Frequency Effects in Language Learning and Processing

Trends in Linguistics
Studies and Monographs 244.1

De Gruyter Mouton

Frequency Effects in Language Learning and Processing

edited by

Stefan Th. Gries
Dagmar Divjak

De Gruyter Mouton

ISBN 978-3-11-027376-2
e-ISBN 978-3-11-027405-9
ISSN 1861-4302

Library of Congress Cataloging-in-Publication Data

A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Typesetting: RoyalStandard, Hong Kong
Printing: Hubert & Co. GmbH & Co. KG, Göttingen
♾ Printed on acid-free paper

Printed in Germany

www.degruyter.com

Table of contents

Introduction

Stefan Th. Gries

The papers in this volume and its companion (Divjak & Gries 2012) were originally part of a theme session 'Converging and diverging evidence: corpora and other (cognitive) phenomena?' at the Corpus Linguistics 2009 conference in Liverpool as well as part of a theme session 'Frequency effects in language' planned for the International Cognitive Linguistics Conference at the University of California, Berkeley in 2009. We are very fortunate to have received a large number of very high-quality submissions to these events as well as to these two volumes and wish to thank our contributors for their contributions and their patience during the time that was taken up by revisions and the preparation of the final manuscript.

Usually, the purpose of an introduction to such an edited volume is to survey current trends in the relevant field(s), provide brief summaries of the papers included in the volume, and situate them with regard to what is currently happening in the field. The present introduction deviates from this tradition because one of the papers solicited for these two theme session was solicited as such an overview paper. Thus, in this volume, this overall introduction will restrict itself to brief characterizations of the paper and a few additional comments – the survey of the field and the identification of current trends and recent developments on the other hand can be found in Ellis's state-of-the-art overview. Ellis discusses the interrelation of frequency and cognition – in cognition in general as well as in (second) language cognition – and, most importantly given current discussions in usage-based approaches to language, provides a detailed account of the factors that drive the kind of associative learning assumed by many in the field: type and token frequency, Zipfian distributions as well as recency, salience, perception, redundancy etc. Just as importantly, Ellis derives a variety of conclusions or implications of these factors for our modeling of learning and acquisition processes, which sets the stage for the papers in this volume.

The other papers cover a very large range of approaches and methods. Most of them are on synchronic topics, but de Vogelaer focuses on frequency effects in language change. Several studies are on native speakers' linguistic behavior, but some involve data from non-adult native language speakers, second language learners, and native speakers (of English) learn-

ing an artificial language. Many studies involve corpus data in the form of various types of frequency data, but many also add experimental or other approaches as diverse as acceptability judgments, shadowing, questionnaires, eye-tracking, sentence-copying, computational modeling, and artificial language learning; most of them also involve sophisticated statistical analysis of various kinds (correlations and different types of linear models, logistic regressions, linear mixed-effect models, cluster analytical techniques). The following summarizes the papers in this volume, which proceeds from phonological topics (Raymond & Brown and Kapatsinksi) via morphological studies (Antić and de Vogelaer) to syntactic/*n*-gram studies (Huang, Wible, & Ko as well as Caines and Mos, van den Bosch, & Berck).

Raymond & Brown explore a range of frequency-related factors and their impact on initial fricative reduction in Spanish. They begin by pointing out that results of previous studies have been inconclusive, in part because many different studies have included only partially overlapping predictors and controls; in addition, the exact causal nature of frequency effects has also proven elusive. They then study data on [s]-initial Spanish words from the free conversations from the New Mexico-Colorado Spanish Survey, a database of interviews and free conversations initiated in 1991. A large number of different frequency-related variables is coded for each instance of an *s*-word, including word frequency, bigram frequency, transitional probability (in both directions), and others, and these are entered into a binary logistic regression to try to predict fricative reduction.

The results show that *s*-reduction is influenced by many predictors, too many to discuss here in detail. However, one very interesting conclusion is that, once a variety of contextual frequency measures is taken into consideration, then non-contextual measures did not contribute much to the regression model anymore, which is interesting since it forces us to re-evaluate our stance on frequency, from a pure repetition-based view to a more contextually-informed one, which in itself would constitute a huge conceptual development (cf. also below).

Kapatsinski's study involves a comparison of product-oriented vs. source-oriented generalizations by means of an artificial-language learning experiment. Native speakers of English are exposed to small artificial languages that feature a palatalization process but differ in terms of whether the sound favoring palatalization is also found attaching to the sound that would be the result of the palatalization. The exposition to the artificial languages (with small interactive video-clips) favors either a source-oriented generalization or a product-oriented generalization. The results as obtained from cluster analyses of rating and production probabilities

provide strong support for product-oriented generalizations (esp. when sources and products are not close to each other).

The paper by Antić studies the productivity of two Russian verb prefixes, *po* vs. *voz*/*vos*/*vz*/*vs* and the morphological decomposition. She first compares the two prefixes in terms of a variety of desiderata of productivity measures (e.g., intuitiveness and hapaxability). She then uses simple linear regressions for both prefixes and shows, on the basis of intercepts and slopes, that *po* is indeed the more productive affix.

In a second case study, Antić reports on a prefix-separation experiment in which subjects' reaction time is the critical dependent variable. On the basis of a linear mixed-effects regression, she identifies a variety of parameters that significantly affect subjects' RTs, such as semantic transparency of the verbs, unprefixed family size, and the difference between the frequency of the base verb and the frequency of the prefixed form. Three different theoretical accounts of the data are discussed, with the final analysis opting for a Bybee/Langacker type of network model of morphological representation.

De Vogelaer studies the gender systems of Dutch dialects. More specifically, he starts out from the fact that Standard Dutch exhibits a gender mismatch of the binary article system and the ternary pronominal system and explores to what degree this historical change is affected by frequency effects. Results from a questionnaire study, in which subjects were put in a position to decide on the gender of nouns, indicate high- and low-frequency items behave differently: the former are affected in particular by standardization whereas the latter are influenced more by resemanticization. However, the study also cautions us that different types of data can yield very different results with regard to the effect of frequency. De Vogelaer compares frequency data from the 9-million-word Spoken Dutch Corpus to age-of-acquisition data from a target vocabulary list. Correlation coefficients indicate that the process of standardization is more correlated with the adult spoken corpus frequencies whereas resemanticization is more correlated with the age-of-acquisition data. As De Vogelaer puts it, "frequency effects are typically poly-interpretable," and he rightly advises readers to regularly explore different frequency measures and register-specific frequencies.

The study by Huang, Wible, & Ko is concerned with transitional probabilities between words at the end of multi-word expressions, or *n*-grams, with the focus being on the contrast between frequent and entrenched cases such as *on the other hand* and less frequent and entrenched cases such as *examined the hand*. A first eye-tracking study tested whether L1

and L2 speakers of English react differently to these different degrees of predictability of *hand*. Results from ANOVAs on fixation probabilities, first-fixation durations, and gaze durations reveal that both speaker groups respond strongly to the difference in predictability that results from the entrenched multi-word expressions.

A follow-up case study explores these results in more detail by investigating some multi-word expression that had not exhibited a sensitivity towards transitional probabilities in the first experiment. L2-speakers were exposed to such expressions during a training phase (with two different types of exposure) and then tested with the same experimental design as before. The results show that frequent exposure during the training phase facilitated their processing, and more so than a less frequent but textually enhanced exposure to the stimuli. Both studies therefore show that the well-researched ability of speakers to detect/utilize transitional probabilities is also observed for L2 speakers and that basic assumptions of usage-based approaches as to how input frequency affects processing are supported.

Caines's study is another one that combines corpus and experimental data. His focus is the zero-auxiliary interrogative in spoken British English (e.g., *you talkin' to me?*). His first case study is based on a multifactorial analysis of nearly ten thousand cases of progressive interrogatives in the spoken part of the BNC. Using a binary logistic regression, he identifies several predictors that significantly affect the probability of zero-auxiliary forms, including, for example, the presence of a subject, second person, as well as the number of the verb.

To shed more light on the construction's characteristics, Caines also reports on two experiments, an acceptability judgment task (using magnitude estimation) and a continuous shadowing task. An ANOVA of the acceptability judgment largely corroborates the corpus-based results, as do restoration and error rates in the shadowing tasks, providing a clear example of how methodological pluralism can shed light on different aspects of one and the same phenomenon.

Mos, van den Bosch, & Berck's study is also devoted to a multi-word expression, namely to what they call the Fixed Adjective Preposition (FAP) construction in Dutch, as exemplified by *de boer is trots op zijn auto* ('the farmer is proud of his car'). They employ a rare and creative way to investigate how speakers partition sentences into parts: subjects are asked to copy a sentence, and the dependent variable is the points at which subjects revisit the sentence they are copying, the assumption being that this will fall between constituents more often than interrupt them. In this case, the subjects – 6th graders – were asked to copy fairly frequent

expressions that could have a FAP or a 'regular' V-PP interpretation. The experimental results were then analyzed and compared to the results of a computational language model trained on approximately 50 m words from newspaper corpora.

The data show that the FAP construction has not been fully acquired by all children and not fully schematically so. However, there are factors, such as verbal collocates, that can enhance the FAP construction's unit status, a finding compatible with the assumption that unithood is in fact not a yes-or-no, but a gradient property. The experimental data were fairly similar to the computational model, but the differences, which may in part just be due to specific properties of the algorithm, still indicate that human processing is much more involved than just based on co-occurrence frequencies.

The papers in this volume, together with those in the companion volume, testify to the richness of contemporary research on frequency effects at the interface of cognitive linguistics and usage-based linguistics on the one hand, and corpus linguistics and psycholinguistics on the other. This is particularly good news for the corpus-linguistic community, parts of which have been resisting a turn towards more cognitively- and psycholinguistically-informed work (cf. Gries (2010) for discussion) in spite of the large amount of compatibility between the disciplines and the possibility that cognitive linguistics and psycholinguistics would breathe some new life into corpus studies. For example, the most interesting conclusions of Raymond & Brown's study above echoes a finding of Baayen (2010: 456), who finds that (my emphases, STG)

> most of the variance in lexical space is carried by a principal component on which *contextual measures* (syntactic family size, *syntactic entropy, BNC dispersion*, morphological family size, and *adjectival relative entropy*) have the highest loadings. Frequency of occurrence, in the sense of pure repetition frequency, explains only a modest proportion of lexical variability.

Findings like these have the potential to bring about no less than a paradigm shift in corpus linguistics such that we finally begin to leave behind simplistic frequencies of (co-)occurrence and take the high dimensionality of our data as seriously as it needs to be taken.

In that regard, it is also cognitive linguistics and psycholinguistics that benefit from the fruitful interdisciplinariness that the papers in this volume already exemplify. Cognitive linguistics in particular has also too long relied on sometimes too simple operationalizations (of, say, frequency), and needs to take the associative-learning literature from cognitive psy-

chology (cf. Ellis's paper), but also new developments in corpus linguistics into consideration, such as association measures (uni- and bidirectional ones), dispersion, entropies of type-token distributions, etc.; cf. Gries 2008, to appear). If more scholars were inspired by the papers in this volume, cognitive and corpus linguistics together will yield a wealth of new findings shedding light on the interplay of language and cognition.

References

Baayen, R. Harald
2010 Demythologizing the word frequency effect: a discriminative learning perspective. *The Mental Lexicon* 5(3): 436–461.

Gries, Stefan Th.
2008 Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4): 403–437.

Gries, Stefan Th.
2010 Corpus linguistics and theoretical linguistics: a love-hate relationship? Not necessarily … *International Journal of Corpus Linguistics* 15(3): 327–343.

Gries, Stefan Th.
to appear 50-something years of work on collocations: what is or should be next … *International Journal of Corpus Linguistics*.

What can we count in language, and what counts in language acquisition, cognition, and use?

Nick C. Ellis

"Everything that can be counted does not necessarily count;
everything that counts cannot necessarily be counted."
—Albert Einstein

"Perception is of definite and probable things"
—William James 1890

1. Frequency and Cognition

From its very beginnings, psychological research has recognized three major experiential factors that affect cognition: frequency, recency, and context (e.g., Anderson 2000; Ebbinghaus 1885; Bartlett [1932] 1967). Learning, memory and perception are all affected by frequency of usage: The more times we experience something, the stronger our memory for it, and the more fluently it is accessed. The more recently we have experienced something, the stronger our memory for it, and the more fluently it is accessed. (Hence your more fluent reading of the prior sentence than the one before). The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization; so a stimulus becomes associated to a context and we become more likely to perceive it in that context. The power law of learning (Anderson 1982; Ellis and Schmidt 1998; Newell 1990) describes the relationships between practice and performance in the acquisition of a wide range of cognitive skills – the greater the practice, the greater the performance, although effects of practice are largest at early stages of leaning, thereafter diminishing and eventually reaching asymptote. The power function relating probability of recall and recency is known as the forgetting curve (Baddeley 1997; Ebbinghaus 1885).

William James' words which begin this section concern the effects of frequency upon perception. There is a lot more to perception than meets the eye, or ear. A percept is a complex state of consciousness in which

antecedent sensation is supplemented by consequent ideas which are closely combined to it by association. The cerebral conditions of the perception of things are thus the paths of association irradiating from them. If a certain sensation is strongly associated with the attributes of a certain thing, that thing is almost sure to be perceived when we get that sensation. But where the sensation is associated with more than one reality, unconscious processes weigh the odds, and we perceive the most probable thing: "all brain-processes are such as give rise to what we may call FIGURED consciousness" (James, 1890, p. 82). Accurate and fluent perception thus rests on the perceiver having acquired the appropriately weighted range of associations for each element of the sensory input.

It is human categorization ability which provides the most persuasive testament to our incessant unconscious figuring or 'tallying' (Ellis 2002). We know that natural categories are fuzzy rather than monothetic. Wittgenstein's (1953) consideration of the concept game showed that no set of features that we can list covers all the things that we call games, ranging as the exemplars variously do from soccer, through chess, bridge, and poker, to solitaire. Instead, what organizes these exemplars into the game category is a set of family resemblances among these members – son may be like mother, and mother like sister, but in a very different way. And we learn about these families, like our own, from experience. Exemplars are similar if they have many features in common and few distinctive attributes (features belonging to one but not the other); the more similar are two objects on these quantitative grounds, the faster are people at judging them to be similar (Tversky 1977). Prototypes, exemplars which are most typical of a category, are those which are similar to many members of that category and not similar to members of other categories. Again, the operationalisation of this criterion predicts the speed of human categorization performance – people more quickly classify as birds sparrows (or other average sized, average colored, average beaked, average featured specimens) than they do birds with less common features or feature combinations like kiwis or penguins (Rosch and Mervis 1975; Rosch et al. 1976). Prototypes are judged faster and more accurately, even if they themselves have never been seen before – someone who has never seen a sparrow, yet who has experienced the rest of the run of the avian mill, will still be fast and accurate in judging it to be a bird (Posner and Keele 1970). Such effects make it very clear that although people don't go around consciously counting features, they nevertheless have very accurate knowledge of the underlying frequency distributions and their central tendencies. Cognitive theories of categorization and generalization show

how schematic constructions are abstracted over less schematic ones that are inferred inductively by the learner in acquisition (Lakoff 1987; Taylor 1998; Harnad 1987). So Psychology is committed to studying these implicit processes of cognition.

2. Frequency and Language Cognition

The last 50 years of Psycholinguistic research has demonstrated language processing to be exquisitely sensitive to usage frequency at all levels of language representation: phonology and phonotactics, reading, spelling, lexis, morphosyntax, formulaic language, language comprehension, grammaticality, sentence production, and syntax (Ellis 2002). Language knowledge involves statistical knowledge, so humans learn more easily and process more fluently high frequency forms and 'regular' patterns which are exemplified by many types and which have few competitors. Psycholinguistic perspectives thus hold that language learning is the implicit associative learning of representations that reflect the probabilities of occurrence of form-function mappings. Frequency is a key determinant of acquisition because 'rules' of language, at all levels of analysis from phonology, through syntax, to discourse, are structural regularities which emerge from learners' lifetime unconscious analysis of the distributional characteristics of the language input. In James' terms, learners have to FIGURE language out.

It is these ideas which underpin the last 30 years of investigations of language cognition using connectionist and statistical models (Christiansen & Chater, 2001; Elman, et al., 1996; Rumelhart & McClelland, 1986), the competition model of language learning and processing (Bates and MacWhinney 1987; MacWhinney 1987, 1997), the investigation of how frequency and repetition bring about form in language and how probabilistic knowledge drives language comprehension and production (Jurafsky and Martin 2000; Ellis 2002; Bybee and Hopper 2001; Jurafsky 2002; Bod, Hay, and Jannedy 2003; Ellis 2002; Hoey 2005), and the proper empirical investigations of the structure of language by means of corpus analysis exemplified in this volume. Corpus linguistics allows us to count the relevant frequencies in the input.

Frequency, learning, and language come together in Usage-based approaches which hold that we learn linguistic constructions while engaging in communication, the "interpersonal communicative and cognitive processes that everywhere and always shape language" (Slobin 1997). Constructions are form-meaning mappings, conventionalized in the speech community,

and entrenched as language knowledge in the learner's mind. They are the symbolic units of language relating the defining properties of their morphological, syntactic, and lexical form with particular semantic, pragmatic, and discourse functions (Croft and Cruise 2004; Robinson and Ellis 2008; Goldberg 2003, 2006; Croft 2001; Tomasello 2003; Bates and MacWhinney 1987; Goldberg 1995; Langacker 1987; Lakoff 1987; Bybee 2008). Goldberg's (2006) Construction Grammar argues that all grammatical phenomena can be understood as learned pairings of form (from morphemes, words, idioms, to partially lexically filled and fully general phrasal patterns) and their associated semantic or discourse functions: "the network of constructions captures our grammatical knowledge *in toto*, i.e. It's constructions all the way down" (Goldberg 2006, p. 18). Such beliefs, increasingly influential in the study of child language acquisition, have turned upside down generative assumptions of innate language acquisition devices, the continuity hypothesis, and top-down, rule-governed, processing, bringing back data-driven, emergent accounts of linguistic systematicities. Constructionist theories of child language acquisition use dense longitudinal corpora to chart the emergence of creative linguistic competence from children's analyses of the utterances in their usage history and from their abstraction of regularities within them (Tomasello 1998, 2003; Goldberg 2006, 1995, 2003). Children typically begin with phrases whose verbs are only conservatively extended to other structures. A common developmental sequence is from formula to low-scope slot-and-frame pattern, to creative construction.

3. Frequency and Second Language Cognition

What of second language acquisition (L2A)? Language learners, L1 and L2 both, share the goal of understanding language and how it works. Since they achieve this based upon their experience of language usage, there are many commonalities between first and second language acquisition that can be understood from corpus analyses of input and cognitive- and psycho- linguistic analyses of construction acquisition following associative and cognitive principles of learning and categorization. Therefore Usage-based approaches, Cognitive Linguistics, and Corpus Linguistics are increasingly influential in L2A research too (Ellis 1998, 2003; Ellis and Cadierno 2009; Collins and Ellis 2009; Robinson and Ellis 2008), albeit with the twist that since they have previously devoted considerable resources to the estimation of the characteristics of another language – the native tongue in which they have considerable fluency – L2 learners' computations and inductions are often affected by transfer, with L1-tuned expectations and

selective attention (Ellis 2006) blinding the acquisition system to aspects of the L2 sample, thus biasing their estimation from naturalistic usage and producing the limited attainment that is typical of adult L2A. Thus L2A is different from L1A in that it involves processes of construction and *re*construction

4. Construction Learning as Associative Learning from Usage

If constructions as form-function mappings are the units of language, then language acquisition involves inducing these associations from experience of language usage. Constructionist accounts of language acquisition thus involve the distributional analysis of the language stream and the parallel analysis of contingent perceptual activity, with abstract constructions being learned from the conspiracy of concrete exemplars of usage following statistical learning mechanisms (Christiansen and Chater 2001) relating input and learner cognition. Psychological analyses of the learning of constructions as form-meaning pairs is informed by the literature on the associative learning of cue-outcome contingencies where the usual determinants include: factors relating to the form such as frequency and salience; factors relating to the interpretation such as significance in the comprehension of the overall utterance, prototypicality, generality, and redundancy; factors relating to the contingency of form and function; and factors relating to learner attention, such as automaticity, transfer, overshadowing, and blocking (Ellis 2002, 2003, 2006, 2008). These various psycholinguistic factors conspire in the acquisition and use of any linguistic construction.

These determinants of learning can be usefully categorized into factors relating to (1) input frequency (type-token frequency, Zipfian distribution, recency), (2) form (salience and perception), (3) function (prototypicality of meaning, importance of form for message comprehension, redundancy), and (4) interactions between these (contingency of form-function mapping). The following subsections briefly consider each in turn, along with studies demonstrating their applicability:

4.1. Input frequency (construction frequency, type-token frequency, Zipfian distribution, recency)

4.1.1. Construction frequency

Frequency of exposure promotes learning. Ellis' (2002a) review illustrates how frequency effects the processing of phonology and phonotactics, reading, spelling, lexis, morphosyntax, formulaic language, language compre-

hension, grammaticality, sentence production, and syntax. That language users are sensitive to the input frequencies of these patterns entails that they must have registered their occurrence in processing. These frequency effects are thus compelling evidence for usage-based models of language acquisition which emphasize the role of input.

4.1.2. Type and token frequency

Token frequency counts how often a particular form appears in the input. Type frequency, on the other hand, refers to the number of distinct lexical items that can be substituted in a given slot in a construction, whether it is a word-level construction for inflection or a syntactic construction specifying the relation among words. For example, the "regular" English past tense *-ed* has a very high type frequency because it applies to thousands of different types of verbs, whereas the vowel change exemplified in *swam* and *rang* has much lower type frequency. The productivity of phonological, morphological, and syntactic patterns is a function of type rather than token frequency (Bybee and Hopper 2001). This is because: (a) the more lexical items that are heard in a certain position in a construction, the less likely it is that the construction is associated with a particular lexical item and the more likely it is that a general category is formed over the items that occur in that position; (b) the more items the category must cover, the more general are its criterial features and the more likely it is to extend to new items; and (c) high type frequency ensures that a construction is used frequently, thus strengthening its representational schema and making it more accessible for further use with new items (Bybee and Thompson 2000). In contrast, high token frequency promotes the entrenchment or conservation of irregular forms and idioms; the irregular forms only survive because they are high frequency. These findings support language's place at the center of cognitive research into human categorization, which also emphasizes the importance of type frequency in classification.

Such effects are extremely robust in the dynamics of language usage and structural evolution: (1) For token frequency, entrenchment, and protection from change, Pagel, Atkinson & Meade (2007) used a database of 200 fundamental vocabulary meanings in 87 Indo-European languages to calculate how quickly the different meanings evolved over time. Records of everyday speech in English, Spanish, Russian and Greek showed that high token-frequency words that were used more often in everyday language evolved more slowly. Across all 200 meanings, word token frequency of usage determined their rate of replacement over thousands of years, with the most commonly-used words, such as numbers, changing

very little. (2) For type and token frequency, and the effects of friends and enemies in the dynamics of productivity of patterns in language evolution, Lieberman, Michel, Jackson, Tang, and Nowak (2007) studied the regularization of English verbs over the past 1,200 years. English's proto-Germanic ancestor used an elaborate system of productive conjugations to signify past tense whereas Modern English makes much more productive use of the dental suffix, '-ed'. Lieberman at al. chart the emergence of this linguistic rule amidst the evolutionary decay of its exceptions. By tracking inflectional changes to 177 Old-English irregular verbs of which 145 remained irregular in Middle English and 98 are still irregular today, they showed how the rate of regularization depends on the frequency of word usage. The half-life of an irregular verb scales as the square root of its usage frequency: a verb that is 100 times less frequent regularizes 10 times as fast.

4.1.3. Zipfian distribution

Zipf's law states that in human language, the frequency of words decreases as a power function of their rank in the frequency table. If p_f is the proportion of words whose frequency in a given language sample is f, then $p_f \sim f^{-b}$, with $b \approx 1$. Zipf (1949) showed this scaling relation holds across a wide variety of language samples. Subsequent research has shown that many language events (e.g., frequencies of phoneme and letter strings, of words, of grammatical constructs, of formulaic phrases, etc.) across scales of analysis follow this law (Ferrer i Cancho and Solé 2001, 2003). It has strong empirical support as a linguistic universal, and, as I shall argue in the closing section of this chapter, its implications are profound for language structure, use, and acquisition. For present purposes, this section focuses upon acquisition.

In the early stages of learning categories from exemplars, acquisition is optimized by the introduction of an initial, low-variance sample centered upon prototypical exemplars (Elio and Anderson 1981, 1984). This low variance sample allows learners to get a fix on what will account for most of the category members. The bounds of the category are defined later by experience of the full breadth of exemplar types. Goldberg, Casenhiser & Sethuraman (2004) demonstrated that in samples of child language acquisition, for a variety of verb-argument constructions (VACs), there is a strong tendency for one single verb to occur with very high frequency in comparison to other verbs used, a profile which closely mirrors that of the mothers' speech to these children. In natural language, Zipf's law (Zipf 1935) describes how the highest frequency words account for the

most linguistic tokens. Goldberg et al. (2004) show that Zipf's law applies within VACs too, and they argue that this promotes acquisition: tokens of one particular verb account for the lion's share of instances of each particular argument frame; this pathbreaking verb also is the one with the prototypical meaning from which the construction is derived (see also Ninio 1999, 2006).

Ellis and Ferreira-Junior (2009, 2009) investigate effects upon naturalistic second language acquisition of type/token distributions in the islands comprising the linguistic form of English verb-argument constructions (VACs: VL verb locative, VOL verb object locative, VOO ditransitive) in the ESF corpus (Perdue, 1993). They show that in the naturalistic L2A of English, VAC verb type/token distribution in the input is Zipfian and learners first acquire the most frequent, prototypical and generic exemplar (e.g. *put* in VOL, *give* in VOO, etc.). Their work further illustrates how acquisition is affected by the frequency and frequency distribution of exemplars within each island of the construction (e.g. [Subj V Obj $Obl_{path/loc}$]), by their prototypicality, and, using a variety of psychological (Shanks 1995) and corpus linguistic association metrics (Gries and Stefanowitsch 2004; Stefanowitsch and Gries 2003), by their contingency of form-function mapping. Ellis and Larsen-Freeman (2009) describe computational (Emergent connectionist) serial-recurrent network models of these various factors as they play out in the emergence of constructions as generalized linguistic schema from their frequency distributions in the input.

This fundamental claim that Zipfian distributional properties of language usage helps to make language learnable has thus begun to be explored for these three verb argument constructions, at least. It remains an important corpus linguistic research agenda to explore its generality across the wide range of the constructicon.

4.1.4. Recency

Language processing also reflects recency effects. This phenomenon, known as priming, may be observed in phonology, conceptual representations, lexical choice, and syntax (Pickering and Ferreira 2008). Syntactic priming refers to the phenomenon of using a particular syntactic structure given prior exposure to the same structure. This behavior has been observed when speakers hear, speak, read or write sentences (Bock 1986; Pickering 2006; Pickering and Garrod 2006). For L2A, Gries and Wulff (2005) showed (i) that advanced L2 learners of English showed syntactic priming for ditransitive (e.g., *The racing driver showed the helpful mechanic*) and prepositional dative (e.g., *The racing driver showed the torn overall* ...)

argument structure constructions in a sentence completion task, (ii) that their semantic knowledge of argument structure constructions affected their grouping of sentences in a sorting task, and (iii) that their priming effects closely resembled those of native speakers of English in that they were very highly correlated with native speakers' verbal subcategorization preferences whilst completely uncorrelated with the subcategorization preferences of the German translation equivalents of these verbs. There is now a growing body of research demonstrating such L2 syntactic priming effects (McDonough 2006; McDonough and Mackey 2006; McDonough and Trofimovich 2008).

4.2. Form (salience and perception)

The general perceived strength of stimuli is commonly referred to as their salience. Low salience cues tend to be less readily learned. Ellis (2006, 2006) summarized the associative learning research demonstrating that selective attention, salience, expectation, and surprise are key elements in the analysis of all learning, animal and human alike. As the Rescorla-Wagner (1972) model encapsulates, the amount of learning induced from an experience of a cue-outcome association depends crucially upon the salience of the cue and the importance of the outcome.

Many grammatical meaning-form relationships, particularly those that are notoriously difficult for second language learners like grammatical particles and inflections such as the third person singular *-s* of English, are of low salience in the language stream. For example, some forms are more salient: '*today*' is a stronger psychophysical form in the input than is the morpheme '*-s*' marking 3rd person singular present tense, thus while both provide cues to present time, *today* is much more likely to be perceived, and *-s* can thus become overshadowed and blocked, making it difficult for second language learners of English to acquire (Ellis 2006, 2008; Goldschneider and DeKeyser 2001).

4.3. Function (prototypicality of meaning, importance of form for message comprehension, redundancy)

4.3.1. Prototypicality of meaning

Some members of categories are more typical of the category than others – they show the family resemblance more clearly. In the prototype theory of concepts (Rosch and Mervis 1975; Rosch et al. 1976), the prototype as an idealized central description is the best example of the category, appropriately summarizing the most representative attributes of a category.

As the typical instance of a category, it serves as the benchmark against which surrounding, less representative instances are classified. The greater the token frequency of an exemplar, the more it contributes to defining the category, and the greater the likelihood it will be considered the prototype. The best way to teach a concept is to show an example of it. So the best way to introduce a category is to show a prototypical example. Ellis & Ferreira-Junior (2009) show that the verbs that second language learners first used in particular VACs are prototypical and generic in function (*go* for VL, *put* for VOL, and *give* for VOO). The same has been shown for child language acquisition, where a small group of semantically general verbs, often referred to as *light verbs* (e.g., *go*, *do*, *make*, *come*) are learned early (Clark 1978; Ninio 1999; Pinker 1989). Ninio argues that, because most of their semantics consist of some schematic notion of transitivity with the addition of a minimum specific element, they are semantically suitable, salient, and frequent; hence, learners start transitive word combinations with these generic verbs. Thereafter, as Clark describes, "many uses of these verbs are replaced, as children get older, by more specific terms.... General purpose verbs, of course, continue to be used but become proportionately less frequent as children acquire more words for specific categories of actions" (p. 53).

4.3.2. Redundancy

The Rescorla-Wagner model (1972) also summarizes how redundant cues tend not to be acquired. Not only are many grammatical meaning-form relationships low in salience, but they can also be redundant in the understanding of the meaning of an utterance. For example, it is often unnecessary to interpret inflections marking grammatical meanings such as tense because they are usually accompanied by adverbs that indicate the temporal reference. Second language learners' reliance upon adverbial over inflectional cues to tense has been extensively documented in longitudinal studies of naturalistic acquisition (Dietrich, Klein, and Noyau 1995; Bardovi-Harlig 2000), training experiments (Ellis 2007; Ellis and Sagarra 2010), and studies of L2 language processing (Van Patten 2006; Ellis and Sagarra 2010).

4.4. Interactions between these (contingency of form-function mapping)

Psychological research into associative learning has long recognized that while frequency of form is important, so too is contingency of mapping (Shanks 1995). Consider how, in the learning of the category of birds, while eyes and wings are equally frequently experienced features in the

exemplars, it is wings which are distinctive in differentiating birds from other animals. Wings are important features to learning the category of birds because they are reliably associated with class membership, eyes are neither. Raw frequency of occurrence is less important than the contingency between cue and interpretation. Distinctiveness or reliability of form-function mapping is a driving force of all associative learning, to the degree that the field of its study has been known as 'contingency learning' since Rescorla (1968) showed that for classical conditioning, if one removed the contingency between the conditioned stimulus (CS) and the unconditioned (US), preserving the temporal pairing between CS and US but adding additional trials where the US appeared on its own, then animals did not develop a conditioned response to the CS. This result was a milestone in the development of learning theory because it implied that it was contingency, not temporal pairing, that generated conditioned responding. Contingency, and its associated aspects of predictive value, cue validity, information gain, and statistical association, have been at the core of learning theory ever since. It is central in psycholinguistic theories of language acquisition too (Ellis 2008; MacWhinney 1987; Ellis 2006, 2006; Gries and Wulff 2005), with the most developed account for second language acquisition being that of the Competition model (MacWhinney 1987, 1997, 2001). Ellis and Ferreira-Junior (2009) use ΔP and collostructional analysis measures (Gries and Stefanowitsch 2004; Stefanowitsch and Gries 2003) to investigate effects of form-function contingency upon L2 VAC acquisition. Wulff, Ellis, Römer, Bardovi-Harlig and LeBlanc (2009) use multiple distinctive collexeme analysis to investigate effects of reliability of form-function mapping in the second language acquisition of tense and aspect. Boyd and Goldberg (Boyd and Goldberg 2009) use conditional probabilities to investigate contingency effects in VAC acquisition. This is still an active area of inquiry, and more research is required before we know which statistical measures of form-function contingency are more predictive of acquisition and processing.

4.5. The Many Aspects of Frequency and their Research Consequences

This section has gathered a range of frequency-related factors that influence the acquisition of linguistic constructions:

- the frequency, the frequency distribution, and the salience of the form types,
- the frequency, the frequency distribution, the prototypicality and generality of the semantic types, their importance in interpreting the overall construction,

– the reliabilities of the mapping between 1 and 2,
– the degree to which the different elements in the construction sequence (such as the Subj V Obj and Obl islands in the archipelago of the VL verb argument construction) are mutually informative and form predictable chunks.

There are many factors involved, and research to date has tended to look at each hypothesis by hypothesis, variable by variable, one at a time. But they interact. And what we really want is a model of usage and its effects upon acquisition. We can measure these factors individually. But such counts are vague indicators of how the demands of human interaction affect the content and ongoing co-adaptation of discourse, how this is perceived and interpreted, how usage episodes are assimilated into the learner's system, and how the system reacts accordingly. We need theoretical models of learning, development, and emergence that takes these factors into account dynamically. I will return to this prospect in sections 7–8 after first considering some implications for instruction.

5. Language Learning as Estimation from Sample: Implications for Instruction

Language learners have limited experience of the target language. Their limited exposure poses them the task of estimating how linguistic constructions work from an input sample that is incomplete, uncertain, and noisy. Native-like fluency, idiomaticity, and selection are another level of difficulty again. For a good fit, every utterance has to be chosen, from a wide range of possible expressions, to be appropriate for that idea, for that speaker, for that place, and for that time. And again, learners can only *estimate* this from their finite experience.

Like other estimation problems, successful determination of the population characteristics is a matter of statistical sampling, description, and inference. There are three fundamental instructional aspects of this conception of language learning as statistical sampling and estimation, and Corpus Linguistics is central in each.

5.1. Sample Size

The first and foremost concerns sample size: As in all surveys, the bigger the sample, the more accurate the estimates, but also the greater the costs. Native speakers estimate their language over a lifespan of usage. L2 and

foreign language learners just don't have that much time or resource. Thus, they are faced with a task of optimizing their estimates of language from a limited sample of exposure.

Corpus linguistic analyses are essential to the determination of which constructions of differing degrees of schematicity are worthy of instruction, their relative frequency, and their best (=prototypical and most frequent) examples for instruction and assessment. Gries (2008) describes how three basic methods of corpus linguistics (frequency lists, concordances, and collocations) inform the instruction of second language constructions.

5.2. Sample Selection

Principles of survey design dictate that a sample must properly represent the strata of the population of greatest concern. Corpus linguistics, genre analysis, and needs analysis have a large role to play in identifying the linguistic constructions of most relevance to particular learners. For example, every genre of English for Academic Purposes and English for Special Purposes has its own phraseology, and learning to be effective in the genre involves learning this (Swales 1990). Lexicographers base their learner dictionaries upon relevant corpora, and these dictionaries focus upon examples as much as definitions, or even more so. Good grammars are now frequency informed. Corpus linguistic analysis techniques have been used to identify the words relevant to academic English (the Academic Word List, Coxhead 2000) and this, together with knowledge of lexical acquisition and cognition, informs vocabulary instruction programs (Nation 2001). Similarly, corpus techniques have been used to identify formulaic phrases that are of special relevance to academic discourse and to inform their instruction (the Academic Formulas List, Ellis, Simpson-Vlach, and Maynard 2008).

5.3. Sample Sequencing

Corpus linguistics also has a role to play in informing the ordering of exemplars for optimal acquisition of a schematic construction. The research reviewed above suggests that an initial, low-variance sample centered upon prototypical exemplars allows learners to get a 'fix' on the central tendency of a schematic construction, and then the introduction of more diverse exemplars facilitates learners to determine the full range and bounds of the category. Although, as explained in section 4.1.3, there is work to-be-done on determining its applicability to particular constructions, and particular learners and their L1s, in second language acquisition, this is

probably a generally useful instructional heuristic. Readings in Robinson and Ellis (2008) show how an understanding of the item-based nature of construction learning inspires the creation and evaluation of instructional tasks, materials, and syllabi, and how cognitive linguistic analyses can be used to inform learners how constructions are conventionalized ways of matching certain expressions to specific situations and to guide instructors in isolating and presenting the various conditions that motivate speaker choice.

6. Exploring what counts

Usage is rich in latent linguistic structure, thus frequencies of usage count in the emergence of linguistic constructions. Corpus linguistics provides the proper empirical means whereby everything in language texts can be counted. But, following the quotation from Einstein that opened this chapter, not everything that we can count in language counts in language cognition and acquisition. If it did, the English articles *the* and *a* alongside frequent morphological inflections would be among the first learned English constructions, rather than the most problematic in L2A.

The evidence gathered so far in this chapter shows clearly that the study of language from corpus linguistic perspectives is a two-limbed stool without triangulation from an understanding of the psychology of cognition, learning, attention, and development. Sensation is not perception, and the psychophysical relations mapping physical onto psychological scales are complex. The world of conscious experience is not the world itself but a perception crucially determined by attentional limitations, prior knowledge, and context. Not every experience is equal – effects of practice are greatest at early stages but eventually reach asymptote. The associative learning of constructions as form-meaning pairs is affected by: factors relating to the form such as frequency and salience; factors relating to the interpretation such as significance in the comprehension of the overall utterance, prototypicality, generality, and redundancy; factors relating to the contingency of form and function; and factors relating to learner attention, such as automaticity, transfer, and blocking.

We need models of usage and its effects upon acquisition. Univariate counts are vague indicators of how the demands of human interaction affect the content and ongoing co-adaptation of discourse, how this is perceived and interpreted, how usage episodes are assimilated into the learner's system, and how the linguistic system reacts accordingly. We

need models of learning, development, and emergence that take all these factors into account dynamically.

7. Emergentism and Complexity

Although the above conclusion is not contentious, the proper path to its solution is more debatable. In these final sections of my introductory review, I outline Emergentist and related approaches that I believe to be useful in guiding future research. Two key motivations of the editors of this volume are those of empirical rigor and interdisciplinarity. Emergentism fits well, I believe, as a general framework in that it is as quantitative as anything we have considered here so far, but more so in its recognition of multivariate, multi-agent, often non-linear, interactions.

Language usage involves agents and their processes at many levels, from neuron, through self, to society. We need to try to understand language emergence as a function of interactions within and between them. This is a tall order. Hence Saussure's observation that "to speak of a 'linguistic law' in general is like trying to lay hands on a ghost... Synchronic laws are general, but not imperative... [they] are imposed upon speakers by the constraints of common usage... In short, when one speaks of a synchronic law, one is speaking of an arrangement, or a principle of regularity" (Saussure 1916). Nevertheless, 100 years of subsequent work in psycholinguistics has put substantial flesh on the bone. And more recently, work within Emergentism, Complex Adaptive Systems (CAS), and Dynamic Systems Theory (DST) has started to describe a number of scale-free, domain-general processes which characterize the emergence of pattern across the physical, natural, and social world:

Emergentism and Complexity Theory (MacWhinney 1999; Ellis 1998; Elman et al. 1996; Larsen-Freeman 1997; Larsen-Freeman and Cameron 2008; Ellis and Larsen-Freeman 2009, 2006) analyze how complex patterns emerge from the interactions of many agents, how each emergent level cannot come into being except by involving the levels that lie below it, and how at each higher level there are new and emergent kinds of relatedness not found below: "More is different" (Anderson 1972). These approaches align well with DST which considers how cognitive, social and environmental factors are in continuous interactions, where flux and individual variation abound, and where cause-effect relationships are non-linear, multivariate and interactive in time (Ellis and Larsen-Freeman 2006, 2006; van Geert 1991; Port and Van Gelder 1995; Spivey 2006; de Bot, Lowie, and Verspoor 2007; Spencer, Thomas, and McClelland 2009; Ellis 2008).

"Emergentists believe that simple learning mechanisms, operating in and across the human systems for perception, motor-action and cognition as they are exposed to language data as part of a communicatively-rich human social environment by an organism eager to exploit the functionality of language, suffice to drive the emergence of complex language representations." (Ellis 1998, p. 657). Language cannot be understood in neurological or physical terms alone, nevertheless, neurobiology and physics play essential roles in the complex interrelations; equally from the top down, though language cannot be understood purely from introspection, nevertheless, conscious experience is an essential part too.

Language considered as a CAS of dynamic usage and its experience involves the following key features:

- The system consists of multiple agents (the speakers in the speech community) interacting with one another.
- The system is adaptive, that is, speakers' behavior is based on their past interactions, and current and past interactions together feed forward into future behavior.
- A speaker's behavior is the consequence of competing factors ranging from perceptual mechanics to social motivations.

The structures of language emerge from interrelated patterns of experience, social interaction, and cognitive processes.

The advantage of viewing language as a CAS is that it provides a unified account of seemingly unrelated linguistic phenomena (Holland 1998, 1995; Beckner et al. 2009). These phenomena include: variation at all levels of linguistic organization; the probabilistic nature of linguistic behavior; continuous change within agents and across speech communities; the emergence of grammatical regularities from the interaction of agents in language use; and stage-like transitions due to underlying nonlinear processes. Much of CAS research investigates these interactions through the use of computer simulations (Ellis and Larsen-Freeman 2009). One reason to be excited about a CAS/Corpus Linguistics synergy is that the scale-free phenomena that are characteristic of complex systems were indeed first identified in language corpora.

