty*. It can be easily seen that the great majority of the top 20 collexemes in Table 10 are nouns which in fact share the same kind of negative nuance that *difficulty* has: *depression*, *pain*, *stress*, *anxiety* etc. And the noun which rises to the first position based on collostructional strength reflects this dominant semantic characteristic. The negative "prosody" evident in Table 10 is not something one could confidently predict from mere reflection on the word. Nor is it a result that is so evident from simply inspecting the most frequent nouns occurring in the EXPERIENCE N construction. The top 20 most frequently nouns in this construction, together with their frequencies, are: *life* 80 *success* 52 *difficulty* 48 *problems* 43, *pain* 40, *music* 34, *feelings* 33, *depression* 32, *things* 32, *stress* 31, *symptoms* 30, *difficulties* 28, *anxiety* 24, *pleasure* 20, *nature* 20, *art* 19, *discrimination* 16, *joy* 15, *frustration* 13. Within this list, almost half the items (*life*, *success*, *music*, *feelings*, *things*, *pleasure*, *nature*, *art*, *joy*) do not show any of the negative prosody so evident in Table 8.

It is also interesting to compare the collostructional profile of the EXPERIENCE N construction with what might appear to be a very similar construction: the EXPERIENCE *of* N construction. Table 9 shows results for a collostructional analysis of this construction, now taking the number of lexical nouns in the corpus as the size of the corpus. In this case, one does not find the same strong tendency towards negative nuances as with the collexemes in Table 8. Instead, the collexemes in the EXPERIENCE *of* N construction include a mixture of negatively nuanced concepts and more abstract, philosophical concepts, e.g., *reality*, *oneness*, *modernity*, *transcendence*, *otherness*. The collostructional profiles in Tables 8 and 9 provide a very convenient way of demonstrating the different types of nouns attracted to what would appear to be, on the surface, similar constructions and lend support to treating the two constructions as objects of study in their own right.

TABLE 9
Collostructional profile of nouns occurring in the EXPERIENCE *of* N construction, based on all genres of COCA, ranked by collostructional strength

| RANK | WORDS | WORD. FREQ | OBS. FREQ | EXP. FREQ | FAITH | RELATION | COLL. STRENGTH |
|---|---|---|---|---|---|---|---|
| 1 | reality | 35688 | 33 | 1.24 | 0.0009 | attraction | 34.4 |
| 2 | oneness | 355 | 11 | 0.01 | 0.0310 | attraction | 28.7 |
| 3 | modernity | 2502 | 15 | 0.09 | 0.0060 | attraction | 28.1 |
| 4 | life | 293108 | 62 | 10.20 | 0.0002 | attraction | 27.5 |
| 5 | depression | 17654 | 17 | 0.61 | 0.0010 | attraction | 18.4 |
| 6 | combat | 12963 | 15 | 0.45 | 0.0012 | attraction | 17.5 |
| 7 | jealousy | 2142 | 9 | 0.07 | 0.0042 | attraction | 15.7 |
| 8 | slavery | 5994 | 11 | 0.21 | 0.0018 | attraction | 15.2 |
| 9 | trauma | 6167 | 11 | 0.21 | 0.0018 | attraction | 15.0 |
| 10 | pain | 40596 | 18 | 1.41 | 0.0004 | attraction | 13.7 |
| 11 | stress | 24492 | 15 | 0.85 | 0.0006 | attraction | 13.5 |
| 12 | giftedness | 1401 | 7 | 0.05 | 0.0050 | attraction | 12.9 |
| 13 | oppression | 3066 | 8 | 0.11 | 0.0026 | attraction | 12.4 |
| 14 | suffering | 16170 | 12 | 0.56 | 0.0007 | attraction | 11.9 |
| 15 | disability | 5911 | 9 | 0.21 | 0.0015 | attraction | 11.8 |
| 16 | transcendence | 967 | 6 | 0.03 | 0.0062 | attraction | 11.7 |
| 17 | reading | 53958 | 18 | 1.88 | 0.0003 | attraction | 11.7 |
| 18 | childbirth | 1290 | 6 | 0.04 | 0.0047 | attraction | 11.0 |
| 19 | otherness | 503 | 5 | 0.02 | 0.0099 | attraction | 10.9 |
| 20 | colonialism | 1524 | 6 | 0.05 | 0.0039 | attraction | 10.5 |

A key point about the interpretation of these tables is that it is the relative order of the nouns which is crucial, more than the precise numerical value. One could make different choices, after all, about how the number of constructions is arrived at which would affect the p-value in the Fisher Exact test. Different choices for the number of constructions, however, would not affect the relative ordering of the degrees of attraction. Indeed, Schmid (2010: 113) opted for a "completely arbitrary number" of 10 million in one such exercise, while Bybee and Eddington (2006) used the total number of words in their corpus (2 million). Note, too, that one could choose from a number of alternative statistical tests on the 2x2 contingency table, the shaded cells in Table 9. Different tests of significance will yield different numbers as collocational strengths.

Two measures used by Schmid (2010) for describing the relationship between words and constructions are worth mentioning: *attraction* and *reliance*.[14] These measures are calculated as in (5).

(5) a. Attraction = $\dfrac{\text{frequency of a word in a pattern x 100}}{\text{total frequency of the pattern}}$

b. Reliance = $\dfrac{\text{frequency of a word in a pattern x 100}}{\text{total frequency of the word in the corpus}}$

Attraction measures the extent to which a particular pattern attracts a word. With respect to the example in Table 7, this would be the equivalent of calculating the frequency of the noun *life* as a percentage of the total (row) frequency of the EXPERIENCE N construction, i.e. (80/4,169)*100 = 1.92%. Reliance measures the extent to which a word appears in one particular pattern versus other patterns. This is the equivalent, in Table 7, of calculating the frequency of the noun *life* as a percentage of the total (column) frequency of *life* in the corpus, i.e. (80/293,108)*100 = 0.03%.

In giving attention above to constructions involving familiar structural units like verb + noun and noun *of* noun, I do not mean to imply that only such units are worthy of interest. On the contrary, many sequences of words which we encounter as n-grams may not have any structure familiar in contemporary linguistic tradition, but are worthy of further study. The very idea that we might learn something of importance from the study of mere sequences of words without giving more prominence to the associated syntactic structure (noun phrases, verb phrases etc.) must seem like an affront to many linguists. It appears to ignore a long tradition within linguistics of

[14] The same measures have also been discussed in Gries, Hampe, and Schönefeld (2005, p. 645-647). Note also that the "faith" score returned in the collostructional analyses shown in Tables 7 and 8 is based on the same proportional calculation as Schmid's reliance measure. The reliance scores are similar to what Janda and Solovyev (2009) incorporate into their constructional profiles.

assigning a hierarchical constituent structure to groupings of words and, in the Chomskyan tradition, seeing rules of language as "structure-dependent". Atkinson, Kilby, and Rocca (1982, p. 149) in a defense of this tradition say: "[…] no serious approach to linguistic analysis looks on sentences merely as sequences of words". I would be inclined to say, rather, that no serious cognitive linguist can afford to ignore the role that sequences of words play, irrespective of what structure one might wish to superimpose on them. Jurafsky, Bell, Gregory, and Raymond (2000), for example, studied different measures of probabilities of occurrence of function words such as *a*, *the*, *in*, *of* etc. and investigated how these measures correlated with phonetic effects such as shortening of the vowel of the function word. They found that a higher conditional probability of the function word given the preceding word predicted vowel shortening in the function word, even in bigrams which are not usually thought of as any kind of structural units such as *them and*, *sometime in*, *where the*, and *fine and.* The study of n-grams along the lines of Jurafsky et al. should interest cognitive linguists, as much as the study of more conventional constructional types.

## 6. A note on corpora

In the preceding sections, I have sketched out some corpus-based methods which can be profitably utilized by cognitive linguists choosing to work with corpora. These methods assume what must now be considered rather traditional kinds of corpora, by which I mean collections of samples of written usage and transcribed spoken usage from the major languages of the world, relying on an orthographic representation of language. The availability of such corpora and their popularity among usage-based language researchers should not distract researchers from the task of developing and analyzing corpora from other domains of usage. Minority languages and most indigenous languages remain understudied linguistically and underrepresented in available corpora. Even within the major languages of the world, regional and social varieties warrant more attention than they have received in corpus-linguistic circles.[15] Corpus collections of varieties of English such as the

[15] Dirk Geeraerts' research on "cognitive sociolinguistics", as in Geeraerts (1994), incorporates the study of sociolinguistic characteristics of the speakers as part of a larger corpus-based approach, and represents a welcome extension of the usual scope of both corpus linguistics and cognitive linguistics.

International Corpus of English (ICE) go some way towards filling this gap for English, but it is relatively rare for corpus-based studies of English to utilize these (free and downloadable!) corpora.[16] It is more common for researchers of English to seek out corpora which are increasingly large in terms of numbers of words, rather than smaller, specialized ones which are designed to reflect specific kinds of language use.

There is, above all, a pressing need to study the more interactive aspects of spoken language, aspects that can only be investigated with fully multimodal corpora which include and integrate video and audio dimensions. One sometimes encounters in the cognitive linguistic literature references to "situated" language as being a desirable focus, contrasting with de-contextualized snippets of language. Consider, for example, the position articulated in Evans and Green (2006, p. 478) who view "situated instances of languages use" as the basic, raw experience from which speakers build up a mental grammar. From this point of view, the study of "situated instances of language use" is a fundamental aspect of the language experience of speakers, not some peripheral, incidental phenomenon. I endorse this view, but I also believe that situated instances of language use must include a great many more aspects of language use than linguists, including linguists from the Conversation Analysis tradition, are accustomed to thinking about. Certainly, we must go beyond the traditional kinds of data represented in the familiar corpora designed along the lines of BNC, COCA etc. Instead, we must look to data in which hand gestures, head movement, gaze, motion, speed of body movements, facial expressions, and bodily stance are all integrated into the data being investigated (cf. WICHMANN, 2007, p. 82-83, in which the author calls for data from all channels of communication to be included in our corpora). I see Charles Goodwin in publications such as Goodwin (1979; 1980; 1981) as an early pioneer of the approach I am advocating. Another forerunner of such an approach would be Harris (1996; 1998) who has argued forcefully for an agenda for the study of language which situates language well and truly in the context of communication, what Harris calls an "integrationist approach". Thorne and Lantolf (2006) call for a "Linguistics of Communicative Activity", the goal of which is to "disinvent language understood as an object and to reinvent language as *activity*..." (THORNE; LANTOLF, 2006, p. 71, italics original). I believe most linguists, including cognitive linguists and

[16] The homepage of ICE is <http://ice-corpora.net/ice/index.htm>.

corpus linguists, are much more comfortable dealing with language as an object rather than as a process and the "disinvention" that Thorne and Lantolf call for is difficult, even troubling, for many linguists, though it is a change which cognitive linguists should welcome and embrace.[17]

My remarks in this section may be taken as an extended footnote to the whole article. I am most concerned, in this overview of corpora and cognitive linguistics, with assembling analytical methods that cognitive linguists may appeal to in working with corpora as we know them now. However, prevailing ideas about linguistic research and the scope of linguistics necessarily influence and constrain the way such corpora have been designed. Current corpus-based methods may help lead us as researchers to fresh insights language usage, but as long as the corpora themselves reflect only a part of our language behavior in the real world, our methods will still not reveal all that is relevant in language activity.

## 7. Summary

If cognitive linguistics is to fully develop as a field of linguistics grounded in actual usage of language, then corpora are not just one more type of data to be considered along with other modes of inquiry such as intuition, experimentally based methods etc. More than any other type of linguistic data, corpora represent usage and are therefore, arguably, the most essential kind of data that a usage-based cognitive linguistics should rely on. As mentioned above, not all linguists who identify themselves as cognitive linguists fully subscribe to the view that language usage is a central component of the whole cognitive linguistic enterprise. For those who do see language usage as central, however, corpora must continue to play a critical role in the development of the field.