8. Zipf, Corpora, and Complex Adaptive Systems

Zipf's (1935) analyses of frequency patterns in linguistic corpora, however small they might seem in today's terms, allowed him to identify a scaling

law that was universal across language usage. He later attributed this law to the Principle of Least Effort, whereby natural languages realize effective communication by balancing speaker effort (optimized by having fewer words to be learned and accessed in speech production) and ambiguity of speech comprehension (minimized by having many words, one for each different meaning) (Zipf 1949). Many language events across scales of analysis follow his power law: phoneme and letter strings (Kello and Beltz 2009), words (Evert 2005), grammatical constructs (Ninio 2006; O'Donnell and Ellis 2010), formulaic phrases (O'Donnell and Ellis 2009), etc. Scale-free laws also pervade language structures, such as scale-free networks in collocation (Solé et al. 2005), in morphosyntactic productivity (Baayen 2008), in grammatical dependencies (Ferrer i Cancho and Solé 2001, 2003; Ferrer i Cancho, Solé, and Köhler 2004), and in networks of speakers, and language dynamics such as in speech perception and production, in language processing, in language acquisition, and in language change (Ninio 2006; Ellis 2008). Zipfian covering determines basic categorization, the structure of semantic classes, and the language form-semantic structure interface (Tennenbaum 2005; Manin 2008). Language structure and usage are inseparable, and scale-free laws pervade both. And not just language structure and use.

Power law behavior like this has since been shown to apply to a wide variety of structures, networks, and dynamic processes in physical, biological, technological, social, cognitive, and psychological systems of various kinds (e.g. magnitudes of earthquakes, sizes of meteor craters, populations of cities, citations of scientific papers, number of hits received by web sites, perceptual psychophysics, memory, categorization, etc.) (Newman 2005; Kello et al. 2010). It has become a hallmark of Complex Systems theory where so-called fat-tailed distributions characterize phenomena at the edge of chaos, at a self-organized criticality phase-transition point midway between stable and chaotic domains. The description and analysis of the way in which items (nodes) of different types are arranged into systems (networks) through the connections (edges) formed between them is the focus of the growing field of network science. The ubiquity and diversity of the systems best analyzed as networks, from the connection of proteins in yeast cells to the close association between two actors who have never been co-stars, has given the study of network typologies and dynamics a place alongside the study of other physical laws and properties (Albert and Barabasi 2002; Newman 2003). Properties of networks such as the 'small world' phenomenon (short path between any two nodes even in massive networks), scale-free degree distribution, and the notion of

'preferential attachment' (new nodes added to a network tend to connect to already highly-connected nodes) hold for networks of language events, structures, and users. Zipfian scale-free laws are universal. They are fundamental too, underlying language processing, learnability, acquisition, usage and change (Ferrer i Cancho and Solé 2001, 2003; Ferrer i Cancho, Solé, and Köhler 2004; Solé et al. 2005). Much remains to be understood, but this is a research area worthy of rich investment, where counting should really count.

Frequency is important to language. Systems depend upon regularity. But not only in the many simple ways. Regular as clockwork proves true in many areas of language representation, change, and processing, as this review has demonstrated. But more is different. In section 7, I argued that the study of language from corpus linguistic perspectives is a two-limbed stool without triangulation from an understanding of the psychology of cognition, learning, attention, and development. Even a three limbed stool does not make much sense without an appreciation of its social use. The cognitive neural networks that compute the associations binding linguistic constructions are embodied, attentionally- and socially- gated, conscious, dialogic, interactive, situated, and cultured (Ellis 2008; Beckner et al. 2009; Ellis and Larsen-Freeman 2009; Bergen and Chang 2003). Language usage, social roles, language learning, and conscious experience are all socially situated, negotiated, scaffolded, and guided. They emerge in the dynamic play of social intercourse. All these factors conspire dynamically in the acquisition and use of any linguistic construction. The future lies in trying to understand the component dynamic interactions at all levels, and the consequent emergence of the complex adaptive system of language itself.

References

Albert, Réka and Albert-László Barabasi
2002 Statistical mechanics of complex networks. *Rev. Mod. Phys* 74: 47–97.

Anderson, John Robert
1982 Acquisition of cognitive skill. *Psychological Review* 89(4): 369–406.

Anderson, John Robert
2000 *Cognitive psychology and its implications (5th ed.)*. New York: W.H. Freeman.

Anderson, Philip Warren
1972 More is different. *Science* 177: 393–396.

Baayen, R. Harald
2008 Corpus linguistics in morphology: morphological productivity. In *Corpus Linguistics. An international handbook*, edited by A. Ludeling and M. Kyto. Berlin: Mouton De Gruyter.
Baddeley, Alan D
1997 *Human memory: Theory and practice.* Revised ed. Hove: Psychology Press.
Bardovi-Harlig, Kathleen
2000 *Tense and aspect in second language acquisition: Form, meaning, and use.* Oxford: Blackwell.
Bartlett, Frederick Charles
[1932] 1967 *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.
Bates, Elizabeth and Brian MacWhinney
1987 Competition, variation, and language learning. In *Mechanisms of language acquisition*, edited by B. MacWhinney. Hillsdale, NJ: Lawrence Erlbaum Associates.
Beckner, Clay, Richard Blythe, Joan Bybee, Morton H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Thomas Schoenemann
2009 Language is a complex adaptive system. Position paper. *Language Learning* 59 Supplement 1: 1–26.
Bergen, Ben K. and Nancy C. Chang
2003 Embodied construction grammar in simulation-based language understanding. In *Construction Grammars: Cognitive grounding and theoretical extensions*, edited by J.-O. Östman and M. Fried. Amsterdam/Philadelphia: John Benjamins.
Bock, J. Kathryn
1986 Syntactic persistence in language production. *Cognitive Psychology* 18: 355–387.
Bod, Rens, Jennifer Hay and Stefanie Jannedy (eds.)
2003 *Probabilistic linguistics*. Cambridge, MA: MIT Press.
Boyd, Jeremy K. and Adele E. Goldberg
2009 Input effects within a constructionist framework. *Modern Language Journal* 93(2): 418–429.
Bybee, Joan
2008 Usage-based grammar and second language acquisition. In *Handbook of cognitive linguistics and second language acquisition*, edited by P. Robinson and N. C. Ellis. London: Routledge.
Bybee, Joan and Paul Hopper (eds.)
2001 *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.
Bybee, Joan and Sandra Thompson
2000 Three frequency effects in syntax. *Berkeley Linguistic Society* 23: 65–85.

Christiansen, Morten H. and Nick Chater (eds.)
2001 *Connectionist psycholinguistics*. Westport, CO: Ablex.
Clark, Eve V.
1978 Discovering what words can do. In *Papers from the parasession on the lexicon, Chicago Linguistics Society April 14–15, 1978*, edited by D. Farkas, W. M. Jacobsen and K. W. Todrys. Chicago: Chicago Linguistics Society.
Collins, Laura and Nick. C. Ellis
2009 Input and second language construction learning: frequency, form, and function. *Modern Language Journal* 93(2): Whole issue.
Coxhead, A.
2000 A new Academic Word List. *TESOL Quarterly* 34: 213–238.
Croft, William
2001 *Radical construction grammar: Syntactic theory in typological perspective.* Oxford: Oxford University Press.
Croft, William and D. Alan Cruise
2004 *Cognitive linguistics*. Cambridge: Cambridge University Press.
de Bot, Kees, Wander Lowie and Marjolyn Verspoor
2007 A dynamic systems theory to second language acquisition. *Bilingualism: Language and Cognition* 10: 7–21.
Dietrich, Rainer, Wolfgang Klein and Colette Noyau (eds.)
1995 *The acquisition of temporality in a second language*. Amsterdam: John Benjamins.
Ebbinghaus, Hermann
1885 *Memory: A contribution to experimental psychology*. Translated by H. A. R. C. E. B. (1913). New York: Teachers College, Columbia.
Elio, Renee and John R. Anderson
1981 The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning & Memory* 7(6): 397–417.
Elio, Renee and John R. Anderson
1984 The effects of information order and learning mode on schema abstraction. *Memory & Cognition* 12(1): 20–30.
Ellis, Nick C.
1998 Emergentism, connectionism and language learning. *Language Learning* 48(4): 631–664.
Ellis, Nick C.
2002 Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2): 143–188.
Ellis, Nick C.
2002 Reflections on frequency effects in language processing. *Studies in Second Language Acquisition* 24(2): 297–339.

Ellis, Nick C.
2003 Constructions, chunking, and connectionism: The emergence of second language structure. In *Handbook of second language acquisition*, edited by C. Doughty and M. H. Long. Oxford: Blackwell.

Ellis, Nick C.
2006 Language acquisition as rational contingency learning. *Applied Linguistics* 27(1): 1–24.

Ellis, Nick C.
2006 Selective attention and transfer phenomena in SLA: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics* 27(2): 1–31.

Ellis, Nick C.
2007 Blocking and learned attention in language acquisition. *CogSci 2007, Proceedings of the Twenty Ninth Cognitive Science Conference. Nashville, Tennesse, August 1–4, 2007.*

Ellis, Nick C.
2008 The dynamics of language use, language change, and first and second language acquisition. *Modern Language Journal* 41(3): 232–249.

Ellis, Nick C.
2008 The Psycholinguistics of the Interaction Hypothesis. In *Multiple perspectives on interaction in SLA: Second language research in honor of Susan M. Gass*, edited by A. Mackey and C. Polio. New York: Routledge.

Ellis, Nick C.
2008 Usage-based and form-focused language acquisition: The associative learning of constructions, learned-attention, and the limited L2 endstate. In *Handbook of cognitive linguistics and second language acquisition*, edited by P. Robinson and N. C. Ellis. London: Routledge.

Ellis, Nick C. and Teresa Cadierno
2009 Constructing a second language. *Annual Review of Cognitive Linguistics* 7 (Special section).

Ellis, Nick C. and Fernando Ferreira-Junior
2009 Construction Learning as a function of Frequency, Frequency Distribution, and function. *Modern Language Journal* 93: 370–386.

Ellis, Nick C. and Fernando Ferreira-Junior
2009 Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, 111–139.

Ellis, Nick C. and Diane Larsen-Freeman
2006 Language emergence: Implications for Applied Linguistics. *Applied Linguistics* 27(4): whole issue.

Ellis, Nick C. and Diane Larsen-Freeman
2006 Language emergence: Implications for Applied Linguistics (Introduction to the Special Issue). *Applied Linguistics* 27(4): 558–589.
Ellis, Nick C. and Diane Larsen-Freeman
2009 Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning* 59 (Supplement 1): 93–128.
Ellis, Nick C. and Diane Larsen-Freeman
2009 Language as a Complex Adaptive System (Special Issue). *Language Learning* 59: Supplement 1.
Ellis, Nick C. and Nuria Sagarra
2010 Learned Attention Effects in L2 Temporal Reference: The First Hour and the Next Eight Semesters. *Language Learning*, 60: Supplement 2, 85–108.
Ellis, Nick C. and Nuria Sagarra
2010 The bounds of adult language acquisition: Blocking and learned attention. *Studies in Second Language Acquisition*, 32(4), 553–580.
Ellis, Nick C. and Richard Schmidt
1998 Rules or associations in the acquisition of morphology? The frequency by regularity interaction in human and PDP learning of morphosyntax. *Language & Cognitive Processes* 13(2&3): 307–336.
Ellis, Nick C., Rita Simpson-Vlach and Carson Maynard
2008 Formulaic Language in Native and Second-Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly* 42(3): 375–396.
Elman, Jeffrey L., Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi and Kim Plunkett
1996 *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.
Evert, Stefan
2005 The Statistics of Word Cooccurrences: Word Pairs and Collocations, University of Stuttgart, Stuttgart.
Ferrer i Cancho, Ramon and Ricard V. Solé
2001 The small world of human language. *Proceedings of the Royal Society of London, B.* 268: 2261–2265.
Ferrer i Cancho, Ramon and Ricard V. Solé
2003 Least effort and the origins of scaling in human language. *PNAS* 100: 788–791.
Ferrer i Cancho, Ramon, Ricard V. Solé and Reinhard Köhler
2004 Patterns in syntactic dependency networks. *Physical Review* E69: 0519151–0519158.
Goldberg, Adele E.
1995 *Constructions: A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Goldberg, Adele E.
2003 Constructions: a new theoretical approach to language. *Trends in Cognitive Science* 7: 219–224.
Goldberg, Adele E.
2006 *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Goldberg, Adele E., Devin M. Casenhiser and Nitya Sethuraman
2004 Learning argument structure generalizations. *Cognitive Linguistics* 15: 289–316.
Goldschneider, Jennifer M. and Robert DeKeyser
2001 Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning* 51: 1–50.
Gries, Stefan Th.
2008 Corpus-based methods in analyses of SLA data. In *Handbook of cognitive linguistics and second language acquisition*, edited by P. Robinson and N. C. Ellis. London: Routledge.
Gries, Stefan Th. and Anatol Stefanowitsch
2004 Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9: 97–129.
Gries, Stefan Th. and Stefanie Wulff
2005 Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3: 182–200.
Harnad, Steven (ed.)
1987 *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press.
Hoey, Michael P.
2005 *Lexical priming: A new theory of words and language*. London: Routledge.
Holland, John H.
1995 *Hidden order: How adaption builds complexity*. Reading: Addison-Wesley.
Holland, John H.
1998 *Emergence: From chaos to order*. Oxford: Oxford University Press.
James, William
1890 *The principles of psychology*. Vol. 1. New York: Holt.
Jurafsky, Daniel
2002 Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In *Probabilistic Linguistics*, edited by R. Bod, J. Hay and S. Jannedy. Harvard, MA: MIT Press.
Jurafsky, Daniel and James H. Martin
2000 *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics.* Englewood Cliffs, NJ: Prentice-Hall.

Kello, Chris T. and Brandon C. Beltz
2009 Scale-free networks in phonological and orthographic wordform lexicons. In *Approaches to Phonological Complexity*, edited by I. Chitoran, C. Coupé, E. Marsico and F. Pellegrino. Berlin: Mouton de Gruyter.
Kello, Chris T., Gordon D. A. Brown, Ramon Ferrer-i-Cancho, John G. Holden, Klaus Linkenkaer-Hansen, Theo Rhodes and Guy C. Van Orden
2010 Scaling laws in cognitive sciences. *Trends in Cognitive Science* 14: 223–232.
Lakoff, George
1987 *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.
Langacker, Ronald W.
1987 *Foundations of cognitive grammar: Vol. 1. Theoretical prerequisites*. Stanford, CA: Stanford University Press.
Larsen-Freeman, Diane
1997 Chaos/complexity science and second language acquisition. *Applied Linguistics* 18: 141–165.
Larsen-Freeman, Diane and Lynne Cameron
2008 *Complex systems and applied linguistics*. Oxford: Oxford University Press.
Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang and Martin A. Nowak
2007 Quantifying the evolutionary dynamics of language. *Nature* 449(7163): 713–716.
MacWhinney, Brian
1987 Applying the Competition Model to bilingualism. *Applied Psycholinguistics* 8(4): 315–327.
MacWhinney, Brian
1987 The Competition Model. In *Mechanisms of language acquisition*, edited by B. MacWhinney. Hillsdale, NJ: Erlbaum.
MacWhinney, Brian
1997 Second language acquisition and the Competition Model. In *Tutorials in bilingualism: Psycholinguistic perspectives*, edited by A. M. B. De Groot and J. F. Kroll. Mahwah, NJ: Lawrence Erlbaum Associates.
MacWhinney, Brian
2001 The competition model: The input, the context, and the brain. In *Cognition and second language instruction*, edited by P. Robinson. New York: Cambridge University Press.
MacWhinney, Brian (ed.)
1999 *The emergence of language*. Hillsdale, NJ: Erlbaum.
Manin, Dmitrii Y.
2008 Zipf's Law and Avoidance of Excessive Synonymy. *Cognitive Science* 32: 1075–1098.

McDonough, Kim
2006 Interaction and syntactic priming: English L2 speakers' production of dative constructions. *Studies in Second Language Acquisition* 28: 179–207.
McDonough, Kim and Alison Mackey
2006 Responses to recasts: Repetitions, primed production and linguistic development. *Language Learning* 56: 693–720.
McDonough, Kim and Pavel Trofimovich
2008 *Using priming methods in second language research*. London: Routledge.
Nation, Paul
2001 *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Newell, Alan
1990 *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
Newman, Mark
2003 The structure and function of complex networks. *SIAM Review* 45: 167–256.
Newman, Mark
2005 Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46: 323–351.
Ninio, Anat
1999 Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language* 26: 619–653.
Ninio, Anat
2006 *Language and the learning curve: A new theory of syntactic development*. Oxford: Oxford University Press.
O'Donnell, Matthew B. and Nick C. Ellis
2009 Measuring formulaic language in corpora from the perspective of Language as a Complex System. In *5th Corpus Linguistics Conference*. University of Liverpool.
O'Donnell, Matthew B. and Nick C. Ellis
2010 Towards an Inventory of English Verb Argument Constructions. *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 9–16) Los Angeles, June 1–6, 2010.
Pagel, Mark, Quentin D. Atkinson and Andrew Meade
2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163): 717–720.
Pickering, Martin J.
2006 The dance of dialogue. *The Psychologist* 19: 734–737.
Pickering, Martin J. and Victor S. Ferreira
2008 Structural priming: A critical review. *Psychological Bulletin* 134: 427–459.

Pickering, Martin J. and Simon C. Garrod
2006 Alignment as the basis for successful communication. *Research on Language and Computation* 4: 203–228.
Pinker, Steven
1989 *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: Bradford Books.
Port, Robert F. and Timothy van Gelder
1995 *Mind as motion: Explorations in the dynamics of cognition.* Boston MA: MIT Press.
Posner, Michael and Steve W. Keele
1970 Retention of abstract ideas. *Journal of Experimental Psychology* 83: 304–308.
Rescorla, Robert A.
1968 Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology* 66: 1–5.
Rescorla, Robert A. and Allen R. Wagner
1972 A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: Current theory and research*, edited by A. H. Black and W. F. Prokasy. New York: Appleton-Century-Crofts.
Robinson, Peter and Nick C. Ellis (eds.)
2008 *A handbook of cognitive linguistics and second language acquisition*. London: Routledge.
Rosch, Eleanor and Carolyn B. Mervis
1975 Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104: 192–233.
Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson and Penny Boyes-Braem
1976 Basic objects in natural categories. *Cognitive Psychology* 8: 382–439.
Saussure, Ferdninand de
1916 *Cours de linguistique générale*. Translated by Roy Harris. London: Duckworth.
Shanks, David R.
1995 *The psychology of associative learning*. New York: Cambridge University Press.
Slobin, Dan I.
1997 The origins of grammaticizable notions: Beyond the individual mind. In *The crosslinguistic study of language acquisition*, edited by D. I. Slobin. Mahwah, NJ: Erlbaum.
Solé, Ricard V., Bernat Murtra, Sergi Valverde and Luc Steels
2005 Language Networks: their structure, function and evolution. *Trends in Cognitive Sciences* 12.
Spencer, John P., Michael S. C. Thomas and James L. McClelland (eds.)
2009 *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered*. New York: Oxford.

Spivey, Michael
2006 *The continuity of mind*. Oxford: Oxford University Press.
Stefanowitsch, Anatol and Stefan Th. Gries
2003 Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8: 209–43.
Swales, John M.
1990 *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
Taylor, John R.
1998 Syntactic constructions as prototype categories. In *The new psychology of language: Cognitive and functional approaches to language structure*, edited by M. Tomasello. Mahwah, NJ: Erlbaum.
Tenenbaum, Joshua
2005 The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science* 29: 41–78.
Tomasello, Michael
2003 *Constructing a language*. Boston, MA: Harvard University Press.
Tomasello, Michael (ed.)
1998 *The new psychology of language: Cognitive and functional approaches to language structure*. Mahwah, NJ: Erlbaum.
Tversky, Alan
1977 Features of similarity. *Psychological Review* 84: 327–352.
van Geert, P.
1991 A dynamic systems model of cognitive and language growth. *Psychological Review* 98(3–53).
Van Patten, Bill
2006 Input processing. In *Theories in second language acquisition: An introduction*, edited by B. Van Patten and J. Williams. Mahwah, NJ: Lawrence Erlbaum.
Wittgenstein, Ludwig
1953 *Philosophical investigations*. Translated by G. E. M. Anscombe. Oxford: Blackwell.
Wulff, Stefanie, Nick C. Ellis, Ute Römer, Kathleen Bardovi-Harlig and Chelsea LeBlanc
2009 The acquisition of tense-aspect: Converging evidence from corpora, cognition, and learner constructions. *Modern Language Journal* 93: 354–369.
Zipf, George K.
1935 *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: The M.I.T. Press.
Zipf, George K.
1949 *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish*

William D. Raymond and Esther L. Brown

The connection between frequency of form use and form reduction in language has been widely studied. After controlling for multiple contextual factors associated with reduction, word frequency, which reflects a speaker's cumulative experience with a word, has been reported to predict several types of pronunciation reduction. However, word frequency effects are not found consistently. Some studies have alternatively reported effects on reduction of the cumulative exposure of words to specific reducing environments or measures of contextual predictability. The current study examines cumulative and contextual effects of reducing environments, as well as non-contextual frequency measures, on the reduction of word-initial /s/ in a corpus of spoken New Mexican Spanish. The results show effects of non-cumulative factors on reduction, argued to occur on-line during articulation. There are also effects of the cumulative exposure of words to specific reducing environments and of contextual predictability, but not of the cumulative experience with a word overall (word frequency). The results suggest representational change in the lexicon through repeated exposure of words to reducing environments and call into question proposals that frequency of use *per se* causes reduction.

1. Introduction

Word frequency can be considered to reflect the relative cumulative experience that speakers have with words. The connection between word fre-

* We would like to thank Stefan Th. Gries for his help with the analysis of our data using R and acknowledge the constructive comments of Dr. Gries and three anonymous reviewers of this paper. We would also like to thank Lise Menn for useful comments on an earlier draft of the paper. This work was supported in part by Army Research Office Grant W911NF-05–1-0153 and Army Research Institute Contract DASW01–03-K-0002 to the University of Colorado.

quency and reduction in language has been widely studied, with the result that word frequency has been implicated in both diachronic change and synchronic production variation. Investigations of the processes of sound change in language going back over a century have noted that more frequent words are shorter and change more quickly than less frequent words (Schuchart 1885; Zipf 1929). In studies of synchronic pronunciation variation, evidence has been offered that higher word frequency is associated with more word reduction in speech production, as measured by both categorical measures of segment reduction or deletion (Bybee 2001, 2002; Krug 1998; Jurafsky et al. 2001; Raymond, Dautricourt, and Hume 2006) and also continuous measures of reduction, including durational shortening (Gahl 2008; Jurafsky et al. 2001; Pluymaekers, Ernestus, and Baayen 2005) and some acoustic parameters (Ernestus et al. 2006; Myers and Li 2007). Given prior results, it is widely assumed that frequency of word use contributes to reductive processes, although the mechanism by which it does so is unclear.

Word frequency is, of course, not the only correlate of reduction. Studies of word frequency effects on pronunciation variation have commonly controlled many factors that contribute to reductive phenomena, including lexical structure and class, extra-lexical phonological context, prosodic environment, speech rate, sociolinguistic factors, and even probabilistic variables other than word frequency. Even after controlling for multiple factors contributing to reduction, word frequency has usually been reported to predict pronunciation reduction by at least some measures; however, frequency effects are not ubiquitous. For example, Pluymaekers et al. (2005) found that word frequency affected reduction of affix form and duration for most but not all of the morphologically complex Dutch words they studied. Similarly, some of the high-frequency function words examined by Jurafsky et al. (2001) had low rates of reduction, despite their high frequency and a control for phonological context. Finally, Cohn et al. (2005) found no effect of word frequency on durational shortening of homophones, although Gahl (2008) did.

Failure to find effects consistently of word frequency on production variation, within studies or between comparable studies, has been attributed to methodological differences, such as sample size (see Gahl 2008) or to the set of factors considered in the study. Indeed, one class of factors that has not commonly been included in reduction studies in conjunction with word frequency is the likelihood that a word occurs in discourse contexts in a phonological environment that promotes reduction. The importance

of this type of cumulative contextual measure has been noted (Bybee 2001, 2002; Timberlake 1978). For example, the rate of word-final t/d deletion in English is lower for words that are more likely to occur in the context of a following vowel in speech (Guy 1991; Bybee 2002). Similarly, reduction rates of word-initial [s] in Spanish are higher for words that are more likely to occur in the context of a preceding non-high vowel in speech (Brown 2004, 2006). However, other probabilistic measures, such as word and phone frequencies and predictabilities, were not always controlled in these studies along with the cumulative reducing context variables. Conversely, studies finding frequency effects have not controlled the likelihood of a word occurring in a reducing environment. For example, Jurafsky et al. (2001) found effects of frequency on segment deletion and durational shortening of final t/d in content words, but their study did not control the likelihood of words occurring before following consonants, an extra-lexical environment promoting deletion. By comparison, in a study of word-internal t/d deletion, thus testing words for which phonological context of t/d is constant, Raymond et al. (2006) found no effect of frequency on deletion after controlling for predictability of the t/d word from the preceding and following words.

If word frequency plays a causal role in reductive processes, how might it affect reduction? In some usage-based theories of language (Bybee 2001, 2002) the effect of word frequency on sound change is explained as the result of automation of production processes. Production automation is claimed to result in more casual, more reduced forms, which will ultimately be registered as change in lexical representation. Automation is signaled by production speed, and frequent words can certainly be accessed more quickly than infrequent words (Balota et al. 2004; Forster and Chambers 1973). It could be that access speed has a direct effect on the articulation of words; however, an explanation based simply on how often a word is used would seem to entail that reductive change should occur uniformly across the word and not merely on certain segments or syllables, contrary to observations of lexical change (see Pluymaekers et al. 2005). The fact that reduction in frequent words is not uniform suggests there is an influence of lexical structure and discourse environments on reductive processes, leading to differential articulatory effects, automation processes, and, ultimately, reduction. Identifying any effect of word frequency on reduced pronunciation at articulation independent of reducing environments thus depends on controlling the factors leading to on-line articulatory reduc-

tion, cumulative measures of exposure to reductive environments, and measures of contextual predictability.

The current study examines the role of word frequency in reduction by examining a specific reduction phenomenon, word-initial [s] (s-) lenition, in the spoken Spanish of New Mexico. Many modern dialects of Spanish exhibit synchronic variation in production of [s] from full /s/ to [h] or even deletion (ø), either syllable finally (Terrell 1979; Lipski 1984; Brown, E. K. 2008; File-Muriel 2009) or syllable initially (Brown and Torres Cacoullos 2002). For syllable initial reduction the segmental context favoring reduction is a neighboring non-high vowel (/a, e, o/). Non-high vowels both preceding and following [s] have been found to increase the likelihood of reduction (Brown 2004, 2006), presumably because the non-high vowels' lower tongue height increases the likelihood that the alveolar target of [s] will be undershot. New Mexican Spanish is one dialect in which syllable-initial [s], including word-initial [s], may undergo lenition to [h] or even be deleted (e.g., *tuve que* [*h*]*alir*, for *tuve que salir*, "I had to leave"). In a study of s- reduction in this dialect, Brown (2006) found that the likelihood with which a word occurs in a non-high vowel environment predicts s- reduction. Interestingly, in her study the rate of reduction was also higher in words with high frequency than in words with low frequency, although other probabilistic measure were not controlled.

In the current study whether word frequency plays an independent role in on-line s- reduction is addressed by controlling both word frequency and frequency of occurrence of a word in phonological environments known to promote articulatory reduction of [s-]. The effects of other probabilistic measures are also assessed, to determine whether they contribute to s- reduction. Both intra- and extra-lexical phonological contexts are controlled, and comparison of their effects is used to determine to what extent reduction can be attributed to lexical representations or on-line articulatory processes.

2. Data and Methods

The data used in this study largely come from the materials of The New Mexico-Colorado Spanish Survey (NMCOSS) (Bills and Vigil 2008). The NMCOSS project, initiated in 1991, documents, via interviews with 350 native speakers, the Spanish language spoken throughout the state of New Mexico and sixteen counties of southern Colorado (Bills and Vigil 1999). The NMCOSS interview corpus was collected by trained field workers

who tape-recorded interviews involving both controlled elicitation and guided conversation (Vigil 1989). Each NMCOSS interview averaged three and a half hours in length, beginning with compilation of personal information regarding the consultant and followed by specific linguistic elicitation and free conversation. The interviews were subsequently orthographically transcribed.

The dataset for this study was created from the free conversation portions of the interviews of a subset of 16 men and 6 women selected at random from the NMCOSS study corpus, as well as two additional interviews with male native speakers of the same New Mexican dialect, for a total of 24 consultants. The data from one of the additional consultants was taken from an unplanned, self-recorded conversation. The data of the other additional consultant was extracted from a recorded conversation making up part of the Barelas study (conversational data of the Spanish spoken in the Barelas neighborhood of Albuquerque, NM), collected in recorded sociolinguistic interviews by students of a Spanish graduate course at the University of New Mexico in 2001. Although all consultants were native speakers of Spanish, most also had English proficiency, and there is a substantial amount of code switching and borrowing in the interviews. About 4% of the interview words were English words. The token dataset analyzed consists of all [s]-initial Spanish words (s- words) extracted from the set of words spoken by consultants in the conversations with the 24 consultants. The final dataset contained 2423 tokens (from 209 types) of [s]-initial Spanish word tokens. The phonetic realization of all /s/ phones in each token of these words was transcribed as perceptually reduced ([h] or ø) or unreduced ([s]) by one of the authors (EB), with reliability checks from native speakers.

The transcribed interview speech of the 22 NMCOSS consultants was used for frequency counts of phone and word units and bigrams. Both interviewer and consultant Spanish utterances in the recorded, transcribed conversations with these consultants were used to calculate unit frequencies. *Word unit* counts were compiled in five categories of units in the corpus subset: (1) whole Spanish word productions (*word*) (2) phrase boundaries (based on punctuation) and utterance (speaker) boundaries (*pause*); (3) partial word productions (*cutoff*); (4) hesitations and fillers (e.g., "uh"; *filler*); and (5) English words (*english*). All backchannel utterances (e.g., "oh" and "uh-huh") and sequences that could not be clearly understood during transcription (and were marked in the transcription as unclear) were excluded from the word unit counts. *Word bigram* counts were

made for all pairs of word units, producing statistics for each word that reflect how often it occurred after each other Spanish word, after a cutoff, after a pause, after a filler, and after an English word. *Phone unit* counts were calculated from word unit and word bigram counts for all phones in Spanish words. The phone unit counts were tallied separately for phones at word boundaries, phones at utterance boundaries, and all other phones within words. Phone bigram counts were tallied separately for phone pairs across word boundaries, phones adjacent to pauses and phone pairs within words. Phones adjacent to cutoffs, fillers, and English words were excluded from the phone and phone bigram counts. After exclusions, counts of word units and phones in the speech of the NMCOSS subset of speakers resulted in frequencies for about 75,000 words units and 280,000 phones of speech.

The word and phone counts from the speech of the corpus subset were used to create a database of word and phone statistics that includes the measures in Table 1.

Table 1. Measures included in the database of word and phone statistics

1. Word unit frequency per million of each word unit in the corpus subset (*word frequency*);
2. Word bigram frequency per million of each word unit and the word unit preceding it (*word bigram frequency*);
3. Predictability of each word unit from the word unit preceding it, calculated as the bigram frequency of the s- word divided by the frequency of the preceding word unit. (*preceding word predictability*, $P(w_i|w_{i-1})$);
4. Predictability of each word unit from the word unit following it (*following word predictability*, $P(w_i|w_{i+1})$);
5. Frequencies of all phone units in words (*phone frequency*);
6. Frequencies of all phone bigrams consisting of a word phone and the phone unit preceding it (*phone bigram frequency*);
7. Predictability of all phone units in words from the phone unit preceding it (*preceding phone predictability*, $P(\varphi_i|\varphi_{i-1})$);
8. Predictability of a phone unit from the phone unit following it (*following phone predictability*, $P(\varphi_i|\varphi_{i+1})$).

Using the interview transcriptions (with phonetic annotation of the realization of initial /s/ phones) and the word and phone statistics described in Table 1, the s- word tokens from the 24 consultants used for the study were coded for the ten variables in Table 2.

Table 2. Variables coded for each token in the s- word dataset

1. Realization of initial /s/ in the consultant's speech ([s] = *unreduced*; [h], ø = *reduced*);
2. Favorability of preceding phone context for s- reduction (*yes* for preceding non-high vowels, *no* for all other preceding phone units);
3. Frequency with which the phone preceding s- occurs before s- in the corpus, that is, the phone bigram frequency of s- (based on the word orthography, with phrase- and utterance-initial words coded as preceded by a pause);
4. Proportion of times in the corpus that the s- word has a preceding context favorable for fricative reduction, which is the proportion of tokens for an s- word type that are preceded by a non-high vowel (Frequency in a Favorable Context, or FFC);
5. Favorability of the phone following s- for s- reduction (*yes* for following non-high vowels, *no* for all other following phones);
6. Log of s- word frequency per million;
7. Identity of tokens of the very frequent clitic *se* used as a 3rd person singular reflexive pronoun, a 3rd person singular indirect object, and in impersonal constructions (*yes* for *se* tokens, *no* for all other tokens);
8. Hapax words in the corpus (*yes* for words with only a single token in the dataset, *no* for all words with more than one token);
9. Stress on s- syllable (*stressed* if lexical stress on primary syllable, *unstressed* if lexical stress on non-initial syllable or if the word is a clitic or function word);
10. Predictability of the s- word from the preceding word unit, $P(w_{s\text{-}}|w_{s\text{-}\,\text{-}1})$.

As an illustration of the measures in Table 2, consider the excerpt from the corpus transcription in (1). The s- word *sobrino* in the token in (1) occurs 11 times in the NMCOSS corpus, giving it a frequency per million of 146 and a log frequency of 2.17. The preceding word bigram in this token is *mi sobrino*, which has a frequency in the corpus statistics of 6, and the frequency of the word preceding the s- word, *mi*, is 485, so that the predictability of *sobrino* from *mi* is 6/485 = .0124. The s- of *sobrino* in this token is followed (word-internally) by the non-high vowel /o/, which is a context hypothesized to favor s- reduction. However, the vowel preceding s- is the high vowel /i/ in *mi*, which is hypothesized not to favor reduction. Overall in the corpus the word *sobrino* occurs after a non-high vowel (/o/, /a/, or /e/) only once, giving *sobrino* a FFC of 1/11 = .091. The frequency with which /i/ precedes /s/ at a word boundary in the corpus is 407, and the log of this frequency per million phones is 2.61.

(1) *. . . a mi sobrino, porque yo . . .*
. . . to my nephew, because I . . .

The variables 2–10 in Table 2 were used in regression analyses as predictors of the outcome variable, the binary variable 1 (phonological realization of s-), coding s- reduction. The predictor variables provide control over many of the factors that have been shown generally to influence reduction in speech. Specifically, reduction is higher in unstressed syllables than in stressed syllables (de Jong 1995). Reduction is also influenced by intra- and extra-lexical phonological contexts, which vary according to the variable under investigation (Brown 2006; Raymond, Dautricourt, and Hume 2006, Rhodes, R. A. 1992). In this study we consider only the phone context preceding s- words and the phone following s-. Non-high preceding vowels in both of these contexts have been shown to encourage s- reduction (Brown 2005). Effects of phone frequencies on s- reduction have not previously been investigated. More likely word and phone combinations are often associated with higher reduction (Jurafsky et al. 2001; Krug 1998). Note that frequency variables are skewed, with a few very high frequency tokens and many low frequency tokens. By taking the log of the frequency, this disparity is lessened (Gries 2009). The word *se* was chosen for identity coding because it is highly frequent (comprising 12.9% of the tokens), unstressed, and has a high reduction rate (.228). It is the only clitic form in the dataset, and clitics are known to behave differently from other words (Gerlach and Grijzenhout 2000). Hapax forms were coded specially because single occurrences of words in the limited speech sample of this corpus may not provide a reliable estimate of the cumulative experience of speakers with factors approximated using the probabilistic variables in Table 2, especially those variables that are calculated across extra-lexical phonological contexts.

The predictor variables for the study thus include four probabilistic measures: (1) s- word frequency; (2) the log frequency of the phone preceding the s-; (3) the predictability of the s- word from the word unit preceding it in context; and (4) the proportion of times that the s- word is preceded in production contexts by a non-high vowel (FFC). The first two measures are simple frequencies, whereas the last two are predictability measures. In addition, FFC and the frequency of the preceding phone consider the preceding phone as the context, whereas $P(w_{s\text{-}}|w_{s\text{-}\,\text{-}1})$ uses as context the preceding word. These variables thus allow us to look separately at the effects of the type of probabilistic measure (frequency or predictability)

Table 3. Independent variables categorized by type of probabilistic measure and context unit size

	Type of probabilistic measure	
Context unit size	*Frequency*	*Predictability*
Word	Word frequency	$P(w_{s-}\|w_{s- -1})$
Phone	Frequency of preceding phone	FFC

and context unit size (word or phone), as summarized in Table 3. Effects of word and phone unit probabilities on reductive processes have been found in previous studies, and FFC, which is a specific measure of exposure of words to a reductive environment, has also been implicated in s- reduction. However, these three measures have not previously been considered together in a study of reduction.

In addition to assessing the effects of frequencies and predictabilities, a second goal of the current study is to examine the contributions to s-reduction of factors that reflect cumulative experience with a word (including word frequency) and those that do not, as well as the effects of variables that refer to the extra-lexical (preceding) production context and those that are purely word internal, and thus are not context dependent. The independent variables chosen can be categorized in terms of these two dimensions, as shown in Table 4. Effects of the variables reflecting cumulative experience are taken to indicate a lexical source of reduction; effects of variables that are context dependent largely reflect an influence of articulation on reduction, either during production or as registered in lexical representation.

Note that other factors that have been shown to influence reduction in speech are not examined in this study, in particular speech rate (Fossler-Lussier and Moran 1999; Jurafsky et al. 2001), syntactic probabilities (Gahl and Garnsey 2004), and predictability from semantic or broader discourse context (Bard et al. 2000; Fowler and Housum 1987). Although these factors may also predict s- reduction, they seem unlikely exclusively to explain any effects of the articulatory and probabilistic variables with which we are concerned in our analyses. Moreover, the focus of the current study is to examine the scope and nature of variables implicated specifically in s- reduction, which can be assessed with the variables chosen for analysis.

Table 4. Independent variables categorized by cumulative experience and context dependence

	Cumulative experience	
Context dependence	*Yes*	*No*
Yes	FFC, $P(w_{s\text{-}}\|w_{s\text{-}\,-1})$, Frequency of preceding phone	Preceding favorable context
No	Word frequency	Stress, Following favorable context

3. Results

The data were analyzed with the R statistical package using logistic regression, with realization of s- as reduced or unreduced as the dependent variable. Analyses were performed on all s- data, as well as on some subsets of s- data, in order to examine the effects of one variable on a particular subcategory of s- words in one case, as described later in this section.

The model likelihood ratio for the model identified in the analysis was 341.64 (d.f. = 17; $p < .0001$). Using the mean of the predicted probabilities as the cutoff, classification accuracy for the model was 83.6%. However, the overall correlation for the model was not high (Nagelkerke $r^2 = .223$). The proportion of reduced tokens is low (.164), so that the null model that assumes no reduction would have comparable accuracy, although little explanatory power for the phenomenon. Additional accuracy would perhaps be achieved with the inclusion of other factors associated with reduction that were not included in the current analysis, especially speech rate.

The results of the analysis of the complete dataset are shown in Table 5, along with the odds ratios for the significant predictors. All measures 2–10 of Table 2 were used in this analysis except the log of the preceding phone frequency, because only three phones can precede s- in tokens preceded by a favorable reducing environment (i.e., the non-high vowels /e, a, o/, all of which have high frequencies), making the continuity of the variable in this environment questionable.

As shown in the table, there was a main effect of both the preceding and following phonological contexts of s-, with non-high vowels predicting higher reduction rates in both environments. When an s-word is preceded by a non-high vowel, it is 2.41 times more likely to be reduced

Table 5. Results of analysis of the complete dataset (N = 2423).

	p	Odds Ratio Effect
Preceding favorable context = *yes*	0.0302	2.41
Following favorable context = *yes*	<.0001	3.03
Syllable stress = *yes*	0.0235	1.59
se = *yes*	0.0019	4.55
$P(w_{s\text{-}}\|w_{s\text{-}\ -1})$	0.0183	1.01
FFC	0.0457	2.33
Preceding favorable context X *se*	0.0310	N.A.
FFC X Log word frequency	0.0036	N.A.

than when it is preceded by a high vowel, a consonant, or a pause. Similarly, when an s- is followed in a word by a non-high vowel, it is 3.03 times more likely to be reduced than when it is followed by a high vowel or a glide (/i, u, j, w/). There was also a main effect of the stress variable on reduction in the complete dataset, with no lexical stress on the initial syllable of an s- word making it 1.59 times more likely to be reduced than if the initial syllable has stress. After controlling for phonological context, there was a main effect of the cumulative contextual variable FFC, reflecting the fact that words in the highest quartile of FFC were 2.33 times more likely to be reduced than words in the lowest quartile. There was also an effect of the cumulative variable word predictability on reduction. Although significant, the effect of s- word predictability from the preceding word was very small, with words in the highest quartile of predictability only 1% more likely to be reduced than words in the lowest quartile of predictability. In addition there was a main effect of *se* word identity, with *se* 4.55 times more likely to be reduced than other words, even after controlling factors of *se* that contribute to its reduction. There was no main effect in the dataset on reduction of the non-contextual cumulative variable word frequency or of hapax words.

There were also two significant interactions in the analysis. The first interaction involved preceding favorable context and *se* word identity. As shown in Figure 1, the phonological context preceding an s- word has a greater effect on reduction for *se* than for other s- words. The interaction suggests a strong influence of articulatory environment on reduction of this clitic.

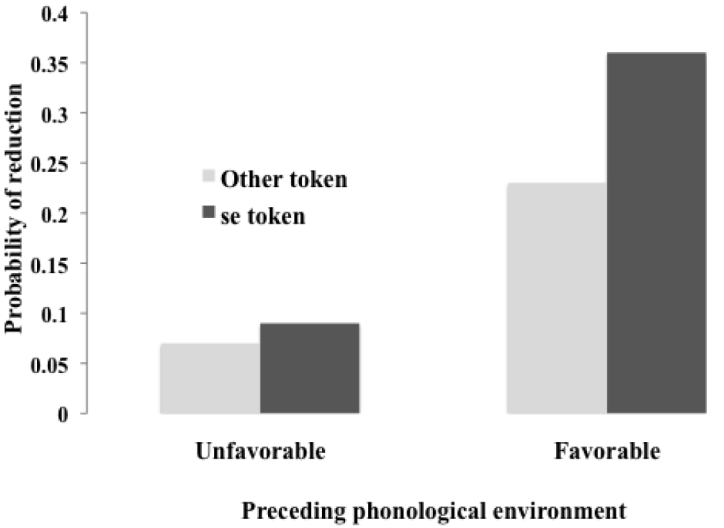


Figure 1. Interaction of preceding favorable context and *se* word identity on reduction of s-

The second interaction involved word frequency and FFC. Note that FFC and word frequency are related measures, because FFC is defined as the number of tokens of a word occurring in reducing environments out of the total number of tokens of the word in the dataset, or its frequency of occurrence. An interaction of FFC and word frequency is suggested by the fact that there is a correlation between FFC and reduction rate for high frequency words ($r = .36$), but not low frequency words ($r = .12$), as shown in Figure 2.

To test the effect of preceding phone frequency on reduction of s-, an analysis was done of a subset of the data that included only tokens that were not preceded by a non-high vowel, that is, the set of tokens not in a discourse environment favorable for reduction. Tokens in an environment favorable to reduction were not included in the analysis because there are only three non-high vowels that comprise a preceding favorable environment. The number of tokens in the reduced dataset was 1177, of which only 7.2% were reduced.

In a logistic regression on this subset of the data including all measures 2–10 of Table 2 92.9% of the tokens were correctly categorized by the model. The analysis showed an effect of a following favorable environment ($p = .0054$, odds ratio effect = 2.46) and a small effect of stress ($p = .0272$,

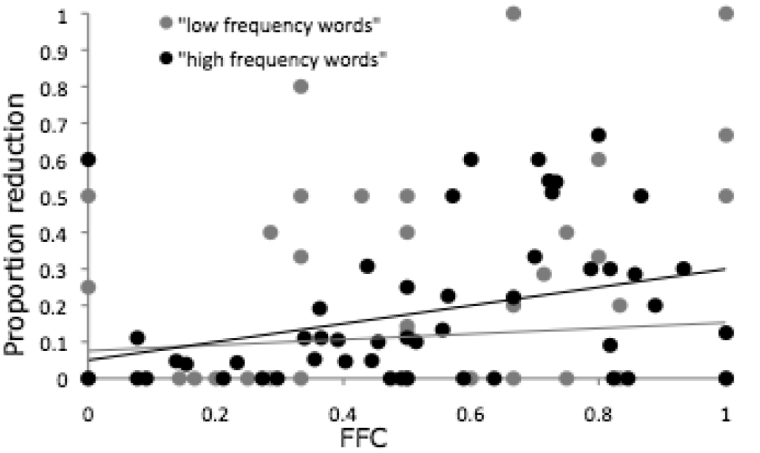


Figure 2. Interaction of FFC and reduction for high and low frequency s- words

odds ratio effect = 1.05). There was also a significant interaction of frequency of preceding phone and predictability of the s- word from the preceding word; however, the interaction was complex and will not be interpreted here. There was no significant effect of preceding phone frequency on reduction.

4. Discussion

The results for the complete dataset show effects on reduction of both extra- and intra-lexical factors, as well as variables that reflect speakers' cumulative experience with words and those that are a function of the context in which words are produced. However, not all variables in the analyses significantly predicted s- reduction, and there were interactions among some variables. The pattern of results allows us to draw conclusions about the sources of reductive influences, and, specifically, the role of word frequency in s- reduction.

There were strong effects on s- reduction of the non-cumulative intra- and extra-lexical variables involving phonological contexts. The tendency for non-high vowels that precede s- in the discourse context or that follow s- in a word's lexical form to encourage reduction suggests an on-line articulatory effect of phonological context. Tongue lowering during the articulation of /s/ before and/or after articulation of a non-high vowel is

a likely explanation for why the /s/ target was sometimes not achieved in these contexts. In the context of high vowels, on the other hand, tongue height in the vowel would facilitate /s/ articulation. Articulations of /s/ were also more likely to be reduced in unstressed syllables (de Jong 1995), which can account for the greater reduction rate for /s/ in lexically unstressed syllables over lexically stressed syllables.

There were also effects of cumulative context variables. The effect of the predictability of the s- word from the preceding word was small. However, the effect of FFC was robust and confirms earlier findings that FFC encourages reduction in studies that did not control other probabilistic factors (Brown 2004, 2006). The effect of FFC indicates that the cumulative experience of words in reducing phonological contexts of non-high preceding vowels results in a greater likelihood of reduction than context of use alone can explain. The effect suggests that reduction of s- reflects changes in the lexical representations of words through cumulative experience with these words in reductive production contexts. Because speakers have more limited experience with low frequency words than with high frequency words, the low frequency words show less of an effect of cumulative experience on their representations than do high frequency words.

The influence of factors defined by cumulative phonological context on s- reduction is compatible with an incipient process of lexicalized reduction of s- to [h] or ø. Lexical changes have not resulted in deterministic allophonic alternation in New Mexican Spanish: Most s- words continue to exhibit production variation in the data examined. The continued cumulative effects of discourse context and lexical structure on individual words may eventually result in a lexical distribution of phone variants, as happened, starting in Medieval Spanish, with words derived from Latin words beginning with /fV/ (FV- words). In Modern Spanish some FV- words begin with [f] (e.g., *fe* "faith" from L. FIDES and *fácil* "easy" from L. FACILIS) and some have an empty onset ([ø], spelled *h*) (e.g., *hablar* "to talk, to speak" from L. FABULARE and *hijo* "son" from L. FILIUS). Brown and Raymond (2010) have shown that the distribution of /f/ ~ [ø] in FV- words in Spanish is predicted by, among other variables, the likelihood that the words occur after a non-high vowel, the FFC variable also shown to predict the reduction of /s/ in s- words in the current synchronic study.

After taking into account the effects on s- reduction of non-cumulative phonological variables and the predictability variables involving word and phone context, there was no influence on reduction in the complete dataset or the subsets tested of preceding phone frequency or s- word frequency.

The failure to find any robust effects of the non-contextual word and phone unit probabilities after controlling the contextual variables suggests that speakers are sensitive to how often a word occurs in environments that encourage reduction, but not measurably to non-contextual probabilistic measures of use. Consequently, an s- word's frequency did not predict /s-/ reduction.

How can the failure to find a significant effect of word frequency on s-reduction in datasets analyzed be reconciled with other studies, in which word frequency effects on a range of reductive processes have been reported? As noted, in most of these studies the likelihood of a word occurring in a reducing environment was not controlled. With respect to phonological context, the environments promoting reduction are generally identifiable, and tests of their importance could be readily made. However, other variables not examined in this study may also promote reduction differentially across the word frequency range. For example, higher rates of speech are associated with reduction, and words may differ in their likelihood of being produced at high speech rates. In addition to a direct effect of speech rate on reduction, higher frequency words may, in particular, be more likely to be produced in contexts with higher speech rates than lower frequency words. Because faster speech rates may encourage reduction, high frequency words would thus have a higher probability of occurring in this reducing environment. In support of this possibility, Gahl (2008) found that, after controlling other factors influencing word duration, a higher speaking rate in the region following target words that are members of homophone pairs predicted shorter durations of the higher frequency member of the pair. Word frequency remained a significant predictor in the Gahl model; however, note that the likelihood of a word occurring in a high speech rate region, a context promoting reduction, was not controlled. It remains to be tested whether variability in the likelihood that words occur in reducing environments defined by non-phonological variables such as speech rate can also predict reduction and eliminate word frequency effects.

References

Balota, David. A., Cortese, Michael J., Sergent-Marshall, Susan D., Spieler, Daniel H. and Yap, Melvin J.
2004 Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133, 382–316.

Bard, Ellen G., Anderson, Anne H., Sotillo, Catherine, Aylett, Matthew, Doherty-Sneddon, Gwyneth and Newlands, Alison
2000 Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, 42, 1–22.

Bills, Garland and Vigil, Neddy
1999 Ashes to ashes: The historical basis for dialect variation in New Mexican Spanish. *Romance Philology*, 53, 43–67.

Bills, Garland and Vigil, Neddy
2008 *The Spanish Language of New Mexico and Southern Colorado: A Linguistic Atlas*. Albuquerque: University of New Mexico.

Brown, Earl K.
2008 A Usage-based Account of Syllable- and Word-final /s/ Reduction in Four Dialects of Spanish. Doctoral Dissertation, University of New Mexico.

Brown, Esther L.
2004 Reduction of syllable-initial /s/ in the Spanish of New Mexico and Southern Colorado: A usage-based approach. Doctoral Dissertation, University of New Mexico.

Brown, Esther L.
2005 Syllable-initial /s/ in Traditional New Mexican Spanish: Linguistic factors favoring reduction 'ahina'. *Southwest Journal of Linguistics*, 24, 1–2, 13–30.

Brown, Esther L.
2006 The effects of discourse context on phonological representation in memory. Paper presented at the 80th Annual Meeting of the Linguistic Society of America (LSA), Albuquerque, NM. January 5–8.

Brown, Esther L. and Raymond, William D.
In press How discourse context shapes the lexicon: Explaining the distribution of Spanish f- /h- words. *Diachronica* 20.2.

Brown, Esther L. and Torres Cacoullos, Rena
2002 ¿Qué le vamoh aher?: Taking the syllable out of Spanish /s/ reduction. In D. Johnson and T. Sanchez (Eds.), *University of Pennsylvania Working Papers in Linguistics: Papers from NWAV 30* (pp. 17–32). Philadelphia: University of Pennsylvania Press.

Bybee, Joan B.
2001 *Phonology and Language Use*. Cambridge, UK: Cambridge University Press.

Bybee, Joan B.
2002 Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14, 261–290.

Cohn, Abigail C., Brugman, Johanna, Crawford, Clifford and Joseph, Andrew
2005 Lexical frequency effects and phonetic duration of English homophones: An acoustic study. *Journal of the Acoustical Society of America*, 118, 2036.

De Jong, Kenneth
1995 The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustic Society of America*, 97, 491–504.
Ernestus, Mirjam, Lahey, Mybeth, Verhees, Femke and Baayen, R. Harald
2006 Lexical frequency and voice assimilation. *Journal of the Acoustical Society of America*, 120, 1040–1051.
File-Muriel, Richard
2009 The role of lexical frequency in the weakening of syllable-final lexical /s/ in the Spanish of Barranquilla, Colombia. *Hispania*, 92, 348–360.
Forster, Kenneth I. and Chambers, Susan M.
1973 Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12, 627–635.
Fosler-Lussier, Eric and Morgan, Nelson
1999 Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29, 137–158.
Fowler, Carol and Housum, Jonathan
1987 Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26, 489–504.
Gahl, Susanne
2008 *Time* and *thyme* are not homophones: The effect of lemma frequency on word duration in spontaneous speech. *Language*, 84, 474–496.
Gahl, Susanne and Garnsey, Susan M.
2004 Knowledge of grammar, knowledge of usage: syntactic probabilities affect pronunciation variation. *Language*, 80, 748–775.
Gerlach, Birgit and Grijzenhout, Janet
2000 Clitics from different perspectives. In Gerlach and Grijzenhout (Eds), *Clitics in Phonology, Morphology and Syntax*. Amsterdam: John Benjamins.
Gries, Stefan Th.
2009 *Quantitative corpus Linguistics with r: A practical introduction.* New York: Routledge.
Guy, Gregory R.
1991 Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change*, 3, 1–22.
Jurafsky, Daniel, Bell, Alan, Gregory, Michelle and Raymond, William D.
2001 Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee and P. Hopper (Eds.), Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins, pp. 229–254.

Krug, Manfred
1998 String frequency: A cognitive motivating factor in coalescence, language processing, and language change. *Journal of English Linguistics*, 26, 286–320.
Lipski, John
1984 On the weakening of /s/ in Latin American Spanish. *Zeitschrift für Dialektologie und Linguistik*, 51, 31–43.
Meyers, James and Li, Yingshing
2007 Lexical frequency effects in Taiwan Southern Min syllable contraction. *Journal of Phonetics*, 37, 21–230.
Pluymaekers, Mark, Ernestus, Mirjam and Baayen, R. Harald
2005 Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustic Society of America*, 118, 2561–2569.
Raymond, William D., Dautricourt, Robin and Hume, Elizabeth
2006 Word-medial /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change*, 18, 55–97.
Rhodes, Richard A.
1992 Flapping in American English. In W. U. Dressler, M. Prinzhorn and J. Rennison Eds.), *Proceedings of the 7th International Phonology Meeting*, pp. 217–232. Rosenberg and Sellier.
Schuchart, Hugo
1885 *Über die Lautgesetze: gegen die Junggrammatiker*. Berlin: R. Oppenheim.
Terrell, Tracy D.
1979 Final /s/ in Cuban Spanish. *Hispania*, 62, 599–612.
Timberlake, Alan
1978 Uniform and alternating environments in phonological change. *Folia Slavica*, 2, 312–28.
Vigil, Neddy A.
1989 Database for a linguistic atlas of the Spanish of New Mexico and southern Colorado. *Computer Methods in Dialectology. Journal of English Linguistics Special Issue*, 22, 69–75.
Zipf, George K.
1929 Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, 15, 1–95.

What statistics do learners track? Rules, constraints and schemas in (artificial) grammar learning*

Vsevolod Kapatsinski

Rule-based grammatical theories hypothesize that learners of morphophonology pay most attention to typical characteristics of mappings between cells in a morphological paradigm, which can be expressed in rules, rather than to typical characteristics of forms belonging to an individual cell. Bybee (2001) makes the opposite suggestion. The present paper reports data from miniature artificial language learning in the lab suggesting that reliance on product-oriented vs. source-oriented generalizations may depend on the presentation conditions. However, Bybee's position is supported even for presentation conditions that were designed to be maximally favorable for extracting rules.

1. Introduction

1.1. Theoretical background

All theories of grammar specify the types of generalizations that a human language user relies on in using language productively and thus restrict the human language learner to pay attention only to certain types of patterns in the data to which s/he is exposed. For instance, Chomsky and Halle (1968) and Albright and Hayes (2003), among others, assume reliance on rules. By contrast, Bybee (2001: 128) writes: "[R]ules express source-oriented generalizations. That is, they act on a specific input to change it in well-defined ways into an output of a certain form. Many, if not all, schemas are product-oriented rather than source-oriented. A product-

* Part of this research was supported by funds from NIDCD Research Grant DC-00111 and NIDCD T32Traning Grant DC-00012 to David Pisoni and the Speech Research Laboratory at Indiana University. Many thanks to Luis Hernandez for technical assistance and to Kenneth deJong, David Pisoni, Robert Port, Linda Smith and the anonymous reviewers for helpful comments.

oriented schema generalizes over forms of a specific category, but does not specify how to derive that category from some other."

The importance of product-oriented generalizations was first influentially pointed out by Kisseberth (1970), who noted that rules conspire to produce certain types of outputs and avoid others. This observation led to a paradigm shift in phonology from generative rules (Chomsky and Halle 1968) towards Optimality Theory (Prince and Smolensky 1993/2004), which allows explicit encoding of a particular kind of product-oriented generalizations (markedness constraints) in the grammar.

By specifying the types of generalizations that can be part of a human language learner's grammar, theories of grammar propose the existence of a hard formal bias on learning, which predisposes the learner to acquire specific types of generalizations and not to acquire others (or, perhaps, to rely on only some types of generalizations that have been acquired in using the language productively). While some theories (in particular, Optimality Theory) constrain the learner even further by endowing him/her with an innate set of generalizations, even theories that do not make this claim (e.g., Bybee 1985, 2001, or Albright and Hayes 2003) assume that the grammar contains only certain types of generalizations. Optimality Theory (along with rules-plus-constraints approaches like Blevins 1997, Paradis 1989, and Roca 1997) assumes that the learner relies on a system combining both product-oriented generalizations (markedness constraints) and source-oriented generalizations (faithfulness or paradigm uniformity constraints). Network Theory (Bybee 1985, 2001), as the quote above indicates, raises the possibility of a completely product-oriented grammar. Finally, the Minimal Generalization Learner (Albright and Hayes 2003) learns only source-oriented generalizations.

1.2. Prior empirical work

The present article tests whether (adult) learners have a bias in favor of either source-oriented or product-oriented generalizations. Much of the prior experimental evidence for product-oriented generalizations is summarized in Bybee (2001: 126–129). In most previous studies (Bybee and Slobin 1982, Bybee and Moder 1983, Köpcke 1988, Lobben 1991, Wang and Derwing 1994, Albright and Hayes 2003), the argument for product-oriented generalizations rests on finding that instead of respecting the input-output mappings present in the lexicon, subjects 'overuse' common output patterns deriving them in ways not attested in the lexicon. Unfortunately, the overuse can also be explained by experiment-internal response priming (cf. also Bickel et al. 2007, Caballero 2010).

More evidence for product-oriented generalizations in natural languages is provided by cases of echolalia, in which a morpheme is not attached to a form if the form sounds like it already has the morpheme (Menn and MacWhinney 1984, Stemberger 1981, Bybee 2001: 128). For instance, *It was thundering and lightning*, not *It was thundering and *lightninging.* Here speakers of English appear to be using the generalization that the progressive should end in -ing, not that one should add -ing to form the progressive. The stability of the no-change class of English verbs and its apparent resistance to overgeneralization is another possible example of this phenomenon (Menn and McWhinney 1984, Stemberger 1981, Bybee 2001: 128). Phonological factors and checking of the output after the application of the -ing-adding rule (Pinker 1999: 61–62) are possible alternative explanations.

Finally, evidence in favor of product-oriented generalizations is provided by Becker and Fainleib (2009) who report a miniature artificial language experiment with native Hebrew speakers. Hebrew prefers to attach the plural -ot, rather than -im, to singulars whose last vowel is [o] but -im to singulars whose final vowel is [i], resulting in oCot and iCim plurals. Becker and Fainleib exposed Hebrew speakers to one of two artificial languages, which the subjects were told were "new kinds of Hebrew". In the "surface" pseudo-Hebrew, iC-final singulars corresponded to oCot-final plurals, while oC-final singulars corresponded to iCim-final plurals. Thus, the language users could transfer product-oriented generalizations from their native language into the artificial language. In the "deep" pseudo-Hebrew, iC-final singulars corresponded to oCim-final plurals, while oC-final singulars corresponded to iCot-final plurals. Thus, the plural forms did not obey the product-oriented generalizations that could be made on the basis of Hebrew. On the other hand, the singular-plural mappings obeyed the Hebrew rule that -ot was to be added to oC-final singulars while -im was to be added to iC-final singulars. The Hebrew learners found the surface pseudo-Hebrew easier to learn than the deep pseudo-Hebrew, suggesting that the product-oriented patterns of Hebrew transferred into the pseudo-Hebrew more easily than the source-oriented patterns, in turn suggesting that Hebrew speakers rely on product-oriented generalizations in plural formation. Becker and Fainleib hypothesize that the product-oriented generalizations are negative product-oriented generalizations that are combined with source-oriented paradigm uniformity constraints in accordance with Optimality Theory.

However, as Becker and Fainleib's (2009) own simulations show, the source-oriented Minimal Generalization Learner (Albright and Hayes

2003) pretrained on Hebrew achieves equal accuracy on both types of pseudo-Hebrew, rather than achieving higher accuracy on deep pseudo-Hebrew. This happens because Hebrew does not feature the singular-plural mappings oC → iCot and iC → oCim found in the deep pseudo-Hebrew. Thus, unless one forces the source-oriented learner to treat the singular-plural mapping as a two-stage process, with one stage selecting the affix and relying on rules shared between Hebrew and deep pseudo-Hebrew, and the other changing the vowel (which is only needed in pseudo-Hebrew), the learner will not rely on the same source-oriented generalizations in Hebrew and deep pseudo-Hebrew. Thus, Becker and Fainleib's results are open to the interpretation that Hebrew plural formation relies largely on source-oriented generalizations, and that product-oriented generalizations are used only when source-oriented generalizations are inapplicable.

1.3. The present experiment

The present experiment exposes native speakers of English to artificial languages that feature a process of velar palatalization before the plural suffix -i (k → tʃ/_i) but differ in whether -i is also shown to attach to [tʃ]]. Examples of tʃ → tʃi exemplify both the product-oriented generalization 'plurals often end in -tʃi', which favors mapping any source (including one ending in [k]) onto [tʃi], and the source-oriented generalizations '0 → i/C_' (which is extracted from the same data by the Minimal Generalization Learner, developed by Albright & Hayes 2003) and 'the stem-final consonant is retained in the plural form' (which is predicted to be active in all languages by Optimality Theory). Thus, if typical characteristics of source-product mapping are more salient than typical characteristics of product forms, examples of tʃ → tʃi should disfavor palatalization, i.e., the addition of such examples to training should favor {k;t;p} → {k;t;p}i over {k;t;p} → tʃi. On the other hand, if product characteristics are more salient than characteristics of source-product mappings, the same examples should favor palatalization, i.e., the addition of such examples to training should favor {k;t;p} → tʃi. Furthermore, across subjects, source-product mappings produced by the same generalization (whether product-oriented or source-oriented) should correlate in productivity. Thus, if examples of tʃ → tʃi primarily exemplify X → tʃi rather than 0 → i/tʃ_, the productivity of tʃ → tʃi for a given subject should correlate with the productivity of {k;t;p} → tʃi for the same subject more than with the productivity of {k;t;p} → {k;t;p}i.

Two different training paradigms were used in the present study. In source-oriented training, the artificial languages are learned under presentation conditions that can be argued to be maximally favorable for noticing relationships between source and product forms: learners are asked to repeat source-product pairs and tested on forming the product when presented with a source form. If typical characteristics of products are more noticeable than typical characteristics of source-product mappings even in this experimental paradigm, resulting in the formation and use of product-oriented generalizations, we would have strong evidence for language learners having a bias in favor of product-oriented generalizations. In product-oriented training, source and product forms sharing the same stem are no longer adjacent, with all wordforms being presented in random order. Comparison of the subjects' behavior following two types of training can shed light on whether the extent to which learners of a language rely on product-oriented vs. source-oriented generalizations depends on the conditions under which the language is learned.

2. Methods

2.1. Languages

2.1.1. The source-oriented paradigm

A given learner was exposed to one of the languages shown in Table 1. Both languages had 30 singular-plural pairs illustrating velar palatalization. Language 1 had no singulars ending in an alveopalatal, while Language 2

Table 1. The languages presented to learners in the source-oriented paradigm

	Language 1	Language 2
{tʃ;dʒ} → {tʃ;dʒ}i	0	20
{k;g} → {tʃ;dʒ}i	30	
{t;d;p;b} → {t;d;p;b}i	16[1]	
{t;d;p;b} → {t;d;p;b}a	16[1]	

1. Half of the subjects were exposed to 24 words taking -i and 8 words taking -a while the other half were exposed to the reversed proportions. See Kapatsinski (2010) for the significance and results of this manipulation.

had 20 singular-plural pairs featuring such a singular. Each singular-plural pair was presented twice during training. The large number of different word types that are presented to subjects and the low token/type ratio are expected to result in generalization across words and lack of memorization of individual wordforms. This feature of the present training paradigm is distinct from the product-oriented paradigm, where subjects are presented with a relatively small number of frequently occurring words that they are asked to memorize.

2.1.2. The product-oriented paradigm

In the product-oriented paradigm, subjects were exposed to individual singular and plural forms in random order. The number of distinct words had to be reduced in order for the subjects to be able to notice the relationship between the two forms of a given word within the same timeframe. The languages are shown in Table 2.

Table 2. The languages presented to learners in product-oriented training

	Language 1	Language 2
tʃ → tʃi	0	4
k → tʃi	4	
{t;p} → {t;p}i	4^2	
{t;p} → {t;p}a	4^2	

Goldberg, Casenhiser and Sethuraman (2004) have shown that the learning of novel argument structure constructions is facilitated if a few of the verbs associated with a construction occur very often while the majority occur infrequently compared to a condition in which all verbs occur equally often. Goldberg (2006: 85–89) reports that the same result also holds for dot pattern classification, suggesting that it is not a peculiarity of syntax (where the meaning of the construction might be gleaned off the meaning of the most frequent verb) and thus may also hold for morphophonology. Therefore, one word exemplifying k → tʃi, one word exemplifying the most frequent p → pV pattern in each language, one word

2. Half of the subjects were exposed to 6 words taking -i and 2 words taking -a while the other half were exposed to the reversed proportions.

exemplifying the most frequent t → tV pattern in each language, and one word exemplifying tʃ → tʃi were presented 42 times each, while the other words were presented 14 times each.

Recchia, Johns and Jones (2008) exposed human learners to an artificial lexicon in which words differed in frequency and the number of different sentences and pictorial scenes they appeared in. They found that frequency of presentation influenced lexical decision only if the word appeared in multiple different contexts, i.e., it had high contextual diversity. Contextual diversity was increased in the present experiment by combining each word with multiple frames: each word could be inserted in the sentences '{That's a; Those are the} ____' and '{I am a; We are the} ____', and also appeared on its own produced in a scared voice, a normal voice, or a touched voice. In addition, a voice was created for each individual creature by manipulating the speed, shifting the formant ratio, the pitch median, and the pitch range of the original speaker (me) using the 'Change gender' function in Praat (Boersma and Weenink 2009). The individual creature voices were used for producing the utterances fitting the schema '{I am a; We are the}___'. In addition, for the frequent words, the isolated word productions were produced in four different creature voices each.

2.2. Tasks

2.2.1. The source-oriented paradigm

The experiment consisted of a training stage, an elicited production test, and a likelihood rating test. During training, participants were asked to learn "how to form plurals in the language". A participant would be presented with a series of trials, each of which began with the presentation of a picture of a novel object on the computer screen. Three hundred milliseconds later, the name of the novel object in one of the four artificial languages was presented auditorily over headphones. Once the sound finished playing, the picture was removed and replaced with a picture of multiple (5–8) objects of the same type. The picture of multiple objects was accompanied by the auditory presentation of the plural form of the previously presented noun. Once the sound file finished playing, the participant repeated the singular-plural pair and clicked a mouse button to continue to the next singular-plural pair. The training task is shown schematically in Figure 1.

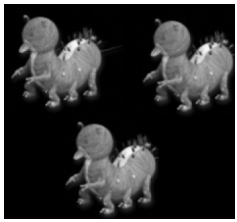
Video:						
Audio:		[boʊk]			[boʊtʃi]	
Learner action:	Watch	Watch and listen		Watch	Watch and listen	Repeat aloud, then click
Dura-tion:	.3		.5	.3		0–10

Figure 1. The source-oriented training task. Durations in seconds

The training stage was followed by the elicited production test, which was exactly like training except instead of hearing the plural form and repeating the singular-plural pair, the learner had to generate the plural and pronounce it aloud. Half of the singulars presented during the testing were novel, i.e., they have not been presented during training. The learner was not required to repeat the singular during the test. The task is shown schematically in Figure 2.

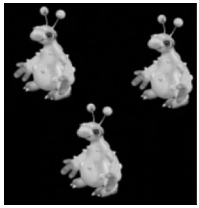
Video:				
Audio:		[vik]		
Learner action:	Watch	Watch and listen		Say the plural aloud, then click
Duration:	.3		.5	0–10

Figure 2. The elicited production test. Durations are in seconds

The elicited production test was followed by the rating task. In the rating task, the subject was presented with a singular-plural pair as s/he would be during training and had to answer "How likely is this plural to be the right plural for this singular?" on a scale from 1 = "impossible" to 5 = "very likely". The scale was displayed on the screen, and the learner responded by clicking a numbered rectangle with the mouse. All of the singular-plural pairs were novel and were presented in random order. Examples of the following mappings were presented for rating: [k] → {k;tʃ}{i;a}, [t] → {t;tʃ}{i;a} and [tʃ] → {tʃ;k}{i;a}. The task is presented schematically in Figure 3.

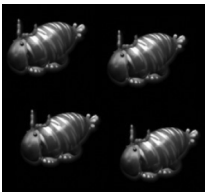
Video:						
Audio:		[fruk]			[fruki]	
Learner action:	Watch	Watch and listen		Watch	Watch and listen	Click on a rating
Dura-tion:	.3		.5	.3		0–10

Figure 3. The ratings task. Durations are in seconds

2.2.2. The product-oriented paradigm

Like in the source-oriented training paradigm, each singular-plural pair was matched with a picture pair. However, pairings of singular nouns with objects and pairings of plural nouns with objects appeared in random order. The learner was asked to learn the names for the objects. The learner repeated the noun forms they were presented with. If the noun appeared in a sentential frame, only the noun needed to be repeated. The training task is shown schematically in Figure 4.

After going through the training set once, the learners were tested on recalling the object names by being presented with an object or a set of identical objects and asked for the corresponding noun form. They were

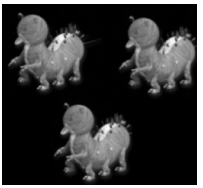
Video:			
Audio:		[boʊtʃi]	
Learner action:	Watch	Watch and listen	Repeat aloud, then click
Duration:	.3		0–10s

Figure 4. The product-oriented training task. Durations in seconds

instructed to produce the right form of the noun (whether singular or plural). The training-recall sequence was repeated twice and then followed by the same generalization and rating tasks used in the source-oriented paradigm.

2.3. Stimulus recording

The auditory stimuli were recorded by the author in a sound-attenuated booth onto a computer. The stimuli were sampled at 44.1 kHz and leveled to have the same mean amplitude. They were presented to the learners at a comfortable listening level of 63 dB. The learners were asked to repeat words they are hearing during training immediately after hearing them. Repetition accuracy was very high (97%). The visual stimuli were a set of made-up creature pictures retrieved from the website http://www.spore.com/sporepedia and are exemplified in Figures 1–4. The number of creatures paired with a plural wordform varied between 5 and 8. All pictures were presented on a black background.

2.4. Procedures

Learners were tested one a time. The learner was seated in a sound-attenuated booth. The audio stimuli were delivered and the learners' speech recorded using a Sennheiser HMD281 headset. The experimenter was seated outside the booth and was able to hear the audio presented to the learner as well as the learner's productions. The learner was unable to see the experimenter. The subject's productions were scored by the experimenter online, as the learner was producing them. The stimuli were presented

and ratings recorded using PsyScript experiment presentation software on Mac OS9.2. The order of presentation of the stimuli was randomized separately for each learner.

2.5. Participants

Participants were assigned to languages in the order they came in (Subject 1 – Language 1, Subject 8 – Language 2, etc.) In the source-oriented paradigm, 22 participants were exposed to each language. Each participant was exposed to only one language. In the product-oriented paradigm, there were also 22 participants assigned to learn each language. However, one participant assigned to Language 2 was subsequently excluded because of forming plurals using a pattern that was not presented in training (adding [tʃa]). One participant assigned to Language 1 was excluded from analyses of ratings because of computer error resulting in his ratings being lost. All of the participants reported being native English speakers with no history of speech, language, or hearing impairments. None reported being fluent in a foreign language. The participants were recruited from introductory psychology classes and received course credit for participation.

2.6. Analyses

All statistical analyses were conducted in R (http://www.cran.r-project.org). Due to severe non-normality of the data distributions, non-parametric statistics were used, i.e., all numerical variables were rank-transformed for the purposes of significance testing. The clustering solution is based on the coordinate matrix of the output of principal components analysis done on the correlation matrix between individual subjects' production probabilities and mean ratings of examples of source-product mappings (with one point per mapping per modality per subject) with centering and scaling. The coordinate matrix contains the locations of various mappings in the multidimensional space defined by the principal components, which are orthogonal dimensions that together accounted for between-subject variance. Clustering was done using Manhattan distance, since subjects are independent non-interacting dimensions, and the Average clustering method; Ward clustering, McQuitty clustering, and Complete clustering yield the same solution.

3. Results

Figure 5 shows a hierarchical clustering solution for correlations of all mappings used or rated in production and perception following source-

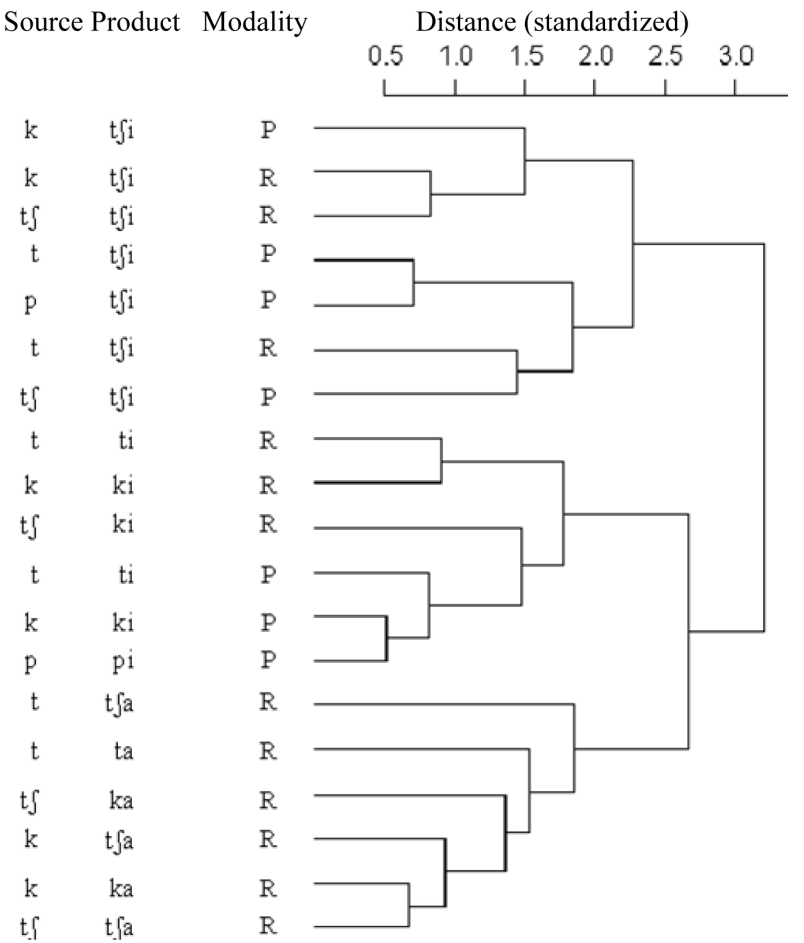


Figure 5. The clustering of the correlation matrix between ratings and production probabilities of various mappings following source-oriented training. 'R' stands for 'ratings', while 'P' stands for production probabilities

oriented training. The basic logic of this analysis is that if the same generalization underlies two source-product mappings, then subjects who assign a high weight to the generalization should consider both mappings acceptable, and subjects who assign a low weight to the generalization should consider both mappings unacceptable (see also Featherston 2007). Thus, we should find that the subjects' ratings and production probabilities for mappings that are produced by the same generalization should show a

positive correlation, and those that are produced by different generalizations should not. In this graph, the further to the right the vertical connection between two singular-plural mappings, the less similarly they were treated by the subjects, i.e., the further from 1 the correlation (r) between the mappings. In the interest of space I am omitting the very similar clustering solution for product-oriented training (the same clusters are formed at the top two branching levels).

Figure 5 shows that even source-oriented training results in a cluster of source-product mappings in which any source is mapped onto [tʃi], a cluster in which any source is mapped onto a product ending in a stop followed by [i], and a cluster in which any source is mapped onto a product ending on [a]. Thus even following source-oriented training, [tʃ] → [tʃi] is unambiguously classified as an instance of C → [tʃi] rather than an instance of 'just add -i' in both perception and production. These results provide support for the overall primacy of product-oriented generalizations over source-oriented generalizations (Bybee 2001) and suggest a similar weighting of source-oriented and product-oriented generalizations in rating and production. Nonetheless, Figure 5 also shows that the cluster of mappings in which some source is mapped onto [tʃi] is further subdivided into a cluster of mappings presented during training and a cluster of mappings that were not presented. This suggests that at the very least the subjects must know which source consonants need to be retained in the product form, a type of source-oriented knowledge describable using faithfulness constraints in Optimality Theory (e.g., Downing et al. 2005).

The clustering analysis in Figure 5 showed us that overall typical characteristics of products are more salient than typical characteristics of source-product mappings, even following source-oriented training, which maximizes the salience of source-product mappings. Thus tʃ → tʃi is classified as primarily X → tʃi rather than 0 → i/C_. Figures 6–7 take a closer look at the data in order to explain what leads to this classification.

When the data from both types of training are combined, examples of tʃ → tʃi significantly favor alveolar palatalization (t → tʃi) ($F(1,78) = 7.7$, $p = .006$ for production, shown in Figure 6; $F(1,78) = 10.9$, $p = .001$ for rating, shown in Figure 7), and there is no significant interaction between training paradigm and whether or not examples of [tʃ] → [tʃi] are presented ($F(1,78) < 1$, $p = .77$ for production; $F(1,78) < 1$, $p = .83$ for rating). This relatively strong effect of exposure to tʃ → tʃi on the productivity of t → tʃi relative to t → ti is the main reason for the clustering algorithm classifying tʃ → tʃi as an instance of X → tʃi following source-oriented training.

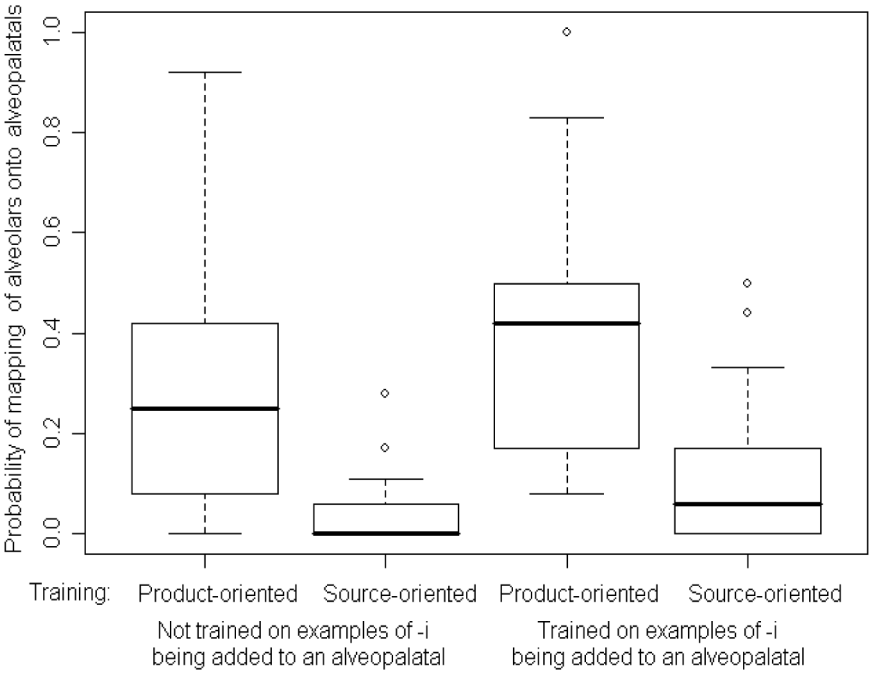


Figure 6. Following either type of training, examples of tʃ → tʃi favor t → tʃi over t → ti in production (notches not shown since they go outside of the box)

Figure 8 shows that the addition of examples of tʃ → tʃi to training has different effects on the productivity of velar palatalization in the two training paradigms. In the source-oriented training paradigm, examples of tʃ → tʃi support k → ki over k → tʃi. In the product-oriented training paradigm, examples of tʃ → tʃi support k → tʃi over k → ki. If we combine the results from both training paradigms (entering training paradigm, whether -i is attached to [tʃ] in training, whether -i often attaches to [p] and [t] in training, and all interactions into a Friedman test (i.e., an ANOVA with a rank-transformed dependent variable) as predictors of production probability the only significant effect is an interaction between experiment and whether or not examples of [tʃ] being mapped onto [tʃi] are presented to the learner ($F(1,79) = 6.25$, $p = .01$). If we take the probability of [k] being mapped onto [ki] as the dependent variable, there is also a significant interaction in the same direction: the additional examples of [tʃ] → [tʃi] presented during training increase the probability of eliciting

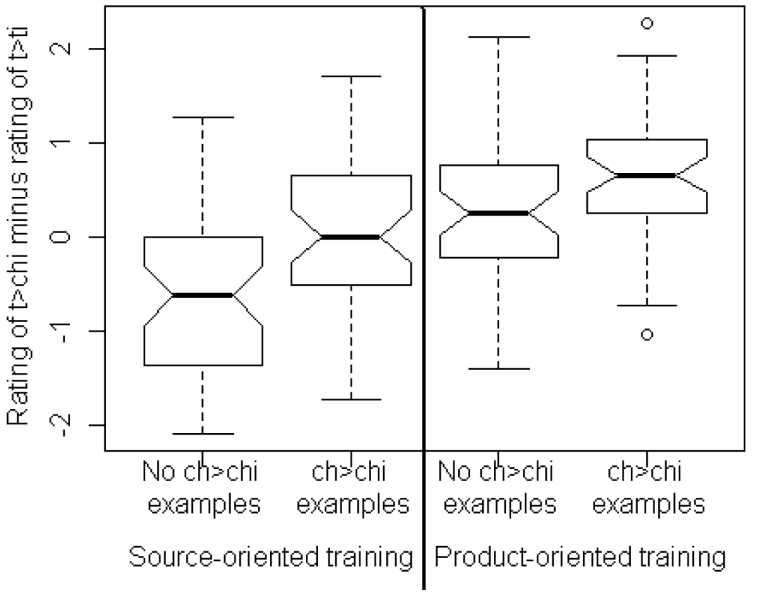


Figure 7. Following either type of training, examples of tʃ → tʃi favor t → tʃi over t → ti in rating

the production of [k] → [ki] in the source-oriented training paradigm while decreasing the probability of eliciting [k] → [ki] in the product-oriented training paradigm ($F(1,79) = 4.02$, $p < .05$). There are no significant effects in the ratings task.

These results suggest that the [tʃ] → [tʃi] examples support 'just add -i' over 'plurals must end in [tʃi]' in source-oriented training while the opposite is true for the source-oriented training paradigm. Thus, the characteristics that distinguish the two training paradigms are able to jointly influence how much the language learner relies on product-oriented vs. source-oriented generalizations in deriving new wordforms, thus extending the lexicon of the language. Interestingly, in the source-oriented paradigm, examples of tʃ → tʃi support k → ki over k → tʃi while supporting t → tʃi over t → ti. We will return to this apparent contradiction in the General Discussion.