Along with a focus on corpus data comes a need for new methods to process the data. It is a reflection of the Information Age in which we live that many corpora which are now becoming available are far, far larger than we can easily cope with simply by casting our eyes over concordance lines or columns

[17] MacWhinney's TalkBank and CHILDES projects contain many examples of corpora integrating audio and video. Large-scale examples of such corpora would be the D64 Multimodal Conversational Corpus (OERTEL; CUMMINS; CAMPBELL; EDLUND; WAGNER, 2010) and Deb Roy's Human Speech Genome project (ROY, 2009). The corpora coming out of these two projects capture video and audio in everyday settings in an extremely intensive manner and pave the way for quite exciting new discoveries about situated language use.

of collocates. The magnitude of the data in many corpora is such that linguists must inevitably turn to methods of analysis which will involve some degree of automatic retrieval and analysis. Linguists have no choice but to appeal to techniques of quantitative analysis which have been more familiar and more accepted in some other areas of social science than in linguistics. For this reason, I have focused my attention on methods in the sections above. At the same time, I do not mean to imply that issues about data, as opposed to methods, are relatively unimportant. On the contrary, the collection of multimodal data, incorporating audio and video, and the development of standards for annotating and accessing such data should be high on the agenda for cognitive linguists. Indeed, rethinking language as an activity realized through "situated instances of language use", studied through multimodal corpora, is potentially of far greater consequence to the field than the development of methods for the analysis of corpus data representing "unsituated instances of language use".

## References

ATKINSON, M.; KILBY, D.; ROCA, I. *Foundations of general linguistics*. London: George Allen and Unwin, 1982.

BAAYEN, R. H. *Analyzing linguistic data*: a practical introduction to statistics using R. Cambridge: Cambridge University Press, 2008.

BENDIXEN, M. A practical guide to the use of Correspondence Analysis in marketing research. *Marketing Bulleting* 14, Technical Note 2, 2003. Available at: <http://marketing-bulletin.massey.ac.nz/V14/MB_V14_T2_Bendixen.pdf>. Retrieved: April 7, 2011.

BOERS, F. When a bodily source domain becomes prominent: the joy of counting metaphors in the socio-economic domain. In: GIBBS, R. W. JR.; STEEN, G. J. (Ed.). *Metaphor in cognitive linguistics*. Amsterdam / Philadelphia: John Benjamins, 1999.

BYBEE, J. *Language, usage and cognition*. Cambridge: Cambridge University Press, 2010.

BYBEE, J.; EDDINGTON, D. A usage-based approach to Spanish verbs of becoming. *Language*, v. 82, n. 2, p. 323-355, 2006.

CAMERON, L. Identifying and describing metaphor in spoken discourse data. In: CAMERON, L.; LOW, G. (Ed.). *Researching and applying metaphor*. Cambridge: Cambridge University Press, 1999.

CAMERON, L.; DEIGNAN, A. Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor and Symbol*, v. 18, n. 3, p. 149-160, 2003.

CHARTERIS-BLACK, J. *Corpus approaches to critical metaphor analysis*. New York: Palgrave Macmillan, 2004.

DĄBROWSKA, E. Words as constructions. In: EVANS, V.; POURCEL, S. (Ed.). *New directions in cognitive linguistics*. Amsterdam and Philadelphia: John Benjamins, 2009.

DAVIES, M. Semantically-based queries with a joint BNC/WordNet database. In: FACCHINETTI, R. (Ed.). *Corpus linguistics 25 years on*. Amsterdam and New York: Rodopi, 2007.

DEIGNAN, A. *Metaphor and corpus linguistics*. Amsterdam and Philadelphia: John Benjamins, 2005.

DOWBOR, D. (ms.). *The polysemy of OVER*: a BP and HCFA investigation. University of Alberta.

EVANS, V.; GREEN, M. *Cognitive linguistics*: an introduction. Edinburgh: Edinburgh University Press, 2006.

FASS, D. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, v. 17, n. 1, p. 49-90, 1991.

FELLBAUM, C. (Ed.). *WordNet*: an electronic lexical database. Cambridge, MA: MIT Press, 1998.

FILLMORE, C. J.; ATKINS, B.T.S. Describing polysemy: the case of *crawl*. In: RAVIN, Y.; LEACOCK, C. (Ed.). *Polysemy*: linguistic and computational approaches. Oxford: Oxford University Press, 2000.

GEERAERTS, D. Methodology in cognitive linguistics. In: KRISTIANSEN, G.; ACHARD, M.; DIRVEN, R.; Ruiz de MENDOZA IBÁÑEZ; F. J. R. (Ed.). *Cognitive linguistics*: current applications and future perspectives. Berlin and New York: Mouton de Gruyter, 2006.

GEERAERTS, D.; GRONDELAERS, S.; BAKEMA P. *The structure of lexical variation*: meaning, naming, and context. Berlin and New York: Mouton de Gruyter, 1994.

GLYNN, D. Multiple Correspondence Analysis: exploring correlations in multifactorial data. In: GLYNN, D; ROBINSON, J. (Ed.). *Polysemy and synonymy*: corpus methods and applications in cognitive linguistics. Amsterdam and Philadelphia: John Benjamins. (In press).

GOODWIN, C. The interactive construction of a sentence in natural conversation. In: PSATHAS, G. (Ed.). *Everyday language*: studies in ethnomethodology. New York: Irvington, 1979.

GOODWIN, C. Restarts, pauses, and the achievement of mutual gaze at turn-beginning. *Sociological Inquiry*, v. 50, n. 3-4, p. 272-302, 1980.

GOODWIN, C. *Conversational organization*: interaction between speakers and hearers. New York: Academic Press, 1981.

GREENACRE, M. *Correspondence Analysis in practice*. 2. ed. Boca Raton: Chapman and Hall/CRC, 2007.

GRIES, St. Th. *Coll.analysis 3. A program for R for Windows 2.x*, 2004a. url: <http://www.linguistics.ucsb.edu/faculty/stgries/>.

GRIES, St. Th. *HCFA 3.2. A program for R*, 2004b. url: <http://www.linguistics.ucsb.edu/faculty/stgries/>.

GRIES, St. Th. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In: GRIES, S. Th.; STEFANOWITSCH, A. (Ed.). *Corpora in cognitive linguistics*: corpus-based approaches to syntax and lexis. Berlin and New York: Mouton de Gruyter, 2006.

GRIES, St. Th. *BehavioralProfiles 1.01*. A program for R 2.7.1 and higher, 2009a. url: <http://www.linguistics.ucsb.edu/faculty/stgries/>.

GRIES, St. Th. *Statistics for linguistics with R*: a practical introduction. Berlin and New York: Mouton de Gruyter, 2009b.

GRIES, St. Th.; DIVJAK, D. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In: EVANS, V.; POURCEL, S. (Ed.). *New directions in cognitive linguistics*. Amsterdam and Philadelphia: John Benjamins, 2009.

GRIES, St. Th.; HAMPE, B.; SCHÖNEFELD D. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, v. 16, n. 4, p. 635-676, 2005.

GRIES, St. Th.; HAMPE, B.; SCHÖNEFELD D. Converging evidence II: more on the association of verbs and constructions. In: RICE, S.; NEWMAN, J. (Eds.), *Empirical and experimental methods in cognitive/functional research*. Stanford, CA: CSLI, 2010.

GRIES, St. Th.; OTANI, N. Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal*, v. 34, p. 121-150, 2010.

GRIES, St. Th.; STEFANOWITSCH, A. (Eds.), *Corpora in cognitive linguistics*: corpus-based approaches to syntax and lexis. Berlin and New York: Mouton de Gruyter, 2006.

HARDIE, A; KOLLER, V.; RAYSON, P.; SEMINO, E. In: DAVIES, M.; RAYSON, P.; HUNSTON, S.; DANIELSSON, P. (Ed.). Corpus Linguistics Conference, CL2007, *Proceedings*... University of Birmingham, UK, 27-30 July 2007. Available at: <http://ucrel.lancs.ac.uk/publications/CL2007/paper/49_Paper.pdf>. Retrieved: April 7, 2011.

HARRIS, R. *Language and communication*: integrational and segregational approaches. London: Routledge, 1996.

HARRIS, R. *Introduction to integrational linguistics*. Oxford: Elsevier Science, 1998.

HILPERT, M. The German mit-predicative construction. *Constructions and Frames*, v. 1, n. 1, p. 29-55, 2009.

JANDA, L. A.; SOLOVYEV, V. D. What constructional profiles reveal about synonymy: a case study of Russian words for SADNESS and HAPPINESS. *Cognitive Linguistics*, v. 20, n. 2, p. 367-393, 2009.

JURAFSKY, D.; BELL, A.; GREGORY, M.; RAYMOND, W. D. Probabilistic relations between words: evidence from reduction in lexical production. In: BYBEE, J; HOPPER, P. (Ed.). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, 2001.

LANDES, S.; LEACOCK, C.; TENGI, R. Building semantic concordances. In: FELLBAUM , C. (Ed.). *WordNet*: an electronic lexical database. Cambridge, MA: MIT Press, 1998.

LAUTSCH, E.; von WEBER, S. *Methoden und Anwendungen der Konfigurationsfrequenzanalyse (KFA)*. Weinheim: Psychologie-Verlags-Union, 1995.

LEWANDOWSKA-TOMASZCZYK, B.; DZIWIREK, K. (Ed.). *Studies in cognitive corpus linguistics*. Frankfurt am Main: Peter Lang, 2009.

MANNING, C. D.; SCHÜTZE., H. *Foundations of statistical natural language processing*. Cambridge, Mass. and London, England: The MIT Press, 1999.

OERTEL, C.; CUMMINS, F.; CAMPBELL, N.; EDLUND, J.; WAGNER, P. D64: A corpus of richly recorded conversational interaction. In: KIPP, M.; MARTIN, J-C.; PAGGIO, P.; HEYLEN, D. (Ed.). LREC 2010 Workshop on multimodal corpora: advances in capturing, coding and analyzing multimodality, *Proceedings*... Valetta, Malta, 2010. p. 27-30. Available at: <http://www.speech.kth.se/prod/publications/files/3433.pdf>. Retrieved: April 7, 2011.

OSTER, U. Using corpus methodology for semantic and pragmatic analysis: what can corpora tell us about the linguistic expression of emotions? *Cognitive Linguistics*, v. 21, n. 4, p. 727-763, 2010.

PETERS, W,; WILKS., Y. Data-driven detection of figurative language use in electronic language resources. *Metaphor and Symbol*, v. 18, n. 3, p. 161-173, 2003.

PHILIP, G. Locating metaphor candidates in specialised corpora using raw frequency and key-word lists. In: MACARTHUR, F.; ONCINS-MARTÍNEZ, J. L.; SÁNCHEZ-GARCÍA, M.; PIQUER-PÍRIZ, A. M. (Ed.). *Metaphor in use*: context, culture, and communication. Amsterdam: John Benjamins. (In press).

PHILIP, G. Metaphorical keyness in specialised corpora. In: BONDI, M.; SCOTT, M. (Ed.). *Keyness in text*. Amsterdam: John Benjamins, 2010.

PRAGGLEJAZ GROUP. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, v. 22, n. 1, p. 1-39, 2007.

RAVIN, Y.; LEACOK, C. (Ed.). *Polysemy*: theoretical and computational approaches. Oxford: Oxford University Press.

RAYSON, P. *Matrix*: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University, 2003.

RAYSON, P. *Wmatrix*: A web-based corpus processing environment. Computing Department, Lancaster University, 2007. Available at: <http://www.comp.lancs.ac.uk/ucrel/wmatrix/>. Retrieved: April 7, 2011.

RAYSON, P.; ARCHER, D.; PIAO, S. L.; MCENERY, T. The UCREL semantic analysis system. In: Workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), *Proceedings*... Lisbon, Portugal, 2004.

ROMESBURG, H. C. *Cluster analysis for researchers*. North Carolina: Lulu Press, 2004.

ROY, D. *New horizons in the study of child language acquisition*. In: Interspeech 2009, *Proceedings*... Brighton, England. 2009. Available at: <http://www.media.mit. edu/cogmac/publications/Roy_interspeech_keynote.pdf>. Retrieved: April 7, 2011.

SCHMID, H.-J. Does frequency in text instantiate entrenchment in the cognitive system? In: GLYNN D.; FISCHER, K. (Ed.). *Quantitative methods in cognitive semantics*. Berlin and New York: Mouton de Gruyter, 2010.

SHIMODAIRA, H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*, v. 32, p. 2616-2641, 2004.

STAMOU, S.; ANDRIKOPOULOS, V.; CHRISTODOULAKIS, D. Towards developing a semantically annotated treebank corpus for Greek. In: NIVRE, J.; HINRICHS, E. (Ed.) Second Workshop on Treebanks and Linguistic Theories, *Proceedings*... Växjö: Växjö University Press, 2003.

STEEN, G. J. *Finding metaphor in grammar and usage*. Amsterdam and Philadelphia: John Benjamins, 2007.