Despite tʃ → tʃi disfavoring velar palatalization in source-oriented training, the clustering solution presented in Figure 5 classifies tʃ → tʃi as an X → tʃi mapping, rather than a 0 → i/C_ mapping even following source-oriented training because the effect of adding examples of tʃ → tʃi on the

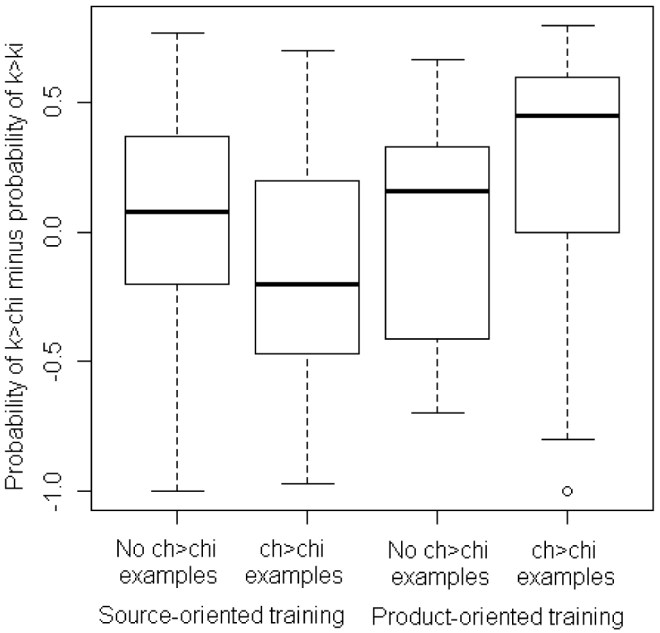


Figure 8. Following product-oriented training, examples of tʃ → tʃi favor k → tʃi over k → ki; the opposite is true for source-oriented training (notches not shown since they go outside of the box)

productivity of velar palatalization is weaker than its effect on alveolar palatalization (the former effect failing to reach significance within task).

Product-oriented training appears to improve ratings of source-product mappings that involve a stem change but result in a good product. Importantly, it does not favor *all* stem changes. In the rating task, learners were asked to rate tʃ → ki, tʃ → ka, k → tʃa, and t → tʃa mappings, which do not result in a good product. Figure 9 shows that the ratings of these mappings are not higher in product-oriented training relative to source-oriented training. There is a significant interaction between type of product and type of training ($F(1,331) = 8.58$, $p = .003$) and a significant interaction between type of source and type of training ($F(1,331) = 5.45$, $p = .02$) with no significant three-way interaction between source, product, and training types ($F(1,331) = 1.02$, $p = .33$). Thus, it appears that product-oriented training increases the productivity/acceptability of stem changes only when those stem changes result in a good product.

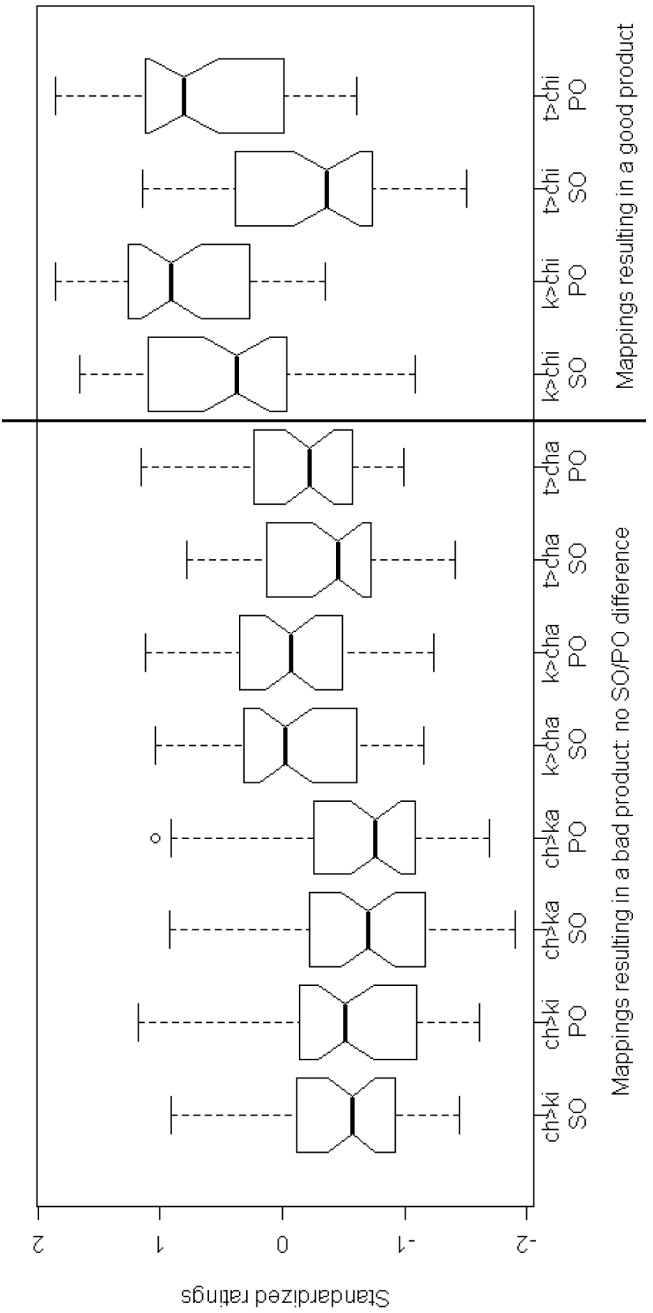


Figure 9. The effect of training type (SO = "source-oriented", PO = "product-oriented) on ratings of stem-changing mappings resulting in the good product (tʃi) and bad products (tʃa, ki, ka)

While typical characteristics of product forms appear to be more salient than typical characteristics of source-product mappings, the learners' behavior is not completely product-oriented even following product-oriented training, as has already been suggested by the finding that the cluster of mappings resulting in [tʃi] is further subdivided into the observed and unobserved mappings. First, overgeneralization of velar palatalization to labial sources following either kind of training is much less likely than overgeneralization to alveolar sources ($p < .00001$ for product-oriented training, $p = .0002$ for source-oriented training, according to the Wilcoxon test). Similarly, after both types of training, k → k{i;a} mappings, which result in an unobserved product, are rated higher than tʃ → k{i;a} mappings, which result in the same unobserved product but also feature a stem change ($p < .0001$ after either type of training). Finally, tʃ → tʃi mappings are rated higher than k → tʃi or t → ti mappings after either type of training ($p < .001$) despite resulting in the same product. In the case of t → tʃi vs. tʃ → tʃi, this result holds even in Languages 1 where examples of tʃ → tʃi are never presented ($p = .0001$ for t → tʃi vs. tʃ → tʃi, $p = .05$ for k → tʃi vs. tʃ → tʃi following source-oriented training; $p = .007$ for t → tʃi vs. tʃ → tʃi, $p = .11$ for k → tʃi vs. tʃ → tʃi after product-oriented training). Thus even following product-oriented training most learners' disprefer stem changes and possess grammars that contain source-oriented generalizations, perhaps, in the form of paradigm uniformity constraints (Becker & Fainleib 2009, Downing et al. 2005, Stemberger & Bernhardt 1999), that allow them to restrict the types of sources that can give rise to a good product; for instance, avoiding mapping [p] onto (p)[tʃi].

4. Discussion

4.1. The influence of the presentation conditions

In the present experiments, the learners were exposed to miniature artificial languages in two different training paradigms. The two training paradigms differ in that:

1. The learner in product-oriented training is presented with one word-form from a paradigm at a time, while the learner in source-oriented training is presented with pairs of words that share the stem.
2. The learner in product-oriented training is exposed to a much smaller number of word types and a much larger number of word tokens per

type than the learner in source-oriented training; one consequence of this difference is that the learner in product-oriented training can, to a large extent (empirically 92% on average), memorize the lexicon exemplifying the grammar, while the learner in source-oriented training acquires a grammar without a lexicon.[3]

When asked to go beyond the acquired lexicon and apply the learned grammar to new words, learners exposed to product-oriented training are found to exhibit stronger reliance on product-oriented generalizations and weaker reliance on source-oriented generalizations than learners exposed to source-oriented training. This manifests itself in two ways:

1. Examples of tʃ → tʃi are taken to support k → ki, a mapping with the same source-product relationship, over k → tʃi, a mapping resulting in the same product, by learners exposed to source-oriented training; the opposite is true for learners exposed to product-oriented training.
2. The learner in product-oriented training overgeneralizes palatalization to alveolar and labial sources much more than does the learner in source-oriented training. That is, the product-oriented learner infers t → (t)tʃi and sometimes p → (p)tʃi based on exposure to k → tʃi while the source-oriented learner does not.

Product-oriented training does not simply influence how much the learners avoid stem changes across the board. Stem changes that do not result in a good product, e.g., tʃ → k{a;i}, do not benefit from product-oriented training. Rather, product-oriented training appears to draw attention away from source-product relationships and towards characteristics of product forms (or perhaps, all individual wordforms), compared to source-oriented training.

Thus the present results support the idea that the types of generalizations that are relied upon by a speaker/hearer in extending his/her lexicon are influenced by the way the speaker/hearer experiences language, and not just by an innate Universal Grammar, suggesting that even formal

3. In addition, in the original experiment subjects in the product-oriented paradigm signed up to 'learn names for objects' while the learners in source-oriented training signed up to learn 'how to make plurals' in a made-up language. However, I have subsequently conducted a product-oriented training experiment in which half of the subjects ($N = 32$) were presented with each type of instruction and found no significant effect of instruction ($F < 1$), thus the difference in instructions is unlikely to lead to the differences in behavior following the two training paradigms.

properties of the grammar may be emergent from patterns of language use (Bybee 2008). As Valian and Coulson (1988: 78) suggested, "Our [...] acquisition of competence is mediated by the performance system. That performance system [...] limits us to acquiring a language only under presentation conditions which are cognitively favorable."

The present results indicate that presentation conditions may bias a learner in favor of source-oriented or product-oriented generalizations. If native speakers of natural languages prefer product-oriented generalizations over rules (Becker and Fainleib 2009, Bybee 2001, Bybee and Slobin 1982, Kőpcke 1988, Lobben 1991, Wang and Derwing 1994), this may be due to the way those languages are experienced by their native speakers, since learners tend not to hear multiple forms of the same lexeme one after another.

At least three predictions for natural languages follow from the observed effect of the learning task. First, reliance on source-oriented generalizations may be more expected in non-native speakers of a language, who experience language through textbooks that explicitly teach the reader to conjugate verbs and decline nouns, than in native speakers, who experience language one wordform at a time. Second, source-oriented generalizations should form when wordforms sharing a stem tend to appear in close temporal proximity. This is, perhaps, the case for noun-adjective pairs of the type 'electric-electricity' in English, for which source-oriented generalizations like k → s/_ʃti (or 'an [l] in the noun corresponds to an [l] in the adjective') appear to be stronger than product-oriented generalizations like '-ity is usually/should be preceded by [l]' (Pierrehumbert 2006). Some support for this hypothesis is provided by Morgan, Meier and Newport (1989) who found that the acquisition of a phrase structure grammar was facilitated when learners were provided with pairs of sentences that could be related by pronominalization or movement rules but were unable to replicate the effect with related pairs of sentences being randomly interspersed with other, unrelated sentences. Finally, product-oriented generalizations may be favored over source-oriented generalizations especially strongly if both have to be acquired over a small set of word types where the inherently lower type frequency of source-oriented generalizations may be of particular importance.

4.2. Task-independent properties of grammar

While the learning situation influences the degree to which the learner relies on source-oriented vs. product-oriented generalizations, and thus the

acquired grammar, there are a number of characteristics of the acquired grammatical systems that hold across the two learning situations.

The learners in both training paradigms learn grammars that contain both product-oriented generalizations, such as 'plurals should end in -tʃi', and source-oriented generalizations that restrict the sources that can be mapped onto a product (cf. also Pierrehumbert 2006). Thus, despite the overall preference for [tʃi]-final plurals, [p]-final sources are less likely to be mapped onto [tʃi]-final plurals than [t]-final or [k]-final sources even after product-oriented training (a finding that mirrors linguistic typology, as shown by Bateman 2007). In addition, stem changes resulting in unobserved products are dispreferred relative to simple addition of an affix resulting in the same unobserved product. Thus, the learned grammar is not purely product-oriented. The product-oriented generalizations need to be supplemented with something analogous to paradigm uniformity constraints, e.g., 'if there is a [k] in the singular, there must be a [k] (in the same position) in the plural'. The present data provide no evidence regarding whether these constraints are learned. It is quite possible that the learners come to the experiment knowing that p → tʃi mappings are worse than t → tʃi mappings. On the other hand, it is also possible that t → tʃi mappings are favored relative to t → ti mappings in a way that p → tʃi mappings are not because t → ti is acoustically more similar to t → tʃi than p → pi is to p → tʃi.

The necessity of supplementing product-oriented generalizations with restrictions on which source forms can be mapped onto a desirable product (i.e., paradigm uniformity constraints, see Becker and Fainleib 2009, Downing, Hall and Raffelsiefen 2005, Stemberger and Bernhardt 1999) is also suggested by Pierrehumbert (2006). Pierrehumbert shows that when a native English speaker is presented with a novel Latinate adjective ending in [k] and produces a noun ending in -ity from it, as in 'interponic' → 'interponicity', the adjective-final [k] is changed into an [s] when followed by -ity. Pierrehumbert argues that English speakers must be using a source-oriented generalization like k → s/_ity and not a product-oriented one like 'Latinate nouns should end in [sɪti]' or 'Latinate nouns should not end in [kɪti]' for two reasons. First, only adjectives ending in [k] are mapped onto nouns ending in [sʃti], This shortcoming of purely product-oriented phonology can be remedied by allowing segment-specific paradigm uniformity constraints like 'a [t] present in the adjective is retained in the noun', which, being made over source-product pairs, are source-oriented generalizations. Second, Pierrehumbert shows that [s] is not the consonant that most commonly precedes -ity in English. Rather, [l] precedes -ity much

more commonly than [s] does. Therefore, a learner generalizing over nouns would be expected to believe that -ity should be preceded by [l] much more often than by [s], Nonetheless, speakers in Pierrehumbert's experiment never changed [k] into [l] when attaching -ity. Generalization over adjective-noun pairs, on the other hand, would yield the observed pattern of [k] being mapped onto [s] and not [l] because adjectives ending in [k] never correspond to nouns ending in [lɪti] but often correspond to nouns ending in [sɪti].

Generalization is not minimal in the present study. This is a violation of the popular Subset Principle (Berwick 1986, Dell 1981, Hale and Reiss 2003). It appears worthwhile to distinguish between two types of overgeneralization. One type of overgeneralization is, I would argue, an inevitable result of perceptual processes. Traditionally, the output of human perception is taken to be a single hypothesis about the identity of the stimulus, thus the only information provided by perception is the identity of the most probable stimulus given the evidence. For instance, Clayards et al. (2008: 804), in a paper arguing for an otherwise Bayesian approach to speech perception, write "the goal of speech perception can be characterized as finding the most likely intended message". Under a purely Bayesian approach, on the other hand, the output of perception is a probability distribution over possible stimuli (Kruschke 2008, Levy 2009). Thus, despite reporting having perceived the most probable stimulus, the perceiver assigns other similar stimuli non-zero probabilities of having been presented. For instance, a subject presented with [ti] may report hearing [ti] but also (subconsciously) consider it possible but less likely that [ki] has just been presented. Note that if the learner intends to maximize the probability of being correct, s/he should always report hearing the stimulus s/he considers to be the most probable one (Norris and McQueen 2008) but should update the probability of each possible hypothesis in proportion to how likely s/he believes it to be given the sensory data (Kruschke 2008, Levy 2009).

Given these assumptions, it appears unsurprising that palatalization is much more likely to be overgeneralized to [t] than to [p] and that palatalization is overgeneralized to [t] despite accurate reporting of hearing t → ti when presented with t → ti. It appears inevitable that a perceiver hearing (and reporting hearing) [t$^{(j)}$i] would assign some probability to having heard [tʃi] and that this estimated probability would be higher when [t$^{(j)}$i] is presented than when [p$^{(j)}$i] is presented. Thus, overgeneralization of palatalization to [t] is predicted to be more likely (perhaps, inevitable) given Bayesian perception, than overgeneralization to [p], which appears

to be 'genuine' overgeneralization due solely to the product-oriented schema 'plurals must end in -tʃi'.

While the learned grammar contains both product-oriented and source-oriented generalizations, learners appear to pay less attention to the source-product relationship than to the shapes of typical products in both training paradigms. Thus, even after source-oriented training, the mapping tʃ → tʃi is treated as more similar to other mappings resulting in the same product (tʃi) than to other mappings featuring the same source-product relationship ([ ] → i). This finding contradicts the assumptions of rule-based models (Chomsky and Halle 1968, Albright and Hayes 2003, Plag 2003) and provides support for the product-oriented Network Theory (Bybee 2001).

An important remaining question is whether the product-oriented generalizations are positive, as in Bybee's Network Theory (Bybee 1985, 2001) and Stemberger and Bernhardt's version of Optimality Theory (Stemberger and Bernhardt 1999) or negative, as in traditional (Prince and Smolensky 1993/2004) and Stochastic Optimality Theory (Boersma 1997, Boersma and Hayes 2001). Interestingly, simulations using the implementation of Stochastic Optimality Theory in Praat (Boersma and Weenink 2009) show that, despite incorporating product-oriented markedness constraints, Stochastic Optimality Theory has problems handling the present data. The fact that learners in the present experiment appear to learn that velars and possibly alveolars become alveopalatals before -i can be modeled by the constraint weighting in (1). Palatalization of a consonant with a certain place of articulation is triggered by the applicable *_i constraint being ranked above the applicable Ident-Place constraint.

(1) *ki, Ident-Labial ≫ *C_{Stop}i, Ident-Alveolar, Ident-Velar, *a

Examples of tʃ → tʃi do not provide evidence on whether Ident-Velar, Ident-Labial, and Ident-Alveolar constraints should be ranked above or below *C_{Stop}i or *ki. Thus, examples of tʃ → tʃi should have no effect on the estimated desirability of [tʃi]-final plurals resulting from [k]-final singulars relative to [ki]-final plurals.

For the examples of tʃ → tʃi to, e.g., favor t → tʃi over t → ti, there must be a *tʃi constraint whose weight is decreased by examples of tʃ → tʃi. Why learners should come to the task with such a constraint (which should be relatively highly-ranked for its demotion to have appreciable effects on behavior) remains a mystery since it is supported neither by training data nor the learners' prior linguistic experience. On the other hand, in Network Theory, [tʃi]-final plurals support other [tʃi]-final

plurals, whatever the source, because of a generalization like 'plurals must end in -tʃi', which is supported by the training data, in which tʃi-final plurals form a large proportion of the lexicon (cf. also Stemberger and Bernhardt 1999: 437–438).

An alternative way to weight a constraint against the unobserved sequence [ki] is to calculate the likelihood that the absence of [ki] is not accidental by taking the difference between how often [ki] is expected to occur and how often it actually occurs based on the frequencies of occurrence of related sequences in plural forms (Frisch, Broe and Pierrehumbert 2004, Pierrehumbert 1993, Stefanowitsch 2008, Xu and Tenenbaum 2007). The actual frequency of occurrence of [ki] is zero across the two artificial languages. However, other [Ci] sequences occur much more often when examples of tʃ → tʃi are presented. Thus, the learner estimating how often [ki] would occur if it were just like the other [Ci] sequences would estimate a higher frequency when exposed to examples of tʃ → tʃi, which would cause him/her to be more confident that [ki] is to be avoided. For example, Xu and Tenenbaum (2007) find that learners presented with three examples of the novel word *fep* infer that *fep* means 'Dalmatian' rather than 'any dog' more often than if only one *fep*-Dalmatian pairing is presented. Xu and Tenenbaum argue that the learners detect a suspicious correlation between *fep* and pictures of Dalmatians, which would be unexpected if *fep* could refer to any dog. Regier and Gahl (2004) and Stefanowitsch (2008: 518) propose that the same mechanism may be used in syntax. If phonology learning worked the same way (as suggested by Frisch, Broe and Pierrehumbert. 2004 and Pierrehumbert 1993 for OCP), we would expect that exposure to examples of tʃ → tʃi would restrain -i from simply attaching to [k]. Thus, the examples of tʃ → tʃi would disfavor palatalization, contrary to the data presented here as well as the data in Kapatsinski (2010), which shows that additional examples of {p;t} → {p;t}i strongly favor k → ki rather than restricting attachment of -i to labial-final and alveolar-final sources. Thus, the present data support reliance on positive, rather than negative, product-oriented generalizations (Bybee 1985, 2001, Stemberger and Bernhardt 1999).

It may be expected that constraints against unobserved combinations of units should be less salient in phonology than in lexical semantics (Xu & Tenenbaum 2007) or syntax (Regier and Gahl 2004, Stefanowitsch 2008) because unobserved unit combinations are usually more similar acoustically to observed combinations in phonology than in syntax or the lexicon. A learner hearing [pa] is expected to assign some probability to having heard [ka], and a learner hearing [t$^{(j)}$i] or [p$^{(j)}$i] may assign some probability to

having heard [k$^{(j)}$i] even if the correct phoneme sequence is reported. Thus exposure to phoneme sequences that are similar to an unobserved phoneme sequence should not necessarily decrease the estimated probability of the unobserved sequence if the similar sequences are similar enough to be confusable with the unobserved sequence (although the observed sequence should benefit from its presentation more than other similar sequences). Perceptual similarity between words, animal pictures (Xu & Tenenbaum 2007), or word sequences (Stefanowitsch 2008) is generally lower than between phoneme sequences, thus an unobserved combination is less likely to benefit from the presentation of a similar combination. Thus, estimation of the reality of a gap based on the frequency of occurrence of related sequences may play a larger role in syntax and word learning than in phonology.

In both training paradigms, examples of tʃ → tʃi support t → tʃi over t → ti and p → tʃi over p → pi. In the source-oriented paradigm, the same examples also support k → ki over k → tʃi. In the product-oriented paradigm, they support k → tʃi over k → ki but not as much as they support t → tʃi over t → ti. One thing that distinguishes t → tʃi, p → tʃi, and k → ki from t → ti, p → pi, and k → tʃi is that the former set of mappings is unobserved during training while the latter is observed. Thus, we may hypothesize that the same amount of extra support increases the strength of a poorly supported mapping (e.g., k → ki) more than it increases the strength of a mapping that is already well supported (e.g., t → ti). That is, the relationship between amount of support from the training data and resulting strength of a source-product mapping or a candidate product form is a decelerating function, like a logarithm (cf. Goldiamond and Hawkins 1958 for the same effect in word recognition; Norris and McQueen 2008 for computational evidence that the decelerating function emerges out of Bayesian inference). An alternative explanation is that source-product mappings involving similar segments support each other and the learners consider [tʃ] to be more similar to [k] than to [t], thus tʃ → tʃi examples provide more support for k → ki than to t → ti and following source-oriented training the increase in support for k → ki happens to be greater than the increase in support for 'plurals end in [tʃi]' but the increase in support for t → ti is not.

In general, the results from elicited production and rating tasks are very similar. The one difference between elicited production and rating observed in the present data is that elicited production appears to disfavor stem changes more than rating does (see also Zuraw 2000 for the same finding in natural language). Thus, only 4/44 learners exposed to source-oriented

training produce more instances of t → tʃi than t → ti but the median difference in standardized ratings between the two mappings is only .13 (standard deviations), and 16/44 learners assign lower ratings to t → ti than to t → tʃi. The median difference in production probability between k → ki and k → tʃi is 0, while k → ki is rated as being somewhat less probable than k → tʃi (.3 standard deviations). Nonetheless, the difference is small and significant only for the velars ($p = .01$, according to the Wilcoxon).

5. Conclusion

The results provide support for a grammar that contains both positive product-oriented generalizations (a.k.a. schemas, Bybee 1985, 2001) and source-oriented paradigm uniformity constraints, a combination proposed by Stemberger and Bernhardt (1999). The learner acquiring the grammar appears to 1) pay more attention to characteristics of the product than to the source-product relationship, especially when sources and products do not occur in close temporal proximity and/or the size of the lexicon exemplifying the grammar is relatively small, and 2) assign some probability mass to percepts other than the most probable one, i.e., the one the learner reports hearing.

References

Albright, Adam and Bruce Hayes
2003 Rules vs. analogy in English past tenses: A computational / experimental study. *Cognition* 90: 119–161.

Bateman, Nicoletta
2007 A crosslinguistic investigation of palatalization. Ph.D. Dissertation, Department of Linguistics, University of California, San Diego.

Becker, Michael and Lena Fainleib
2009 The naturalness of product-oriented generalizations. *Rutgers Optimality Archive*, 1036–0609. Available at http://roa.rutgers.edu/files/1036–0609/1036-BECKER-0–0.PDF.

Berwick, Robert C.
1986 *The acquisition of syntactic knowledge*. Cambridge, M.A.: M.I.T. Press.

Bickel, Balthasar, Goma Banjade, Martin Gaenszle, Elena Lieven, Netra Paudyal, Ichchha Purna Rai, Manoj Rai, Novel Kishor Rai and Sabine Stoll
2007 Free prefix ordering in Chintang. *Language* 83: 1–31.

Boersma, Paul
1997 How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21: 43–58.
Boersma, Paul and Bruce Hayes
2001 Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45–86.
Boersma, Paul and David Weenink
2009 Praat: doing phonetics by computer (Version 5.1.07). [Computer program]. Retrieved May 19, 2009, from http://www.praat.org/
Bybee, Joan L.
1985 *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.
Bybee, Joan L.
2001 *Phonology and language use*. Cambridge: Cambridge University Press.
Bybee, Joan L.
2008 Formal universals as emergent phenomena: The origins of structure preservation. In Jeff Good, ed. *Linguistic universals and language change*, 108–24. Oxford: Oxford University Press.
Bybee, Joan L. and Carol Lynn Moder
1983 Morphological classes as natural categories. *Language* 59: 251–270.
Bybee, Joan L. and Dan I. Slobin
1982 Rules and schemas in the development and use of the English past. *Language* 58: 265–289.
Caballero, Gabriela
2010 Scope, phonology and templates in an agglutinating language: Choguita Rarámuri (Tarahumara) variable suffix ordering. *Morphology* 20: 165–204.
Chomsky, Noam and Morris Halle
1968 *The sound pattern of English*. New York, Evanston, London: Harper and Row.
Clayards, Meghan, Michael K. Tanenhaus, Richard N. Aslin and Robert A. Jacobs
2008 Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108: 804–809.
Dell, François
1981 On the learnability of optional phonological rules. *Linguistic Inquiry* 12: 31–37.
Downing, Laura J., Tracy A. Hall and Renate Raffelsiefen, eds
2005 *Paradigms in phonological theory*. Oxford: Oxford University Press.

Featherston, Sam
2007 Data in generative grammar: the stick and the carrot. *Theoretical Linguistics* 33: 269–318.
Frisch, Stephen A., Janet B. Pierrehumbert and Michael B. Broe
2004 Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22: 179–228.
Goldberg, Adele E.
2006 *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Goldberg, Adele E., Devin Casenhiser and Nitya Sethuraman
2004 Learning argument structure generalizations. *Cognitive Linguistics* 14: 289–316.
Goldiamond, Israel and William F. Hawkins
1958 Vexierversuch: The log relationship between word-frequency and recognition obtained in the absence of stimulus words. *Journal of Experimental Psychology* 56: 457–463.
Hale, Mark and Charles Reiss
2003 The Subset Principle in phonology: Why the *tabula* can't be *rasa*. *Journal of Linguistics* 39: 219–244.
Kapatsinski, Vsevolod
2010 Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory Phonology* 1: 361–393.
Kisseberth, Charles
1970 On the functional unity of phonological rules. *Linguistic Inquiry* 1: 291–306.
Köpcke, Klaus-Michael
1988 Schemas in German plural formation. *Lingua* 74: 303–335.
Kruschke, John K.
2008 Bayesian approaches to associative learning: From passive to active learning. *Learning and Behavior* 36: 210–226.
Levy, Roger
2009 With uncertain input, rational sentence comprehension is good enough. Paper presented at LSA Annual Meeting, San Francisco, CA.
Lobben, Marit
1991 Pluralization of Hausa nouns, viewed from psycholinguistic experiments and child language data. M.Phil. Thesis, University of Oslo.
Menn, Lise and Brian MacWhinney
1984 The repeated morph constraint: Toward an explanation. *Language* 60: 519–541.
Morgan, James L., Richard P. Meier and Elissa L. Newport
1989 Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language* 28: 360–74.

Norris, Dennis and James M. McQueen
2008 Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115: 357–395.
Paradis, Carole
1989 On constraints and repair strategies. *The Linguistic Review* 6: 71–97.
Pierrehumbert, Janet B.
1993 Dissimilarity in the Arabic Verbal Roots. *Proceedings of the North East Linguistics Society* 23: 367–381.
Pierrehumbert, Janet B.
2006 The statistical basis of an unnatural alternation. In Louis Goldstein, D. H. Whalen and Catherine T. Best, eds. *Laboratory Phonology VIII, Varieties of phonological competence*, 81–107. Berlin: Mouton de Gruyter.
Pinker, Steven
1999 W*ords and rules: The ingredients of language*. New York: Harper Collins.
Plag, Ingo
2003 *Word formation in English*. Cambridge: Cambridge University Press.
Prince, Alan and Paul *Smolensky*
2004 [1993] *Optimality Theory: Constraint interaction in generative grammar*. Malden, MA: Blackwell.
Recchia, Gabriel, Brendan T. Johns and Michael N. Jones
2008 Context repetition benefits are dependent on context redundancy. *Proceedings of the Annual Conference of the Cognitive Science Society* 30: 267–272.
Regier, Terry and Susanne Gahl
2004 Learning the unlearnable: The role of missing evidence. *Cognition* 93: 147–155.
Roca, Iggy
1997 Derivations of constraints, or derivations and constraints? In I. Roca, ed. *Derivations and constraints in phonology*, 3–41. Oxford: Clarendon Press.
Stefanowitsch, Anatol
2008 Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics* 19: 513–531.
Stemberger, Joseph P.
1981 Morphological haplology. *Language* 57: 791–817.
Stemberger, Joseph P. and Barbara H. Bernhardt
1999 The emergence of faithfulness. In B. MacWhinney, ed. *The emergence of language*, 417–446. Mahwah, London: Erlbaum.
Valian, Virginia and Seana Coulson
1988 Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language* 27: 71–86.

Wang, H. Samuel and Bruce L. Derwing
1994 Some vowel schemas in three English morphological classes: Experimental evidence. In Matthew Y. Chen and Ovid C. L. Tseng, eds. *In honor of Professor William S.-Y. Wang: Interdisciplinary studies on language and language change*, 561–575. Taipei: Pyramid Press.

Xu, Fei and Joshua B. Tenenbaum
2007 Word learning as Bayesian inference. *Psychological Review* 114: 245–272.

Zuraw, Kie
2000 Patterned exceptions in phonology. Ph.D. Dissertation, Department of Linguistics, University of California, Los Angeles.

Relative frequency effects in Russian morphology

Eugenia Antić

1. Introduction

A variety of experiments has been carried out in the last four decades aimed to answer the question of which factors influence morphological processing. Proponents of decomposition (e.g., Taft (1985)) argue that every word is decomposed into its constituent morphemes and a lexical search is carried out on the root. On the other hand, proponents of whole-word processing argue that all words are accessed as one whole entity (e.g., Butterworth, 1983). The latest view is that both routes of processing, whole-word and decomposition, exist. In different models each encountered word is processed using both routes, and the faster one prevails (Frauenfelder and Schreuder, 1992), one is employed for known words, the other for novel (Caramazza et al., 1988), or both operate at the same time (Wurm, 1997).

In race models, determining which route prevails in a particular item is usually done by manipulating cumulative root frequency and surface frequency of that item. In this experimental paradigm, cumulative root frequency is defined as the combined frequency of the root of the word and surface frequency is the frequency of the word as a whole. For example, Taft (1985) cites the following frequencies for the following words: *approach* 123, *reproach* 3, *persuade* 17, *dissuade* 3. These are the surface frequencies for those words. Cumulative root frequencies of *proach* and *suade* are 126 (123 + 3) and 20 (17 + 3), respectively. In an experiment, if the surface frequency is held constant and the cumulative root frequency is manipulated, faster reaction times for higher base frequency items is taken as evidence that that set of words is accessed via the morphological decomposition route. If, on the other hand, cumulative root frequency is held constant and surface frequency is manipulated, and more frequent items elicit a faster response, it is taken as evidence that the direct route is favored for the set of words in question. Usually, the set of words tested includes words with a certain affix and the findings are assumed to apply to all words with that affix. Several studies used this methodology in prefix stripping experiments. Cole et al. (1989) argue against prefix stripping and for suffix stripping. In a set of French lexical decision experiments, the authors find

differences between processing of prefixes and suffixes. They find a significant difference in reaction time between suffixed words with high versus low cumulative root frequency, but no significant difference in reaction time between prefixed words with high versus low cumulative root frequency. They explain this effect by proposing that, since words are processed with the prefix first, then root and then suffix, root effects only appear in suffixed words, where the root is processed first. Also in French, Giraudo and Grainger (2003) find opposite results. In masked priming experiments, they find prefix priming, but not suffix priming. Based on three English experiments, Taft and Forster (1975) propose a model of word recognition based on the root where the prefix is stripped first. In a theoretical investigation, Schreuder and Baayen (1994) show that such a model would be highly inefficient and thus improbable.

Other studies show that an important factor in morphological processing not taken into account in the above experiments is relative frequency (Cole et al., 1997; Hay, 2001, 2002; Burani and Thornton, 2003; Zuraw, 2009). Relative frequency is the difference between the frequency of the derived word and the frequency of its base. Using the English words *approach* and *inaccurate* I illustrate these terms:

1. Derived frequency: frequency of the word itself. For both *approach* and *inaccurate* that would be the word frequency.
2. Base frequency: the frequency of the unprefixed word. Since **proach*, the base of *approach*, is a bound root, the base frequency of *approach* is zero. On the other hand, *accurate*, the base of *inaccurate*, exists as a separate word, and thus the base frequency of *inaccurate* is the frequency of *accurate*.

Hay (2002) predicts that words that are more frequent than the bases they contain are accessed via the direct route, and that words that are less frequent than the bases they contain are accessed via decomposition. For example, a word like *inaccurate* (frequencies are from (Hay, 2002)) should be accessed via decomposition, since the derived frequency of *inaccurate* (53) is less than its base frequency (377). A word like *unleash*, on the other hand, should be processed as a whole word, since its derived frequency (65) is larger than its base frequency (16). These predictions were borne out in (Hay, 2001), where she asked subjects to provide judgments on relative complexity of pairs of words. She asked subjects to rate which word in a pair was more complex, one that is more frequent than its base or the one where the base is more frequent. Consistently, subjects rated words that are less frequent than their bases as more complex. Results of Hay's experiments in English are corroborated by Burani and

Thornton's (2003) results in Italian and Zuraw's (2009) results in Tagalog. According to Hay, relative frequency and derived frequency are highly correlated, and previous experimental results might be inconsistent because of this. Additionally, relative frequency plays a role in determining affix productivity. Hay and Baayen (2002) show that relative frequency is one of the most important factors in determining affix productivity, where affixes that are associated with more words whose derived frequency is less than their base frequency (and thus these words are presumed to be decomposed) are more productive. What this means for the dual route models is that access to the morphological route might depend on relative frequency of base and derived words. It is plausible that the prefix stripping results described above are contradictory because relative frequency was not taken into account in the design of those experiments.

In this paper I present an analysis of productivity of two Russian prefixes, *po-*, a very productive prefix, and *voz*/*vos*/*vz*/*vs-*, an unproductive prefix. This analysis shows that the correlation of base and derived frequencies and the slope and intercept of the regression line on these two variables are all important predictors of productivity of these two prefixes. In addition, results of a prefix separation experiment described below show that relative frequency is an important factor in morphological processing of Russian verbs, suggesting that the decomposition of the prefix out of the words depends on the relative frequency of the derived and base words. This evidence, together with previous results in English, Italian and Tagalog, suggest a universal principal of organization and should be taken into account by models of morphological processing.

One morphological model that is consistent with these results is the network theory (Bybee 1988, Langacker 2002), where the two units of storage are words and connections between them. Different factors affect the strength of those connections, including relative frequencies of words and their bases. Thus, for example, a word like *inaccurate* would have strong connections with its base, *accurate*, since *inaccurate* is less frequent than *accurate*. On the other hand, *unleash* would have weaker connections to its base, *leash*, since *unleash* is more frequent than *leash*.

2. *Po-* and *voz-* productivity analysis

2.1. Prefix descriptions

The two prefixes I chose for the productivity analysis are *po-* and *voz*/*vos*/*vz*/*vs-*. *Po-* only has one form, while *voz*/*vos*/*vz*/*vs-* has four allomorphs: *voz-*, *vos-*, *vz-*, and *vs-*. *Voz-* and *vz-* occur before vowels and voiced con-

sonants, while *vos-* and *vs-* occur before voiceless consonants. In the rest of the paper, I refer to the latter prefix as just *voz-* for simplicity. Intuitively, the two prefixes are different in their numeric characteristics and also in their meanings. Townsend (1975) lists the following meanings for the two prefixes:

Voz-:

1. Up: physical or abstract.
   vs-prygnut' 'to jump up'
   vos-pitat' 'to bring up'
2. Intensity or suddenness.
   vs-kriknut' 'to utter a sudden shriek'
   vz-boltat' 'to shake up'
3. Back.
   voz-vratit' 'to return'
   voz-obnovit' 'to renew'

Po-:

1. Begin to.
   po-nesti 'to start carrying'
   po-ljubit' 'to become fond of'
2. Do for a short time.
   po-sidet' 'to sit for a while'
   po-govorit' 'to have a talk'
3. Do somewhat, to some extent.
   po-lečit' 'to cure a little bit'
   po-veselit' 'to amuse somewhat'
4. Do from time to time and/or with diminished intensity.
   po-kurivat' 'to smoke from time to time'
   po-čityvat' 'to read a little bit from time to time'

In addition to the meanings listed above, the prefixes also have a 'pure' aspectual meaning, where it only adds perfective aspect to a verb (e.g., *slat'* 'to send' (impf.) and *po-slat'* 'to send' (pf.), *pomnit'* 'to remember' (impf.) and *vs-pomnit'* 'to remember' (pf.)).

We see that both prefixes have several well-defined meanings. However, my intuition is that there are more words where the meaning of the prefix is not clear for *voz-* than for *po-*. This intuition is confirmed in the next section where I analyze the numeric characteristics of these two prefixes.

2.2. Productivity analysis

In this section I perform a numeric analysis of the productivity of the prefixes *po-* and *voz-*. I find that relative frequency, along with other factors, is a good predictor of prefix productivity.

In his discussion of quantifying productivity of morphological units, Baayen (1992) lists the criteria of a good productivity measure: it should provide productivity rankings that correspond to linguistic intuitions (*intuitiveness*), it should reflect how well the morphological particle combines with new words (*hapaxability*), words with idiosyncratic properties should lower the productivity value (*idisyncraticness*) and it should reflect the fact that productivity does not simply equal the number of types associated with that morphological unit (*going beyond types*). In addition, as Hay and Baayen (2002) argue, the number of decomposed forms, or forms whose base frequency is larger than derived frequency, associated with an affix affects its productivity as well: the higher the number of those forms, the more productive the affix (*decomposed forms*).

I compare the two prefixes on a few of these criteria. First, intuitively, *po-* is much more productive than *voz-*. There are many more words with *po-* (4278) than with *voz-* (1236). Next, there are many more new words used with *po-* than with *voz-*. This is a notable characteristic, since one of the most important indicators of productivity of an affix is how readily it enters into new formations. In order to show that *po-* is used with new words more, I selected 47 verbs that entered the Russian language in the past two decades, mostly computer terms from the English language, such as *fludit'* 'to flood' and *frendit'* 'to friend' (on Facebook, Livejournal, etc.). Then I entered those verbs plus the prefix *po-* and *voz-* into a Russian search engine, Yandex[1], to see if any results appear. Since I was only interested in whether the prefixed neologisms exist in usage, I only needed to make sure that the words were not misspellings when there was a small number of returned results. The actual count of the occurrences was not important as long as it was above zero. The complete list of words used for this test is in Table 5. Out of the 47 verbs used for the test, 46 words, or 98%, are also used with *po-*, as evidenced by results of a Yandex search. In contrast, *voz-* is only used with 6 verbs out of 47 (or 13%). This is further evidence that *po-* is productive, while *voz-* is not.

Next, I studied the *po-* and *voz-* prefixed words based on relative frequency of the derived and base words. Relative frequency of base and

1. http://www.yandex.ru

derived words is important not only for morphological processing, as Hay (2001) argues, but also for affix productivity, as is discussed in Hay and Baayen (2002). Hay and Baayen analyzed 80 English affixes and plotted log derived versus log base frequency for them. Several factors are important for those affixes: correlation between the two variables, the slope of the resultant line, and its intercept. They argue that a positive and significant correlation between two variables, a higher intercept and steeper slope of the resultant line are all characteristics of a more productive affix. A positive and significant correlation is important, since the more transparent the relationship between bases and corresponding derived words, the more predictable the relationship between frequencies should be. A higher intercept and steeper slope of the resultant line effect in more points being above the x = y line, meaning more words where the derived form is less frequent than the base, i.e. words that are more prone to morphological decomposition.

To test whether relative frequency of the prefixed words is reflective of the prefixes' productivity, I plotted the words with existing unprefixed bases using their derived and base frequency. The calculations were done as follows: all words starting with the relevant letter sequence (*po*, *voz*, *vos*, *vz* or *vs*) were selected from the Russian orthographic dictionary[2], only prefixed words with those sequences were selected, and their frequencies were calculated using the main subcorpus of the Russian National Corpus[3]. Then the base frequency was calculated by stripping of the prefix and querying the RNC with the result. Overall, 70% (1944 out of 2755) of words used with *po-* are less frequent than their bases, and 54% (431 out of 787) of words used with *voz-* are less frequent than their bases.

To determine the correlation, intercept and slope for *po-* and *voz-*, I plotted derived versus base frequency for all the *po-* and *voz-* prefixed words, excluding the words with zero base or zero derived frequency. The resulting plots are shown in Figure 1 and Figure 2. For *po-*, the correlation between log base and log derived frequency is 0.18 ($p = 0$). The intercept of the regression line is 4.74 and the slope is 0.21. Thus, there is a positive and significant correlation between log base and log derived frequency of words with *po-*, the resulting regression line has a high intercept and a positive slope.

On the other hand, for *voz-* there is a positive, but not significant, correlation for log base and log derived frequency, 0.02 ($p = 0.50$). The intercept of the regression line is 4.37 and the slope is 0.04. Thus, the

2. http://ru.wikisource.org/wiki/Орфографический_словарь_русского_языка
3. http://www.ruscorpora.ru

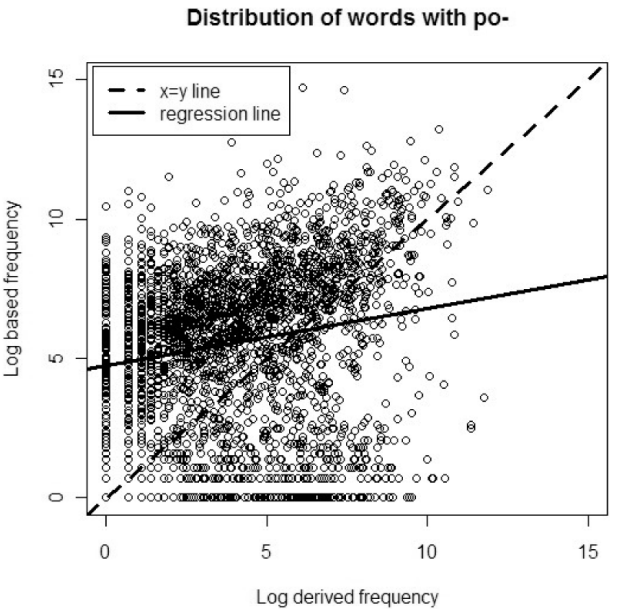


Figure 1. Plot of base versus derived frequency for *po-*

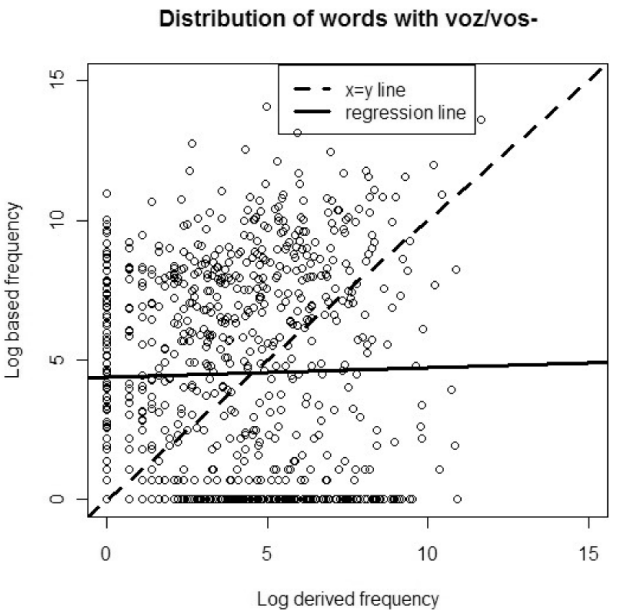


Figure 2. Plot of base versus derived frequency for *voz-*

correlation is not significant, the slope is much lower than for *po-*, although positive, and the intercept is also lower than for *po-*.

These data show that the proportion of words less frequent than their bases used with a particular prefix is a good predictor of prefix productivity.

To summarize, *po-* and *voz-* differ by all relevant parameters. There are many more words used with *po-* than with *voz-*, the correlation for base and derived frequency for *po-* is positive and significant, while it is positive but insignificant for *voz-*, borrowings combine freely with *po-* and almost not at all with *voz-*. Overall, this confirms the intuition that *po-* is productive, while *voz-* is not. In addition, we see that all measures we selected for the productivity analysis are well-suited: there is a difference between the prefixes in the expected direction in the overall number of words used, in the ratio of words more frequent than their bases to words less frequent than their bases and in the number of neologisms used with that particular prefix. Thus, we can conclude that these measures, and in particular relative frequency of words and their bases, are reliable in informing us of prefix productivity. Next I report the results of a prefix separation experiment with verbs with the prefix *po-* that show a difference in processing words that are more frequent than their bases and words that are less frequent than their bases.

3. *Po-* experiment

The purpose of this experiment was to establish whether or not relative frequency effects are present in Russian, using words with the very productive prefix *po-*. The task in the experiment was prefix separation. Participants were presented with verbs starting with *po*, both prefixed and not and words without the prefix *po-* and their task was to press 'yes' or 'no', depending on whether the prefix *po-* was present or not. There are several predictions about reaction times in this experiment.

Since this task required separating the prefix out of the word, the prediction is that the reaction time will be longer for those words that are generally *not* decomposed into constituent parts. Words that have a greater than derived base frequency are hypothesized to be decomposed, while words with smaller than derived base frequency are hypothesized to be processed as whole words. That means that the words whose base frequency is smaller than their derived frequency, are predicted to have longer reaction times than the words, whose base frequency is larger than their derived frequency. However, if, as Cole et al. (1989) argue, prefixes are never decomposed out of words containing them, there should be no

difference in reaction times between these two groups of words. Thus, a difference in reaction times would demonstrate the validity of two hypotheses: that prefixes are separated out of some morphologically complex words and that relative frequency is an important factor in morphological processing.

3.1. Methods

3.1.1. Participants

Forty-one native speakers of Russian from the San Francisco Bay area and New York City greater area participated in the experiment in exchange for payment.

3.1.2. Materials

The materials included 20 data items with the prefix *po*-, 25 fillers that start with *po*, but do not have the prefix and 25 fillers that do not contain *po* at all. Out of the 20 words with the prefix, 10 words were more frequent than their bases, and 10 were less frequent. The range of logarithm of frequency of words more frequent than their bases was from 2 to 4.4, of words less frequent than their bases 1.8 to 3.7. Data items are presented in Table 1 and fillers are presented in Table 6.

Table 1. Data items used in the experiment

Word	**Gloss**	**Base frequency**	**Derived frequency**
posramit'	'to disgrace'	253	367
pobagrovet'	'to redden'	242	486
pogubit'	'to ruin'	1,637	3,135
poxoronit'	'to bury'	2,333	3,218
pogrustnet'	'to become sad'	40	115
poprobovat'	'to try'	5,211	12,272
postupit'	'to act' (pf.)	492	2,894
postupat'	'to act' (impf.)	6,654	22,596
potušit'	'to put out'	1,439	941
pobrezgovat'	'to disdain'	533	144
počmokat'	'to give smacking kisses'	271	61
poxlopat'	'to clap'	2,936	1,031
podobret'	'to become kinder'	273	131
povesit'	'to hang'	6,653	4,961
porožđat'	'to give birth'	13,251	4,541

3.1.3. Design

One word was presented at a time, with a new randomized sequence of presentation for each subject.

3.1.4. Procedure

Participants were tested individually on a laptop computer. Both the instructions and the experiment were exclusively in Russian. The instructions explained what constitutes a prefix and how to answer questions. The concept of prefix was explained by showing that the word *izbegat'* 'to avoid' contains the prefix *iz-* and the root *-beg-*.[4] Two sample questions were shown before going on to the experiment. Participants were then presented with the stimuli, one word at a time, and were asked to press *da* 'yes' if the word contained the prefix *po-* and *net* 'no' if it did not, and to do it as fast and as accurately as possible. A new random order of stimuli was shown to each participant. The word stayed on the screen until the participant pressed 'yes' or 'no'.

3.2. Results

Reaction times higher than three standard deviations from the mean were discarded. Four items were excluded. These items contained the reflexive suffix *-sja*, which made the morphological structure of those words more complex and thus harder to analyze. One item (*poumnet'* 'to become smarter') was excluded because it was the only item whose base started with a vowel, and contained a V-V transition between the prefix and the root, an extremely unlikely within-morpheme transition. This item's average reaction time was 1546 ms, almost 500 ms less than the average of all the other items. This is according to expectations; an item containing an extremely unlikely within-morpheme phonotactic transition is expected to be decomposed easier than other items (Hay, 2002). Thus, this item was excluded. After this exclusion, there were 7 items less frequent than their bases and 8 items more frequent than their bases. Results of three subjects were excluded because of high error rates (more than 25%).

4. A reviewer asks whether the instructions give the purpose of the experiment away. While I did explain what a prefix is, and ask the subjects to separate it out of the word, I could have not possibly influenced their reaction times in the experiment, if it is dependent on frequency.

Table 2. *Po-* experiment results, mean analysis

	Mean RT		**p-value**
	Word freq > base freq	**Base freq > word freq**	
By subject	2373 ms	2097 ms	<0.001
By item	2375 ms	2093 ms	0.048

Two analyses were performed on *po-* prefixed data, a mean analysis and a mixed regression analysis, in order to evaluate which other factors might have influenced the RT. In the mean analysis results were analyzed by item and by subject. The mean analysis results are summarized in Table 2. They are represented graphically in Figure 3 and Figure 4.

What we see from the mean analysis is that there is a difference between the two sets of words, significant both by subject and by item (the borderline *p*-value of the by-item analysis might be attributed to the small number of items). To investigate further, I carried out a multiple regression statistical analysis with the logarithm of reaction time as the dependent variable and subject and item as random effects. I performed the analysis according to (Crawley 2007), and retained all the factors

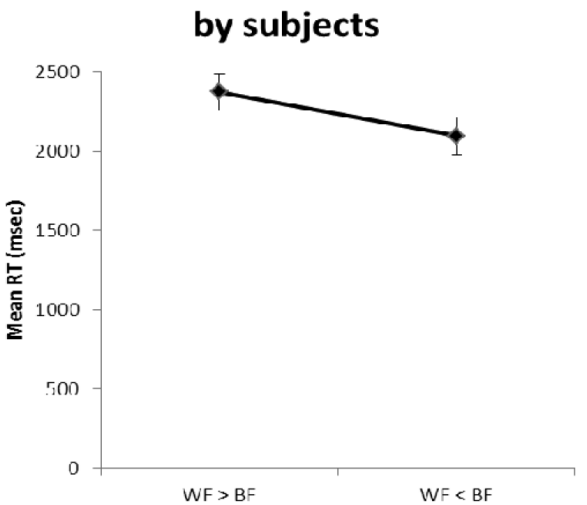


Figure 3. Mean RT by subjects

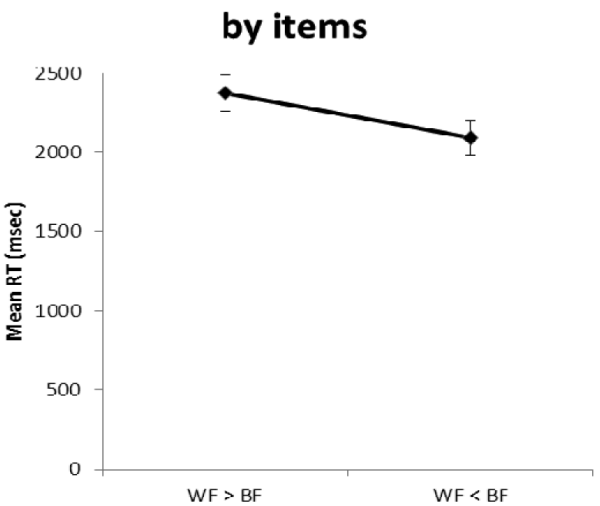


Figure 4. Mean RT by items

whose p-values were under 0.1. While designing the model, I input factors that could have affected the response time, including information about frequency, family size, semantic transparency, interactions of these factors with accuracy of response, trial number and phonological and orthographic information. Hay (2002) showed that phonological transitions can affect morphological decompositionality, thus word length (in letters and syllables) and the prefix-root transition (VCC or VCV) were included as possible influencing factors in the model. Word frequency was included in the model, as it has been shown to be an important factor in morphological processing (Bybee 2007). Family size has been shown to affect morphological processing of English words (e.g., Baayen, Lieber and Schreuder 1997), even monomorphemic ones, and thus was included as a possible influencing factor. Semantic transparency has also been shown to affect morphological processing (Wurm 1997), and thus it was included in the model. Semantic transparency was calculated as follows: after inputting each word into the dictionary on http://www.gramota.ru, I counted the number of unprefixed words with the same root appear in the definition. This procedure is similar to the calculation of semantic transparency in (Hay 2001), and the reasoning is that a word that is more semantically transparent should include more words with the same root in its definition than a word that is less semantically transparent. Finally, to test whether relative frequency of derived and base words is important in morphological proc-

essing, I included the difference between the logarithms of derived and base frequencies as a possible influencing factor.

The factors that are included in the model are accuracy (accurate answers were faster), semantic transparency (semantically transparent items were faster), trial number (the later in the experiment the item was, the faster was the reaction time), unprefixed family size (a small inhibitory effect), and relative frequency (words that are more frequent than their bases were reacted to slower than words less frequent than their bases). The interactions of the included factors with accuracy did not turn out to be significant. Relative frequency and unprefixed family size are marginally significant ($p = 0.08$), and that might be again attributed to a small number of items. The resulting model is shown in Table 3 (fixed effects) Table 4 (random effects).

Table 3. Fixed effects of the mixed effects regression model

Factor	**Estimate (Log RT)**	**p-value**
Intercept	7.6743	$p = 0.0000$
Accuracy	0.3441	$p = 0.0014$
Unprefixed family size	0.0024	$p = 0.0757$
Semantic transparency	−0.1114	$p = 0.0050$
Base-derived frequency difference	−0.0376	$p = 0.0788$
Trial number	−0.0020	$p = 0.0018$

Table 4. Random effects of the mixed effects regression model

Random effect	**Variance**	**Standard deviation**
Subject number	0.0805	0.2838

The R^2 of this model is 0.39, compared to the R^2 of the null model, which is 0.38. Although the increase in R^2 is relatively small, the variance for the adjustment by item is reduced by 100% to 0 (and thus is taken out of the model), while the variance for the adjustment by subject is reduced by 6%. This means that the fixed effects model explains a little more variance than the null model, but now the same amount of variance is explained with fixed effects instead of the random effects of subject and item number.

4. Discussion

Overall, both the mean analysis and the mixed-effects regression models confirm that there are differences between how words that are more frequent than their bases and words that are less frequent that their bases are processed, although the experiment needs to be replicated with more items. It is easier to separate the prefix from words that are less frequent than their bases. A better predictor is the difference between the frequency of the base and the derived word, where the larger the difference, the easier it is to separate the prefix. I will take this as evidence that the relative frequency effect is present. Thus, the prediction that there are relative frequency effects in Russian is borne out.

Another factor that turned out to be significant in the model was unprefixed family size, with a small inhibitory effect. We might hypothesize that that stems from a strategy by subjects to make a lexical decision on the unprefixed base: the word starts with *po-*, and if the unprefixed base is a word, there is a prefix in that word. Usually, a facilitatory family size effect is observed (e.g. Baayen, Lieber and Schreuder 1997), and Wurm (1997) cites U-shaped family size effects in lexical decision tasks, where initially large family size is to the advantage, and inhibits reaction time later on. The reasoning underlying this effect is as follows. In the early stages of lexical decision, a large family size is facilitatory, as it raises the probability that the string is a word, while in later stages, where the exact identification of the word is necessary, a large family size makes the probability of that particular word low, and thus is inhibitory. Here we see an inhibitory family size effect, and we might hypothesize that it is due to the fact that the unprefixed base is already very word-like, since it is a part of another word, and only the exact identification of the string is necessary, where a large family size is inhibitory. This is an interesting question for a future more thorough investigation.

The last important factor in this prefix separation experiment is semantic transparency. If the word was semantically transparent (or its unprefixed relatives appeared in the definition in the dictionary on http://www.gramota.ru), it was easier to decompose the prefix out of it. A clear semantic connection makes lexical connections between words stronger and the word parts easier to discern.

The experimental results clearly show that some words are reacted to faster than others, and that is evidence for morphological decomposition of prefixes out of words, at least in some cases. Since in the experiment the participants were asked to answer ‘yes’ or ‘no’ to the question ‘Does

this word contain the prefix *po*?', the difference in reaction times suggests that in some words the prefix is separated out more easily than in others. This finding goes against previous findings by Cole et al. (1989), where they found that suffixes are decomposed out of words, while prefixes are not. It is possible that previous prefix separation experiment results would be reinterpreted if relative frequency were to be taken into account.

These results agree with Cole et al.'s (1997) findings for French, Hay's (2001) findings for English, Zuraw's (2009) findings for Tagalog and Burani and Thornton's (2003) findings for Italian: relative frequency of base and derived word is an important factor in morphological processing. This experiment, using Russian and a different experimental paradigm, adds cross-linguistic evidence to the previous results.

As Stemberger and MacWhinney (1988) argue, frequency effects arise from storage. The result we see in this experiment is the larger the difference between the base and the derived frequency, the easier it is to separate the prefix. Thus, there must be a difference in storage of words that are more frequent than their bases and words that are less frequent than their bases. There are at least three possibilities as to how the words might be stored. One possibility is that words that are less frequent than their bases are stored decomposed, while words that are more frequent than their bases are stored as whole words. Another possibility is that there are two representations of a word, a whole-word one and a decomposed one, and the one accessed is the more frequent one. The last possibility is that there is only one representation of a word, a whole-word one, and that the stronger the links to the unprefixed base, the easier the decomposition. The difference in storage in this third option is the strength of the connections between words. The difference between base and derived frequencies was a better predictor in the mixed-effects model than the dichotomous division into two sets of words, one where the base frequency was larger than derived frequency, and another where base frequency was smaller than derived frequency. Thus, the first option is not optimal. The difference between word storage should reflect the continuous difference between base and derived frequency, and not be a dichotomy where all words more frequent than their bases are stored as whole words, while all words less frequent than their bases are stored decomposed. The two other options, on the other hand, reflect the continuous change in frequency difference. Consider the details of the two options, using the Russian word *pohlopat'* 'to clap' (pf.) and *pohoronit'* 'to bury' (pf.) as examples.

Pohlopat' has the frequency of 1031 and *hlopat'* 'to clap' (impf.) has the frequency of 2936. Thus, *pohlopat'* is less frequent than its base and is

more likely to be decomposed. On the other hand, *pohoronit'* has the frequency of 3218, while *horonit'* 'to bury' (pf.) has the frequency of 2333. Thus, *pohoronit'* is more likely to be processed as a whole word. In the two representations option, *pohlopat'* has two representations: *pohlopat'*, with a frequency of 1031, and a decomposed representation. If the decomposed representation is just the prefix *po-* and *hlopat'* then its frequency is the same as the frequency of *hlopat'*. However, *hlopat'* is not monomorphemic and can be further decomposed into the root *-hlop-*, the theme vowel *-a-* and the infinitive suffix *-t'*. Many additional questions arise, such as is the word stored exhaustively decomposed and why the difference between the base frequency of *hlopat'* and the derived frequency of *pohlopat'* is a factor in processing of *pohlopat'*, and not the frequencies of the individual parts.

Several other studies shed light on this question. Antić (2007) performed a prefix separation experiment, where stimuli were used that did not necessarily exist unprefixed, and base frequency was calculated in two ways: one, as above, the frequency of the base as a standalone word, and another, named 'stem frequency', the frequency of the base as a standalone word plus its frequency as it appears in other words. For example, the word *poručat'* 'to commission' does not exist unprefixed, and thus its base frequency is 0, but it appears with other prefixes (e.g. *vyručat'* 'to rescue'), and thus its stem frequency is the sum of frequencies of all words where *-ručat'* appears as the base. There was no significant difference in reaction time between words whose stem frequency was higher than word frequency and words whose stem frequency was lower than word frequency. From this we can conclude that the word is not stored just as prefix and the base in decomposed form. If it is not the prefix and the base in decomposed form, then the decomposed form must be an exhaustive breakdown of all the parts. Continuing with the example above, *pohlopat'* 'to clap' (pf.) must be stored in two representations, *pohlopat'* and *po-hlop-a-t'*. I suggested above that which representation is chosen depends on the frequencies of the two representations. It is, however, not clear how to calculate the frequency of the decomposed representation. If it is the frequency of the base *hlopat'*, what is the connection between *po-hlop-a-t'* and *hlopat'*? Another possibility is some combination of frequencies of the constituent morphemes, such as the cumulative root frequency of *hlop-* and the frequency of *po-*. However, in the present experiment neither cumulative root frequency nor the difference between cumulative root frequency and word frequency were predictors in the mixed effects models. This is in accordance with other studies, such as (Baayen, Liber and

Shreuder 1997), where the authors find that the processing of monomorphemic words depends on the family size, but not cumulative root frequency. In addition, Wurm (1997) in an auditory gating lexical identification experiment found that prefix frequency had an inhibitory relationship with the time it took to identify the word. Thus, words with a more frequent prefix were identified later than words with a less frequent prefix. If prefix frequency were a significant factor in decomposed morphological processing, it would have a facilitatory influence on reaction time. Thus, both cumulative root and prefix frequency are poor predictors of decomposition. Hence I conclude that the option where there are two representations of a word, one decomposed and one as a whole word, is not viable.

The final option is the option of one whole word representation. I assume that this representation contains phonological (and, presumably, orthographic) and semantic information about the word. The question then arises, what is the difference in storage between words that are more frequent than their bases versus words that are less frequent than their bases? Framing the experiment results in the Network theory of morphology (Bybee 1988, Langacker 2002) gives a fitting answer. In this theory lexical entries are word-based, there are form and semantic connections between words, and the strength of these connections depends on frequency. Examples are illustrated in Figure 5 and Figure 6 (*ponižat'* 'to lower', *nižnij* 'lower' (adj.), *nanizyvat'* 'to string', *povyšat'* 'to raise'). The larger the difference between base and word frequency, the stronger the connections between the word and its unprefixed base. Thus, in the examples above, the connections between *pohlopat'* and *hlopat'* would be strong, while the

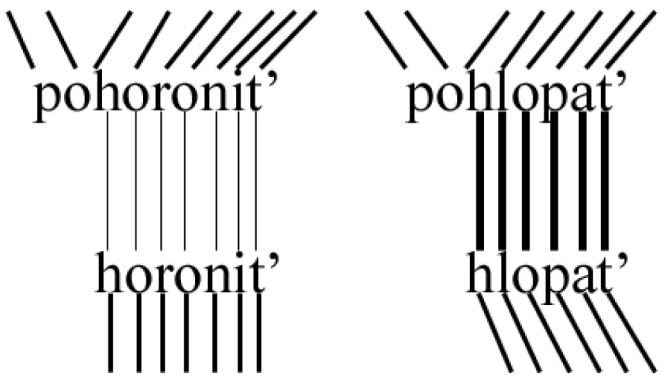


Figure 5. Difference in strength of connection

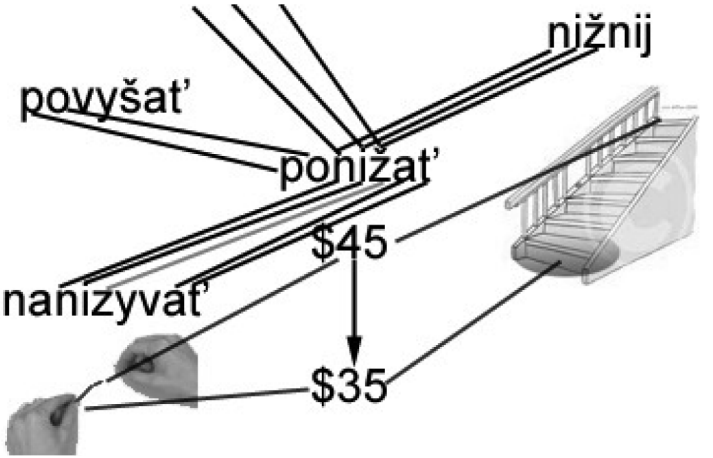


Figure 6. Lexical connections between words in the Network theory of morphology

connections between *pohoronit'* and *horonit'* would be weaker, accounting for the difference in reaction time in the prefix separation experiment. This is shown in Figure 5, while Figure 6 shows how several words might be organized in the network model.

Representing the results in this theoretical framework is consistent with the finding that the difference between base and word frequency was a better predictor in the mixed effects regression model than the simple dichotomous distinction between words more or less frequent than their bases, since difference in frequency affects the lexical connections directly. In addition, this theoretical framing is also consistent with previous findings of the role of family size in morphological processing (e.g., Wurm 1997, Baayen, Lieber and Schreuder 1997). Family size is an important factor in morphological processing, which means that morphological (and orthographic; Rastle, Davis and New 2004; McCormick, Rastle and Davis 2008) neighbors of a word are activated when that word is processed. A model where there are lexical connections between words that depend on frequency of the words on the two ends of a connection assumes precisely this. When a word such as *pohlopat'* is processed, its base *hlopat'* is activated because of a strong connection between the two words, and it is easier to decompose the word. On the other hand, when a word such as *pohoronit'* is processed, its base *horonit'* is activated less easily because of a weaker connection between the two words.

Apart from the question of how the words are stored, the issue of processing is also important. As described in the beginning of the paper, there are several dual route processing models. Caramazza et al. (1988) suggest a model where the decompositional route is accessed only by novel words, while the whole-word route by known words. However, even if the word is accessed via the decompositional route, its morphological representation is activated. For example, *walked* is accessed via the whole-word route, and it activated the representations of its morphemes, *walk*$_V$- and *-ed*. In the light of the current results, the model would need to be modified to take into account relative frequency effects for known words. The race model of Frauenfelder and Schreuder (1992), where the two routes race, would need to be modified, where the likelihood of activation of the decompositional route for morphologically complex words would depend on the relative frequency of the word and its base. Finally, Wurm's (1997) model, where there is an obligatory whole-word route and a decompositional route that is selective about which words it considers, would also need to be modified, and the decompositional route might be accessed only when the relative frequency difference is large enough.

5. Conclusion

To summarize, in this paper I presented the results of an experiment that confirm the existence of relative frequency effects in Russian, where the prefix was separated more easily from words that were less frequent than their bases. The larger was the difference between the base and the derived frequency, the easier it was to separate the prefix. These results need to be replicated in future studies with more items. Relative frequency effects are also confirmed to be an important factor in determining affix productivity, where an affix that is associated with more words that are less frequent than their bases was more productive. These results add to cross-linguistic evidence of relative frequency effects and suggest a universal principle of lexical organization.

Different options of word representation were considered, and the one whole word representation option, couched in the Network morphology framework, was found to be more theoretically plausible. The results described in the paper call for a morphological processing model where relative frequency is taken into account, and a decompositional route is more or less likely in a dual-route model depending on the relative frequency of the word and its base.

References

Antić, Eugenia
2007 Relative and absolute frequency in Russian verbal morphology and processing. Unpublished manuscript, University of California-Berkeley.

Antić, Eugenia
Forthcoming *The Representation of Morphemes in the Russian Lexicon*. Ph.D. dissertation, Department of Linguistics, University of California-Berkeley.

Baayen, Harald
1992 Quantitative aspects of morphological productivity. In: Geert E. Booij and Jaap van Marle (eds.), *Yearbook of Morphology*, 109–149. Kluwer Academic Publishers.

Baayen, Harald, Rochelle Lieber and Robert Schreuder
1997 The morphological complexity of simplex nouns. *Linguistics* 35: 861–877.

Burani, Cristina and Anna M. Thornton
2003 The interplay of root, suffix and whole-word frequency in processing derived words. In: Harald Baayen and Robert Schreuder (eds.), *Morphological structure in language processing*, 157–207. Mouton de Gruyter.

Butterworth, Brian
1983 Lexical representation. In: Brian Butterworth (ed.), *Language production*, 257–294. Academic Press.

Bybee, Joan
1988 Morphology as lexical organization. In M. Hammond and M. Noonan (eds.), *Theoretical Morphology*, 119–140. Academic Press.

Bybee, Joan
2007 *Frequency of use and the organization of language*. Oxford University Press.

Caramazza, Alfonso, Alessandro Laudanna and Cristina Romani
1988 Lexical access and inflectional morphology. *Cognition* 28: 297–332.

Cole, P., C. Beauvillain and J. Segui
1989 On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language* 28:1: 1–13.

Cole, P., J. Segui and M. Taft
1997 Words and morphemes as units for lexical access. *Journal of Memory and Language* 37: 312–330.

Crawley, Michael J.
2007 The R Book. John Wiley and Sons Ltd.

Feldman, Laurie
1994 Beyond orthography and phonology: Differences between inflections and derivations. *Journal of Memory and Language* 33: 442–470.

Frauenfelder, Uli H. and Robert Schreuder
1992 Constraining psycholinguistic models of morphological processing and representation: the role of productivity. In: Geert Booij and Jaap van Marle (eds.), *Yearbook of morphology*, 165–183. Kluwer Academic Publishers.

Giraudo, Hélène and Jonathan Grainger
2003 On the role of derivational affixes in recognizing complex words: Evidence from masked priming. In Harald R. Baayen and Robert Schreuder (eds.), *Morphological structure in language processing*, 209–232. Mouton de Gruyter.

Hay, Jennifer
2001 Lexical frequency in morphology: is everything relative? *Linguistics* 39: 1041–70.

Hay, Jennifer
2002 From speech perception to morphology: affix ordering revisited. *Language* 72: 527–555.

Hay, Jennifer and Harald Baayen
2002 Parsing and productivity. In: Geert Booij and Jaap van Marle (eds.), *Yearbook of morphology*, 203–235. Kluwer Academic Publishers.

Langacker, Ronald
2002 *Concept, image and symbol: the cognitive basis of grammar*. Berlin: Mouton de Gruyter.

Losiewicz, Beth
1992 The effect of frequency on linguistic morphology. Ph.D. dissertation, Department of Linguistics, University of Texas-Austin.

Marslen-Wilson, William D., Mike Ford, Lianne Older and Xiaolin Zhou
1996 The combinatorial lexicon: Priming derivational affixes. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*. La Jolla, California: Lawrence Erlbaum Associates.

McCormick, Samantha F., Rastle, Kathleen and Davis, Matthew H.
2008 Is there a 'fete' in 'fetish'? Effects of orthographic opacity on morpho-orthographic segmentation in visual word recognition. *Journal of Memory and Language* 58: 307–326.

Rastle, Kathleen, Davis, Matthew H. and New, Boris
2004 The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review* 11(6): 1090–1098.

Reid, Agnieszka Anna and William David Marslen-Wilson
2003 Lexical representation of morphologically complex words: evidence from polish. In: Harald Baayen and Robert Schreuder

(eds.), *Morphological structure in language processing*, 287–336. Mouton de Gruyter.

Schreuder, Robert and Harald Baayen
1994 Prefix stripping re-visited. *Journal of Memory and Language* 33: 357–375.

Stemberger, Joseph Paul and Brian MacWhinney
1988 Are inflected forms stored in the lexicon? In: M. Hammond and M. Noonan (eds.), *Theoretical morphology*, 101–116. Academic Press.

Taft, Marcus
1985 Recognition of affixed words and the word frequency effect. *Memory and Cognition* 7: 263–272.

Taft, Marcus and Kenneth I. Forster
1975 Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior* 14: 638–647.

Townsend, Charles E.
1975 Russian word-formation. Columbus, OH: Slavica Publishers.

Wurm, Lee H
1997 Auditory processing of prefixed english words is both continuous and decompositional. *Journal of Memory and Language* 37: 438–491.

Zuraw, Kie
2009 Frequency influences on rule application within and across words. In *Proceedings of the 43rd Annual Meeting of the Chicago Linguistic Society*.

Appendix A. Table of neologisms

Words used for the productivity test

Word	Gloss	Number of Yandex results
avangardit'	'to do something vanguard'	38
burokratit'	'to red tape'	297
gangsterit'	'to be a gangster'	35
gejmit'	'to game'	2,680
guglit'	'to google'	162,000
dajvit'	'to SCUBA dive'	2,349
developit'	'to develop'	7,099
diversificirovat'	'to diversify'	795,000
dizajnit'	'to design'	20,000
imejlit'	'to e-mail'	43
investit'	'to invest'	976
indeksit'	'to index'	1,769
insajdit'	'to earn using inside information'	368
integrirovat'	'to integrate'	9,000,000
kastigovat'	'to cast'	3,222
kvotirovat'	'to impose a quota'	69,000
klonirovat'	'to clone'	1,000,000
kommercializirovat'	'to commercialize'	92,000
kompilit'	'to compile'	93,000
konsaltit'	'to consult'	472
kreativit'	'to do something creative'	103,000
kserit'	'to copy' (on a copy machine)	55,000
lizingovat'	'to lease'	6,400
liftingovat'	'to do face lifting'	2,187
pilingovat'	'to do face peeling'	7,987
piratit'	'to pirate'	20,000
pressit'	'to pressure'	1,050
provajdit'	'to provide'	865
programmit'	'to program'	146,000
rejtit'	'to rate'	234
rekrutit'	'to recruit'	1,081
roumit'	'to roam'	657
servisit'	'to service'	456
skanit'	'to scan'	67,000
skrabit'	'to do body scrubbing'	12,000
spamit'	'to spam'	477,000
tʲuningovat'	'to tune up'	650,000
fludit'	'to flood'	2,000,000

frilansit'	'to freelance'	33,000
šejpingovat'	'to do fitness'	2,278
esemesit'	'to sms'	2,895
kommentit'	'to comment'	476,000
frendit'	'to friend'	40,000
daunlodit'	'to download'	1,153
apgrejdit'	'to upgrade'	207,000
juzat'	'to use'	1,000,000
jandeksit'	'to search using Yandex.ru'	3,890

Appendix B. Filler items used in the experiment

Filler items used in the experiment

Word	**Gloss**
pozvolit'	'to allow, permit'
pozirovat'	'to pose'
pozorit'	'to disgrace'
pokoit'sja	'to rest'
polzat'	'to crawl'
polirovat'	'to polish'
polnet'	'to gain weight'
poloskat'	'to rinse'
polosnut'	'to slash'
polučat'	'to receive'
polyhat'	'to blame'
polzovat'sja	'to use'
porot'	'to whip'
poročit'	'to defame'
portit'sja	'to become spoiled'
poseščat'	'to visit'
potet'	'to sweat'
potčevat'	'to entertain'
podbadrivat'	'to cheer on'
podbegat'	'to run to'
podbit'	'to line with'
podkaraulit'	'to be on watch'
podkatit'	'to roll up'
podkačat'	'to pump'
podkinut'	'to toss'
skazat'	'to say'

razuznat'	'to find out'
videt'	'to see'
stojat'	'to stand'
sprosit'	'to ask'
smotret'	'to watch'
ponjat'	'to understand'
vysidet'	'to sit out'
sdelat'	'to do'
kazat'sja	'to seem'
ostanovit'sja	'to stop'
iskat'	'to look for'
razuverit'	'to dissuade'
zabežat'	'to run in'
priexat'	'to come'
nakričat'	'to yell'
otkryt'	'to open'
proizojti	'to happen'
sjezit'	'to go and come back'
prijti	'to come'
sobrat'sja	'to pack, prepare'
uslyšat'	'to hear'
slučit'sja	'to happen'
starat'sja	'to try'
kupit'	'to buy'

Frequency, conservative gender systems, and the language-learning child: Changing systems of pronominal reference in Dutch

Gunther De Vogelaer

Language change is well-known to show frequency effects. Depending on the mechanism of change that is observed, frequent items may lead a change or lag behind in it (see, e.g., Bybee and Hopper 2001: 10–19 for discussion). This chapter discusses shifts in the gender system of East and West Flemish dialects of Dutch. It is shown that at least two mechanisms of change are at work, viz. standardisation, which causes lexical items to adopt the gender of their Standard Dutch counterpart, and resemanticisation, i.e. a tendency in Dutch to replace the 'grammatical' system of pronominal reference with a system operating on a semantic basis, in which highly individuated nouns trigger the use of etymologically masculine pronouns (*hij* 'he', *hem* 'him'), whereas weakly individuated nouns are referred to with neuter *het* 'it' (Audring 2006). It is investigated to what extent frequency data can be used to disentangle the effects of standardisation and resemanticisation. Data from a questionnaire survey show that standardisation affects high-frequency items, whereas resemanticisation affects low-frequency items. In addition, differences are found with respect to the type of frequency data that provide the best match for the data. For standardisation, frequency data extracted from the Spoken Dutch Corpus (CGN) provide the best results, whereas resemanticisation is better predicted using a frequency measure capturing age of acquisition and usage frequencies in child language. This underscores that frequency effects often merely reflect some deeper property of language patterns rather than being a conclusive explanation in their own right. In this chapter, frequency effects in standardisation reflect the intensity to which dialect speakers are exposed to nouns' standard language gender, whereas the frequency effects in resemanticisation reveal different ages at which nouns are acquired by children, which appears to influence the odds that these nouns' grammatical gender can be learned successfully.

1. Introduction: change and variation in Dutch gender, and frequency

Present-day Standard Dutch differs from historical varieties of the language in that the difference between the marking of masculine and feminine gender is levelled out, yielding a dyadic gender distinction between so-called 'common' and neuter gender. For instance, Standard Dutch has only two definite articles (common *de* vs. neuter *het*) and only distinguishes between common and neuter nouns in adjectival inflection in indefinite NPs (e.g. *een mooi-e man/vrouw* 'a beautiful man/woman' vs. *een mooi kind* 'a beautiful child'). This creates a mismatch between the (dyadic) adnominal system and pronominal gender, where the three-way distinction between masculine, feminine and neuter pronouns is preserved. This mismatch seems to have given rise to a reshuffle of pronominal gender, especially in reference to inanimates: while pronominal gender traditionally matched the grammatical gender of the antecedent inanimate noun, northern varieties of Dutch, including Standard Dutch as spoken in the Netherlands, seem to be shifting towards a semantic system of pronominal gender, operating along the lines of the Individuation Hierarchy (Siemund 2002; Audring 2006, 2009): highly individuated nouns (including neuter words such as *masker* 'mask' and *apparaat* 'device'; cf. Audring 2009: 86) increasingly trigger the use of masculine pronouns such as *hij* 'he' or *hem* 'him', weakly individuated ones (including common nouns such as *spinazie* 'spinach' and *wol* 'wool'; cf. Audring 2009: 98) combine with neuter *het* 'it'.

Significantly, contemporary varieties of Dutch display variation with respect to resemanticisation: while the process has advanced considerably in some varieties, other varieties by and large maintain a grammatical system of pronominal reference. This chapter focuses on pronominal gender in a number of varieties of Dutch in which the grammatical gender system still stands strong, more specifically on West and East Flemish dialects. In these dialects any instances of semantically-motivated pronouns are highly ambiguous with respect to the mechanism of language change explaining them: these instances may exemplify ongoing change within these varieties, but they may also be adopted from varieties of Dutch in which semantic agreement occurs more often. In addition, not all changes in the choice of a pronoun referring to an antecedent noun are due to resemanticisation. Apart from resemanticisation there is also variation in that many nouns have a different gender in the traditional dialects than in the standard language. In more recent times, extensive levelling is causing these dialects to converge to Standard Dutch, so it is likely that many nouns having a different gender in the dialect than in Standard Dutch are under pressure

to switch gender. Such gender shifts, of course, make an investigation of the dialects even more interesting, since this adds a dimension of variation that is not present in varieties of the standard language. But it also makes the task of disentangling which mechanisms of change are operating in these dialects very challenging.

In cases such as this, where different mechanisms of change interact, frequency effects may cast light on which part of the changes is explained by which mechanism of change. Indeed depending on the type of change that is observed, high vs. low frequency items are affected first. One well-known hypothesis regarding frequency is that conservative features in language are preserved longer in high frequency items (see, e.g., Bybee and Hopper 2001: 17–18; Corbett, Hippisley, Brown, and Marriot 2001; Smith 2001). According to Phillips (2006: 87), this characterisation holds for all changes that are implemented in cases 'when memory fails', for instance in sound changes affecting words of which the phonetic word form is not well entrenched in memory, which drives speakers to choose pronunciations motivated by surface phonetics, pronunciations analogous to other patterns in the language, or, in general terms, innovations requiring "access to generalisations that have emerged from word forms" (Phillips 2006: 157). Changes directly involving the production of word forms, however, affect the most frequent words first (e.g., deletion, assimilation, . . .).

From the hypothesis that infrequent items are likely to be affected by innovations motivated by generalisations that have emerged from word forms, it follows that Phillips' generalisation typically holds in situations where the innovation originates within a speech community. Thus in situations of innovation diffusion through contact with other varieties, other regularities may be at work. At the moment there are contradictory opinions in the literature, however, as to whether dialect contact leads to change especially in high or low-frequency items. The most widely held opinion seems to be that exposure, and hence high frequency, increases the likelihood of change. For instance, Trudgill (1986) considers change through dialect contact to be a kind of 'long-term accommodation' which basically patterns like accommodation in conversation. It is claimed that in accommodation between adults, salient properties of the donor dialect are more likely to be adopted than non-salient ones, since accommodating salient items is a more effective means to achieve accent convergence in conversation. As factors contributing to a pattern's salience, Trudgill (1986: 11–21) lists, among others, phonetic distance, the relation between a variant and orthography, or whether a variant is involved in a change in progress. In addition, structural (e.g., phonotactics) and functional (e.g.,

homonymy avoidance) factors play a role. Other factors being equal, high frequency obviously increases a pattern's salience, which raises the hypothesis that high-frequency items are more likely to be involved in processes of (short and long term) accommodation. However, the reverse correlation has been proposed as well, viz. that in accommodation "words learned at the mother's knee, so to speak, would be the most conservative, while the least frequent words would be affected first" (Bybee 2000: 82), simply because the latter are less entrenched in the mind of the speaker. In an attempt to reconcile the two positions, Phillips (2006: 141), following L. Milroy (2003), distinguishes between ideologically motivated and ideologically free changes. Depending on attitudinal factors, ideologically motivated changes typically affect words from a certain register (e.g., formal or rather informal vocabulary) rather than high or low-frequency items. Ideologically free changes behave as changes emerging within a speech community, i.e. whether they affect high or low frequency items first is determined by the nature of the change: changes directly involving the production of word forms affect the most frequent words first; changes being implemented 'as memory fails' first affect low frequency items (Phillips 2006: 157).[1]

Given that there is at least some agreement on the role of frequency in different types of language change, the first goal of this chapter is to explore to what extent frequency effects reveal which mechanisms of language change are observed in the gender system of West and East Flemish dialects. Second, this chapter aims to provide insight into which frequency data need to be used to obtain an optimal 'fit' between frequency and its role for language change. In a paper on different mechanisms of language change, Labov (2007) claims that different mechanisms of language change

1. Phillips (2006) reaches this conclusion in an inductive manner, by generalising over a large set of examples of language change. She does not, however, provide a principled account of why high-frequency items are more liable to change involving mere word forms, even though they are allegedly more entrenched in language users' minds. One principled reason could be that types of contact that do not affect high-frequency items are too weak to have any effect at all. Alternatively, it could also be the case that effects are observed, but not on the community level. Thus, some low frequency items may be affected by the contact situation in the language of a number of individual language users, but these effects disappear if the data for individuals are pooled in larger data sets. It would take an experimental setting to verify whether such an account is plausible.

are pushed forward by different cohorts of language users. Change independently originating within a certain variety is typically due to the imperfect transmission of language from one generation to another, whereas dialect contact predominantly takes place between adults. This hypothesis yields important predictions with respect to the type of frequency data that need to be used to reveal the role of frequency in a particular diachronic development. If language change is due to imperfect transmission between generations, it may be worthwhile to draw frequency data from corpora of non-adult language usage, whereas this would be less useful in cases of dialect contact taking place between adults.