STEEN, G. J.; DORST, A. G.; HERRMANN, J. B.; KAAL, A. A. *A method for linguistic metaphor identification*: from MIP to MIPVU. Amsterdam / Philadelphia: John Benjamins, 2010.

STEFANOWITSCH, A. Corpus-based approaches to metaphor and metonymy. In: STEFANOWITSCH, A.; GRIES, St. Th. (Ed.). *Corpus-based approaches to metaphor and metonymy*. Berlin / New York: Mouton de Gruyter, 2006.

STEFANOWITSCH, A.; GRIES, St. Th. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, v. 8, n. 2, p. 209-243, 2003.

STEFANOWITSCH, A.; GRIES, St. Th. (Ed.). *Corpus-based approaches to metaphor and metonymy*. Berlin and New York: Mouton de Gruyter, 2006.

STUBBS, M. *Words and phrases*: corpus studies of lexical semantics. Oxford: Blackwell, 2001.

SUZUKI, R.; SHIMODAIRA, H. pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, v. 22, n. 12, p. 1540-1542, 2006.

TAYLOR, J. Polysemy and the lexicon. In: KRISTIANSEN, G.; ACHARD, M.; DIRVEN, R.; de MENDOZA IBÁÑEZ, F. J. R. (Ed.). *Cognitive linguistics*: current applications and future perspectives. Berlin and New York: Mouton de Gruyter, 2006.

*THE R PROJECT for Statistical Computing*. <http://www.r-project.org/>.

THORNE, S. L.; LANTOLF, J. P. A linguistics of communicative activity. In: MAKON, I., S.; PENNYCOOK, A. (Ed.). *Disinventing and reconstituting languages*. Clevedon: Multilingual Matters, 2006.

VALENZUELA, J. A psycholinguists view on cognitive linguistics: an interview with Ray W. Gibbs. *Annual Review of Cognitive Linguistics*, v. 7, p. 301-317, 2009.

von EYE, A. *Introduction to configural frequency analysis*: the search for types and anti-types in cross-classification. Cambridge: Cambridge University Press, 1990.

von EYE, A.; LAUTSCH, E. Charting the future of configural frequency analysis: the development of a statistical method. [Introduction to a special issue devoted to configural frequency analysis.] *Psychology Science*, v. 45, n. 2, p. 217-222, 2003.

VOSSEN, P. Introduction to EuroWordNet. In: IDE, N.; GREENSTEIN, D.; VOSSEN, P. (Ed.). *Computers and the Humanities*, v. 32, n. 2-3, p. 73-89, 1998. Special issue on EuroWordNet.

WICHMANN, A. Corpora and spoken discourse. In: FACCHINETTI, R. (Ed.), *Corpus linguistics 25 years on*. Amsterdam and New York: Rodopi, 2007.

## Appendix. R scripts used for statistical calculations and plots

```
#Data for reanalysis of Dąbrowska's data in Table 3:
> data # a dataframe constructed in a spreadsheet and imported into R
          INJURED  MALE  PLURAL  in_the_room  from_the_pub  home
stagger         5    65      10            5            40    40
hobble         15    55       0            0             5     0
limp           40    60       0            0             0    10
trudge          0    40      45            0             0    20
plod            0    45      25            0             0    20
amble           0    20      70            0             0     0
saunter         0    60       5           25             0     0
sidle           0    75       5            0             0     0
slink           0    25       5            0             0     0
```

```
# For Correspondence Analysis of Dąbrowska's data in Figures 1-3:
> library(ca)
> plot(ca(data))
> summary(plot(ca(data))
> plot(ca(x),dim=c(1,2), map="rowprincipal", mass=c(TRUE, TRUE),
xlim=c(-3.5, 3.5), ylim=c(-4, 4))¶
> plot(ca(x),dim=c(1,3), map="rowprincipal", mass=c(TRUE, TRUE),
xlim=c(-3.5, 3.5), ylim=c(-4, 4))¶?
> plot(ca(x),dim=c(2,3), map="rowprincipal", mass=c(TRUE, TRUE),
xlim=c(-3.5, 3.5), ylim=c(-4, 4))¶
```

```
# For Dendogram of Dąbrowska's data with probability values in Figure 2:
> data.trans = t(data)
> data.trans
              stagger  hobble  limp  trudge  plod  amble  saunter  sidle  slink
INJURED             5      15    40       0     0      0        0      0      0
MALE               65      55    60      40    45     20       60     75     25
PLURAL             10       0     0      45    25     70        5      5      5
in_the_room         5       0     0       0     0      0       25      0      0
from_the_pub       40       5     0       0     0      0        0      0      0
home               40       0    10      20    20      0        0      0      0
> library(pvclust)
> plot(pvclust(data.trans, method.dist = "canberra", method.hclust = "ward"),
cex.pv = 1.0, col.pv = c(1, 0, 1)) # to suppress a "BP" probability value and
set number font a little higher than the default
```

```
#For Hierarchical Configural Frequency analysis in Table 6:
> source(“file.path”) # run Stefan Gries’ hcfa_3-2.R script and follow prompts
# OR for similar results
> library(cfa)
> cfa(dataframe)

#For Fisher Exact test on EXPERIENCE life in Table 7:
> data<-matrix(c(80, 293028, 4089, 47263480), nrow = 2)
> fisher.probability<-fisher.test(data)$p.value
> fisher.probability
[1] 6.035451e-18
> round(-log10(fisher.probability),2)
[1] 17.22

#For collostructional analysis in Tables 8-9:
> source(“file.path”) # run Stefan Gries’ coll.analysis.R (version 3) script and follow prompts
```

Recebido em 29/06/2010. Aprovado em 08/05/2011.

# Holistic corpus-based dialectology

## *Dialetologia holística baseada em corpus*

Benedikt Szmrecsanyi*
Christoph Wolk**
Freiburg Institute for Advanced Studies
Freiburg / Germany

ABSTRACT: This paper is concerned with sketching future directions for corpus-based dialectology. We advocate a holistic approach to the study of geographically conditioned linguistic variability, and we present a suitable methodology, 'corpus-based dialectometry', in exactly this spirit. Specifically, we argue that in order to live up to the potential of the corpus-based method, practitioners need to (i) abandon their exclusive focus on individual linguistic features in favor of the study of feature aggregates, (ii) draw on computationally advanced multivariate analysis techniques (such as multidimensional scaling, cluster analysis, and principal component analysis), and (iii) aid interpretation of empirical results by marshalling state-of-the-art data visualization techniques. To exemplify this line of analysis, we present a case study which explores joint frequency variability of 57 morphosyntax features in 34 dialects all over Great Britain.

KEYWORDS: corpus-based dialectology; holistic approach; corpus-based dialectometry; feature aggregates; multivariate analysis; visualization techniques.

RESUMO: Este artigo debruça-se sobre o esboço propositivo de futuras direções para a dialetologia baseada em corpus. Defendemos uma abordagem holística para o estudo da variabilidade linguística geograficamente condicionada, e apresentamos uma metodologia adequada para tal – a dialetometria baseada em corpus. Mais especificamente, defendemos que para que se obtenham todos os resultados esperados da metodologia de corpus, pesquisadores devem: (i) abandonar seu foco exclusivo em traços linguísticos individuais em favor do estudo dos agregados de traços, (ii) amparar-se em métodos computacionais avançados de técnicas de análise multivariada (tais como escalagem multidimensional, análise de clusters, e análise de componente principal), e (iii) auxiliar a interpretação de resultados empíricos através da utilização do estado da arte em técnicas de visualização. A fim de exemplificarmos essa linha de análise, apresentamos um estudo de caso que explora a variabilidade da frequência agregada de 57 traços morfossintáticos de 34 dialetos da Grã-Bretanha.

PALAVRAS-CHAVE: dialetologia baseada em corpus; abordagem holística; dialetometria baseada em corpus; agregados de traços; análise multivariada; técnicas de visualização.

* bszm@frias.uni-freiburg.de

** christoph.wolk@frias.uni-freiburg.de

## 1. Introduction

The customary data sources in traditional dialectology are dialect dictionaries, dialect atlases, and assorted other competence-centered materials. In the past couple of decades, however, more and more dialect corpora have been coming on-line, and corpus-linguistic methodologies have increasingly found their way into the dialectological toolbox (see ANDERWALD; SZMRECSANYI, 2009 for an overview). This is good news, for compared to survey material corpora arguably yield a more realistic and performance-based linguistic signal. Yet, on the empirical-analytical plane corpus-based approaches to dialectology are still a far cry from the rigor and sophistication customary in survey-based dialectology. This is particularly galling since corpora as a data type offer a host of exciting research opportunities not available otherwise. In this paper, we shall argue that corpus-based dialectologists would be well advised to abandon their customary reliance on single-feature studies in favor of holistic, computational approaches that explore joint variability of feature aggregates. In short, we will be advocating a methodology that we have referred to as CORPUS-BASED DIALECTOMETRY (CBDM) elsewhere (cf. SZMRECSANYI, 2008; SZMRECSANYI, 2011).

As a case study to explore CBDM's analytical potential and to highlight the benefits of holistic analysis, we shall tap the *Freiburg Corpus of English Dialects* (FRED) (HERNÁNDEZ, 2006; SZMRECSANYI; HERNÁNDEZ, 2007). FRED contains 368 individual texts and spans approximately 2.4 million words of running text, consisting of samples (mainly transcribed so-called 'oral history' material) of naturalistic, dialectal speech from a variety of sources. Most of these samples were recorded between 1970 and 1990; in most cases, a fieldworker interviewed an informant about life, work, etc. in former days. The 431 informants sampled in the corpus are typically elderly people with a working-class background. The interviews were conducted in 156 different locations (that is, villages and towns) in 34 different pre-1974 counties in Great Britain including the Isle of Man and the Hebrides. The level of areal granularity investigated in the present study will be the county level. This leaves us with 34 dialect objects that will be exemplarily subjected to dialectometrical analysis in the subsequent sections.

This paper is structured as follows. In section 2, we present a number of arguments in favor of holistic analysis. Section 3 defines corpus-based dialectometry. Section 4 sketches some methodical preliminaries. Section 5 draws on a measure of aggregate morphosyntactic distance to present a number of ways in which dialectological datasets can be analyzed holistically: cartographic projections to geography (Section 5.1.), network diagrams (Section 5.2.), and correlational quantitative techniques (Section 5.3.). Section 6 utilizes Principal Component Analysis to identify linguistic structure in the dataset. Section 7 offers some concluding remarks.

## 2. Holistic analysis – why?

AGGREGATE DATA ANALYSIS (also known as DATA SYNTHESIS) is concerned not with the distribution of individual features, properties, or measurements, but with the joint analysis of multiple characteristics. Aggregation is a methodical cornerstone in many academic disciplines. Taxonomists, for instance, typically categorize species not on the basis of a single morphological or genetic criterion, but holistically on the basis of many. By contrast, in linguistics and particularly in corpus linguistics, we find a long and strongly entrenched tradition of looking at individual features in isolation, which is partly a legacy of the discipline's philological origins, and partly a convenience issue. In any event, the one-feature-at-a-time line of analysis – exceptions such as the multidimensional register studies in the spirit of Biber (1988) notwithstanding – has yielded a corpus-based dialectology literature dominated by an abundance of what Nerbonne (2008) has referred to as 'single-feature-based studies'. We will refrain from citing actual studies here (but see the survey in ANDERWALD; SZMRECSANYI, 2009), though fictitious titles such as 'Verbal complementation in West Yorkshire English' or 'The KIT vowel in Appalachian English' are entirely realistic. Now, single-feature studies like this are completely fine, of course, when it is really the features themselves (verbal complementation, the KIT vowel) that are of analytic interest. The approach, however, is uninformative and, in fact, woefully inadequate when single-feature analysts endeavor to characterize multidimensional objects such as dialects and the relationships between them, along the lines of research questions such as 'How does (the grammar and/or phonology and/ or … of) Yorkshire English relate to (the grammar and/or phonology and/or … of) Appalachian English?'. In fact, for addressing questions like these the single-feature approach is about as uninformative and inadequate as a car comparison test whose only criterion is, e.g., the number of cup holders installed.