Empirically speaking, this article compares the results of a late 19th century survey on gender in the dialects (Pauwels 1938) with recent data from the Belgian provinces of East and West Flanders. The article is structured as follows: after a description of the data and a number of methodological preliminaries in section 2, section 3 identifies the two most important developments in the gender systems of East and West Flemish dialects: a) influence from Standard Dutch; and b) resemanticisation along a similar pathway as observed in present-day northern varieties of Dutch. Section 4 focuses on the role of 'frequency' in both developments, and argues that both cases require other methods of establishing usage frequencies, in line with the 'locus' of language change in either case. Section 5 concludes this article.

2. Investigating gender in East and West Flemish dialects of Dutch

2.1. Gender in Dutch: the progressive north vs. the conservative south

The Dutch gender system has been undergoing change for centuries, thereby gradually decreasing the number of exponents of the grammatical three-gender system observed in the oldest documented varieties of the language: while Middle Dutch case inflection of articles, adjectives and nouns themselves revealed whether a given noun was masculine, feminine or neuter, present-day varieties of Dutch have dispensed with most of their adnominal morphology. Thus, case marking has gone and little gender agreement is left (cf. Geerts 1966). The processes of change have unevenly affected different varieties of Dutch. More particularly, they have resulted in massive geographical variation in the domain of gender marking at the level of the dialects (as described most recently in De Schutter et al. 2005), and also in smaller differences between varieties of the standard language. Dutch dialects can be categorised as two- or three-gender dialects: in the

former, masculine and feminine gender have completely collapsed to form the category of 'common' gender; in the latter some remnants linger on of the distinction between masculine and feminine nouns. For instance, most Belgian dialects of Dutch still show masculine, feminine and neuter forms of articles and adjectives. This can be derived from table 1, in which the Standard Dutch paradigm merely distinguishes between common and neuter gender. In the East Flemish dialect of Sint-Niklaas, however, masculine forms of articles and adjectives can clearly be distinguished from their feminine counterparts since they take a final *-n* that is lacking on feminine articles and adjectives. In addition, the masculine indefinite article *ne*(*n*) (*man*) 'a (man)' clearly differs from feminine *een* (*vrouw*) 'a (woman)'.

Table 1. Adnominal gender in two varieties of Dutch

	Standard Dutch	East Flemish (e.g. Sint-Niklaas)
masculine		
definite:	*de* grot-*e* man	*de*(*n*) grot-*e*(*n*) man
indefinite:	*een* grot-*e* man	*ne*(*n*) grot-*e*(*n*) man
	'the/a tall man'	'the/a tall man'
feminine		
definite:	*de* grot-*e* vrouw	*de* grot-*e* vrouw
indefinite:	*een* grot-*e* vrouw	*een* grot-*e* vrouw
	'the/a tall woman'	'the/a tall woman'
neuter		
definite:	*het* grot-*e* kind	*het* groot kind
indefinite:	*een* groot kind	*e*(*en*) groot kind
	'the/a tall child'	'the/a tall child'

It needs to be added, however, that even in the most conservative varieties there are quite a few noun phrases revealing no gender information at all. This is mainly due to the process of *n*-deletion rendering masculine definite articles and adjectives identical to feminine forms in a number of phonological circumstances (more precisely whenever the *-n* is not followed by a vowel, /h/, /b/, /t/ or /d/, see Taeldeman 1980).

In correspondence with the conservative nature of their adnominal gender system, southern varieties of Dutch have by and large preserved the traditional system of pronominal reference: anaphoric pronouns may be masculine, feminine and neuter, and are chosen on the basis of a noun's

grammatical gender. Hence pronominal gender in these varieties differs from northern varieties of Dutch, especially in reference to inanimates, in that the vast majority of pronominal references in the south of the language area are still in line with the triadic distinction between masculine, feminine and neuter nouns (see Geeraerts 1992 for figures). This is no longer the case in areas where two-gender dialects of Dutch are spoken. Varieties spoken in two-gender areas have dispensed with grammatically feminine gender in pronominal reference: feminine pronouns such as *zij* 'she' or *haar* 'her' are only used to refer to female persons and animals, but never to refer to traditionally feminine inanimate nouns.[2] Consequently, most reference grammars of Dutch (e.g., Haeseryn et al.'s 2002 *Algemeen Nederlandse Spraakkunst*) describe these varieties as having not only a two-gender system adnominally, but also a grammatical two-gender system of pronominal reference, in which common nouns trigger the use of masculine pronouns such as *hij* 'he' and *hem* 'him', and neuter nouns are referred to with *het* 'it'. Unlike for adnominal gender, where only two-gender systems are considered part of the standard language, little or no normative pressure exists to adopt a three- or a two-gender grammatical system for pronominal reference (see, e.g., Haeseryn et al. 2002: 161–162).

One other development in pronominal gender does not (yet?) seem to be endorsed by normative sources, however: many varieties of Dutch appear to be engaging in a more far-reaching development in which a noun's grammatical gender becomes unimportant in the choice of the pronoun referring to it. Audring (2006, 2009) has investigated pronominal reference in informal registers of the Spoken Dutch Corpus (CGN), thereby focusing on varieties spoken in the west of the Netherlands (which is considered the centre of the Dutch language area). According to her, contemporary spoken varieties of Dutch tend to base their use of pronouns on the semantics of the antecedent noun rather than on its grammatical gender: highly individuated nouns are increasingly referred to using masculine pronouns such as *hij* 'he' and *hem* 'him', weakly individuated nouns are referred to with the neuter pronoun *het* 'it'. This is shown in (1):

2. To be more precise, the use of strong forms such as *zij* 'she' and *haar* 'her' is in most varieties restricted to human or animate antecedents. For non-humans, three-gender varieties prefer weak pronouns, particularly *ze* 'she/her'.

(1) Semantic gender in Dutch (examples from Audring 2006: 87, 95)

a. about *het apparaat* 'the device' (neuter, but count noun: masculine pronoun):
. . . ik wil ***'m*** *opwaarderen.*
I want.1SG him top up
'I want to top it up.'

b. about *olijfolie* 'olive oil' (common/feminine, but mass noun: neuter pronoun):
. . . hoe ***'t*** *geconserveerd wordt.*
. . . how it preserved.PART become.3SG
'. . . how it is preserved.'

Example (1) illustrates the different behaviour of a concrete count noun (*apparaat* 'device') and a concrete mass noun (*olijfolie* 'olive oil'), exemplifying the different behaviour of count nouns and mass nouns. Yet the system does not appear to be operating on the basis of a (relatively) clear distinction such as the count-mass-distinction. Rather the degree of individuation also depends on other parameters such as concreteness vs. abstractness, and boundedness vs. unboundedness (Audring 2009: 123–129). The development by which pronominal gender is reorganized in terms of individuation is termed 'resemanticisation'. The process can be described as involving a change in the nature of the antecedent-pronoun relationship, which is considered a syntactic property. But the changes in pronominal gender also reflect the problematic nature of the categorisation of Dutch nouns in three morphological classes, viz. masculine, feminine and neuter gender. Especially in two-gender varieties, these class distinctions, more particularly the distinction between masculine and feminine, is no longer recoverable from the language input. Hence the resemanticisation of pronominal gender can also be considered an instance of morphological regularisation, in which a system that has grown opaque is brought in line with a number of transparent rules (see also De Vos and De Vogelaer 2011).

One of the results of these diachronic developments are clear geographical differences in the pronoun that is used to refer back to nouns for which grammatical gender yields a different pronoun than the innovative semantic system. An example is the noun *tafel* 'table', which, as a historically feminine noun, triggers the use of feminine *ze* 'she' in southern varieties, but which is commonly referred to with *hij* 'he' in the north, in line with the fact that it has a clearly individuated referent.

(2) Reference to *tafel* 'table' in two varieties of Dutch
   a. northern Dutch:
      Die tafel? Hij heeft slechts 3 poten
      that.common table he has only 3 legs
      'That table? It (lit. 'he') has only got 3 legs.'
   b. southern Dutch:
      Die tafel? Ze heeft slechts 3 poten.
      that.common table she has only 3 legs
      'That table? It (lit. 'she') has only got 3 legs.'

De Vogelaer and De Sutter (2011) not only observe the geographical correlation between the loss of the three-gender system in the adnominal domain and resemanticisation of pronominal gender, they also causally relate the conservative nature of pronominal gender in the south, even in the southern standard, to the maintenance of the three-gender system in the dialects' adnominal gender system.[3] Indeed it seems reasonable to assume that the visibility of grammatical gender on, for instance, articles and adjectives helps speakers to determine a noun's grammatical gender, thereby reducing the need to rely on semantic rules for pronominal reference. Thus the degree to which speakers engage in resemanticisation reflects the transparency of the masculine-feminine distinction in grammar, or, put differently, the frequency with which these speakers are exposed to non-standard gender agreement markers unambiguously distinguishing masculine and feminine gender (see also Hoppenbrouwers 1983).

2.2. A complication: changing lexical gender

In addition to variation with respect to the way grammatical gender maps onto the use of gendered pronouns, extensive differences have been reported in the grammatical gender that is assigned to individual nouns. Thus, Pauwels (1938) discusses the gender of a large number of nouns in

3. This implies that southern speakers' use of the grammatical-three gender system in pronominal reference, while considered Standard Dutch, depends on their knowledge of a dialect. Up until a few decades ago, this entailed that virtually everyone was able to acquire the three-gender system, since dialects had preserved very well in the relevant area. See Hoppenbrouwers (1983) for discussion. Nowadays, especially the south of the Netherlands has seen extensive dialect levelling, and the grammatical three-gender system tends to be confined to Belgian varieties of Dutch. Belgium has witnessed dialect levelling too, but the use of dialectal adnominal morphology is one of the dialect features that resists levelling, since it is also use in supraregional, substandard language (Plevoets 2008).

Belgian Dutch dialects as documented in the late 19th century, including many East and West Flemish dialects. It appears that all these dialects at the time had preserved the grammatical three-gender system, but there is a lot of variation on the lexical level: nouns that are masculine in one dialect may be feminine or neuter elsewhere. For instance, *bos* 'forest' is masculine in some dialects, but neuter in others; *kraag* 'collar' is feminine in some dialects, masculine in others, etc. Some nouns, like *suiker* 'sugar', can even be masculine, feminine, and neuter, depending on the dialect in which they are used. Since this variation has emerged in the history of Dutch, it appears that nouns may change gender in the course of history (see Geerts 1966 for examples).

Such shifts in lexical gender clearly testify to the unstable nature of the Dutch gender system. More importantly, the fact that massive variation is observed between a noun's grammatical gender in different dialects makes it likely that situations of dialect contact will result in changes of a noun's gender in one dialect, under the influence of another. Given the fact that East and West Flemish dialects have witnessed extensive dialect levelling in recent decades, it seems especially likely that deviations from Standard Dutch are progressively levelled out, through dialects adopting a noun's Standard Dutch gender. Such changes in lexical gender are, of course, interesting in their own right and hence they deserve to be studied. But they also pose serious methodological problems for any research into pronominal gender in the Dutch dialects, since the use of a pronoun not in line with a noun's historical grammatical gender may not only reflect changes in the pronominal gender system as a whole, but also a mere shift in the relevant noun's lexical gender.

2.3. Methodological preliminaries

This investigation addresses developments in pronominal gender in a number of East and West Flemish dialects of Dutch. Since the investigation also aims to take stock of processes of diffusion through contact, more precisely of gender shifts under the influence of dialect contact, data are needed from a large number of locations. In the absence of extensive dialect corpora for the relevant region, the investigation has adopted questionnaires as a method for data collection, allowing to gather information on a restricted number of lexemes in a large number of villages and towns.

As a result of the gender variation encountered at the lexical level, investigating changes in pronominal reference requires that the historical gender of the nouns under investigation is known for all the locations

where the investigation is carried out. This can be done by taking the data from Pauwels' (1938) survey on gender variation as a starting point. From Pauwels' list, 50 nouns were selected for which gender information was gathered in a large sample of dialects spoken in the Belgian provinces of West and East Flanders, in the course of 2006. In selecting the questionnaire items, mainly nouns were chosen that occur in all dialects under investigation. Nevertheless, in addition to providing the pronoun, the informants were asked to translate the relevant noun. All answers in which a lexical alternative was given rather than a phonological variant of the word from the example sentence are left out of consideration, since in these cases it cannot be excluded that the informants referred to the alternative lexeme rather than to the word in the example sentence. The questionnaire nouns showed variation with respect to their semantics (high vs. weak individuation), their gender in both the traditional dialects and Standard Dutch, and their usage frequency (cf. *infra*), all of which are factors that are believed to be operating in the choice of an anaphoric pronoun. These factors by no means constitute an exhaustive list: previous research on the topic adopting a corpus method, most notably Audring (2009), has yielded a number of factors which cannot easily be operationalized using questionnaires, including syntactic factors such as the distance between antecedent and noun, anaphoric vs. cataphoric reference and the syntactic function of the pronoun, and discourse factors such as the thematic status of the referent. Such factors cannot be included in a questionnaire survey without decreasing the odds to obtain robust results for the parameters involving the antecedent nouns' semantics, and hence it has been attempted to neutralize the effects of these factors by keeping them constant (cf. *infra* on the design of the test sentences).

The questionnaire that was used only takes into account pronominal gender, and it consisted of sentence completion tasks of the type shown in (3): the informants had to fill in a subject pronoun referring to a (boldfaced) noun that was used in the preceding sentence. The preceding sentence did not contain any elements marking the gender of the noun (such as a definite article or an adjective). The informants were instructed to fill in the subject pronouns *hij* 'he', *ze* 'she' or *het* 'it' or a variant of these pronouns used in their regional variety of Dutch. Many such variants are indeed attested for the masculine pronoun *hij* 'he', for which Flemish dialects show extensive morphological variation (e.g. weak forms such as *en* or *ie*, or strong *em*), none for feminine *ze* 'she' and neuter *het* 'it'. The pronouns that had to be filled in referred to an activated referent used in the preceding sentence, in most cases as the rightmost noun, and with a

high degree of focality. While Dutch, in general, shows quite a prolific use of demonstratives in subject position, it is well-known that this context is the preferred one for personal pronouns to be used (Gundel, Hedberg, and Zacharski 1993; see also Comrie 2000 on Dutch). In addition, the second sentence was constructed in a way that only the bold-faced noun could logically be referred to. These measures sufficed to elicitate personal pronouns rather than demonstratives, since no demonstratives are found in the informants' answers.

(3) Example sentence from the 2006 questionnaire
Er is veel ***sneeuw*** *gevallen maar* _____ *is* gesmolten.
There is much **snow** fallen but _____ is melted.
'A lot of **snow** has fallen but in the mean time ____ has melted.'

The questionnaire was administered to the informants of the Dictionary of the Flemish dialects, and 138 of them were returned, from 103 different locations. The informants of the Dictionary of the Flemish dialects are all required to be L1 speakers of their local dialects. Since the network was established in the 1970s, nearly all informants are aged 50 or older. As dialects are exclusively spoken varieties in Belgium, written questionnaires are generally not considered the most reliable source for dialectological investigations, but most informants have several years of experience in filling out questionnaires, and the information they provide has proven a reliable source of information (for the methodology of the Dictionary of Flemish dialects, see Van Keymeulen 2003).

3. Mechanisms of gender change

3.1. The overall stability of Flemish gender

In total, 5515 data tokens have been gathered, which were entered in an SPSS-database, as were the expected answers on the basis of Pauwels' (1938) investigation. Overall, the results of the 2006 questionnaire correspond quite well to grammatical gender in the 19th century, with 64.66% of the answers (3566/5515 tokens) being inferable from Pauwels' (1938) results. All dialects still show at least some instances of grammatically feminine nouns referred to with feminine pronouns, leading to the conclusion that the grammatical three-gender system still survives in present-day East and West Flemish dialects of Dutch.

With respect to the answers that did not correspond to the expected gender, two developments are expected to be observable, viz. standardisation (see section 2.2) and resemanticisation (section 2.1). These two processes of change need to be disentangled. In order to do this, it was decided to focus on nouns for which deviations of grammatical gender unambiguously reveal the effect of one of the processes under investigation. This means that shifts that may be attributed both to standardisation and resemanticisation are not taken into consideration here. More precisely, table 2 illustrates that the investigation focuses on count nouns which are neuter in Standard Dutch, and mass nouns which are common in Standard Dutch. In the former category, a switch towards *hij* 'he' can safely be interpreted as a result of resemanticisation, whereas switches to *het* 'it' are likely to be due to standardisation. In the latter category, switches toward *hij* 'he' exemplify standardisation, whereas switches towards *het* 'it' must be explained as resemanticisation. Other types of change are ambiguous with respect to their interpretation, since in those cases Standard Dutch and resemanticisation 'conspire' to obtain the same effect.

Table 2. Conflicts between standardisation and resemanticisation

switch towards:	St. Dutch common	St. Dutch neuter
high individuation (*concrete count nouns*)	semantic = grammatical gender	*hij* 'he' = resemanticisation *het* 'it' = standardisation
weak individuation (*mass nouns and abstracts*)	*hij* 'he' = standardisation *het* 'it' = resemanticisation	semantic = grammatical gender

In principle, then, four types of deviations from the traditional gender described by Pauwels (1938) can safely be attributed to one of the two mechanisms of change allegedly at work in present-day Flemish dialects. Only two of these types are actually found, however, viz. the use of *het* 'it' for common, concrete count nouns being neuter in Standard Dutch, and the use of *het* 'it' for masculine/feminine mass nouns and abstract nouns being common in Standard Dutch. Resemanticisation with *hij* 'he' simply does not occur in present-day Flemish dialects: for none of the nouns liable to this development, a significant tendency is observed to use *hij* 'he'. While this presents a clear difference with resemanticisation

patterns in the north of the Dutch language area as described by Audring (2006, 2009), this finding is in line with descriptions of pronominal usage by southern Dutch children. Thus, De Paepe and De Vogelaer (2008) do not observe the use of *hij* 'he' for non-masculine concrete count nouns in East Flemish 7-year-olds, whereas these children display massive usage of *het* 'it' for mass nouns and abstracts. The fact that no standardisation with *hij* 'he' is observed, at first sight appears to be due the design of the questionnaire: the questionnaire did not contain mass nouns and abstracts which are neuter in the dialect but common in Standard Dutch. Some observations on count nouns suggest that this should not just be interpreted as a methodological shortcoming, however: closer inspection reveals that nouns with neuter gender in Flemish dialects and common gender in Standard Dutch are extremely rare. Among the rare examples are *fabriek* 'factory' and *machine* 'machine', two nouns that were included in the questionnaire but for which gender shifts cannot be unambiguously attributed to either standardisation or resemanticisation (cf. table 2). The fact that it is rare for a noun to be neuter in Flemish dialect but common in Standard Dutch is also mentioned in De Gruyter's ([1907] 2007) description of the East Flemish dialect of Ghent, where a handful of examples are given, but where almost 80 instances are given of nouns being masculine or feminine in the dialect and neuter in Standard Dutch.

One of the results of restricting the analysis to gender shifts for which the relevant mechanism of change can be determined unambiguously is that not all the 50 nouns from the questionnaire are taken into consideration. Apart from unambiguous shifts, this chapter only takes into account changes that can be attributed to one of the tendencies under investigation with a sufficient degree of accuracy. More precisely, the effects of resemanticisation are too weak in comparison to those of standardisation to be visible in nouns that are neuter in Standard Dutch (cf. *infra*). Thus, when addressing frequency effects in section 4, only data concerning 31 nouns are discussed, totalling 3514 tokens.

3.2. Standardisation effects

Many dialects of Dutch suffer from large-scale dialect loss and levelling (see, e.g., Hoppenbrouwers 1991, Taeldeman 1991), and the Flemish dialects are no exceptions to this, even though they are considered to be among the most conservative ones in the Dutch language area (Taeldeman 2005: 89–102 for East Flanders, Devos and Vandekerckhove 2005: 142–148 for West Flanders). The overall stability of the Flemish gender system discussed

in section 3.1 implies that the gender system by and large resists to the pressure to converge to the standard language. To disentangle standardisation from resemanticisation, this investigation focuses on traditionally masculine and feminine count nouns which are neuter in Standard Dutch. In this case, there is indeed a strong tendency to take over neuter gender: a ratio of 42.1% of the answers show neuter gender (528/1254 answers). If the number of neuter answers is contrasted with the amount of shifts for nouns not neuter in Standard Dutch (384/3544 or 10.84%), a chi square test reveals that the effect is highly significant (chi square = 588, d(f) = 1; $p < .001$; OR = 5.98). Given the fact that resemanticisation is very weak in comparison to standardisation, mass nouns have not been excluded from the analysis, i.e. both the nouns neuter in Standard Dutch and the nouns not neuter in Standard Dutch contain a few mass nouns. But the effect remains highly significant if they are not taken into consideration.

The most conspicuous examples that are undergoing this shift include *artikel* 'article', for which 80 informants were expected to provide a masculine pronoun if they followed the norms of their local dialect, but 74 used the neuter *het* 'it', totalling a ratio of 92.5%. High proportions of shifts are also obtained for *bos* 'forest' and *boek* 'book' (which are masculine in many dialects but neuter in Standard Dutch), and for *feest* 'party' and *dozijn* 'dozen' (which are feminine in the dialects but neuter in Standard Dutch). All these nouns show more than 70% shifts. The other nouns show smaller standardisation ratios (ranging from *bureau* 'desk' 34.2% via *vernis* 'polish' 20.8% and *nest* 'nest' 19.8% to *horloge* 'watch' 9.9% and *lak* 'polish' 8.54%; see appendix 1 for a complete overview of the results). It is obvious that standardisation must be considered an instance of diffusion, which takes effect as dialect speakers adapt to another variety under social pressure, in this case Standard Dutch.

The most striking difference between standardisation and the resemanticisation process discussed in section 2.1 is that standardisation does not appear to be sensitive to the semantics of the noun: among the nouns shifting to neuter gender both highly and weakly individuated nouns are found (cf. the examples above, which include both concrete count nouns such as *horloge* 'watch' and abstract mass nouns such as *vernis* 'polish'; see also appendix 1). In addition, data from other sources show that the standardisation effect is not restricted to pronominal gender. For instance, the database of the Syntactic Atlas of the Dutch Dialects (Barbiers et al. 2006) contains dialectal equivalents to Standard Dutch sentences with nouns such as *boek* 'book' or *feest* 'party'. In both cases, a few examples surface in Flanders in which such a noun is combined with neuter adnominal morphology (e.g. *dat boek* 'that boek', *het feest* 'the party').

3.3. Resemanticisation?

In present-day Dutch as spoken in the north of the Dutch language area, mass nouns are predominantly referred to with the neuter pronoun *het* 'it' and count nouns with masculine *hij* 'he' (Audring 2006). The feminine pronoun *ze* 'she' is only used to refer to female human beings and animals. Hence the traditional grammatical system of gender marking in pronominal reference is given up in favour of a semantic system. As shown in table 2, clear tendencies towards resemanticisation can only be found in nouns for which the alleged semantically-driven pronoun differs from the pronoun to be expected on the basis of the noun's grammatical gender in both the traditional dialect and Standard Dutch. It appears that in the Flemish dialects there is indeed a statistically significant effect to use the neuter pronoun *het* 'it' to refer to mass nouns and abstracts, whether they are grammatically neuter or not: the ratio of *het* 'it' answers is higher for non-neuter mass nouns and abstracts than for non-neuter concrete count nouns: 16.3% (286/1752 answers) vs. 5.5% (98/1792); all nouns neuter in Standard Dutch have been kept out of the analysis. This effect is statistically significant (chi square = 108, d(f) = 1; $p < .001$; OR = 3.39). Examples of nouns from the questionnaire with a strong tendency towards resemanticisation, i.e. reference with *het* 'it', are *achterdocht* 'suspicion' with 42.5% *het* 'it', *beet* 'bite' 37.8%, *pels* 'fur (mass noun)' 24.6%, *olie* 'oil' 23.2%, and *kalk* 'lime' 21.7%; examples with weak resemanticisation rates include *peper* 'peppar' 3.0% and *chocolade* 'chocolate' 3.0%. As in Standard Dutch, resemanticisation seems to affect pronominal gender only (cf. similar tendencies in other Germanic varieties, as described by Siemund 2002 and Audring 2006). Quite surprisingly, as was already noted in section 3.1, no tendency is observed to extend masculine *hij* 'he' to all concrete count nouns.

The question should be addressed whether resemanticisation in East and West Flemish presents a spontaneous development or an adoption from northern varieties. Some preliminary arguments can be given in favour of the former view. First, the ongoing change in West and East Flanders is not completely parallel to Audring's (2006) scenario for spoken northern Dutch. The change nevertheless boils down to resemanticisation, since the use of *het* 'it' is restricted to weakly individuated items. Rather than showing a tendency towards pronominalisation with masculine *hij* 'he', however, highly individuated nouns tend to preserve grammatical gender. Second, the geography of the phenomenon does not point in the direction of language contact as an important cause for resemanticisation in the south. The tendency towards resemanticisation appears to be

stronger in the western dialects (De Vogelaer and De Sutter 2011), the area which is more isolated and known to be the more conservative dialect area (Devos and Vandekerckhove 2005: 142–148). In section 4, frequency data will be discussed corroborating the claim that resemanticisation in East and West Flemish is not adopted from northern varieties.

According to Labov (2007), two main types of language change need to be distinguished, viz. 'diffusion', i.e. change through (dialect) contact, and 'imperfect transmission', i.e. change that is incrementally implemented by successive generations of language users. Given the fact that resemanticisation is not adopted from northern varieties, a characterisation of the change as an instance of, in Labov's (2007) terms, imperfect transmission seems warranted. In such processes of change an innovative variant is gradually replacing the older variant, through a process of incrementation whereby each generation advances the relevant change beyond the level of the preceding generation. Labov (2007: 346) has pointed out that language acquiring children play an important role in this process. Similarly, with respect to morphological change Bybee and Slobin (1982) claim that innovations in older school children (in their case aged 8½ to 10) may give rise to language change. For the resemanticisation of Dutch pronominal gender, it is relevant that children appear to start from semantically-motivated systems of pronominal gender, which are given up in favour of a grammatical system as they grow older. According to De Houwer (1987), who investigates a child acquiring a southern variety of Standard Dutch, pronominal reference in three-year old children mainly operates on the basis of the animate-inanimate distinction: animate entities are referred to with *hij* 'he'; for inanimate entities both *hij* 'he' and *het* 'it' are found. The motivation to use *hij* 'he' vis-à-vis *het* 'it' remains unclear in De Houwer's account, but given the amount of deviations from the adult system, grammatical gender hardly plays a role. At the age of 7, noun semantics are still the main factor underlying pronominal reference (De Vogelaer 2010). Not only the animate-inanimate distinction but also mass-count and concrete-abstract play a crucial role: both mass nouns and abstracts tend to trigger the use of the neuter pronoun *het* 'it' even when they are not grammatically neuter. Nevertheless, substantial proportions of pronominal reference are in line with grammatical gender (De Paepe and De Vogelaer 2008). Significantly, during adolescence the semantically-driven usage of *het* 'it' further decreases in favour of pronominal reference in line with grammatical gender, but even at the age of 18–20 the adolescents do not quite attain the same proportions of grammatical gender as previous generations (De Vos 2009; De Vos and De Vogelaer 2011).

These results on Dutch pronominal gender are all the more striking since in other languages in which pronouns agree in gender with their antecedent nouns, the grammatical system appears to be mastered already at a very young age, to the extent that deviations from grammatical gender are extremely rare. Thus, German children of six hardly deviate from a noun's grammatical gender in pronominal reference (Mills 1986: 92), and the same holds for French-speaking children (Maillart 2003; Van der Velde 2003: 328, 340). This is likely due to the arbitrariness of the Dutch gender system: gender of nouns referring to inanimates is not motivated semantically in Dutch, nor are there any clues in the form of (monomorphemic) nouns that allow to determine gender (Durieux, Daelemans and Gillis 1999). Hence children acquiring Dutch can only derive nouns' gender from the form of adnominal modifiers and pronouns, not from the form and/or meaning of the noun itself. This situation contrasts sharply with German and French, where gender assignment is at least partly motivated by semantic and/or formal regularities (see, e.g., Mills 1986 and Köpcke and Zubin 1996 on German, and Tucker, Lambert and Rigault 1977 on French). Such regularities minimise memory load, and are well-known to contribute to the acquirability of gender systems (Frigo and McDonald 1998; Gerken, Wilson and Lewis 2005).

Given the way gender is acquired in Dutch, there is little doubt that the Dutch resemanticisation process is indeed pushed forward by children acquiring their mother tongue but never reaching the same level of proficiency in grammatical gender as their predecessors did. More precisely, the acquisition of grammatical gender in pronominal reference should be conceived of as a process of 'un-learning' to use semantically motivated pronouns. At present, grammatical gender still stands strong enough in pronominal reference to motivate children to adopt the system, at least in southern varieties, but it seems likely that, in the long run, the semantic system will overtake the grammatical system in all parts of the Dutch language area.

4. Frequency effects

4.1. Frequency, and mechanisms of language change

Thus in the varieties of Dutch under investigation, several processes of linguistic change are in progress. The role of frequency in linguistic change has been investigated extensively with respect to phonological change (see,

e.g., Hooper 1976; Bybee 1995, 2001; Phillips 1984, 2001, 2006). The relevance of word frequency has been highlighted repeatedly, e.g. by Hooper (1976), who discusses two different frequency effects: on the one hand, processes of phonetic reduction are first visible in highly frequent items, whereas, on the other hand, processes of regularisation typically affect low-frequency items. In a survey of potential frequency effects in grammar, Bybee and Hopper (2001: 10–19) mention several types of frequency effects relating to language change, among which effects boiling down to a tendency in high-frequency patterns to engage in innovations (grammaticalization, lexicalization of multi-word-patterns, formal reduction, . . .), but also conservative effects in high-frequency patterns, such as the retention of certain morphological properties. Phillips (2006: 157) proposes that innovations implemented as speakers memory fails to provide the traditional variant typically affect low frequency items, whereas changes directly involving the production of word forms as stored in memory affect the most frequent words first.

In addition to playing a role in sound change and other processes of 'regular' linguistic change, frequency is found to play a role in dialect contact. Thus, Trudgill (1986: 11–21, 43–53) describes processes of long-term accommodation of one dialect towards another, and observes that salient features are adopted more easily.[4] It is rather obvious that, all other properties being equal, highly frequent features are more salient than infrequent features, and thus Trudgill's observations lead to suggest that frequent items of the donor variety will be easily borrowed by the target variety. According to Phillips (2006: 141), however, such contact-induced changes will only affect high frequency items provided that there are no 'ideological' reasons for doing otherwise (cf. also Trudgill 1986: 17–19, 125 on 'extra-strong salience') and if the relevant change directly involves the production of the relevant word form.

In section 3 it was claimed that the gender shifts observed in the data are due to two phenomena: 1. standardisation, and 2. resemanticisation, taking effect independently of the resemanticisation of pronominal gender in northern varieties. The effect of frequency on standardisation depends on the classification of the phenomenon as ideologically motivated vs. ideologically free, and, in the latter case, on the nature of the ongoing change. Since differences in a noun's gender are not overtly stigmatised in

4. Dialect contact is understood here in a broad sense, i.e. as including contact between dialects and prestige varieties such as Standard Dutch.

the area and even tend to go unnoticed most of the time, they are most likely an ideologically free change. In addition, a gender shift under the influence of standardisation directly involves the production of word forms (especially in a language like Dutch, in which gender is essentially arbitrary; cf. *supra*). Hence, given Phillips' (2006) hypothesis on frequency in language change, it is expected that standardisation will mainly affect high-frequency items: language users are more likely to adopt a noun's Standard Dutch gender the more they are exposed to the relevant noun. It should be noted, however, that this hypothesis is only valid for speakers for whom the dialect is their 'base variety', and for whom the standard language is their second dialect. While this situation is common across older speakers in the Dutch language area, recent investigations into dialect usage in younger generations (especially Rys 2007) have revealed that it is nowadays probably more accurate to consider Standard Dutch to be the native dialect of children growing up in Belgium, since the acquisition of the dialect primarily takes place in adolescence, and is characterised by overgeneralisations typically found in second dialect acquisition. Such younger speakers, however, did not take part in this survey. The role of frequency for resemanticisation is somewhat more difficult to determine. On the one hand, dialect geographical data indicated that resemanticisation in Flemish is likely not taken over from other varieties (De Vogelaer and De Sutter 2011). Hence resemanticisation can be characterised as a development taking place within a speech community, more specifically as a type of regularisation, a kind of innovation being implemented as speakers' memories fail (Phillips 2006: 157). This characterisation leads to believe that the phenomenon will be found primarily in low-frequency items. On the other hand, it remains at least theoretically possible that resemanticisation is diffused from northern Dutch on a word-by-word basis. Assuming that the liability to semantically motivated pronominalisation may be lexically specific (cf. Smith 2001: 365, 373–374 and Poplack 2001: 411–414 on the potential of morpho-syntactic traits to be lexically specific), and that patterns of semantic agreement may be diffused from one variety to another, this creates a potential for highly frequent items to trigger semantic agreement more easily.

In the remainder of this article, it is tested which of these hypotheses are borne out by the data, by correlating the questionnaire data already discussed in section 3 with two frequency measures. The frequency data are taken from two different sources. The first source is one of the frequency lists of the Spoken Dutch Corpus (CGN), which provides raw data on the

frequency with which a noun occurs in the CGN, a corpus of spoken Dutch of approximately 9 million words (see Oostdijk 2000 for a description). More precisely, the word form list drawn from the Belgian part of the CGN has been consulted. The use of this list rather than, for instance, the Celex database or frequency data drawn from larger, written corpora is motivated by the fact that dialects tend to be exclusively spoken languages, and it is expected that frequency data drawn from written corpora will yield less favourable results (see Clark, *to appear* for an illustration of how frequency data drawn from written, standard language corpora may yield wrong predictions). The usage frequencies of the questionnaire items in the list for the Belgian part of the CGN range from 0 (for *dozijn* 'dozen' and *zink* 'zinc') to 1005 (for *boek* 'book'). The second source for frequency information is chosen to reflect the age of acquisition of the relevant nouns. 'Age of acquisition' is a popular parameter in psycholinguistic work, which is believed relevant in processes such as lexical access, word naming, and visual word recognition (Caroll and White 1973; Gilhooly 1984; Brysbaert, Lange, and Van Wijnendaele 2000). Since it is extremely time consuming to determine the age at which words are acquired with naturalistic data, the age of acquisition of words is typically investigated with questionnaire surveys among adults, who are asked to estimate at which age they have acquired the relevant noun. The results of such surveys are found to be both very robust, i.e. different surveys among different informants yield almost identical results, and valid, i.e. the results of the questionnaire results correspond very well to objective descriptions of children's vocabulary (Caroll and White 1973, Gilhooly and Gilhooly 1980; see also Morisson and Ellis 2000).

At present, the most elaborate source for age of acquisition in Dutch is the 'target vocabulary list for 6-year-old children' (Schaerlaekens, Kohnstamm, and Lejaegere 1999). This list provides the proportion of investigated caretakers that considered a given word to be known by most 6-year-olds. The present investigation used the %-score attributed by Belgian caregivers. Vervoorn (1989) has investigated which factors determine a noun's score in the target list (thereby making use of an older version of the list). Very strong correlations are obtained between the target list score and age of acquisition as estimated by adults for two random samples of nouns, a 44-word sample (with $r = .92$, Vervoorn 1989: 40), and a 300-word sample (with $r = .93$, Vervoorn 1989: 42). In addition, all nouns appearing high in Beyk and Aan de Wiel's (1978) frequency list on 3 and 4-year-olds' language production also have a score of $>90\%$ in the target

list, leading to the conclusion that the target list scores closely match frequency counts in child production data (Vervoorn 1989: 46–47).[5] Hence the target vocabulary list is a good measure of both age of acquisition and frequency in child language.

With respect to the questionnaire items that are further analysed in this section, the correlation between the target list score and frequency turns out to be remarkably low. There are only 23 nouns for which both scores are available. Although recent investigations have revealed that input frequency plays an important role in vocabulary acquisition, at least in content words such as nouns (Goodman, Dale, and Li 2008), Vervoorn (1989: 24–26, 64) observes that frequency in adult corpora correlates rather weakly with the target list score (with r-values equalling approximately .35). For the nouns included in the present investigation, even weaker correlations are found: CGN frequency and the target list score show a correlation of no more than $r = .182$, which fails to be statistically significant ($p = .405$). This means that the questionnaire items are biased towards nouns for which strong discrepancies are observed between their frequency and their target list score. Thus, infrequent nouns such as *limonade* 'lemonade', *spinazie* 'spinach' or *horloge* 'watch' have relatively high target list scores, whereas nouns such as *artikel* 'article' and *vlucht* 'flight' are frequently used without belonging to young children's vocabulary. Given the nature of the present investigation, in which strict selection criteria needed to be imposed on the questionnaire nouns (conflict between semantic and grammatical gender, information available on the noun's grammatical gender in the 19th century dialects), this can hardly be remedied. But it needs to be borne in mind in the analysis that both frequency measures might correlate more strongly than appears from the present data, and that the present selection of nouns increases the likelihood that only one of the two frequency measures will correlate with the degree of diachronic change that is observed.

Depending on the type of change to be discussed, it is expected that one of the frequency sources will provide a better match for the changes that are observed, allowing to draw conclusions on the locus of language change.

5. Since the data on age of acquisition and/or frequency in 3–4-year-olds have been drawn from very restricted sets of data, they could not be used as frequency measures in this investigation, simply because most of the nouns on the questionnaire were lacking from them. For instance, the frequency data used by Vervoorn are calculated on the basis of her own 54000-word corpus, which does not contain instances of many of the nouns under investigation.

For instance, processes of dialect contact and, consequently, standardisation, typically take place in adults. Hence standardisation is expected to correlate with CGN-frequency. By contrast, there are reasons to believe that the resemanticisation process in Dutch is pushed forward by language acquiring children never attaining the same level of proficiency in the grammatical system of pronominal reference as their parents (De Vos 2009; De Vos and De Vogelaer, to appear; cf. *supra*). The target vocabulary list better captures the degree to which children are familiar with certain nouns, which is, given the resemanticisation scenario laid out in section 3, likely to contribute to these nouns' susceptibility to resemanticisation.

In a way, both the adoption of Standard Dutch gender and the acquisition of a noun's grammatical gender (which makes the noun less susceptible to resemanticisation) can be described as learning processes. Hence it is expected that the influence of frequency on both phenomena is best described by means of a 'learning curve': the first instances of a Standard Dutch noun will contribute stronger to the standardisation process than any succeeding ones, whereas the first instances of a noun will also be more crucial for children to determine the noun's grammatical gender during acquisition (cf. also Hay and Baayen 2002: 208, who observe that differences amongst lower frequencies often are more salient than equivalent differences amongst higher frequencies). Therefore, rather than testing for correlations between the observed changes and raw frequency data, a logarithmic transformation has been applied on the frequency data (which indeed yields better fits).

4.2. Dialect contact affects high frequency items

In order to investigate the role of frequency, for each word on the questionnaire the strength was calculated with which it is affected by each of the investigated tendencies. For instance, for the noun *bos* 'forest' 92 answers are available from regions where *bos* is traditionally a masculine noun, whereas it is neuter in Standard Dutch. In 74 cases, the neuter pronoun *het* 'it' was given as an answer. This means that *bos* 'forest' shows a standardisation ratio of 74/92 or 80%. This figure can then be correlated with the frequency data, i.e. both with (the logarithmic transformations of) the noun's score on Schaerlaekens, Kohnstamm, and Lejaegere's (2000) Target Vocabulary List and the noun's frequency in the Spoken Dutch Corpus. Table 3 shows the correlations for the relevant nouns (see appendix 1 for the raw data).

Table 3. Correlation between standardisation and frequency[6]

		% acquired in Schaerlaekens et al. 2000 (Log)	Attestations in the Spoken Dutch Corpus (Log)
Shifts to Standard Dutch neuter	Pearson Correlation	−.205	.442
	Sig. (1-tailed)	.285	.057
	N	10	14

No significant correlations are obtained for the Target Vocabulary List. CGN-frequency does appear to have an effect on standardisation, with $r = .442$. With $p = .057$, the effect is borderline significant. Rank correlation measures yield somewhat more favourable p-values for the correlation between standardisation and CGN-frequency: Kendall's tau-b is .367 with $p = .035$; Spearman's rho is .546 with $p = .022$. From this it can be concluded that standardisation, at least in gender change, mainly affects highly frequent items: highly frequent items tend to shift towards Standard Dutch gender more easily.[7] While correlations exceeding .40 are generally considered strong in the social sciences, Figure 1 reveals that the fit between the logarithmic transformation of usage frequency and standardisation is far from perfect.

6. Note that the informants for this study are L1 speakers of their dialect, who are reporting about the use of their dialect. The effect of frequency will probably be different in other circumstances, e.g. in cases where a dialect speaker tries to accommodate to the standard, or in situations in which the dialect is no longer learned as an L1. For instance, in a study of children acquiring dialect as a second language, Rys (2007: 236–240) shows that highly frequent dialect features are picked up better than less frequent ones. This is especially relevant as second dialect learning seems to become the norm in the Dutch language area: parents typically talk (sub)standard Dutch to their children; to the extent that children still learn a dialect, it is picked up at a later age, starting from nursery school with dialect proficiency increasing until deep in adolescence (see Rys 2007 for discussion). Thus it may very well be the case that findings such as these cannot be replicated in younger dialect speakers.
7. When used linearly rather than logarithmically, CGN frequency yields a correlation of $r = .422$ with standardisation, with $p = .066$. Thus the difference with the data in Table 3 is small, and it remains theoretically possible that the effect of usage frequency on standardization is better described as a linear function. It can be expected that an investigation with a larger sample of nouns will cast more light over this issue.

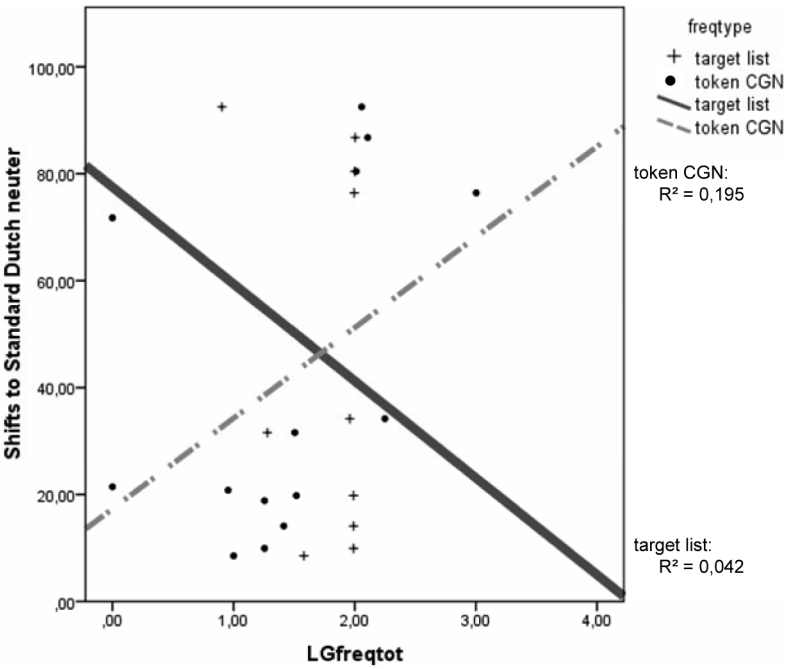


Figure 1. Standardisation and two frequency measures (logarithmically transformed)

There may be several explanations for this: for one, the present investigation abstracts away from contact between the dialects themselves. Thus, it is not taken into consideration whether Standard Dutch gender is also found in dialects neighbouring a certain dialect under investigation or not, whether this can theoretically have an influence on the speed with which standardisation takes effect. Another possible explanation is that certain nouns may be more specific to registers in which dialect and/or Standard Dutch is used, which would inhibit or stimulate the odds of a gender shift. For instance, a noun like *boek* 'book' is likely to be associated more strongly with registers in which Standard Dutch is used, than a noun like *zink* 'zinc'. In order to investigate whether this explanation is plausible, frequency lists would be needed for the different dialects under investigation, or at least from different registers of the standard language.

4.3. Transmission and low frequency items

In 4.1 it was hypothesized that resemanticisation correlates negatively with frequency, i.e. infrequent items are affected more strongly by resemanticisa-

tion than items ranking high on the frequency lists. In addition, it was expected that the Target Vocabulary List would yield a stronger correlation than the raw frequency data from the Spoken Dutch Corpus. The data are shown in table 4 (see also appendix 2).

Table 4. Correlation between resemanticisation and frequency[8]

		% acquired in Schaerlaekens et al. 2000 (Log)	Attestations in the Spoken Dutch Corpus (Log)
% HET 'it' for mass nouns & abstracts	Pearson Correlation	−.729**	−.161
	Sig. (1-tailed)	.002	.269
	N	13	17

The hypothesis is borne out. In addition, rank correlation measures yield highly similar results for the Target Vocabulary List. Kendall's tau-b is −.416 with p = .025; Spearman's rho is −.579 with p = .019. No significant correlations are observed between resemanticisation and raw frequency in the Spoken Dutch Corpus. Figure 2 shows a scatter plot with the results.

Both the table and the scatter plot indicate that items high on the target vocabulary list resist resemanticisation. The very same elements are believed to be acquired early and to be the most frequent items in young children's speech (Vervoorn 1989: 40, 46; cf. section 4.1). The fact that the target vocabulary list yields much clearer results adds support to the idea that resemanticisation relates to the language acquisition process, providing an extra argument to consider it change through 'imperfect transmission'. In addition, the fact that the target list score, which correlates strongly with age of acquisition and with usage frequency at the age of 3–4, provides such a powerful predictor for resemanticisation suggests that the ability to pick up a noun's grammatical gender declines with age. It is, for instance, well-known that second language learners experience much more difficulties in acquiring gender systems than first language acquirers

8. When used linearly rather than logarithmically, the Target List score yields a correlation of r = .672 with resemantisation, with p = .006. Again, this difference is far from spectacular.

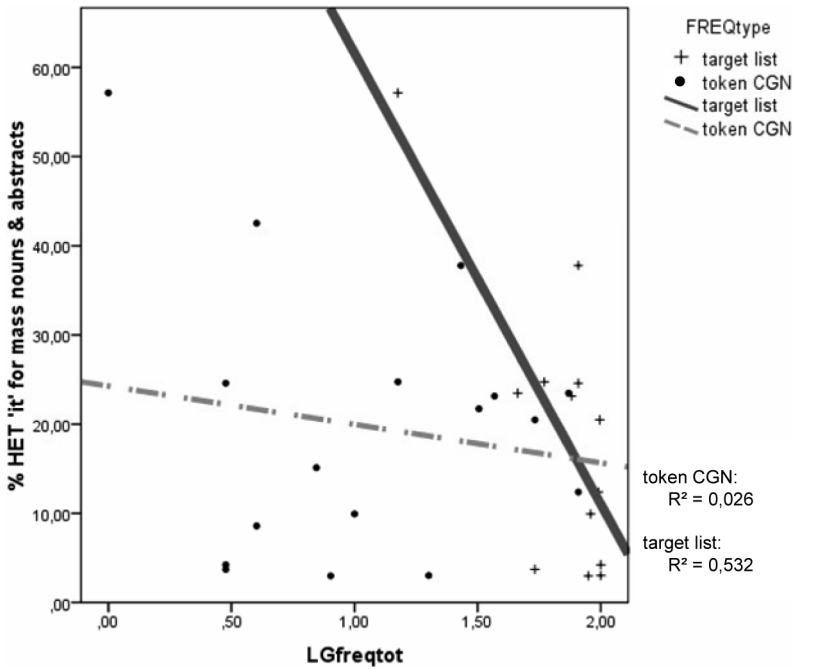


Figure 2. Resemanticisation and two frequency measures (logarithmically transformed)

(cf. Cornips and Hulk 2006 on Dutch gender). However, further investigation is needed to substantiate this relation between the correlation shown in table 4 and possibly critical age effects.

Significantly, the frequency data from the CGN do not correlate with resemanticisation. This may in part be due to the fact that the investigation only targeted a limited number of nouns, for which corpus frequency and target list score correlate less strongly than for most nouns. On the basis of the stronger correlations calculated by Vervoorn (1989: 64–65), it can be expected that large-scale investigations will reveal statistically significant correlations between resemanticisation and frequency data drawn from adults (such as CGN frequency). Indeed De Vos (2009) detects clear frequency effects with respect to the proportion of pronominal references in line with grammatical gender, using frequency data from adults rather than children. This, in turn, underscores the poly-interpretability of frequency effects: within the domain of diachronic research, frequency effects may reflect liability on a language pattern's part to engage in

processes of routinization (grammaticalization, phonetic reduction, ...), different degrees of entrenchment in grammar, different ages of acquisition, etc. Hence researchers should be very explicit on the nature of frequency effects in their data, and on the underlying explanation. In many cases, frequency effects will merely reflect some deeper property of language patterns rather than being a conclusive explanation in their own right. The data in this chapter are a case in point: during processes of standardisation, frequency effects reflect the intensity with which dialect speakers are exposed to nouns' standard language gender; in resemanticisation, frequency effects reveal different ages at which nouns are acquired by children, which appears to influence the odds that these nouns' grammatical gender can be learned successfully.

5. Conclusions

Like the northern Standard Dutch system, the gender system in present-day southern Dutch dialects is undergoing change. At least in the provinces of East and West Flanders, 1. originally non-neuter words are shifting to neuter gender under the influence of Standard Dutch; and 2. a tendency towards resemanticisation of pronominal gender is witnessed, mainly in West Flanders (cf. Audring 2006 for (northern) Standard Dutch). The former development involves both adnominal and pronominal gender, the latter development is restricted to pronominal gender. The tendencies differ with respect to the underlying mechanism of change too (cf. Labov 2007): standardisation is the result of diffusion; resemanticisation appears to be an instance of 'imperfect transmission', which constitutes a spontaneous development in the region under investigation rather than a borrowing from Standard Dutch.

This classification of resemanticisation as a spontaneous development is supported by the fact that the phenomenon shows different frequency effects than standardisation: while standardisation typically affects highly frequent items, resemanticisation is observed more frequently in infrequent items. Different frequency measures yield different results, in each case corroborating the alleged 'locus of language change' for the relevant change. Standardisation is the result of accommodation in adult speech, and thus frequency data extracted from the Spoken Dutch Corpus provide a better match with the diachronic changes than the Target Vocabulary List for six-year-olds (Schaerlaekens, Kohnstamm, and Lejaegere 2000).

For resemanticisation, an instance of change through imperfect transmission, the reverse holds.

On a more general note, these data corroborate the importance of frequency data for our understanding of processes of linguistic change. But they also call for scrutiny as to the use of different types of frequency measures. In comparison to larger samples of nouns investigated by Vervoorn (1989), the nouns in the present study showed a remarkably weak correlation between target list score and usage frequency. This may have skewed the results somewhat, to the effect that both for standardisation and for resemanticisation only one frequency measure yields statistically significant correlations, whereas in investigations targeting nouns for which both frequency measures correlate more strongly, both measures may correlate significantly with the investigated phenomenon. This illustrates that frequency effects are typically poly-interpretable, which emphasizes the need for explicitness on the nature of any frequency effects that can be found in linguistic data.

In addition, it can be considered good practice to check for correlations with different frequency measures rather than to focus on just one type of frequency. In that respect it should be mentioned that even more frequency measures can be taken into account than the two used here. This investigation has focused, on the one hand, on frequency drawn from a corpus of spoken language (the Spoken Dutch Corpus), and, on the other hand, on a target list score that is found to reflect both age of acquisition and usage frequency by 3–4-year old children (Vervoorn 1989). Alternatives are conceivable for both measures. With respect to frequencies in adult language, it would be interesting to incorporate the role of register effects in the investigation, for instance by contrasting the results obtained with frequency measures drawn from different language varieties and/or different media. For instance, frequency in the Spoken Dutch Corpus, a corpus designed to contain only Standard Dutch, could be replaced with frequency in dialect corpora. Alternatively, frequency in a written, larger corpus than the Spoken Dutch Corpus might provide more robust estimates of usage frequency and also better reflect usage frequency in more formal registers, in which most of dialect speakers' contacts with standard languages take place. With respect to the target list score, the fact that it correlates strongly with adult estimates of age of acquisition does not render a test with naturalistic data concerning age of acquisition obsolete, however difficult this may be to measure. The same applies to the correlation between age of acquisition and usage frequency in 3–4-year old children, which is in need of testing against data drawn from larger corpora than the 54000-

word corpus used by Vervoorn (1989). In addition, age-of-acquisition is known to correlate with a range of factors, including input frequency (see Brysbaert and Ghyselinck 2006 for discussion) and semantic factors such as imageability (Masterson and Druks 1998). All of these factors can be correlated with diachronic data of the type described in this chapter.

Unfortunately, even for a relatively well investigated language such as Dutch not all of these measures are readily available, and for some measures there are even insufficient resources to develop them. However, since we have the technological tools to store and exploit large corpora, even of spoken (child) language, it is to be expected that some of these frequency measures will be developed in the not too distant future. Thus, investigations such as this one present only the beginning of an interesting research line exploring the role of frequency in language change.

References

Audring, Jenny
2006 Pronominal gender in spoken Dutch. *Journal of Germanic Linguistics* 18: 85–116.

Audring, Jenny
2009 *Reinventing pronoun gender*. Utrecht: LOT.

Barbiers, Sjef et al.
2006 *Dynamische Syntactische Atlas van de Nederlandse Dialecten*. Amsterdam: Meertens Instituut. http://www.meertens.knaw.nl/sand/

Beyk, E.M. and M.C. Aan de Wiel
1978 *Woordfrequenties in de spontane taal van enkele groepen Haarlemse peuters en kleuters in de schoolsituatie*. Report Universiteit van Amsterdam.

Brysbaert, Marc, Marielle Lange and Ilse Van Wijnendaele
2000 The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology* 12: 65–85.

Brysbaert, Marc and Mandy Ghyselinck
2006 The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition* 13: 992–1011.

Bybee, Joan
1995 Regular morphology and the lexicon. *Language and Cognitive Processes* 10: 425–455.

Bybee, Joan
2000 The phonology of the lexicon: evidence from lexical diffusion. In: Michael Barlow and Suzanne Kemmer (eds.), *Usage-based Models of Language*, 65–85. Stanford: CSLI.

Bybee, Joan
2001 *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan and Paul Hopper
2001 Introduction to frequency and the emergence of linguistic structure. In: Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 1–26. Amsterdam/Philadelphia: John Benjamins.

Bybee, Joan and Dan Slobin
1982 Why small children cannot change language on their own: suggestions from the English past tense. In: A. Ahlqvist (ed.), *Papers from the 5th International Conference on Historical Linguistics*, 29–37. Amsterdam: John Benjamins.

Carroll, John and Margaret White
1973 Age of acquisition norms for 220 picturable nouns. *Journal of Verbal Learning & Verbal Behavior* 12: 563–576.

Clark, Lynn
to appear Dialect data, lexical frequency and the usage-based approach. In: Gunther De Vogelaer and Guido Seiler (eds.), *The dialect lab: dialects as a testing ground for theories of language change*. Amsterdam/Philadelphia: John Benjamins.

Comrie, Bernard
2000 Pragmatic binding: Demonstratives as anaphors in Dutch. *Proceedings of the Annual Meeting of the Berkeley Linguistic Society* 23: 50–61.

Corbett, Greville, Andrew Hippisley, Dunstan Brown and Paul Marriot
2001 Frequency, regularity and the paradigm: a perspective from Russian on a complex relation. In: Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 201–228. Amsterdam/Philadelphia: John Benjamins.

Cornips, Leonie and Aafke Hulk
2006 External and internal factors in bilingual and bidialectal language development: grammatical gender of the Dutch definite determiner. In: Claire Lefebvre, Lydia White and Christine Jourdan (eds.), *L2 Acquisition and Creole Genesis. Dialogues*, 355–378. Amsterdam/Philadelphia: John Benjamins.

De Gruyter, Jan-Oscar
[1907] 2007 *Het Gentse dialekt: klank- en vormleer*. Edited by Johan Taeldeman. Gent: KANTL.

De Houwer, Annick
1987 Nouns and their companions, or how a three-years-old handles the Dutch gender system. In: Annick De Houwer and Steven Gillis (eds.), *Perspectives on child language*, 55–74. Belgian journal of linguistics 2. Amsterdam/Philadelphia: John Benjamins.

De Paepe, Jessie and Gunther De Vogelaer
2008 Grammaticaal genus en pronominale verwijzing bij kinderen. Een taalverwervingsperspectief op een eeuwenoud grammaticaal probleem. *Neerlandistiek.nl* 08.02: 1–23.
De Schutter, Georges, Boudewijn van den Berg, Ton Goeman, Thera de Jong
2005 *Morphological atlas of the Dutch dialects. Volume 1*. Amsterdam: Amsterdam University Press.
De Vogelaer, Gunther
2010 (Not) acquiring grammatical gender in two varieties of Dutch. In: Dirk Geeraerts, Gitte Kristiansen and Yves Peirsman (eds.), *Advances in cognitive sociolinguistics*, 167–190. Berlin: Mouton de Gruyter.
De Vogelaer, Gunther and Gert De Sutter
2011 The geography of gender change: pronominal and adnominal gender in Flemish dialects of Dutch. *Language Sciences* 33: 192–205.
De Vos, Lien
2009 De dynamiek van hersemantisering. In: *Taal en Tongval*, Theme issue 22: 82–110.
De Vos, Lien and Gunther De Vogelaer
2011 Dutch gender and the locus of morphological regularisation. *Folia Linguistica* 45: 245–281.
Durieux, Gert, Walter Daelemans and Steven Gillis
1999 On the arbitrariness of lexical categories. In: Frank van Eynde (ed.), *Computational linguistics in the Netherlands 1998*, 19–35. Amsterdam: Rodopi.
Devos, Magda and Reinhild Vandekerckhove
2005 *West-Vlaams*. Tielt: Lannoo.
Frigo, Lenore and Janet McDonald
1998 Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language* 39: 218–245.
Geeraerts, Dirk
1992 Pronominale masculiseringsparameters in Vlaanderen. In: Hans Bennis and Jan de Vries (eds.), *De binnenbouw van het Nederlands: een bundel artikelen voor Piet Paardekooper*, 73–84. Dordrecht: ICG Publications.
Geerts, Guido
1966 *Genus en geslacht in de Gouden Eeuw*. Brussel: Belgisch Interuniversitair Centrum voor Neerlandistiek.
Gerken, Louann, Rachel Wilson and William Lewis
2005 Infants can use distributional cues to form syntactic categories. *Journal of Child Language* 32: 249–68.
Gilhooly, Kenneth and Mary Gilhooly
1980 The validity of age-of-acquisition rating. *British Journal of Psychology* 71: 105–110.

Gilhooly, Kenneth
1984 Word age-of-acquisition and residence time in lexical memory as factors in word naming. *Current Psychological Research* 3: 24–31.
Goodman, Judith, Philip Dale and Ping Li
2008 Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language* 35: 515–531.
Gundel, Jeanette, Nancy Hedberg and Ron Zacharski
1993 Cognitive status and the form of referring expressions in discourse. *Language* 69: 274–307.
Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jaap de Rooij and Maarten van den Toorn
2002 *Elektronische Algemene Nederlandse Spraakkunst*. http://www.let.ru.nl/ans/e-ans/
Hay, Jennifer and R. Harald Baayen
2002 Parsing and productivity. In: Geert Booij and Jaap Van Marle (eds.), *Yearbook of morphology 2001*, 203–235. Dordrecht: Kluwer.
Hooper, Joan
1976 Word frequency in lexical diffusion and the source of morphophonological change. In: W. Christie (ed.), *Current progress in historical linguistics*, 347–357. Amsterdam: North Holland.
Hoppenbrouwers, Cor
1983 Het genus in een Brabants regiolect. *Tabu* 13: 1–25.
Hoppenbrouwers, Cor
1991 *Het regiolect. Van dialect tot Algemeen Nederlands*. Muiderberg: Coutinho.
Köpcke, Klaus-Michael and David Zubin
1996 Prinzipien für die Genuszuweisung im Deutschen. In: Ewald Lang and Gisela Zifonun (eds.), *Deutsch-typologisch. Institut für deutsche Sprache Jahrbuch 1995*, 473–491. Berlin: Walter de Gruyter.
Labov, William
2007 Transmission and diffusion. *Language* 83: 344–387.
Maillart, Christelle
2003 Origine des troubles morphosyntaxiques chez les enfants dysphasiques. Ph.D-dissertation University Louvain-la-Neuve.
Masterson, Jackie and Judit Druks
1998 Description of a set of 164 nounsand 102 verbs matched for printed word frequency, familiarity and age-of-acquisition. *Journal of Neurolinguistics* 11: 331–354.
Mills, Anne
1986 *The acquisition of gender. A study of English and German*. Berlin: Springer.
Milroy, Lesley
2003 Social end linguistic dimensions of phonological change: fitting the pieces of the puzzle together. In: David Britain and Jenny

Cheshire (eds.), *Social Dialectology, in Honour of Peter Trudgill*, 155–171. Amsterdam: John Benjamins.

Morisson, Catriona and Andrew Ellis
2000 Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology* 91: 167–180.

Oostdijk, Nelleke
2000 Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde* 5: 280–284.

Pauwels, Jan L.
1938 *Bijdrage tot de kennis van het geslacht der substantieven in Zuid-Nederland*. Tongeren: Michiels.

Phillips, Betty
1984 Word frequency and the actuation of sound change. *Language* 60: 320–342.

Phillips, Betty
2001 Lexical diffusion, lexical frequency, and lexical analysis. In: Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 123–136. Amsterdam/Philadelphia: John Benjamins.

Phillips, Betty
2006 *Word frequency and lexical diffusion*. Basingstoke: Palgrave Macmillan.

Plevoets, Koen
2008 Tussen spreek- en standaardtaal: Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfo-syntactische verschijnselen uit het gesproken Belgisch-Nederlands. Ph.D-dissertation KULeuven.

Poplack, Shana
2001 Variability, frequency, and productivity in the irrealis domain of French. In: Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 405–430. Amsterdam/Philadelphia: John Benjamins.

Rys, Kathy
2007 Second dialect acquisition: a study into the acquisition of the phonology of the Maldegem dialect by children and adolescents. Ph.D-dissertation Ghent University.

Schaerlaekens, Annemarie, Dolf Kohnstamm and Marilyn Lejaegere
1999 *Streeflijst woordenschat voor zesjarigen* [third, revised edtion]. Lisse: Swets & Zeitlinger.

Siemund, Peter
2002 Mass versus count: Pronominal gender in regional varieties of Germanic languages. *Language Typology and Universals* (STUF) 55: 213–233.

Smith, Aaron
2001 The role of frequency in the specialization of the English anterior. In: Joan Bybee and Paul Hopper (eds.), *Frequency and the emer-*

gence of linguistic structure, 361–382. Amsterdam/Philadelphia: John Benjamins.

Taeldeman, Johan
1980 Inflectional aspects of adjectives in the dialects of Dutch-speaking Belgium. In: Wim Zonneveld et al. (eds.), *Studies in Dutch Phonology*, 265–292. Dutch Studies Vol. 4. Den Haag: Martinus Nijhoff.

Taeldeman, Johan
1991 Dialect in Vlaanderen. In: Herman Crompvoets and Ad Dams (eds.), *Kroesels op de bozzem: het dialectenboek 1*, 34–52. Waalre: Stichting Nederlandse Dialecten.

Taeldeman, Johan
2005 *Oost-Vlaams*. Tielt: Lannoo.

Treffers-Daller, Jeanine
1994 *Mixing two languages: French-Dutch contact in a comparative perspective*. Berlin/New York: Mouton de Gruyter.

Trudgill, Peter
1986 *Dialects in Contact*. Oxford/New York: Blackwell.

Tucker, Richard, Wallace Lambert and Andre Rigault
1977 *The French speaker's skill with grammatical gender: An example of rule-governed behavior*. The Hague: Mouton.

Van der Velde, Marlies
2003 Déterminants et pronoms en néerlandais et en français: syntaxe et acquisition. Ph.D-dissertation Paris 8.

Van Keymeulen, Jacques
2003 Compiling a dictionary of an unwritten language. A non corpus-based approach. *Lexikos* 13: 183–205.

Vervoorn, Willemina
1989 *Eigenschappen van basiswoorden*. Amsterdam/Lisse: Swets & Zeitlinger.

Appendix 1: Standardisation and frequency (cf. table 3)

	Target list score (*Schaerlaekens et al*).		*Token frequency* (*CGN*)		Standard-isation ratio
	%	Log	raw	Log	
lak 'polish'	37	1,58	9	1	8,54
horloge 'watch'	97	1,99	17	1,26	9,92
venster 'window'	97	1,99	25	1,41	14,1
marmer 'marble'	no data		17	1,26	18,84
nest 'nest'	97	1,99	32	1,52	19,78
vernis 'polish'	no data		8	0,95	20,78
zink 'zinc'	no data		0	0	21,43
gram 'gram'	18	1,28	31	1,51	31,58
bureau 'desk'	90	1,96	176	2,25	34,17
dozijn 'dozen'	no data		0	0	71,74
boek 'book'	98	2	1005	3	76,42
bos 'forest'	99	2	102	2,01	80,43
feest 'party'	100	2	127	2,11	86,76
artikel 'article'	7	0,9	113	2,06	92,5

Appendix 2: Resemanticisation and frequency (cf. table 4)

	Target list score (*Schaerlaekens et al*).		*Token frequency* (*CGN*)		Reseman-ticisation ratio
	%	Log	raw	Log	
peper 'pepper'	88	1,95	7	0,9	3,01
chocolade 'chocolate'	99	2	19	1,3	3,03
vangst 'catch'	53	1,73	2	0,48	3,73
limonade 'lemonade'	99	2	2	0,48	4,24
jenever 'gin'	no data		3	0,6	8,59
spinazie 'spinach'	90	1,96	9	1	9,92
sneeuw 'snow'	97	1,99	80	1,91	12,39
waarborg 'guarantee'	no data		6	0,85	15,12
suiker 'sugar'	98	2	53	1,73	20,48
kalk 'limestone'	no data		31	1,51	21,74
olie 'oil'	75	1,88	36	1,57	23,15
vlucht 'flight'	45	1,66	73	1,87	23,48
pels 'fur'	80	1,91	2	0,48	24,59
diamant 'diamond'	58	1,77	14	1,18	24,74
beet 'bite'	80	1,91	26	1,43	37,8
achterdocht 'suspicion'	no data		3	0,6	42,53
stijfsel 'starch'	14	1,18	0	0	57,14

Frequency Effects and Transitional Probabilities in L1 and L2 Speakers' Processing of Multiword Expressions

Ping-Yu Huang, David Wible and Hwa-Wei Ko

Abstract

In this study we used eye tracking techniques to investigate whether English L1 and L2 speakers were sensitive to transitional probabilities between linguistic items and whether such sensitivity was shaped by frequency of input. We focused on the linguistic construct "multiword expressions" (MWEs) (e.g. *on the other hand*) within which their final words had extremely high forward transitional probability in such syntagmatic contexts. Those final words were also embedded in non-MWE contexts which did not provide the high contextual predictability as the MWEs for comparison. According to the eye movement data we collected, both English L1 and L2 readers were sensitive to co-occurrences between words in the MWEs; both our L1 and L2 subjects showed significantly lower fixation probability and shorter fixation duration for the final words in the MWEs. Next, we conducted an input training task, which targeted seven MWEs for which one subgroup of the L2 subjects did not show the transitional probability sensitivity. In the training task we compared two types of input treatments: frequent input (providing multiple instances of exposure to the MWEs) and textually enhanced input (using textual enhancement to highlight the MWEs). The L2 subjects' eye-movement behaviors were then collected again. Based on the results of the second experiment, frequency of input more effectively reduced the L2 subjects' processing time on the final words of the MWEs. Our findings in general support the claims of usage-based or frequency-based models of language processing and learning and provide some preliminary results which confirm the effects of input frequency on language learners' exploitation of forward probabilistic relations in word sequences.

1. Introduction

Usage-based or frequency-based models of language learning claim that linguistic knowledge is mainly shaped by linguistic experience; frequency

of events in input largely determines how linguistic knowledge is represented and processed in the mind. These models suggest that humans are able to detect frequencies of events that they perceive and experience, and the information concerning the frequencies is then stored mentally and affect how the perceived events are processed and organized. The ability to detect the frequency information in input has been exemplified and evidenced by studies which demonstrated that humans, either at college levels or at young ages, could accurately measure relative frequencies with which English words occur in normal text (e.g. Hasher and Chromiak 1977; Shapiro 1969). This view does not assume, however, that human beings consciously count the events surrounding them. A statistical learning mechanism, as Saffran et al. (1997) suggest, seems to operate whenever humans perceive linguistic forms or structures in input. This mechanism enables humans to implicitly receive and record statistical information, which in turn determines how linguistic units will be integrated into the mental linguistic system and be processed in real-time language processing. Work in psycholinguistics and cognitive linguistics has shown that frequency exerts effects in various aspects of language processing or storage; for example, frequency has been confirmed to play a significant role in recognizing spoken words (Lively, Pisoni, and Goldinger 1994), creating past tense forms of irregular verbs (Seidenberg and Bruck 1990), disambiguating verb senses and sentences (MacDonald, Pearlmutter, and Seidenberg 1994), acquiring form-meaning mappings, i.e. linguistic constructions (Goldberg 2006), etc. In language acquisition, learners are claimed to implicitly detect how frequently one linguistic item is associated with one particular function. Language learning in this sense is taken as a process of acquiring such form-function associations (Bybee and Hopper 2001; Croft 2000).[1] In the present study, we would like to contribute to the literature on frequency effects upon linguistic processing by employing eye tracking techniques to investigate whether and how frequent exposure to input affects L2 learners' sensitivity to transitional probabilities or co-occurrences between words.

In psycholinguistics, some researchers have demonstrated that humans are capable of detecting co-occurrences of linguistic items presented frequently in input. Aslin, Saffran, and Newport (1998), for example, showed that 8-month infants were able to segment speech sounds into words based solely on conditional probability statistics between sounds. The discovery

1. See also Ellis (2002) for a complete review of frequency effects on language processing and acquisition.

of patterns of sounds is of great significance for language acquisition because it allows language learners to detect what orderings of sounds constitute words and what constitutes word boundaries. Most of the empirical evidence for the mental analysis of co-occurrence distributions, however, has been restricted to phonological or phonotactic studies. The only exceptions, to our knowledge, were McDonald and Shillcock (2003a, 2003b), in which sensitivity to transitional probabilities between lexical items was investigated. Specifically, in their experiments, McDonald and Shillcock hypothesized and confirmed that L1 readers tended to detect co-occurrences between words in written text, and thus spent less time fixating on a word when this word was preceded by a high forward transitional probability word (e.g. *rely* → *on*). In our research, we also explored whether readers would perceive transitional probabilities between lexical items while we focused on a different linguistic construct: multiword expressions (MWEs). According to Sag et al. (2002), multiword expressions such as *on the other hand* and *as a matter of fact* are defined as word sequences whose meaning and use cannot be completely derived from component lexemes and grammatical rules. Furthermore, words which appear in multiword expressions tend to bear strong transitional probability or co-occurrence relationships. This linguistic construct thus serves as an ideal context to study whether readers are sensitive to co-occurrences between words.

This study consisted of two main experiments. In Experiment 1, we used several corpus-derived recurrent sequences as visual stimuli in an eye tracking experiment and hypothesized that, if our subjects (both English L1 and L2 speakers) implicitly perceived the strong relationships between words in MWEs, they would spend less time on the final word of an MWE (e.g. *as a matter of fact*) which frequently co-occurs with its preceding words than they would on the same word appearing in a non-MWE context (e.g. . . . . *whether this is a fact or just*. . . . .). In addition to simply verifying humans' sensitivity to lexical transitional probabilities, we intended to further explore whether this sensitivity was indeed shaped and affected by input frequency. To do this, in Experiment 2 we conducted an input training task within which we focused on certain MWEs for which one subgroup of our L2 subjects did not show the transitional probability sensitivity in Experiment 1. In the training task, we compared effects of frequent input (providing multiple instances of exposure to the MWEs) with textually enhanced input (using text enhancement to highlight the MWEs). The L2 subjects' eye movement patterns were tracked before and after the training period. If sensitivity to conditional probability information is increased by frequency effects, as certain researchers such as

Bybee (2002) and Ellis (2002) suggest, frequent input was expected to exert stronger effects on reducing the L2 subjects' processing time on the final words of the MWEs than textually enhanced input. Basically, the L2 subjects tested in our study were a group of college students who had learned English as a foreign language for around six years in Taiwan. Given that Mandarin Chinese (the students' L1) and English are substantially different in terms of structures, we assume that the learners' processing and learning of English MWEs would hardly be effected by their L1. Nevertheless, our expectation was that our results concerning the L2 subjects can be generalized to all L2 learners since, according to frequency-based models, language learning in general is a process involving "the gradual strengthening of associations between co-occurring elements of the language" no matter whether a learner's L1 and L2 are structurally similar or different (Ellis 2002: 173).[2]

2. Previous research on sensitivity to transitional probabilities in language

Before a discussion of our eye tracking experiments, we first briefly review some studies relevant to our research. Most of the previous studies, as indicated above, were related to phonotactics. Jusczyk et al. (1993), for example, reported a series of experiments looking into whether and when sensitivity to native language sound patterns developed during the first-year human life. Their intention was to test a fundamental assumption in L1 acquisition: the innate ability to discriminate nonnative-language phonetic contrasts would be lost in the second half of the first year (Werker and Tees 1984), and infants during this period would move from language-universal to language-specific stages with respect to sound perception. Jusczyk et al. used lists of English and Dutch words and observed whether infants of the two languages would attend to the sound patterns of their native languages longer. As their results demonstrated, both American and Dutch infants at 9 months of age as expected did listen to the phonetic patterns of their native languages longer, and this sensitivity disappeared as phonotactic information of the sound strings was removed and only prosodic information was left intact. It suggests that the infants were sensitive to the phonotactic patterns of the words rather than the prosodic

2. We thank an anonymous reviewer of this article for requesting clarification of this issue.

cues. Interestingly enough, this sensitivity was not detected in 6-month infants. These findings indicate that infants tend to discover which sounds are more likely to co-occur with which to form sound patterns in their first languages during 6–10 months.

In another experiment (Jusczyk, Luce, and Charles-Luce 1994), Jusczyk and his colleagues attempted to replicate their earlier finding and further investigate whether this sensitivity was frequency-driven. The rationale behind this experiment was that infants have to learn to differentiate sounds which occur in their native language from those which do not and to recognize which strings of sounds constitute words. The assumption to be tested was that frequent input would make such learning possible. Infants after listening to and perceiving sufficient aural input of their L1 would discover which combinations of phonotactics were probable in their first language. The stimuli used in this experiment contained both highly probable patterns (e.g. [kik]) and lowly probable ones (e.g. [giθ]). Both kinds of patterns were possible sound sequences in English; they differed only in different frequencies with which they occurred in language use. Similar to the results gathered in the earlier experiment, Jusczyk, Luce, and Charles-Luce found that 9-month infants attended to high-frequency syllables significantly longer than low-frequency ones, while 6-month infants did not show this syllable differentiation. These results not only indicate *when* infants would begin to discover patterns of sounds of their native languages (i.e. 6–10 months), but suggest *how* the discovery develops in language acquisition (i.e. frequency).

Similar evidence for the implicit detection of phonotactic patterns has been reported by Saffran et al. (1997), which demonstrated how humans learned words based solely on transitional probabilities between sounds. Learning a language, as Saffran et al. claimed, entails segmentation of phonotactics into words, and the segmentation depends largely on transitional probabilities between sounds. That is, in language, sounds appearing in words side by side often bear stronger co-occurrence relationships (e.g. [r] and [aI]) while sounds across word boundaries tend to have weaker relationships (e.g. [r] and [θ]). In addition to the use of prosodic information and pauses to detect word segmentation in a speech stream, it was hypothesized that humans rely on computing transitional probabilities between sounds as well. To test this hypothesis, Saffran et al. created an artificial language which consisted of sequences of syllables. High transitional probability tri-syllable strings were assumed to be acquired by subjects as words and low transitional probability ones as sounds across

word boundaries. Saffran et al. collected strong evidence for their hypothesis; after being exposed to the speech streams for around 20 minutes, both adults and children discriminated words from nonwords. This discrimination, furthermore, occurred when the subjects focused on an illustration task rather than attend to the aural input. These findings taken together confirm the researchers' assumption that a statistical learning mechanism is at work in the mind computing and recording statistical information, and the computation basically is implicit and incidental rather than intentional.

In addition to the experiments discussed above, in the research literature, some studies report that articulatory reduction tends to occur in highly frequent word pairs (Bush 2001; Bybee and Scheibman 1999). Specifically, these studies imply that humans are able to perceive that certain words tend to co-occur with each other (e.g. *I don't* and *did you*) and reduction often occurs in these pairs in order to expedite speech processing and contribute to speech fluency. Based on these observations, Bybee (2002) indicates that co-occurring words seem not to be stored as individual words but as multiword sequences in the mental lexicon.

McDonald and Shillcock's (2003a, 2003b) experiments were the only research which studied the mental computation of transitional probabilities between linguistic items other than phonotactics. In the two eye movement experiments that they conducted, specifically, they used two different types of stimuli attempting to observe whether their subjects during reading were implicitly computing transitional probabilities between lexical items. The evidence for such computation would be shown by, for example, shorter fixation time on a word which was more probable from its preceding word. First, in McDonald and Shillcock (2003a), the researchers embedded tightly controlled verb-noun pairs in sentences where the target words in high transitional probability conditions (e.g. *avoid* <u>*confusion*</u>) were expected to attract shorter fixation durations than the ones in low transitional probability conditions (e.g. *avoid* <u>*discovery*</u>). The transitional probabilities between words were determined by statistical information from the British Nation Corpus (BNC), with the high-probability pairs sharing a significantly higher mean value (.1011) than the low-probability ones (.00038). Furthermore, to ensure that the differences in fixation durations were affected by transitional probabilities rather than by higher-level discourse information which has been found to affect eye movements (Ehrlich and Rayner 1981; Rayner and Well 1996), a cloze task was performed to make sure that the discourse information of both types of target words

was not strong enough to influence eye movements.[3] The results of this experiment showed that the effects of lexical transitional probabilities emerged clearly in the measure of initial-fixation duration. The mean initial-fixation duration for the high-probability words was significantly shorter than that for the low-probability ones. This finding led the researchers to conclude that humans rapidly estimate the transitional probabilities of words in reading, guiding the eyes to fixate for a shorter time on those words which are probable from preceding texts. In the other experiment, McDonald and Shillcock (2003b) used different stimuli to investigate the sensitivity to lexical transitional probability. Materials included in this experiment were ten newspaper articles and the researchers' intention was to see whether for all the words appearing in those articles their subjects would show a tendency to spend less time processing words which statistically tended to co-occur with their preceding words than those did not. Similar to the previous study, McDonald and Shillcock in this follow-up experiment once again collected strong evidence that transitional probabilities between words significantly affected eye fixations or cognitive processing. Although the effects observed in this experiment looked small, they were statistically significant and showed up in several eye-movement measures, including initial-fixation duration and gaze duration. Additionally, the effects of transitional probability appeared not only in forward reading, but in backward reading as well. McDonald and Shillcock found that words which were followed by high transitional probability words also tended to attract shorter fixations (e.g. *rely on*). This

3. Traditionally eye movement researchers investigating the effects of context on lexical processing use cloze tasks to determine predictability values of words. Subjects who do the tasks are presented a non-complete text or a sentence fragment and required to suggest a word which they think follows the text or fragment-string naturally (e.g. *I always thought the trip in a foreign* _____). The predictability of words therefore is influenced by the discourse-level information provided by the preceding text or fragment. In McDonald and Shillcock's (2003a) experiment, the predictability values of their high-probability and low-probability bi-grams were 7.96% and 0.79%, respectively. The difference between the two percentages was much smaller than those manipulated in previous experiments (e.g. 86% and 41% vs. 4% in Rayner and Well 1996) and McDonald and Shillcock assumed that the small difference would not affect eye movements. McDonald and Shillcock's assumption and design, however, were criticized by Frisson, Rayner, and Pickering (2005), who indicated that the effects from transitional probability found by McDonald and Shillcock might still be caused by contextual predictability. We discuss Frisson, Rayner, and Pickering's criticism and experiment in more detail in Section 4.

latter finding suggests that parafoveal preview plays an important part in reading in that readers spend less time fixating on words if in parafoveal preview they detect that the following words are congruent in context.

As indicated above, in the present study we also examined whether readers perceived co-occurrence between lexical items. Our research, however, differs from McDonald and Shillcock's (2003a; 2003b) in several respects. For example, the linguistic construct that we investigated was multiword expressions, rather than bi-grams, and, besides testing L1 readers, we further explored whether L2 learners were able to show sensitivity to lexical transitional probability. The great majority of the research in this literature tests only L1 subjects and rarely addresses whether L2 learners are able to psychologically connect adjacent linguistic items as L1 speakers. Our results hopefully will help answer this question. In addition to these issues, another important characteristic of our research was that we provided some input containing MWEs to L2 learners and examine the input effects. We tracked our L2 subjects' eye movements on and processing of selected MWEs before and after the input exposure and attempted to directly examine whether frequent input was beneficial to human beings' estimate of transitional probability in language, as usage-based and frequency-based theories claim.

3. The present study

This study consisted of two eye-movement experiments. In the first experiment, our purpose was to understand whether both L1 and L2 speakers were able to show sensitivity to transitional probabilities between words in multiword expressions. In the second experiment, we investigated whether such sensitivity was indeed increased and affected by frequency of input. The methodology and results of the two experiments are detailed below.

3.1. Experiment 1: Sensitivity to lexical transitional probability in L1 and L2

The goal of Experiment 1 was to check whether the initial words of multiword expressions would enable readers to accurately predict the final words (e.g. *as a matter of* → *fact*). In the past research on MWEs, there has been little consensus on how MWEs are defined and what kind of strings constitutes MWEs (e.g. Cruys and Moirón 2007; Sag et al. 2002). In our research the MWEs that we tested were recurrent and frequent word sequences in normal text, regardless of whether the sequences can be inter-

preted compositionally or not. Those sequences were extracted from the BNC and the words in them basically would bear strong co-occurrence relationships.

3.1.1. Method

3.1.1.1. Participants

Fourteen English L1 speakers participated in this experiment as the L1 subjects. They came from several different English-speaking countries, including America, Canada, England, New Zealand, and South Africa. To ensure that they were proficient readers of English, they were required to read four short reading passages with their eye fixations and movements monitored before they read the research materials. On average, they read around 245 words per minute and their mean fixation duration was 210 milliseconds. The two figures were comparable to the L1 reading data reported in previous eye tracking experiments (e.g. Just and Carpenter 1987; McConkie et al. 1991).

The L2 subject group was comprised of thirty freshman students in the National Central University in Taiwan. They majored in technology or engineering in the National Central University and had learned English as a foreign language for more than six years. As the L1 subjects, the L2 learners' English reading proficiency level was assessed. As their eye movement data on the four reading passages showed, in general the learners read much slower than the L1 subjects. The learners' WPM (words per minute) was 129 and their average fixation was 253 milliseconds.

3.1.1.2. Materials

To prepare a set of multiword expressions, we used a computational chunking tool to search a 20 million word proportion of the BNC. This chunking tool was created by Wible et al. (2006) to automatically generate a list of word sequences or patterns for a target word. Here we use the word *fact* as an example to illustrate the search procedures. First, after *fact* was fed into the system, the chunking tool began looking for words which tended to co-occur with it (e.g. *matter…fact*) in the BNC. Then, those word pairs or bi-grams were used as key words (or phrases) to search the corpus again. At this time, the third words which statistically were more likely to appear close to those bi-grams were the targets (e.g. *matter of fact*). The search procedures would be iterated continuously until no other words were found. The results of these procedures were a list of word sequences or patterns containing the target word (e.g. *as a matter of fact*). Basically, the chunking tool used several corpus-based

measures such as mutual information and conditional probability to decide word association strengths. In this regard, we were convinced that the words in the sequences found by the chunking tool did have strong co-occurrence relationships.