The problem with single-feature studies – in dialectology as well as everywhere else – is that feature selection is ultimately arbitrary (VIERECK, 1985), and that the next feature down the road may or may not contradict the characterization suggested by the previous feature. For example, Yorkshire English may be progressive in regard to verbal complementation, but conservative as far as verbal agreement is concerned. Thus, there is no guarantee that some dialect or variety will exhibit the same distributional behavior in regard to different features. In addition, individual features may have fairly specific quirks to them that are irrelevant to the big picture and which create noise (NERBONNE, 2009). For instance, the KIT vowel in Appalachian English may very well be a stark outlier in that dialect's phonology, a possibility that we cannot rule out unless we proceed holistically and also look at other features.

In sum, we offer that holistic data analysis is indispensable whenever the analyst's attention is turned to the forest ('dialects'), not the trees ('dialect features'). Data synthesis and aggregation mitigate the problem of feature-specific quirks, irrelevant statistical noise, and the problem of inherently subjective feature selection, and can thus unearth a more robust, objective, and realistic linguistic signal.

## 3. Corpus-based dialectometry

The shortcomings of non-holistic analysis have been known since at least the 1930s (cf., for example, BLOOMFIELD, 1984 [1933]: chapter 19). Starting in the 1970s, computationally inclined dialectologists have addressed these worries by developing a methodology known as DIALECTOMETRY. Dialectometry is defined as the branch of geolinguistics concerned with measuring, visualizing, and analyzing aggregate dialect similarities or distances as a function of properties of geographic space (for seminal work, see SÉGUY, 1971; GOEBL, 1982; GOEBL, 1984; NERBONNE; HEERINGA; KLEIWEG, 1999; HEERINGA, 2004; NERBONNE, 2005; GOEBL, 2006; NERBONNE; KLEIWEG, 2007). Dialectometrical inquiry marshals computational approaches to identify "general, seemingly hidden structures from a larger amount of features" (GOEBL; SCHILTZ, 1997, p. 13) and puts a strong emphasis on quantification, cartographic visualization, and exploratory data analysis to infer patterns from feature aggregates.

Orthodox dialectometry relies on digitized dialect atlases as its primary data source. By contrast, the present contribution outlines a variety of dialectometry that we call CORPUS-BASED DIALECTOMETRY (henceforth: CBDM). The atlas-based method has undeniable advantages – in particular, a fairly widespread availability of data sources and superb areal coverage. By contrast, dialect corpora are in somewhat shorter supply, and their areal coverage is typically inferior to dialect atlases. Having said that, as a data source, corpora have interesting advantages over dialect atlases. First and foremost, the atlas signal is categorical, exhibits a high level of data reduction, and may hence be less accurate than the corpus signal, which can provide graded frequency information. While the exact cognitive status of text frequencies is admittedly still unclear – for example, we do not currently know about the precise extent to which corpus frequencies correlate with psychological entrenchment (ARPPE; GILQUIN; GLYNN; HILPERT; ZESCHEL, 2010) – we do claim that text frequencies match better with the reality of the input perceived by hearers than discrete atlas classifications. Second, we note that the atlas signal is non-naturalistic and, basically, meta-linguistic in nature. It typically relies on elicitation and questionnaires, and is analytically twice removed (via fieldworkers and atlas compilers) from the analyst. By contrast, text corpora – and, by extension, CBDM – provide more direct access to language form and function, and may thus yield a more realistic and trustworthy picture. Furthermore, corpus material is more easily extensible in two ways. On the one hand, it is easier to supplement corpus databases with additional material; for example, oral history recordings comparable to the ones used in FRED are easier to come by than informants that are equally comparable to the ones that completed some atlas questionnaire decades ago. On the other hand, the analysis of atlas data is constrained by the design of the questionnaire, allowing only in a limited way for the study of research questions not originally envisaged. The corpus-based analyst, by contrast, is at more liberty to approach new questions, given that the corpus is of sufficient size.

The well-known major intrinsic drawback of the corpus-based method is that it is unable to deal with textually infrequent phenomena (see, e.g., PENKE; ROSENBACH, 2004, p. 489), and data sparsity is a particular concern when the focus is on syntax and lexis; in this case, a questionnaire study may indeed be the more appropriate research design. Nonetheless, one may justifiably wonder if phenomena that are so infrequent that they cannot be described on the basis of a major text corpus should have a place in an aggregate analysis at all.

## 4. Methodical preliminaries

The first step in CBDM calls for defining the *feature catalogue* as the empirical basis for the data synthesis endeavor. In keeping true to the spirit of dialectometrical analysis and for the sake of avoiding the subjectivity inherent in feature selection, the goal is to base the analysis on as many features as possible. In the case study at hand, we surveyed the dialectological, variationist, and corpus-linguistic literature on morphosyntactic variability in varieties of English for suitable phenomena. This resulted in a list of $p = 57$ features, which overlaps with but is not identical to recent comparative English morphosyntax surveys (cf. KORTMANN; SZMRECSANYI, 2004; SZMRECSANYI; KORTMANN, 2009) and the battery of morphosyntax features covered in the *Survey of English Dialects* (for example, ORTON; SANDERSON; WIDDOWSON, 1978). The Appendix lists the features in the catalogue; for a detailed discussion of the selection criteria, the reader is referred to Szmrecsanyi (2011).

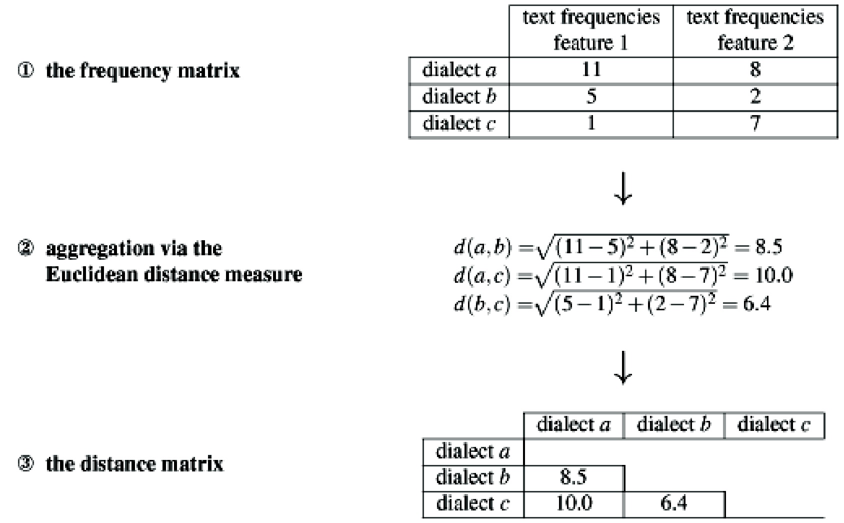
Next, the analyst extracts feature frequencies from the corpus according to best corpus linguistic practice. Szmrecsanyi (2010) details the technicalities of the extraction process in terms of our CBDM case study. Once feature frequencies are extracted, the analyst will normalize text frequencies, and possibly apply a *log*-transformation to de-emphasize large frequency differentials and to alleviate the effect of frequency outliers. Lastly, an $N \times p$ frequency matrix is created in which the $N$ objects (that is, dialects or varieties) are arranged in rows and the $p$ features in columns, such that each cell in the matrix specifies a particular (normalized and *log*-transformed) feature frequency. Our CBDM case study thus yields a 34 × 57 frequency matrix: 34 British English dialects, each characterized by a vector of 57 (normalized and *log*-transformed) text frequencies. The matrix yields a Cronbach's $\alpha$ (cf. NUNNALLY, 1978) value of .77, a score that indicates acceptable reliability.

## 5. Analyzing dialect relationships in the aggregate perspective

The first line of holistic analysis that we shall explore in this paper converts an $N \times p$ frequency matrix into an $N \times N$ distance matrix. This transformation is radically aggregational, in that the resulting distance matrix abstracts away from individual feature frequencies and specifies pairwise distances between the objects. Given the continuous nature of corpus-derived frequency vectors, we advocate usage of the well-known and fairly straightforward *Euclidean distance measure* (ALDENDERFER; BLASHFIELD, 1984, p. 25), which is also known as 'ruler distance'. Based on the Pythagorean theorem, the measure defines the distance between two dialect objects $a$ and $b$ as the square root of the sum of all $p$ squared frequency differentials:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_p - b_p)^2} = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

where $p$ is the number of features, $a_1$ is the frequency of feature 1 in object $a$, $b_1$ is the frequency of feature 1 in object $b$, $a_2$ is the frequency of feature 2 in object $a$, and so on.



**① the frequency matrix**

| | text frequencies feature 1 | text frequencies feature 2 |
|---|---|---|
| dialect *a* | 11 | 8 |
| dialect *b* | 5 | 2 |
| dialect *c* | 1 | 7 |

↓

**② aggregation via the Euclidean distance measure**

$$d(a,b) = \sqrt{(11-5)^2 + (8-2)^2} = 8.5$$
$$d(a,c) = \sqrt{(11-1)^2 + (8-7)^2} = 10.0$$
$$d(b,c) = \sqrt{(5-1)^2 + (2-7)^2} = 6.4$$

↓

**③ the distance matrix**

| | dialect *a* | dialect *b* | dialect *c* |
|---|---|---|---|
| dialect *a* | | | |
| dialect *b* | 8.5 | | |
| dialect *c* | 10.0 | 6.4 | |

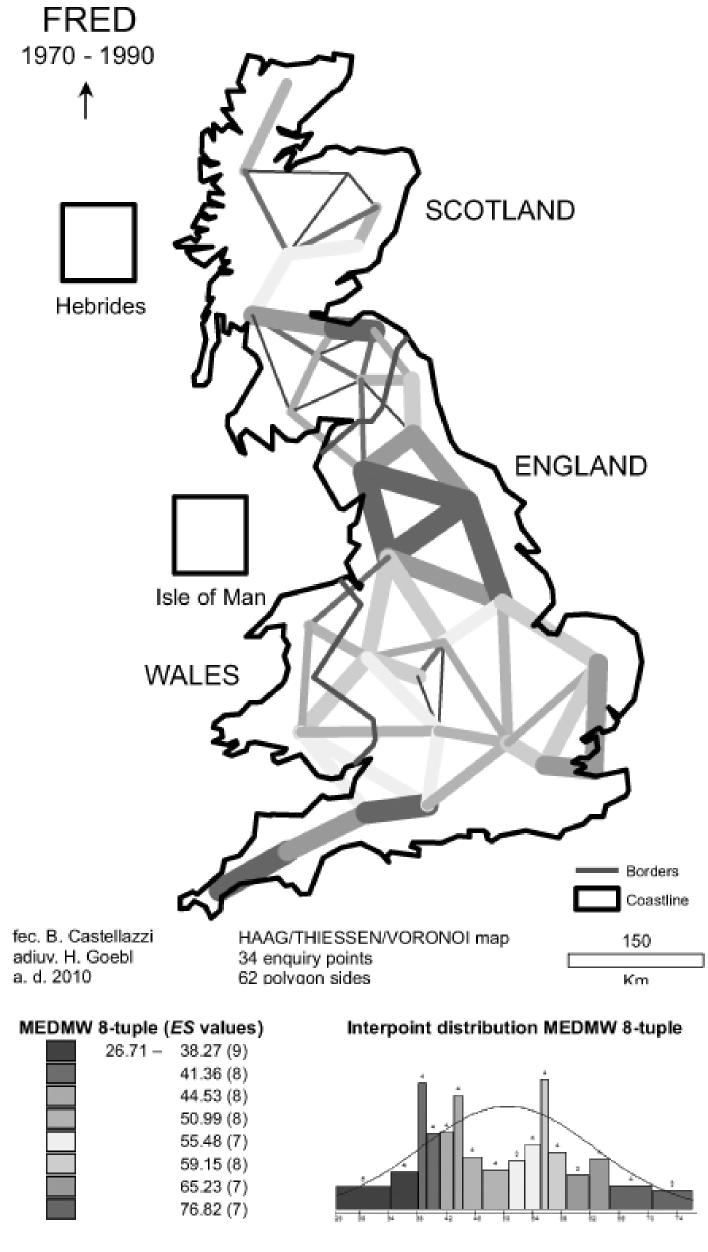
The chart in Figure 1 illustrates the aggregation process. In step①, we start out with a fictional 3 × 2 frequency matrix, which has 6 cells specifying frequencies of 2 features in 3 dialects. In step ② we calculate three distances: the distance between dialects a and b (which we commonsensically define as identical to the distance between dialects b and a), the distance between dialects a and c, and the distance between dialects b and c. In step ③, we enter these distances into a 3 × 3 distance matrix.

Distance matrices can be analyzed in a myriad ways – numerically, cartographically, and diagrammatically. Our cbdm case study's 34 × 57 frequency matrix yields a 34 × 34 distance matrix which describes 34 × 33/2 = 561 pairwise distances between the dialect objects under study. The mean morphosyntactic distance is 5.41 Euclidean distance points. As for the dataset-internal dispersion around the mean, we are dealing with a standard deviation of 1.11. This is another way of saying that roughly two thirds of the 561 dialect pairings score a distance within 1.11 points of the mean, and that 95% of all pairwise distances do not deviate more than 2.22 points from the mean. The minimum observable distance in the dataset is 2.32 points (this happens to be the morphosyntactic distance between the dialects spoken in the county of Somerset and the county of Wiltshire, two neighboring counties located in the Southwest of England). The maximum observable distance in the dataset is 8.14 points, which is the distance between the dialects spoken in the county of Denbighshire in Wales and the county of Kincardineshire in the Scottish Lowlands. The distance matrix comes with a skewness value of -.06, which indicates a very slight negative skew. The kurtosis value is -.37, which is another way of saying that the distribution of distances is a bit flatter than it would be in a perfectly normal distribution.

### 5.1. Cartography

This section will introduce three fairly customary map types that can be utilized to project (aspects of) the information provided in distance matrices to geography: beam maps, continuum maps, and cluster maps. On a technical note, all maps presented in this section were created using freely available dialectometry software: the *Visual DialectoMetry* (VDM) package developed in Salzburg (HAIMERL, 2006), and the Groningen linguist Peter Kleiweg's *Ru*G/*L04* dialectometry software package (available online at <http://www.let.rug.nl/~kleiweg/L04/>).

### 5.1.1. Beam maps



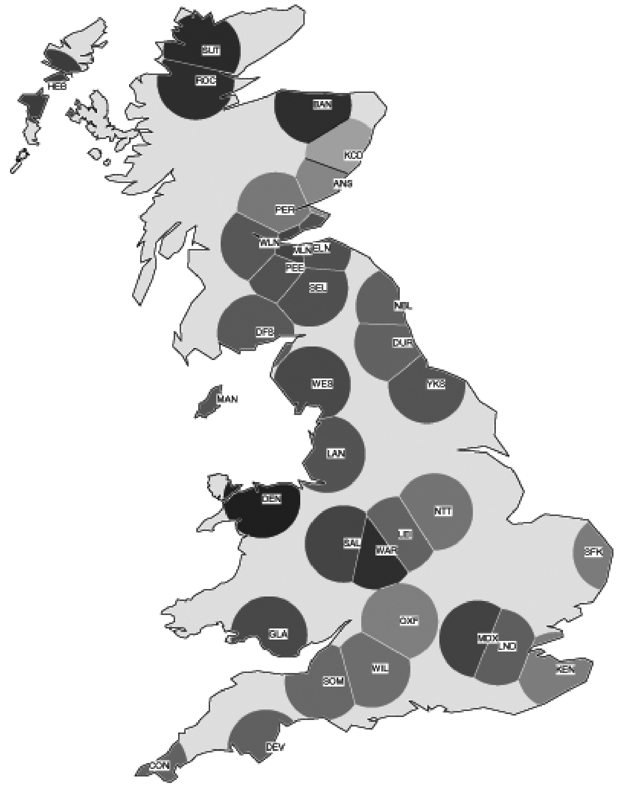
MAP 1. Beam map. Morphosyntactically distant neighbors are connected by cold and thin beams; neighbors that are close morphosyntactically are connected by warm and heavy beams

Beam maps are comparatively straightforward maps that project distance matrices to geography without much statistical ado. They are easy to read because the map type restricts attention to so-called 'interpoint' (i.e. neighbor) relationships (GOEBL, 1982, p. 51). In this spirit, we now turn to Map 1, which features a beam map visually depicting interpoint relationships in our 34 × 34 distance matrix. As for the color coding, note that morphosyntactically distant neighbors are connected by cold (blueish) and thin beams; neighbors that are close morphosyntactically are connected by warm (reddish) and heavy beams. Visual inspection of Map 1 suggests four hotbeds of neighborly similarity in Great Britain. These highlight a very crucial dialect division well-known from the literature – the split between dialects spoken (i) in the Southwest of England, (ii) dialects spoken in the Southeast of England, (iii) dialects spoken in the North of England, and (iv) Scots dialects:

- In the Southwest of England, there is a comparatively marked axis of interpoint similarities running from Cornwall via Devon and Somerset all the way to Wiltshire.
- In the Southeast of England, we note a triangle of relatively modest morphosyntactic similarities connecting Kent, London, and Suffolk.
- In the Northern Midlands and the North of England, we find a web of strong interpoint similarities encompassing Nottinghamshire, Lancashire, Westmorland, Yorkshire, and Durham.
- The Central Scottish Lowlands exhibit a bolt of interpoint similarities involving parts of the urbanized 'Central Belt'.

### 5.1.2. Continuum maps

Many geolinguists intuitively assume that geographic proximity predicts dialectal similarity (cf. NERBONNE; KLEIWEG, 2007, p. 154). This section utilizes more advanced cartography – specifically, so-called continuum maps (HEERINGA, 2004) – to map the extent to which linguistic distance is directly proportional to geographic distance such that there are "no real boundaries, but only gradual transitions" (BLOOMFIELD, 1984 [1933], p. 341). We set the scene by utilizing customary Voronoi tesselation (VORONOI, 1907) to assign each dialect site on the map a convex polygon such that each point within the polygon is closer to the generating dialect site than to any other dialect site (note that as our CBDM case study covers Great Britain with just $N = 34$ sampling points, we will prefer to limit the radius of the Voronoi polygons to approximately 50km in order to do visual justice to the areal coverage of the dialect corpus). The next step is a computational one and subjects the 34 × 34 distance matrix to *Multidimensional Scaling* (MDS) (KRUSKAL; WISH, 1978; EMBLETON, 1993), an exploratory statistical technique to reduce a higher-dimensional dataset to a lower-dimensional representation which is more amenable to visualization. We thus scale down our high-dimensional distance matrix to a three-dimensional representation, in which each object (i.e. dialect) has a coordinate in three artificial MDS dimensions. These coordinates are then mapped to the red-green-blue color scheme, giving each of the Voronoi polygons a distinct hue. Interpetationally, smooth color transitions between dialect polygons emphasize the continuum-like nature of the dialect landscape; abrupt color transitions point to the necessity of alternative explanations.

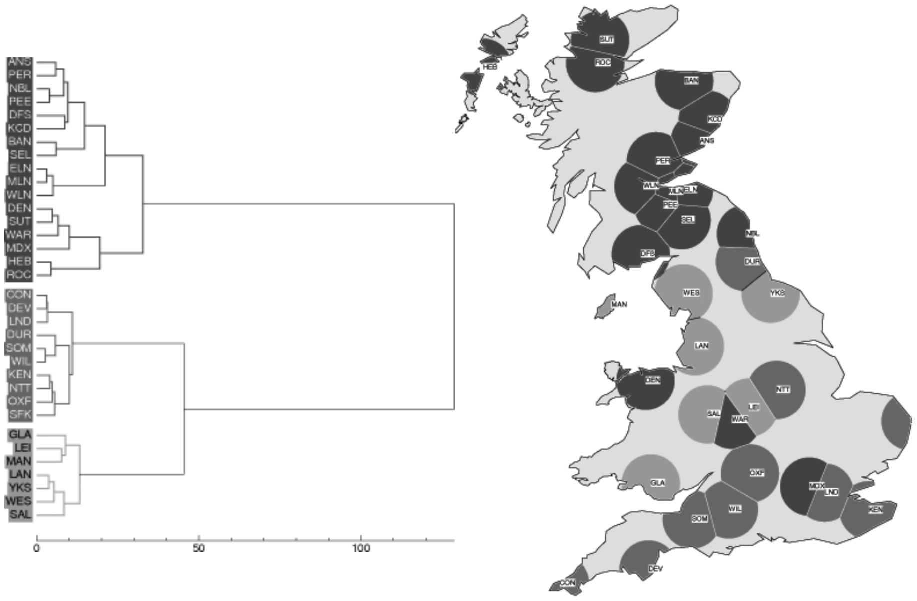


MAP 2. Continuum similarity. Correlation with distances in the original distance matrix: *r* = .95. Map labels are three-letter Chapman county codes (see <http://www.genuki.org.uk/big/Regions/Codes.html> for a legend)

Consider, now, the continuum map in Map 2. The MDS solution depicted is a very accurate one, in that the distances in the three artificial MDS dimensions correlate highly ($r$ = .95) with the distances in the original 34 × 34 distance matrix. In all, the mosaic pattern in the continuum map suggests that the morphosyntactic dialect landscape in Great Britain is in all not exceedingly continuum-like. For sure, there are some fairly nice micro-continua (in, say, the Southwest of England and in the Central and Northern Scottish Lowlands); notice also how nicely dialects spoken in the North of England fade into Southern Scottish Lowlands dialects. But we also observe rather abrupt transitions, for example between the Central Scottish Lowlands and Southern Scottish dialects (Peebleshire and Selkirkshire). In England, the dialects spoken in Middlesex and Warwickshire are outliers. In Wales, it is Denbighshire that does not fit into the picture.

### 5.1.3. Cluster maps

The assumption guiding the discussion in the previous section was that linguistic similarity between dialects is inversely proportional to geographic distance between dialects, and we have seen that this assumption does not necessarily mesh well with the empirical facts. There is, however, an alternative view, according to which dialect landscapes may be geographically organized along the lines of geographically cohesive and linguistically homogeneous "areas within which similar varieties are spoken" (HEERINGA; NERBONNE, 2001, p. 375).



MAP 3. Hierarchical agglomerative cluster analysis (matrix updating algorithm: Ward's method). Left: dendrogram. Right: cluster map

The dialect area scenario may be cartographically explored using cluster maps, a map type which projects the outcome of cluster analysis to geography (HEERINGA, 2004; GOEBL, 2007). As with continuum maps, the starting point is a Voronoi tesselation of map space. Subsequently, the $N \times N$ distance matrix is subjected to *Hierarchical Agglomerative Cluster Analysis* (JAIN; MURTY; FLYNN, 1999), a technique designed to group a number of objects (in this study, dialects) into a smaller number of discrete clusters. While there are many different clustering algorithms, we prefer 'Ward's Minimum Variance Method' (WARD, 1963), an algorithm that tends to create small and even-sized clusters.[1] Cluster analysis can be used to generate a so-called 'dendrogram'

[1] Observe that simple clustering can be unstable, which is why we utilize a technique known as 'clustering with noise' (NERBONNE; KLEIWEG; MANNI, 2008): The original distance matrix is clustered repeatedly, adding some random amount of noise ($c = \sigma/2$) in each run. Then, the collection of resulting treesis examined for groupings that appear in a majority of the individual trees, and from these a new tree with average branch lengths is constructed. This exercise yields a so-called 'cophenetic' distance matrix which provides consensus (and thus more stable) cophenetic distances between dialects, i.e. distances as implied by a tree depicting taxonomic resemblances.

(cf. Map 3), which depicts cophenetic distances between the clustered objects. The optimal number of clusters can be determined by 'elbowing', i.e. diagramming the number of clusters against the fusion coefficient and spotting the 'elbow' in the resulting graph (ALDENDERFER; BLASHFIELD, 1984, p. 54). Finally, each of the clusters is assigned a distinct color hue and the Voronoi polygons are colorized accordingly. Map 3 projects clusters in our CBDM dataset to geography. Despite some geographic incoherence, cluster analysis does detect an areal pattern:

- We find a geographically modestly coherent red cluster comprising most Southern English measuring points (Middlesex being the exception) plus Nottinghamshire in Central England, Suffolk in East Anglia, and Durham in Northern England.
- We also obtain a geographically fairly coherent green group encompassing the majority of measuring points in Northern England (Westmorland, Yorkshire, Lancashire), the Isle of Man, Shropshire and Leicestershire in Central England, and Glamorganshire in Southern Wales.
- Lastly, we are faced with a blue mixed-bag cluster uniting all measuring points in Scotland plus Northumberland in Northern England plus Denbighshire in Northern Wales plus Warwickshire in Central England plus Middlesex in Southern England.

### 5.2. Network diagrams

The previous section introduced agglomerative clustering as a classification method based on dissimilarity, and dendrograms as one way of visualizing its results. Many variants of this general approach have been developed, most of which yield a strictly hierarchical output. Their representation of sub-cluster structure allows interpretation in terms of diachronic development, which is used to great effect in bioinformatics for inferring evolutionary history. In that field, the need to represent uncertainty in the resulting phylogenies as well as mixed evolutionary paths resulting from reticulate effects such as genetic recombination has led to the development of 'splits graphs' for representing non-hierarchical classification (DRESS; HUSON, 2004). One method for constructing such networks, *NeighborNet* (BRYANT; MOULTON, 2004), has found a following in linguistics for historical (McMAHON; McMAHON, 2005), dialectological (McMAHON; HEGGARTY; McMAHON; MAGUIRE, 2007), and typological (CYSOUW,

2007) purposes, thanks to NeighborNet's ability to detect conflicting signals and to represent the effects of language contact. The majority of current applications of NeighborNet in linguistics are restricted to the analysis of categorical atlas-type data. In this section we seek to sketch some of the promises the technique holds for frequency-based analyses.

Let us begin by briefly sketching the algorithm that generates the network diagrams. NeighborNet has the same starting point as the previous analyses – a distance matrix.[2] As with hierarchical agglomerative cluster analysis, the distance matrix is searched for the pair of points with the shortest distance. Instead of immediately fusing these points to a single cluster, however, they are just marked, and this procedure is repeated until the same point is marked twice. Then, these points are replaced with two clusters, each representing the doubly marked point in relation to one of its marked neighbors. This process is repeated until only three clusters are left. The fusion sequence can subsequently be used to generate a network-like diagram. This procedure has some beneficial properties. First, the result will not be needlessly complex. For cases where a segment of the data can be adequately represented as a hierarchical tree, the corresponding segment of the network will be tree-shaped. Second, the method will always produce graphs that are planar, i.e. without crossing lines, which aids visual interpretation.

Figure 2 depicts a network based on the FRED 34 × 34 distance matrix; broad *a priori* dialect areas are indicated via colored labels. The graph was created using the freely available *SplitsTree* package (HUSON; BRYANT, 2006). When interpreting such networks, the equivalents of edges connecting two tree nodes in a dendrogram are either individual lines, or sets of parallel lines. In this network, we only find individual lines directly at the leaf nodes, and many sets of parallel lines, combining to the boxy shapes that form the body of the network. Each represents a way of splitting the total set of dialects into exactly two groups. The longer a given line or set of lines, the greater the difference between the groups. To give an example, the comparatively large vertical set of lines directly below the point where Durham joins the network divides the dialects into the following two groups: one group that consists of Nottinghamshire as well as all Southwestern and Southeastern dialects except Middlesex, and another group that contains all other dialects. When two such divisions are not

[2] On a technical note, NeighborNet relies on observed distances to create a new matrix which takes the net divergences of the involved objects into account.

representable as strictly hierarchical, the resulting lines form boxy shapes. Comparing the network to the strict clustering provided by the dendrogram in Map 3, we find that the network shows considerable amounts of incompatible groupings, indicating that a simple hierarchical classification structure does not entirely adequately capture the uncertainty present in the data.

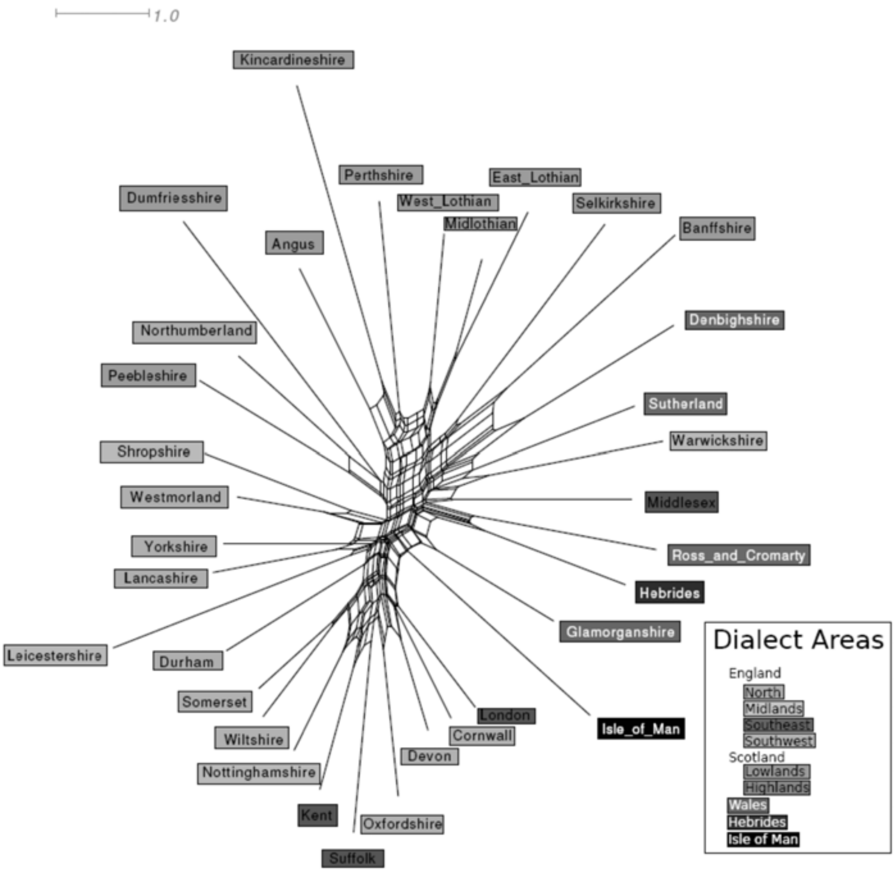


FIGURE 2. Network representation of morphosyntactic distances.
Colors indicate *a priori* dialect areas.

We now turn to the actual networks presented in Figure 2. Overall, the match between dialect areas and placement along the graph seems quite good, as there are several regions on the graph that map to large-scale geographic areas: most Southern dialects can be found at the lower end of the diagram, progressing clockwise through Midlands and Northern dialects toward the Scottish dialects at the top. Most of the non-Scottish components of the 'mixed-bag' cluster discussed in Section 5.1.3., as well as the Scottish Highlands and the Hebrides, are distributed along the right-hand side. The correspondence between geography and the network is certainly not perfect, as some distinctions – such as the difference between Southeastern and Southwestern English dialects – do not materialize in the graph, and the Midlands are mostly intermingled with either Southern or Northern English dialects. Closer inspection of the individual groupings shows that while some of the large-scale areas, such as the (mostly) Southern group mentioned above, are actually represented by an individual split, others (such as the North of England) are not really a single group, but a collection of smaller ones with interlocking resemblances. As one moves from the center of the network toward the individual dialects, such structure becomes apparent throughout the graph, and it is here where the advantage of networks over tree representations is easiest to see. For example, the sub-tree of the dendrogram in Map 3 that connects the rather central Oxfordshire to Nottinghamshire, Kent and Suffolk is still present in the network. Nonetheless, there is also a new, incompatible grouping of Oxfordshire with Devon to the West. This suggests that individual similarities in both directions exist, beyond those that can be explained by the fact that each is a member of the group of (mostly) Southern English dialects. A similar case can be found in the Scottish Lowlands, where the measuring points East Lothian and Midlothian form a rather treelike subgroup. West Lothian, by contrast, is notably removed toward the northern Lowlands. Again, a geographic interpretation is possible, as West Lothian is closer to the northern areas by land and the fjord that separates them from the Lothians, the Firth of Forth, widens considerably to its east.

Network representations are well-suited for finding such suggestive patterns. Compared to the other methods presented in the current paper, though, they are still rather new – especially as applied to dialectological data – and we anticipate future scholarship to further enhance their interpretational utility in the realm of (corpus-based) geolinguistics. Fruitful topics may include context-appropriate validation techniques to increase classification stability in a principled way, projection of non-hierarchical clusters to geography, and techniques for folding network structures back on the individual features from which they originate.

### 5.3. Quantifying the effect of language-external predictors

CBDM is intrinsically quantitative, yet it is fair to say that the foregoing discussion has relied heavily on interpreting cartographic projections to geography and other diagrammatic representations. However, the analyst may also correlate language-external parameters with linguistic distances to precisely quantify the extent to which dialect distances are predictable from language-external factors in the aggregate perspective. Starting out with an $N \times N$ linguistic distance matrix, the name of the game is creating parallel language-external distance matrices, one for each predictor to be tested. In the simplest case, each of these language-external distance matrices is then correlated with the linguistic distance matrix by calculating, e.g., a Pearson product-moment correlation coefficient. The language-external predictor that scores the highest coefficient is the best predictor of linguistic distances (more sophisticated research designs may marshal regression analysis or similar techniques).

To exemplify, let us revisit our dataset on dialect variability in Great Britain. We will correlate the 34 × 34 morphosyntactic distance matrix with three language external distance matrices:

- *As-the-crow-flies distances*. Using a trigonometry formula on the FRED county coordinates, it is computationally trivial to calculate pair wise as-the-crow-flies distances. A proxy for the likelihood of social contact, as-the-crow-flies distance is the most popular geographic distance measure in the dialectometry literature (for example, GOEBL, 2001; GOOSKENS; HEERINGA, 2004; SHACKLETON, 2007).
- *Least-cost travel times*. Speakers do not actually have wings, so we may presume that what really matters for dialect distances is how much time it would take a human traveler to get from point A to point B (GOOSKENS, 2005; SZMRECSANYI to appear). To calculate this measure, we turned to Google Maps (<http://maps.google.co.uk/>), which has a route finder tool that allows the user to enter longitude/latitude pairings for two locations to obtain a least-cost travel route and, crucially, an estimate of the total travel time. We queried Google Maps for all 34 × 33/2 = 561 dialect pairings in our dataset, thus obtaining pair wise least-cost-travel time estimates.[3]

[3] We fully acknowledge that matching linguistic data sourced from speakers born around the beginning of the twentieth century with travel estimates based on twenty-first century transportation infrastructure is convenient but clearly suboptimal.

- Linguistic gravity indices. Trudgill (1974) suggested a Newtonian gravity model to account for geographic diffusion of linguistic features, conjecturing that “the interaction (M) of a centre i and a centre j can be expressed as the population of i multiplied by the population of j divided by the square of the distance between them” (1974, p. 233). Using this formula, we calculated log-transformed (to mitigate the effect of population outliers) linguistic gravity values for each of the 561 data pairings in our database, feeding in least-cost travel time as geographic distance measure and early twentieth century population figures[4] (in thousands) as a proxy for speaker community size.

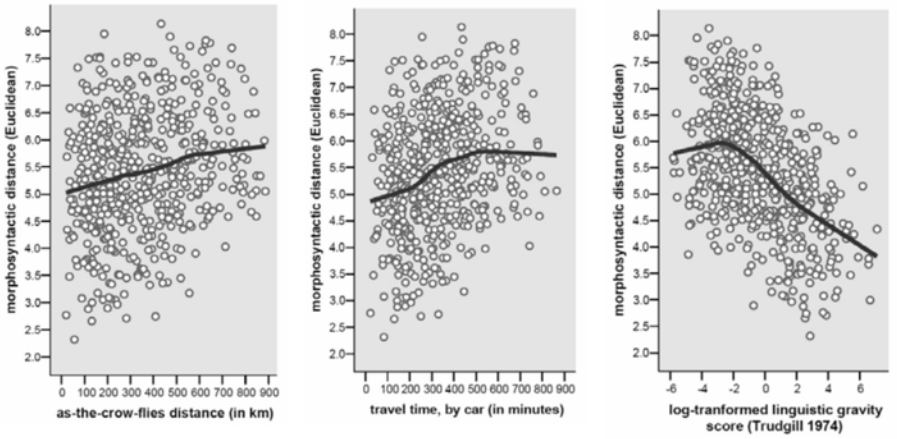


FIGURE 3. Correlating distance matrices: morphosyntactic distances versus (i) as-the-crow-flies distance (left) (r = .21, p < .001, R2=4.4%), (ii) least-cost travel time (middle) (r = .27, p < .001, R2=7.4%), and (iii) log-transformed linguistic gravity indices (right) (r = -.49, p < .001, R2=24.1%). Each dot represents one unique dialect pairing. Solid lines are LOESS curves estimating the overall nature of the relationship.

Nonetheless, we submit that the procedure is not fatally flawed, as modern infrastructure can be argued to actually follow, to a large extent, historical travel routes, trade patterns, and avenues of social contact.

[4] Specifically, we used 1901 population figures, as published in the Census of England and Wales, 1921 and the Census of Scotland, 1921. These documents are available online at <http://histpop.org/>.

Figure 3 provides three scatterplots that graph morphosyntactic distances against the language-external distance measures listed above. The direction of the effect is the theoretically expected one throughout. Increasing as-the-crow-flies distance and increasing least-cost travel time predict increasing morphosyntactic distance; conversely, increasing linguistic gravity indices predict decreasing morphosyntactic distance. The R2 values suggest that as-the-crow flies distance accounts for a meager 4.4% of the morphosyntactic variance, least-cost travel time for 7.4%, and linguistic gravity – and this is a share that one can start writing home about – for 24.1%. Hence, by factoring in speaker community size in addition to geographic distance, we can explain up to a quarter of the aggregate variance in morphosyntactic dialect distances. This does not mean, of course, that cartographic projections to geography – which, after all, inherently draw on as-the-crow-flies distances – are somehow 'wrong'; but we do have an explanation here why, say, the cluster map in Map 3 is not maximally homogeneous geographically.

## 6. Towards identifying linguistic structure

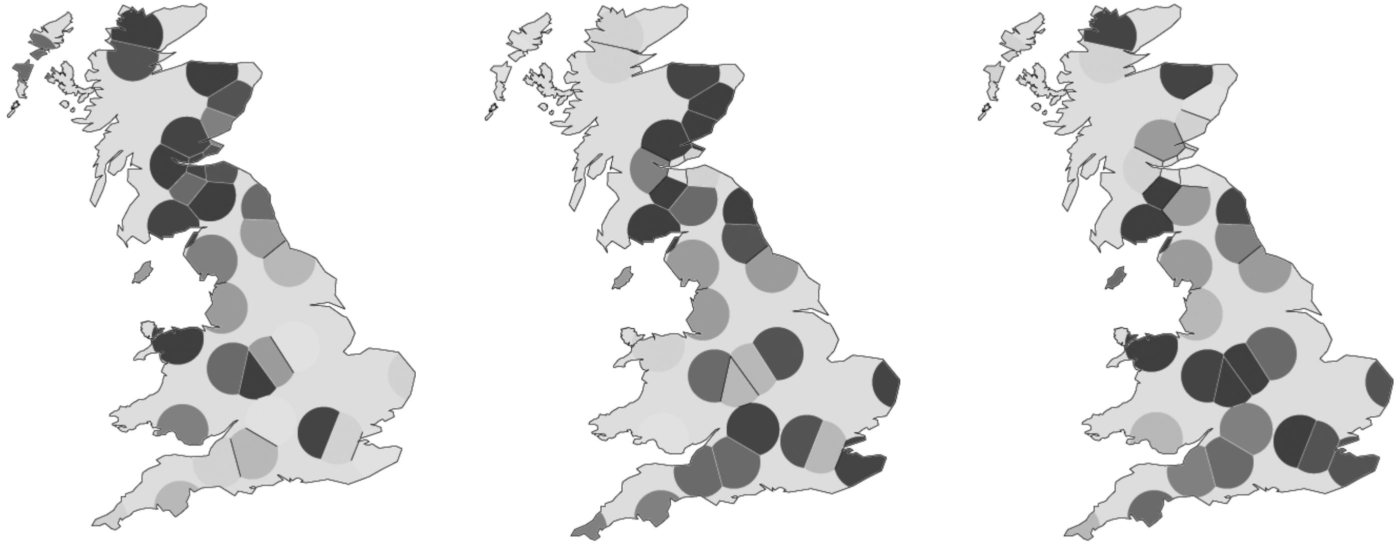
By virtue of analyzing distance matrices which are derived from feature frequencies but which, once the derivation is complete, are completely agnostic of frequencies, the analyses presented in the previous sections were uncompromisingly holistic. However, it is possible to link aggregate patterns of variability to the distribution of individual features, and in so doing detect linguistic structure in aggregate comparison (cf. NERBONNE, 2006). To showcase this approach, we will now utilize *Principal Component Analysis* (PCA) for the sake of addressing two questions: First – on the linguistic/structural level – to what extent do high text frequencies of some feature predict high or low text frequencies of other features? Second – on the geographical plane – how do features thus gang up to create areal patterns?

PCA is a multivariate dimension-reduction technique that transforms a set of high-dimensional vectors (in our case, 57-dimensional feature frequency vectors) into a set of lower-dimensional vectors (so-called 'principal components', which we will interpret as feature bundles) that preserve as much information in the original dataset as possible (DUNTEMAN, 1989, p. 7). PCA is a fairly popular exploratory analysis method; in linguistics, PCA and related techniques are customary in multidimensional studies of register variation (cf. BIBER, 1988). In dialectology, PCA (and a close cousin, factor analysis) have been utilized quite widely as well (SHACKLETON, 2005;

NERBONNE, 2006; WIELING; HEERINGA; NERBONNE, 2007; LEINONEN, 2008). We started out by subjecting the 34 × 57 frequency matrix specifying 57 normalized and *log*-transformed feature frequencies for each of the 34 FRED dialects (cf. Section 4) to PCA.[5] As output, PCA generates two sets of statistics: *component loadings*, which measure the importance of individual linguistic features in particular principal components; and *component scores*, which measure the strength of particular components in particular dialect objects as a function of each feature's frequency value in that dialect object and the feature's component loading in a given component.

PCA extracted 15 components from our case study dataset, of which we will discuss the first three – accounting collectively for about 37% of the morphosyntactic variance – in some detail. The first principal component (PC1), which captures the main dimension of variation, accounts for 17.2% of the variance in the dataset. Adopting a common practice in PCA interpretation (DUNTEMAN, 1989, p. 51), we will select one feature with a particularly high loading to label the principal component in question. The feature loading highest on PC1 is feature [33] (multiple negation, as in *don't you make no damn mistake* [FRED CON005]), with a component loading of .85. This is why we consider PC1 the 'multiple negation component'. The component is associated with a variety of other broad dialect features loading highly on PC1, such as the negator *ain't* (feature [32], as in *people ain't got no money* [FRED NTT013]), *don't* with 3rd person singular subjects (feature [40], as in *this man don't come up to it* [FRED SOM032]), and *as what* or *than what* in comparative clauses (feature [49], as in *we done no more than what other*

[5] We would like to emphasize that like most statistical analysis techniques, PCA does not like small sample sizes, which may lead to overfitting. The 34 × 57 FRED frequency matrix we use here as input to PCA has a subject-to-item ratio that is clearly less than fully satisfactory. In an attempt to increase this ratio, we experimented with excluding 'crossloaders' (i.e. features that load comparatively high on more than one component) and 'non-loaders' (i.e. features that do not load high on any component) from the analysis, the rationale being that crossloaders and non-loaders do obviously not partake in straight feature bundling anyway. This roughly halved the number of features and so doubled the subject-to-item ratio, though the results (that is, component loadings and component scores) stayed overwhelmingly the same. We shall thus proceed in what follows with analyzing the full 34 × 57 FRED frequency matrix, though we would like to caution the reader that the analysis, while accurately describing interdependencies in the FRED dataset, may have a generalizability issue.



Map 4. Component score maps. Left: principal component 1 (variance explained: 17.2%). Middle: principal component 2 (variance explained: 11.1%). Right: principal component 3 (variance explained: 8.9%).. Yellowish hues indicate higher component score.

*kids used to do* [FRED LEI002]). The leftmost projection in Map 4 projects component scores of PC1 to geography. The projection makes amply clear that the multiple negation component has, despite some outliers (Warwickshire, Middlesex) a very nice South-North distribution: the component is very characteristic of dialects in the South of Great Britain, and becomes increasingly less important as one moves North. In fact, component scores exhibit over 40% of shared variance ($r$ = .64, $p$ < .001) with geographic latitude scores.

PC2 seeks to explain as best as it can the variation left over in the dataset after the variance explained by PC1 is taken out of the picture, and in this endeavor it manages to capture 11.1% of the variance. Features loading high on PC2 are typically features that are close to the standard and which would have non-standard alternatives, which we typically also check in the feature catalogue. Consider feature [11] (cardinal number + *years,* as in *ten years later* [FRED HEB006]) – in many dialects, one would hear *ten year later*, which we investigate via feature [12]. Feature [11] is a strong loader on PC2 (.71), and so is feature [46] (*wh*-relativization, as in *the man who has the boat* [FRED HEB028]) and feature [2] (standard reflexives, as in *they was all for theirselves* [FRED NTT002]). We thus choose to label PC2 the '*wh*-relativization component'. Areally, PC2 has not nearly as nice a geographical distribution as PC1, exhibiting as it does a mosaic pattern (cf. the middle projection in Map 4). It is clear, though, that those dialects in which the *wh*-relativization component is particularly popular include all of the comparatively 'young' dialects in Northern Wales (Denbighshire) and the Scottish Highlands (the Hebrides, Ross and Cromarty, and Sutherland). These, in other words, are dialects that are especially close to Standard English.

PC3 accounts for 8.9% of the left-over variance. We dub PC3 the '*-nae* component', as the negative suffix *–nae* (feature [31], as in *I cannae mind of that* [FRED NBL003]) loads high on the component (.59), as does archaic *ye* (feature [4], as in *ye'd dancing every week* [FRED ANS001]). The connoisseur will notice immediately that these are stereotypical Scots features – and indeed, the rightmost projection in Map 4 (which projects PC3 component scores to geography) highlights the *-nae* component's popularity in the Scottish Lowlands. In fact, the component creates a North-South distribution such that geographic latitude scores overlap with PC3 component scores to 13% ($r$ = .37, $p$ = .033). PC3 thus is a Scots component.

## 7. Conclusions and future directions

This paper has advocated an approach – CORPUS-BASED DIALECTOMETRY (CBDM) in short – to the study of geographically conditioned linguistic variability that holistically focuses on the wood and not on the trees. In this spirit, we have argued that corpus-based dialectologists

- would be well-advised to abandon their exclusive focus on individual linguistic features in favor of the study of feature aggregates;
- should reap analytical benefits from utilizing computationally advanced[6] multivariate analysis techniques (multidimensional scaling, cluster analysis, principal component analysis);
- ought to aid interpretation of their results by drawing on various advanced visualization techniques (cartographic projections to geography, network diagrams, and so on).

In this spirit, we hope to have demonstrated that the study of many features in many dialects, coupled with advanced computational analysis methods and sophisticated visualization techniques, can yield insights and generalizations that must remain elusive to the analyst who is beholden to the philologically inspired study of a particular feature in maybe a couple of dialects. For example, our case study on British English dialects has indicated, among other things, that aggregate morphosyntactic variability in Great Britain is, on the whole, not consistently organized along the lines of a dialect continuum, and that we are dealing with some fairly cohesive dialect areas. The layered perspective afforded by principal component analysis subsequently identified those linguistic features that have a continuous geographic distribution (such as features associated with the 'multiple negation component'), and those that don't. We think it is fair to say that the breadth of these statements would be hard to come by in any single-feature study, no matter how interesting the feature.

The methodology sketched in this contribution is, of course, not limited to morphosyntactic phenomena. Phonology, lexis, and even pragmatics are all in principle amenable to dialectometrical analysis using a

[6] By 'computationally advanced' we mean analysis techniques that – unlike e.g. eyeballing the data, simple crosstabulation etc. – cannot be normally conducted without computer-aided processing.

corpus-based approach. There is even the intriguing possibility of aggregating not 'surfacy' feature frequencies but 'deep' feature conditionings (e.g. via probabilistic regression weights), a feat that is simply not possible on the basis of decontextualized survey data. Basing future extensions to the CBDM tool set on a probabilistic basis would furthermore allow taking variation on the level of the speaker into account, concerning both how the independent effects of other factors such as gender and speaker age influence language variation and how homogeneous individual counties really are. Also note that CBDM can be applied to any corpus in which we find geographic variability. This includes not only dialect corpora in the traditional sense, but also corpora sampling geographically non-contiguous regional language varieties (such as the *International Corpus of English*) or corpora concerned with variation in written, not spoken, language (such as the letters-to-the-editor corpus analyzed in Grieve 2009).

## References

ALDENDERFER, M. S.; BLASHFIELD, R. K. *Cluster Analysis*. Newbury Park, London, New Delhi: Sage Publications, 1984.

ANDERWALD, L.; SZMRECSANYI, B. Corpus linguistics and dialectology. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus Linguistics.* An International Handbook. Handbücher zur Sprache und Kommunikationswissenschaft/ Handbooks of Linguistics and Communication Science. Berlin / New York: Mouton de Gruyter, 2009.

ARPPE, A.; GILQUIN, G.; GLYNN, D.; HILPERT, M.; ZESCHEL, A. Cognitive Corpus Linguistics: Five points of debate on current theory and methodology. *Corpora*, v. 5, n. 2, p. 1-27, 2010.

BIBER, D. *Variation across Speech and Writing*. Cambridge: Cambridge University Press, 1988.

BLOOMFIELD, L. *Language*. Chicago: University of Chicago Press, 1984 [1933].

BRYANT, D.; MOULTON, V. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Mol. Biol. Evol.*, v. 21, n. 2, p. 255-265, 2004.

CYSOUW, M. New approaches to cluster analysis of typological indices. In: KÖHLER, R.; GRZBEK, P. (Ed.). *Exact Methods in the Study of Language and Text*. Berlin, New York: Mouton de Gruyter, 2007.

DRESS, A. W. M.; HUSON, D. H. Constructing Splits Graphs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, v. 1, n. 3, p. 109-115, 2004.

DUNTEMAN, G. H. *Principal components analysis*. Newbury Park: Sage Publications, 1989.

EMBLETON, S. Multidimensional scaling as a dialectometrical technique: Outline of a research project. In: KÖHLER, R.; RIEGER, B. (Ed.). *Contributions to quantitative linguistics*. Dordrecht: Kluwer, 1993.

GOEBL, H. *Dialektometrie:* Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie. Wien: Österreichische Akademie der Wissenschaften, 1982.

GOEBL, H. *Dialektometrische Studien:* Anhand italoromanischer, rätroromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. Tübingen: Niemeyer, 1984. 3 v.

GOEBL, H. Arealtypologie und Dialektologie. In: HASPELMATH, M.; E. KÖNIG, E.; OESTERREICHER, W.; RAIBLE, W. (Ed.). *Language Typology and Language Universals / La typologie des langues et les universaux linguistiques / Sprachtypologie und sprachliche Universalien: An International Handbook / Manuel international / Ein internationales Handbuch*. Berlin, New York: Walter de Gruyter, 2001. v. 2.

GOEBL, H. Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing*, v. 21, n. 4, p. 411-435, 2006.

GOEBL, H. A bunch of dialectometric flowers: a brief introduction to dialectometry. In: SMIT, U.; DOLLINGER, S.; HÜTTNER, J.; KALTENBÖCK, G.; LUTZKY, U. (Ed.). *Tracing English through time:* Explorations in language variation. Wien: Braumüller, 2007.

GOEBL, H.; SCHILTZ, G. A dialectometrical compilation of CLAE 1 and CLAE 2: Isoglosses and dialect integration. In: VIERECK, W.; RAMISCH, H. (Ed.). *Computer developed linguistic atlas of England (CLAE)*. Tübingen: Max Niemeyer Verlag, 1997. v. 2.

GOOSKENS, C. Traveling time as a predictor of linguistic distance. *Dialectologia et Geolinguistica*, v. 13, p. 38-62, 2005.

GOOSKENS, C.; HEERINGA, W. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, v. 16, n. 3, p. 189-207, 2004.

GRIEVE, J. *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English*. 340f. 2009. PhD (Dissertation) – Northern Arizona University.

HAIMERL, E. Database Design and Technical Solutions for the Management, Calculation, and Visualization of Dialect Mass Data. *Literary and Linguistic Computing*, v. 21, n. 4, p. 437-444, 2006.

HEERINGA, W. *Measuring dialect pronunciation differences using Levenshtein distance*, 2004. 312f. PhD (Dissertation) – University of Groningen.

HEERINGA, W.; NERBONNE, J. Dialect areas and dialect continua. *Language Variation and Change*, v. 13, n. 3, p. 375-400, 2001.

HERNÁNDEZ, N. *User's Guide to FRED*. URN: urn:nbn:de:bsz:25-opus-24895, URL: http://www.freidok.uni-freiburg.de/volltexte/2489/. Freiburg: University of Freiburg, 2006.

HUSON, D. H.; BRYANT, D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology Evolution*, v. 23, n. 2, p. 254-267, 2006.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys*, v. 31, n. 3, p. 264-323, 1999.

KORTMANN, B.; SZMRECSANYI, B. Global synopsis: morphological and syntactic variation in English. In: KORTMANN, B.; SCHNEIDER, E.; BURRIDGE, K.; MESTHRIE, R.; UPTON, C. (Ed.). *A Handbook of Varieties of English*. Berlin/New York: Mouton de Gruyter, 2004. v. 2.

KRUSKAL, J. B.; WISH, M. *Multidimensional Scaling*. Newbury Park, London / New Delhi: Sage Publications, 1978.

LEINONEN, T. Factor Analysis of Vowel Pronunciation in Swedish Dialects. *International Journal of Humanities and Arts Computing*, v. 2, n. 1-2, p. 189-204, 2008.

MCMAHON, A.; HEGGARTY, P.; MCMAHON, R.; MAGUIRE, W. The sound patterns of Englishes: representing phonetic similarity. *English Language and Linguistics*, v. 11, n. 1, p. 113-142, 2007.

MCMAHON, A. M. S.; MCMAHON, R. *Language classification by numbers*. Oxford New York: Oxford University Press, 2005.

NERBONNE, J. Computational Contributions to Humanities. *Linguistic and Literary Computing*, v. 20, n. 1, p. 25-40, 2005.

NERBONNE, J. Identifying Linguistic Structure in Aggregate Comparison. *Literary and Linguistic Computing*, v. 21, n. 4, p. 463-475, 2006.

NERBONNE, J. Variation in the aggregate: an alternative perspective for variationist linguistics. In: DEKKER, K.; MACDONALD, A.; NIEBAUM, H. (Eds.); *Northern Voices:* Essays on Old Germanic and Related Topics offered to Professor Tette Hofstra. Leuven: Peeters, 2008.

NERBONNE, J. Data-driven dialectology. *Language and Linguistics Compass*, v. 3, n. 1, p. 175-198, 2009.

NERBONNE, J.; HEERINGA, W.; KLEIWEG, P. Edit Distance and Dialect Proximity. In: SANKOFF, D.; KRUSKAL, J. (Ed.). *Time Warps, String Edits and Macromolecules:* The Theory and Practice of Sequence Comparison. Stanford: CSLI Press, 1999.

NERBONNE, J.; KLEIWEG, P. Toward a Dialectological Yardstick. *Journal of Quantitative Linguistics*, v. 14, n. 2, p. 148-166, 2007.

NERBONNE, J.; KLEIWEG, P.; MANNI, F. Projecting dialect differences to geography: bootstrapping clustering vs. clustering with noise. In: PREISACH, C.; SCHMIDT-THIEME, L.; BURKHARDT, H.; DECKER, R. (Ed.). *Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society*. Berlin: Springer, 2008.

NUNNALLY, J. C. *Psychometric Theory*. McGraw-Hill, 1978.

ORTON, H.; SANDERSON, S.; WIDDOWSON, J. D. A. *The Linguistic Atlas of England*. London, Atlantic Highlands, N.J.: Croom Helm, 1978.

PENKE, M.; ROSENBACH, A. What counts as evidence in linguistics? An introduction. *Studies in Language*, v. 28, n. 3, p. 480-526, 2004.

SÉGUY, J. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, v. 35, p. 335-357, 1971.

SHACKLETON, R. G. J. English-American Speech Relationships: A Quantitative Approach. *Journal of English Linguistics*, v. 33, n. 2, p. 99-160, 2005.

SHACKLETON, R. G. J. Phonetic variation in the traditional English dialects: a computational analysis. *Journal of English Linguistics*, v. 35, n. 1, p. 30-102, 2007.

SZMRECSANYI, B. Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing*, v. 2, n. 1-2, p. 279-296, 2008.

SZMRECSANYI, B. *The morphosyntax of BrE dialects in a corpus-based dialectometrical perspective: feature extraction, coding protocols, projections to geography, summary statistics.* URN: urn:nbn:de:bsz:25-opus-73209, URL: http://www.freidok.uni-freiburg.de/volltexte/7320/. Freiburg: University of Freiburg, 2010.

SZMRECSANYI, B. Corpus-based dialectometry – a methodological sketch. *Corpora*, v. 6, n. 1, 2011.

SZMRECSANYI, B. Geography is overrated. In: HANSEN, S.; SCHWARZ, C.; STOECKLE, P.; STRECK, T. (Ed.). *Dialectological and folk dialectological concepts of space*. Berlin, New York: Walter de Gruyter, to appear.

SZMRECSANYI, B.; HERNÁNDEZ, N. *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S").* URN: urn:nbn:de:bsz:25-opus-28598, URL: http://www.freidok.uni-freiburg.de/volltexte/2859/. Freiburg: University of Freiburg, 2007.

SZMRECSANYI, B.; KORTMANN, B. The morphosyntax of varieties of English worldwide: a quantitative perspective. *Lingua*, v. 119, n. 11, p. 1643-1663, 2009.

TRUDGILL, P. Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society*, v. 2, p. 215-246, 1974.

VIERECK, W. Linguistic atlases and dialectometry: The survey of English dialects. In: KIRK, J. M.; SANDERSON, S.; WIDDOWSON, J. D. A. (Ed.). *Studies in linguistic geography:* The dialects of English in Britain and Ireland. London: Croom Helm, 1985.

VORONOI, G. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik*, v. 133, p. 97-178, 1907.

WARD, J. H. J. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, v. 58, p. 236-244, 1963.

WIELING, M.; HEERINGA, W.; NERBONNE, J. An aggregate analysis of pronunciation in the Goeman-Taeldeman-van Reenen-Project data. *Taal en Tongval*, v. 59, n. 1, p. 84-116, 2007.

## Appendix: the feature catalogue

NOTE: see Szmrecsanyi (2010) for a version of the feature catalogue annotated with linguistic examples

A. Pronouns and determiners

[1] non-standard reflexives

[2] standard reflexives

[3] archaic *thee/thou/thy*

[4] archaic *ye*

[5] *us*

[6] *them*

B. The noun phrase

[7] synthetic adjective comparison

[8] the *of*-genitive

[9] the *s*-genitive

[10] preposition stranding

[11] cardinal number + *years*

[12] cardinal number + *year*-Ø

C. Primary verbs

[13] the primary verb TO DO

[14] the primary verb TO BE

[15] the primary verb TO HAVE

[16] marking of possession – HAVE GOT

D. Tense and aspect

[17] the future marker BE GOING TO

[18] the future markers WILL/SHALL

[19] WOULD as marker of habitual past

[20] *used to* as marker of habitual past

[21] progressive verb forms

[22] the present perfect with auxiliary BE

[23] the present perfect with auxiliary HAVE

E. Modality

[24] marking of epistemic and deontic modality: MUST

[25] marking of epistemic and deontic modality: HAVE TO

[26] marking of epistemic and deontic modality: GOT TO

F. Verb morphology

[27] a-prefixing on *-ing*-forms

[28] non-standard weak past tense and past participle forms

[29] non-standard past tense *done*

[30] non-standard past tense *come*

G. Negation

[31] the negative suffix *-nae*

[32] the negator *ain't*

[33] multiple negation

[34] negative contraction

[35] auxiliary contraction

[36] *never* as past tense negator

[37] WASN'T

[38] WEREN'T

H. Agreement

[39] non-standard verbal *-s*

[40] *don't* with 3rd person singular subjects

[41] standard *doesn't* with 3rd person singular subjects

[42] existential/presentational *there is/was* with plural subjects

[43] absence of auxiliary BE in progressive constructions

[44] non-standard WAS

[45] non-standard WERE

I. Relativization

[46] *wh*-relativization

[47] the relative particle *what*

[48] the relative particle *that*

J. Complementation

[49] *as what* or *than what* in comparative clauses

[50] unsplit *for to*

[51] infinitival complementation after BEGIN, START, CONTINUE, HATE, and LOVE

[52] gerundial complementation after BEGIN, START, CONTINUE, HATE, and LOVE

[53] zero complementation after THINK, SAY, and KNOW

[54] that complementation after THINK, SAY, and KNOW

K. Word order and discourse phenomena

[55] lack of inversion and/or of auxiliaries in *wh*-questions and in main clause *yes/no*-questions

[56] the prepositional dative after the verb GIVE

[57] double object structures after the verb GIVE

Recebido em 31/08/2010. Aprovado em 08/05/2011.