Our research materials were: (1) sentences which contained twenty-five MWEs and (2) sentences which included only the final words of the MWEs. The MWEs (e.g. *the sort of things, on the other hand, all over the world*) were extracted from the BNC with the chunking tool and each was put into three sentences. The mean length of the MWEs was 4.22 words long and their final words were 5.28 letters on average. As for the other set of sentences, they contained only the last words of the MWEs and we created these sentences with special care that the target words were not part of any recurrent lexical patterns. Each of the last words of the MWEs appeared in two sentences.[4] Example sentences for the two sets of research materials are listed below. The MWEs are underlined and the target words are in bold here for convenience; the strings were shown to the subjects normally without the underlining or boldface in the eye movement experiment. A complete list of the sentences is provided in the Appendix.

a. *These are not the sort of **things** that any student at this school should be carrying in their bag or pocket.*
b. *A wide variety of people from all over the **world** are united in the fight against continuing environmental pollution.*
c. *He said that he would come and pick up all his **things** before noon but then he called and said he was running a little late.*
d. *It is hard to imagine finding a **world** with no living creatures, but this may be what space exploration reveals.*

3.1.1.3. Research apparatus and procedures

We collected the subjects' eye movement behaviors individually with an SR Research EyeLink Eye Tracker System. The eye tracker consisted of

4. In Experiment 1, in addition to the two sets of sentences, one more set of sentences (36 ones) was also read by our subjects with which we intended to investigate some other research issues. All of the sentences were randomized as they were shown to the subjects. About the two sets targeted here, although they involved different numbers of sentences, this design was not found to affect how our subjects perceived the target words in either MWEs or non-MWE contexts. More specifically, both the L1 and L2 subjects' reading time for the target words in the MWEs was not significantly shorter as their eyes landed on the words for the third time.

two P4-1.8 GHz computers and an eye-movement detector that tracked the subjects' eye fixations and movements from the right eye (but the reading was binocular). Fixation position was measured and determined every four milliseconds and every velocity over 30 degree/second was marked as a saccade (i.e. an eye movement).

As a subject arrived at our eye tracking lab, a calibration procedure was performed to make sure the equipment correctly tracked where the subject fixated. The procedure often took 5 minutes. Subsequently, the subjects were required to read the four reading passages which checked their reading proficiency and the sentences which included the MWEs or the last words of the MWEs. The subjects were informed that the purpose of the experiment was to understand their normal reading speed so they were encouraged to read as they did in daily life. Most of the subjects took one or two five-minute break(s) during the experiment and the calibration procedure was carried out after each break. Additionally, the experimenter monitored the subjects' fixations closely and suggested some subjects take a short break whenever he found that the eye tracker did not accurately track where the subjects fixated.

3.1.2. Results and discussion

A small set of the collected eye tracking data was removed from the analysis due to track losses. Fixation durations which were either too short (less than 100 milliseconds) or too long (over 800 milliseconds) were also eliminated because they tended to reflect physical programming or were treated as blinks rather than revealing cognitive processing (Morris 1994). Here we report three eye-movement measures with respect to the processing of the target words: fixation probability (i.e. the probability of fixating at target words), first-fixation duration (i.e. the average duration of the first fixations on target words), and gaze duration (i.e. the average of all fixations on target words including both initial fixations and re-fixations in first-pass reading). If the subjects were sensitive to lexical transitional probabilities in the tested MWEs, they would show: (1) lower fixation probability and (2) shorter first-fixation and gaze durations on the final words of the MWEs than the same words in non-MWE contexts.

Both the L1 and L2 subjects' eye-movement data concerning the three measures are displayed in Table 1. The data were subjected to a two-way analysis of variance (ANOVA), with word context (MWE vs. non-MWE) and participant group (L1 vs. L2) as the main factors. First, concerning fixation probability, the ANOVA indicated that the L1 subjects showed

a significantly lower probability than the L2 subjects ($F(1,42) = 23.808$, $p < .01$, $\eta 2 = .362$) and the words in the MWEs were fixated less frequently than in non-MWE contexts ($F(1,42) = 9.875$, $p < .01$, $\eta 2 = .190$). There was no significant interaction between the two factors ($F(1,42) = 2.935$, $p = .094$). About first-fixation duration, the L1 subjects demonstrated shorter fixations than the L2 learners ($F(1,42) = 112.227$, $p < .01$, $\eta 2 = .728$), with the words in the MWEs gaining significantly shorter fixations ($F(1,42) = 13.081$, $p < .01$, $\eta 2 = .237$). The interaction between the factors statistically was not reliable ($F(1,42) = .238$, $p = .629$). Finally, for the gaze durations, the ANOVA revealed significant effects of both participant group ($F(1,42) = 116.220$, $p < .01$, $\eta 2 = .735$) and word context ($F(1,42) = 19.729$, $p < .01$, $\eta 2 = .320$). An analysis of the interaction, again, suggested that it did not exist between the two main effects ($F(1,42) = 2.640$, $p = .112$). Taken together, the statistical analyses confirm the strong effects of transitional probability in MWEs on L1 and L2 readers' eye movements and cognitive processing. When readers encounter MWEs in reading, they tend to be sensitive to the statistical relations between words in the MWEs. After they read over the initial words of a multiword expression, they will mentally predict the last word of the MWE and then either skip this word or process it in a short time. This also indirectly explains why the L1 and L2 subjects were more likely to fixate at the words in non-MWE contexts or gazed at the words longer. Even though the words were plausible in the non-MWE contexts, they were relatively less predictable and thus gained higher fixation probability as well as longer processing time.

Table 1. Average fixation probability, first-fixation duration, and gaze duration for L1 and L2 target word processing

	L1 Subjects			**L2 Subjects**		
	Fixation Probability	**First-Fixation Duration**	**Gaze Duration**	**Fixation Probability**	**First-Fixation Duration**	**Gaze Duration**
Contexts						
MWE	72%	189.81 (14)	201.40 (15)	89%	269.48 (24)	324.67 (27)
Non-MWE	77%	198.73 (17)	210.26 (20)	91%	278.76 (22)	352.32 (43)

Note. The first-fixation and gaze durations are in milliseconds and the figures in parentheses are standard deviations.

The results of Experiment 1 confirmed McDonald and Shillcock's (2003a, 2003b) findings that transitional probability between lexical items has strong effects on word identification or processing in reading. Readers implicitly compute the statistical information concerning which word tends to follow which and such computation can be demonstrated clearly by eye tracking. Our results further indicate that the sensitivity to lexical transitional probability holds not only in L1, but in L2 readers' mental processing as well. In language development, it seems that L2 learners as L1 learners are able to acquire the statistical information from reading experiences and integrate the information into their L2 system which enables the learners to read more proficiently and fluently. The L2 acquisition of the statistical information based on input, however, has rarely been investigated and requires more empirical evidence. Experiment 2 of the present study was therefore conducted to examine whether the sensitivity to transitional probability between words was indeed affected by frequency of input. We designed an input training task which compared the effects of frequent input and textually enhanced input and studied which type of input was more effective and useful.

3.2. Experiment 2: Frequency effects on processing multiword expressions in L2

In Experiment 2, we investigated whether frequency of input allowed L2 learners to detect the likelihood that a word tends to occur following a string of words in written text. A subgroup of the L2 subjects tested in Experiment 1 was targeted and certain multiword expressions were selected for which the subgroup of L2 subjects did not show the transitional probability sensitivity. In the subsections below, we will first discuss the selected L2 subjects and MWEs, and indicate how the MWEs were presented to the L2 subjects in the input training task. Then, we will describe the method of the second eye-movement experiment and report its results.

3.2.1. The input training task

Seven multiword expressions and sixteen L2 subjects were chosen for the input training task. For the seven MWEs, the sixteen L2 subjects did not show the transitional probability sensitivity in them and even processed the target words more quickly in non-MWE contexts than in MWEs in Experiment 1. The seven MWE were: *spend a great deal of time, a large sum of money, cause and effect, from the point of view, face to face, for the same reason,* and *vary from person to person*. Table 2 presents the sixteen

Table 2. Average first-fixation duration and gaze duration for the 16 L2 learners' target word processing in the seven MWEs/non-MWE contexts

	Sixteen L2 Subjects	
	First-Fixation Duration	**Gaze Duration**
Contexts		
MWEs	264.35 (32)	318.41 (59)
Non-MWE	255.04 (28)	307.15 (55)

L2 learners' processing time (first-fixation and gaze durations) on the target words.

One important feature displayed in Table 2 was the large standard deviations of the gaze durations. They appear to imply that the sixteen learners might be at different proficiency levels of English reading; some of the L2 subjects processed the target words with relatively shorter gaze durations (200–250 milliseconds) while some spent a long time recognizing those words (more than 400 milliseconds). It suggests that higher-proficiency L2 learners might not be familiar with more MWEs than lower-proficiency learners (Dörnyei, Durow, and Zahran 2004). Regarding the sixteen L2 learners targeted here, although they were at different levels of reading proficiency in English, they all shared a characteristic that in the seven MWEs they were not sensitive to the transitional probability between the lexical items.

The seven multiword expressions were presented to the sixteen L2 learners in three lessons through an online language learning platform, IWiLL (Intelligent Web-based Interactive Language Learning).[5] We divided the seven sequences into two groups to create two types of input materials. The frequent input materials contained four MWEs: *a large sum of money*, *face to face*, *for the same reason*, and *vary from person to person*, with each of them appearing in each lesson five times without any special marking in example sentences. The textually enhanced input materials included the other three MWEs. Specifically, for the three MWEs, we used underlining to highlight them.[6] The three MWEs appeared only once

5. See Wible et al. (2001) for a clear description of IWiLL.
6. We created the textually enhanced input materials partly based on the research design of Bishop (2004) who demonstrated that underlining led L2 learners to notice multiword units more.

in example sentences in each lesson. The sixteen L2 subjects were required to get access to the lessons anytime in an eight-day period. Each of the lessons included about ten pages; following a webpage which showed three to five example sentences including the MWEs, a question page was presented which asked a comprehension question to test understanding of one of the sentences showing up on the previous page in order to ensure that the learners did read the example sentences attentively and carefully. If frequency of input as frequency-based or usage-based models claim is beneficial to language learners' detection of co-occurrence of linguistic units, the frequent input materials would lead the L2 subjects to process the final words of the MWEs more quickly than the same words in non-MWE contexts more effectively than the textually enhanced input materials.

3.2.2. Method of the second eye-movement experiment

The sixteen L2 subjects were asked to participate in an eye-movement task again right after they completed the input training task. The visual stimuli of the second experiment were the sentences including the target words in the two different contexts used in Experiment 1. The only difference was that the seven MWEs appeared in only two sentences rather than three. The second eye-movement experiment was performed around one and a half months following the first experiment, a period which we assumed long enough for the L2 subjects not to recall what they had read in Experiment 1.

The procedures and experimental apparatus were generally the same as the first experiment. The calibration was conducted in a much shorter time since the L2 subjects were rather familiar with eye tracking procedures. After the calibration, the L2 subjects began to read a practice sentence and then the research materials. Unlike Experiment 1, most of the L2 subjects in Experiment 2 did the eye tracking experiment very smoothly without taking any breaks. All of the subjects finished the experiment within twenty-five minutes. After the second experiment finished, three subjects were briefly interviewed, and it appeared that they had no idea of the main purposes of the input training task and Experiment 2.

3.2.3. Results and discussion

The second experiment was run more smoothly and therefore, compared with Experiment 1, a much smaller proportion of data was excluded from analysis. In total only 2% data which either were un-trackable due to track losses or involved extremely long or short fixations were removed. We conducted two separate one-way ANOVAs to examine the effects of the

Table 3. Average first-fixation duration and gaze duration for the 16 L2 subjects' target word processing in Experiment 2

	Sixteen L2 Subjects			
	Frequent Input		**Textually Enhanced Input**	
	First-Fixation Duration	**Gaze Duration**	**First-Fixation Duration**	**Gaze Duration**
Contexts				
MWE	234.69 (35)	292.27 (68)	263.98 (59)	317.88 (87)
Non-MWE	268.51 (44)	319.34 (60)	253.32 (37)	322.23 (78)

two types of input treatments, and, as in Experiment 1, the evidence for the sensitivity to transitional probability between words in the multiword expressions would be shown by the shorter reading time on the target words appearing in MWEs. Table 3 shows the L2 subjects' average first-fixation duration and gaze duration data in terms of the two input treatment types.

At first sight, the data, especially those about the gaze durations, seem to indicate that the frequent input and textually enhanced input materials were equally effective in guiding the learners to predict the final words of the tested MWEs. The two ANOVAs, however, revealed that the effects of textually enhanced input might not be comparable to those of frequent input. Specifically, regarding the frequent input MWEs, the ANOVA indicated that the first-fixation duration and gaze duration differences were highly or marginally significant ($F(1,15) = 9.043$, $p < .01$, $\eta 2 = .376$; $F(1,15) = 4.182$, $p = .059$, $\eta 2 = .218$). Similar statistical significance was not observed in the results concerning the textually enhanced input MWEs. As the ANOVA revealed, although the average gaze duration on the final words of the three MWEs treated through textually enhanced input was shorter, the 4–5 millisecond difference was not statistically significant ($F(1,15) = .039$, $p = .847$). As for the first-fixation durations, the mean duration for the target words in the MWEs was even longer. Taking all the results and findings together, Experiment 2 demonstrated that frequent input was rather effective in facilitating L2 learners' processing of multiword expressions. Compared with textually enhanced input, frequent input better led the sixteen L2 subjects to detect the forward transitional probability of the words in the multiword expressions, and con-

sequently these subjects showed expedited processing on the last words of the sequences.

4. Discussion

In this study we addressed two main issues, i.e. whether L1 and L2 readers were sensitive to forward transitional probability between lexical items in reading and whether such sensitivity was shaped by frequency and experiences of input. We found that both L1 and L2 subjects recognized the last words of multiword expressions significantly faster than the same words embedded in non-MWE contexts and that frequent input effectively facilitated targeted L2 subjects' sensitivity to transitional probabilities in multiword units, whereas the second finding was based on few tested items and requires further experimental evidence. Below we discuss these findings in relation to some relevant research.

First, the results of Experiment 1 confirmed that human beings tend to detect co-occurrence probabilities of linguistic items (Aslin, Saffran, and Newport 1998; Jusczyk et al. 1993; Jusczyk, Luce, and Charles-Luce 1994; McDonald and Shillcock 2003a, 2003b; Saffran et al. 1997). Specifically, in Experiment 1 we put multiword expressions or only their final words in sentences and found that both the L1 and L2 subjects recognized that the target words in MWEs were the candidates which statistically often follow their preceding words in normal text and thus processed these words more quickly. McDonald and Shillcock (2003b) claim that the sensitivity to and implicit computation of transitional probability between words is an important component of proficient reading. As human beings are exposed to written texts, their language processor would automatically compute the likelihood that one word appears following another word or word string. In the cases where a two-word sequence shows a strong transitional probability relationship, the second word of the sequence would be processed more quickly than it is when preceded by a lower transitional probability word. McDonald and Shillcock demonstrate that transitional probability between lexical units in a two-word window (i.e. bi-gram) does affect readers' eye-movement patterns and cognitive processing while in the present study we focus on a different linguistic construct, multiword expressions, and show that readers indeed are sensitive to lexical transitional probabilities during reading. One of our major contributions to the relevant research literature is that we found the lexical probability sensitivity is not only enjoyed by L1 speakers, but by L2 learners as well. Although the

L2 subjects in Experiment 1 relative to the L1 subjects read much slower and had much longer fixation durations, the L2 data did indicate that the learners detected the strong forward transitional probability between words in MWEs and clearly differentiated their processing of the target words in MWEs and in non-MWE contexts. L2 acquisition in this regard is approximately similar to L1 acquisition since both involve implicit computation of transitional probability of linguistic units which has been assumed a key aspect of language acquisition (Saffran et al. 1997). Although both the experiments by McDonald and Shillcock and the present research confirm the effects of lexical transitional probability on eye movements, it is important to note that the effects are questioned and challenged by some eye-movement researchers. In Frisson, Rayner, and Pickering (2005), for instance, the researchers criticized that the effects of transitional probability in actuality were caused by contextual predictability. In a strictly controlled experiment in which the contextual predictability values of both higher- and lower-transitional probability bi-grams were held constant, Frisson, Rayner, and Pickering found no effects of transitional probability on eye movements. Transitional probability between words alone seems not to influence eye movement patterns and lexical processing. Here we intend not to get into the argument over whether transitional probability effects exist independently from contextual predictability. We did not separately test the effects of transitional probability and contextual predictability and thus could not show which exerted stronger effects or whether it was likely that transitional probability alone did not have effects on lexical processing. To date, the experiment by Frisson et al. has been the only study which reported such a finding which still needs more empirical evidence. Our suggestion is that transitional probability between lexical units should be regarded as one type of contexts like discourse information or previous mentions which affect predictability and recognition of words. Our results basically clearly demonstrate that transitional probability of words in MWEs has strong facilitative effects on the recognition of the last words of the MWEs.[7]

7. One anonymous reviewer of this paper indicated that the MWEs we used were rather heterogeneous and they might be processed by different mechanisms. While we acknowledge that the processing speed of MWEs, as Gibbs and Gonzales (1985) showed, was influenced by degree of fixedness, we would like to point out that one of the main purposes of our experiments was to demonstrate that the initial words of an MWE expedited the recognition of the final word and it was necessary for us to employ a wide range of MWEs.

In addition to confirming readers' detection of lexical transitional probability in processing multiword sequences, we further investigated whether such detection or sensitivity was affected by input frequency. According to usage-based models of language acquisition, linguistic knowledge and processing are claimed to be shaped and determined by input. The more often two linguistic units are encountered together in input, the more likely that the two units will be associated in future processing or in mental representations (Bybee 2002; Ellis 2002). Our purpose in Experiment 2 was to empirically test such claims by usage-based models, examining whether being exposed to multiword expressions frequently in written input would help L2 learners detect that the words in MWEs tend to co-occur with each other. The results of Experiment 2 in general were consistent with such claims; for the four MWEs treated through frequent input, the targeted sixteen L2 learners processed their final words significantly faster than the same words in non-MWE contexts. Similar effects were not observed in the textually enhanced input MWEs; the average first-fixation duration and gaze duration on the last words of the textually enhanced input MWEs were not significantly shorter than those for the same words appearing in non-MWE contexts. These results taken together suggest that encountering a word sequence frequently in input is much more facilitative than paying temporary attention to the sequence for leading L2 learners to mentally associate words in the sequence. Experiment 2 provided some clear evidence that frequency has beneficial effects on the sensitivity to co-occurrence between lexical items.

Since it has been shown that sensitivity to transitional probability of linguistic units develops based on input frequency, how does this finding inform language acquisition research? In the literature review in Section 2, we noted that transitional probability in phonotactic patterns provides

---

Those MWEs might vary in terms of fixedness, but they did share a characteristic that the words within them had strong co-occurrence relationships which made their final words highly predictable and easier to be processed, as evidenced in our results. The reviewer also questioned whether the results of our Experiment 2 were due to different treatment methods or due to dissimilarity of test items, as the frequent input materials included two semi-fixed sequences (*vary from person to person* and *face to face*) among the tested four MWEs. Basically, we were convinced that our frequent input treatment was effective because, for the other two MWEs, it was found that the processing time of their final words was significantly shorter than that of the same words in non-MWE contexts. We appreciate the reviewer for raising the questions which allow us to provide clarifications.

useful information for L1 acquirers to discover sound sequences as words and word boundaries (Saffran et al. 1997). In fact, such discovery and analysis of statistical distribution in language input have been considered important learning mechanisms for language acquisition in the field of cognitive linguistics (Croft and Cruse 1999) in which researchers claim that segmentation of linguistic elements is not restricted to phonotactics. Mental processing of lexical transitional probability and segmentation of lexical items, for example, have been assumed to play a central part in the acquisition of grammatical knowledge, which is seen as a collection or system of linguistic constructions in the mind (Ellis 2003). Below we summarize important views of constructivist approaches to language acquisition, and indicate some of our contributions to them.

Constructivists of language learning hold that language learning involves the operations of simple learning mechanisms. Structures of language are complex, but the mechanisms underlying language learning might be simple (Saffran 2003). Constructivists dismiss the view that language comprises a system of universal and language-specific grammatical rules, instead claiming that language acquisition basically involves learning of several thousands of form-function mappings, i.e. constructions. According to Goldberg (2006), constructions refer to form-meaning correspondences which members of a speech community use as conventions. Constructions may be short and rather simple, as a noun phrase (Det Noun), or be long and complex, as a ditransitive structure (Subj V Obj1 Obj2). To integrate these constructions into mental linguistic systems, constructivists and cognitive linguists believe that a language learner needs to encounter the structures or patterns frequently in input. A statistical learning mechanism is presumed to operate which implicitly connects smaller linguistic units to form larger sequences with the sequences later being abstracted into constructions. Ellis (1996, 2002, 2003) points out that the development of constructions might consist of three steps, beginning from memorized formulas to low scope patterns and then to constructions. Formulas, as the multiword expressions such as *on the other hand* and *as a matter of fact* that we used in our research, are memorized and fixed sequences of lexical items (see also Weinert 1995 and Wray 2002 for a review of formulaic language in L2). Those sequences are claimed to be acquired and processed as fixed expressions rather than compositionally, and frequent exposure to input is required to make mental associations of the words possible. As a collection of formula is represented and stored in a learner's mind, the formulas would be analyzed into limited scope patterns. Ellis (2003: 70) specifies that limited scope patterns are short

"slot-and-frame patterns" in which some words are fixed and some word positions are open for learners to fill words in (e.g. *I can't* ___). Finally the patterns will be further abstracted into constructions such as the ditransitive structure mentioned above or the conditional construction (*The Xer, the Yer*). In the present study, we offer some preliminary evidence for the first-stage of the learning sequence. Specifically we demonstrated that frequent input allowed the L2 subjects to psychologically associate words in the tested MWEs and thus process the MWEs as fixed expressions. Whether L2 learners do abstract the fixed strings of words into low scope patterns or constructions is an interesting research issue which needs to be addressed by future studies. As Ellis (2003) suggests, the data from large-scale L2 corpora or longitudinal studies are especially useful for L2 researchers to explore whether the learning sequence does accurately predict or account for L2 acquisition of grammatical knowledge.

Finally, we would like to indicate some limitations of the present research. Those limitations basically concern the design and method of our eye-movement experiments which may be overcome in future studies. Considering the participants and items tested in our Experiment 2, for example, obviously the numbers of subjects and multiword expressions were fairly small. We focused on only sixteen L2 learners and seven MWEs in the second eye tracking experiment. A natural next question is whether the same effects produced by frequent input can be replicated by a study which involves a larger group of subjects and more word sequences. A study like this would provide more conclusive evidence concerning whether human beings indeed implicitly compute transitional probability between lexical items in their linguistic input. Another limitation of our research was that the effects of frequent input that we observed might be short-term and not last over time. We examined the effects of the frequent/textually enhanced input immediately following the input training task and could not verify whether the input would have lasting effects. These issues were limitations of the present study, but they also raise interesting possibilities for further studies. Our experiments in general found that L2 readers, as L1 readers, are able to show sensitivity to transitional probability of words during reading, and the computation involved basically is frequency-driven. Although this study was limited in certain ways, we believe it offers evidence for the effects of frequency and transitional probability on both L1 and L2 speakers' processing. We also expect to see more psycholinguistic research techniques such as eye tracking exploited to examine the claims of cognitive linguists.

Appendix: Sentences containing the tested MWEs or only the last words of the MWEs

1a) We are often forced to spend a great deal of time on things that aren't very interesting to us.
1b) All doctors must spend a great deal of time reading new medical textbooks and papers during their vacations.
1c) The naturally gifted persons don't need to spend a great deal of time practicing their chosen discipline in order to reach a high level.
2a) I couldn't really put it another way without avoiding the whole truth but I am sorry to have upset you.
2b) If it was possible to put it another way then I would certainly consider doing that to avoid offending anybody.
2c) Maybe you could put it another way since at the moment it sounds like that you are blaming me for all the problems that have occurred.
3a) These are not the sort of things that any student at this school should be carrying in their bag or pocket.
3b) Those are not the sort of things you should say to somebody unless you are sure that it is the truth.
3c) Are you sure that those are the sort of things she had in mind when she asked you to buy some paintings for her?
4a) A wide variety of people from all over the world are united in the fight against continuing environmental pollution.
4b) We can see this happening all over the world from Asia and Africa to Europe and the Americas.
4c) This decision affects people from all over the world and it will also affect the lives of their children and grandchildren.
5a) It appears that finding the solution to the problem isn't as easy as we thought it was.
5b) That isn't really the best solution to the problem but it will just have to do for now.
5c) The process of looking for a solution to the problem will often teach you more than the solution itself.
6a) If I find on the other hand that you have been lying, then I won't be very happy.
6b) He said that on the other hand it was a well paid job with regular hours.
6c) It was reported that on the other hand the army had restored order to the city center despite the deaths.

7a) It is easier to travel from country to country now than it ever has been.
7b) The disease spread quickly from country to country leaving thousands of sick people in its wake.
7c) The movement of information from country to country has been speeded up by computer networks.
8a) In response to your query the answer to the question is not something that I can tell you at the moment.
8b) His face told me that the answer to the question was something he would not enjoy giving to me.
8c) I think that the answer to the question should be clear from what we have read already.
9a) They can't be expected to work seven days a week every week all year; it's not reasonable.
9b) This store will remain open seven days a week even during the Chinese New Year vacation.
9c) We want to practice seven days a week so that we are ready for the big event.
10a) What I'm trying to say in other words is that I will not give you permission to go on this trip.
10b) What the letter says in other words is that you will have to pay quite a lot of money to get your scooter back.
10c) I think what he means in other words is that that there is nothing he can do to help you at the moment.
11a) What he meant as a matter of fact was that nothing can travel faster than the speed of light.
11b) What I'm saying as a matter of fact has nothing to do with what happened yesterday.
11c) I think that as a matter of fact scooters are much more environmentally friendly than most cars.
12a) City lights are so bright that even on a dark night with no moon we can see the other side of the mountain very clearly.
12b) It's not very pleasant on a dark night to be out walking alone in the countryside.
12c) You can hardly see it on a dark night but during the day it looks really close.
13a) Even things that appear perfectly smooth to the naked eye are quite rough when viewed through a microscope show themselves to be quite rough.

13b) Relatively few stars are visible to the naked eye at night so astronomy is only made practical by the use of powerful telescopes.
13c) Snowflakes might all appear identical to the naked eye but in fact each one is unique.
14a) Someone just put a large sum of money into my bank account and I have no idea who it was.
14b) The terrorists demanded a large sum of money for the safe return of their hostages.
14c) The band demanded a large sum of money to play for only thirty minutes so we decided not to hire them after all.
15a) The government announced that one way or the other they are determined to reduce the number of people without jobs.
15b) I'm sure that one way or the other you will achieve your goals if you really put your mind to it.
15c) The manager warned his staff that one way or the other the department was going to have to work more efficiently.
16a) The students arranged themselves in a straight line while waiting to go into the cafeteria.
16b) The scooters were parked in a straight line along the side of the road from the junction all the way down to the bridge.
16c) The prisoners were made to stand in a straight line for several hours while their cells were being cleaned.
17a) It's a very simple cause and effect connection; if you eat too much and don't exercise then you will get fat.
17b) It is true that cause and effect is sometimes difficult to be sure of but in this case I think there is no doubt.
17c) If you don't understand simple cause and effect relationships then you will not get very far with physics.
18a) Mark made sure his studies were put to good use as soon as he joined the company.
18b) I know my money will be put to good use if I make a donation to that animal charity.
18c) More volunteers can always be put to good use visiting the elderly and giving them information on how to keep warm in winter.
19a) If we look at this from the point of view of the customer I don't think they will see any real improvements in our service.
19b) It seems clear that from the point of view of the school, people playing basketball in the evenings just create security problems.

19c) We tried to consider this from the point of view of everyone who has worked on this project even if they disagree with the committee members.
20a) The new school swimming pool was completed bit by bit as new funds became available.
20b) I learned the child's complete story bit by bit as he came to trust me more and more.
20c) The snow melted bit by bit as the weather warmed up and it began to rain lightly.
21a) Finally meeting my pen pal face to face after three years of communicating by letter was great.
21b) The Italian and the French champions met face to face after a long period of anticipation.
21c) I find it easier to deal face to face rather than trying to do things via email or fax.
22a) Please, can I borrow your secretary for half an hour just to type up this urgent proposal?
22b) I need to talk to you for half an hour to arrange the schedule for the trip to Japan next week.
22c) I'll lend this to you for half an hour but after that I really need it back.
23a) I can't go this time for the same reason I couldn't go last time; it's just too expensive.
23b) You need to do it for the same reason as everyone else is doing it; it is part of the course.
23c) I did the job for the same reason you did it; I needed the money.
24a) The effectiveness of drugs will vary from person to person depending on a variety of factors including body weight and age.
24b) Speed of learning does vary from person to person but it is largely dependent on how much practice you put in.
24c) Tolerance to alcohol does vary from person to person so if you want to drink at all, please don't drive.
25a) Speaking slowly or in a loud voice doesn't help someone understand you if they don't speak your language.
25b) Conducting an unnecessary conversation in a loud voice somewhere quiet like a library is generally considered to be impolite.
25c) Talking on your cell phone in a loud voice in a public place is starting to become unacceptable.
26a) I really don't have the time to do all the things you asked me to do before tomorrow.

26b) I'm not sure of the time because I left my watch at home and my mobile phone has run out of power.
27a) I would really like to find a way to get from my house to my office faster because I spend too long commuting.
27b) You will find that there is usually a way to reach a solution to almost every problem.
28a) He said that he would come and pick up all his things before noon but then he called and said he was running a little late.
28b) It seems that things have been a little crazy here since you became the world champion.
29a) We all tend to expect a world just like the one we grew up in when we return home.
29b) It is hard to imagine finding a world with no living creatures, but this may be what space exploration reveals.
30a) If there is going to be a problem to get it done by tomorrow then I can come back at a more convenient time.
30b) I'm sorry, there has been a problem with the server so I'm afraid you won't be able to download anything for the next few hours.
31a) The inspectors carefully examined the hand and decided it did not belong to the victim.
31b) You can consider yourself very lucky that your hand wasn't more badly damaged in that accident.
32a) I always find that a trip in a foreign country is the best way to recover from a tiring semester at school.
32b) He was the first person from our country to qualify for the final of any Olympic event.
33a) Solving most problems is a question of using the right negotiation techniques to ensure that everybody involved ends up feeling satisfied.
33b) Do you have time to respond to a question before you leave to take a taxi to the airport?
34a) It has been a week since I heard from her, I hope that she is OK.
34b) It looks like they might need a week to finish the project.
35a) Even though I couldn't see his face, his words came to me very clearly through the dark sky.
35b) History shows us that words can change the way that people think and feel should never be underestimated.
36a) You need to decide whether this is a fact or just something made up by the gossip columnists.

36b) After all of this speculation can you at least give me a fact or two that I can use in the report.
37a) Before a big test a lot of my classmates study for a night or two in an attempt to improve their scores.
37b) My roommates consider that it has been an early night if they return home before 4am.
38a) I wasn't looking forward to dissecting the human eye that was waiting for me in the biology laboratory.
38b) We were surprised and a bit scared when we found an eye along the path during our hike.
39a) I was hoping that I would have the money for six weeks in Europe but I wasn't very confident.
39b) I honestly think that all of that money will not really make them any happier.
40a) Do you think that we could watch the other channel because a program is on that I really want to see?
40b) I'd like to try the other shirt on again because this one is a little too big.
41a) When I arrived there was a line of people standing outside the theater waiting for tickets.
41b) The heavily polluting line of slowly moving cars on the freeway stretched all the way to the horizon.
42a) You can achieve an interesting effect in photographs if you allow the camera shutter to remain open for a long time.
42b) A serious disease not only has an effect on the person with the disease but also the family members and friends around them.
43a) I don't think that I really understand the use of this device that my friends bought me.
43b) I really need to think of a use for all the dozens of spare cables I have in my house.
44a) There is an unbelievable view of the mountains from the roof of my new house in the suburbs.
44b) From the cheaper seats the view of the stage was half-obscured by a giant concrete column.
45a) If I have another bit of this delicious cake I think I will probably be sick.
45b) I am sure that I lost every bit of information on my computer when it crashed early this morning.
46a) I didn't see a single face that I could recognize out of the more than one hundred people.

46b) From where I was standing the face of the cliff looked perfectly sheer and very difficult to climb.
47a) The entire group was more than an hour late this morning because of a crash just outside a tunnel.
47b) It is widely known that an hour with a lawyer asking for advice can cost you more than 10000NT.
48a) If that is the best reason you can think of for being late then I have no choice but to punish you.
48b) I cannot see any good reason for you to not be able to complete this on time.
49a) I think the kind of person you need for this job is someone who enjoys children.
49b) By the time I finished work there wasn't another person anywhere in the office building.
50a) Even with my music playing I could notice her voice as she shouted at my brother for forgetting to buy something.
50b) I wanted to run from the bear but a voice in my head told me that it would be safer if I stood perfectly still.

References

Aslin, Richard N., Jenny R. Saffran and Elissa L. Newport
1998 Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9(4): 321–324.

Balota, David, Alexander Pollatsek and Keith Rayner
1985 The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology* 17(3): 364–390.

Bush, Nathan
2001 Frequency effects and word-boundary palatalization in English. In: Joan Bybee and Paul Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*, 255–280. Philadelphia, PA: John Benjamins.

Bybee, Joan
2002 Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition* 24(2): 215–221.

Bybee, Joan and Paul Hopper
2001 *Frequency and the Emergence of Linguistic Structure*. Philadelphia, PA: John Benjamins.

Bybee, Joan and Joanne Scheibman
1999 The effect of usage on degree of constituency: The reduction of *don't* in American English. *Linguistics* 37(4): 575–596.

Croft, William
2000 *Explaining Language Change: An Evolutionary Approach*. London: Longman.
Croft, William and D. Alan Cruse
1999 *Cognitive Linguistics*. Cambridge: Cambridge University Press.
Cruys, Tim Van de and B. Villada Moirón
2007 Semantics-based multiword expression extraction. In Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, 25–32.
Dörnyei, Zoltán, Valerie Durow and Khawla Zahran
2004 Individual differences and their effects on formulaic sequence acquisition. In: Norbert Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing, and Use*, 97–106. Amsterdam: John Benjamins.
Ehrlich, Susan F. and Keith Rayner
1981 Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior* 20(6): 641–655.
Ellis, Nick C.
1996 Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition* 18(1): 91–126.
Ellis, Nick C.
2002 Frequency effects in language processing. *Studies in Second Language Acquisition* 24(2): 143–188.
Ellis, Nick C.
2003 Constructions, chunking, and connectionism: The emergence of second language structure. In: Catherine Doughty and Michael H. Long (eds.), *Handbook of Second Language Acquisition*, 33–68. Oxford: Blackwell.
Frisson, Steven, Keith Rayner and Martin J. Pickering
2005 Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(5): 862–877.
Gibbs, Raymond W. and Gayle P. Gonzales
1985 Syntactic frozenness in processing and remembering idioms. *Cognition* 20(3): 243–259.
Goldberg, Adele E.
2006 *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
Hasher, Lynn and Walter Chromiak
1977 The processing of frequency information: An automatic mechanism? *Journal of Verbal Learning and Verbal Behavior* 16(2): 173–184.

Jusczyk, Peter W., Angela D. Friederici, Jeanine M. Wessels, Vigdis Y. Svenkerud and Ann Marie Jusczyk
1993 Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language* 32(3): 402–420.

Jusczyk, Peter W., Paul A. Luce and Jan Charles-Luce
1994 Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* 33(5): 630–645.

Just, Marcel Adam and Patricia A. Carpenter
1987 *The Psychology of Reading and Language Comprehension*. Boston: Allyn and Bacon.

MacDonald, Maryellen C., Neal Pearlmutter and Mark S. Seidenberg
1994 The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101(4): 676–703.

McConkie, George W., David Zola, John Grimes, Paul W. Kerr, Nancy R. Bryant and Phillip M. Wolff
1991 Children's eye movements during reading. In J.F. Stein (ed.), *Vision and Visual Dyslexia*, 251–262. London: The Macmillan Press.

McDonald, Scott A. and Richard C. Shillcock
2003a Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science* 14(6): 648–652.

McDonald, Scott A. and Richard C. Shillcock
2003b Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research* 43(15): 1735–1751.

Morris, Robin K.
1994 Lexical and message level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory and Cognition* 20(1): 92–103.

Lively, Scott, David B. Pisoni and Stephan D. Goldinger
1994 Spoken word recognition. In: Morton Ann Gernsbacher (ed.), *Handbook of Psycholinguistics*, 265–318. San Diego, CA: Academic Press.

Rayner, Keith
1979 Eye guidance in reading: Fixation locations within words. *Perception* 8(1): 21–30.

Rayner, Keith and Arnold D. Well
1996 Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin and Review* 3(4): 504–509.

Saffran, Jenny R.
2003 Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science* 12(4): 110–114.

Saffran, Jenny R., Elissa L. Newport, Richard N. Aslin, Rachel A. Tunick and Sandra Burrueco
1997 Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science* 8(2): 101–105.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger
2002 Multiword expressions: A pain in the neck for NLP. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002), Mexico City, Mexico, 1–15.

Seidenberg, Mark S. and Maggie Bruck
1990 Consistency effects in the generation of past tense morphology. Paper presented at the 31st meeting of the Psychonomic Society, New Orleans, LA.

Shapiro, Bernard J.
1969 The subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behavior* 8(2): 248–251.

Weinert, Regina
1995 The role of formulaic language in second language acquisition: A review. *Applied Linguistics* 16(2): 180–205.

Werker, Janet F. and Richard C. Tees
1984 Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7(1): 49–63.

Wible, David, Chin-Hwa Kuo, Meng Chang Chen, Nai-Lung Tsao and Tsung Fu Hung
2006 A computational approach to the discovery and representation of lexical chunks. Paper presented at TALN 2006, Leuven, Belgium.

Wible, David, Chin-Hwa Kuo, Feng-yi Chien, Anne Liu and Nai-Lung Tsao
2001 A web-based EFL writing environment: Exploiting information for learners, teachers, and researchers. *Computers and Education* 37(3–4): 297–315.

Wray, Alison
2002 *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

You talking to me? Corpus and experimental data on the zero auxiliary interrogative in British English[1]

Andrew Caines

Abstract

The zero auxiliary, as exemplified by the utterance – *you talking to me?*, is an under-reported feature in descriptions of English grammar. Evidence from the set of all progressive aspect interrogative in the spoken section of the British National Corpus shows that it occurs with a frequency of one in every five. This ratio increases to one-in-three when we constrain this set further to second person subject interrogatives only. Evidence from two experiments suggest that the high frequency zero auxiliary construction is cognitively entrenched in some way, since it is rated as more acceptable and shadowed more accurately than a low frequency zero auxiliary construction – the first person singular interrogative. This research not only confirms that the zero auxiliary is widely in use, but also provides support for the usage-based linguistic approach, according to which "grammar is the cognitive organization of one's experience with language" (Bybee 2006: 711).

1. Introduction

This paper presents corpus and experimental evidence on the 'zero auxiliary interrogative' in English, a construction exemplified in (1)–(3):

1. This research was supported by the Arts and Humanities Research Council. I am grateful to Dagmar Divjak and Stefan Th. Gries, the organizers of the ICLC theme session for which this paper was prepared. I am grateful also to the anonymous reviewers who gave comments for improvement based on an earlier version of this paper. Finally, I thank Paula Buttery for her help, support, and valuable advice on this work.

(1) you been exercising? (KD8 9785)[2]

(2) so what we doing tonight then? (KC2 3134)

(3) who they playing next week then? (KD6 2711)

These examples contradict the standard assumption that where the auxiliary is required it is supplied (Leech 1987: 18; Greenbaum 1991: 52; Gramley and Pätzold 2004: 111), the zero auxiliary must therefore be considered a non-standard construction. Under this assumption, the interrogatives in (1)–(3) would have been expected to take the form given in (4)–(6):

(4) have you been exercising? *Or*, you have been exercising.

(5) so what are we doing tonight then?

(6) who are they playing next week then?

To date, there has been little discussion of the zero auxiliary: it has received incidental mention as a dialect marker of African American Vernacular English (Labov 1969; Rickford 1998) and as a feature of early child language (Rizzi 1993; Theakston et al. 2005). There has been no substantial corpus or experimental investigation of the construction. This study fills a gap in the literature.

I present a corpus study that focuses on the progressive interrogatives in the spoken section of The British National Corpus (2001). The interrogatives were extracted exhaustively (there are 9950 in total) and have been annotated by subject type (person, number, pronoun, zero) and for auxiliary realization: supplied or zero. Analysis of the resulting annotated corpus showed that the presence of a zero auxiliary is dependent on subject type. These results influenced the design of two experimental studies: (i) an acceptability judgment task, which requires comparison of the interrogatives with a range of filler items; (ii) a continuous shadowing task, in which subjects imitate pre-recorded dialogues.

The comparison of empirical evidence from two sources (corpus and experiment data) can provide us with a broader understanding of a research

2. A quotation followed by an alphanumeric code in brackets indicates an extract from the British National Corpus. The code is a unique identifier in which the first part represents the document and the second part refers to the sentence number. The distributor is Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

question (de Mönnink 1997; Nordquist 2004, 2009; Kepser and Reis 2005; Gries and Gilquin 2009). Conclusions drawn from a single data source can often leave an incomplete picture. By also drawing on data from a second source we may converge on a clearer solution to the question. In our case, corpus analysis provides statistics from transcriptions of production data; experimental evidence can tell us more about cognitive processing. The importance of this dual-method approach to our understanding of language and cognition should not be underestimated. As Divjak and Gries state –

> Despite the importance attributed to frequency in contemporary linguistics, the relationship between frequencies of occurrence in texts on the one hand, and status or structure in cognition as reflected in experiments on the other hand has not been studied in great detail, and hence remains poorly understood (Divjak and Gries 2009)

Corpus and experimental data are complementary in the sense that the former can give a large scale perspective of language use sampled from and extrapolated to the speech community as a whole whereas the latter indicates how given language forms are processed by individual speakers. This dual source method is a research paradigm whose popularity has grown rapidly in recent years. Some studies offer support for a convergence between frequency and cognitive structure (Gries, Hampe and Schönefeld 2005, 2010; Hoffmann 2006; Ellis and Simpson-Vlach 2009; Wulff 2009). Others show divergence, reminding us that any such relationship is not a straightforward one (Roland and Jurafsky 2002; Arppe and Järvikivi 2007).

The purpose of the present study is to investigate whether the zero auxiliary has a more concrete cognitive status than occasional *ad hoc* omission, and secondly whether that status differs according to constructional form. The auxiliary, as with the copula cross-linguistically, is known to be omitted as an efficiency measure. It is considered obligatory in grammatical terms and yet "comparatively insignificant" (Jespersen 1933: 100), both in a semantic and phonological sense. It is for this reason that it is prone to omission, given that the "principle of least effort" (Wells 1982: 94) states that the effort required to produce a linguistic item should to some extent be justified by the significance of that particular item. This would suggest that historically the zero auxiliary was an *ad hoc* reduction of the full form.

However, I propose that now in certain linguistic contexts among certain individuals the zero auxiliary has become cognitively entrenched,

and this is most likely the case for those types of zero auxiliary which occur with high frequency across the general population. An important consequence is that an entrenched zero auxiliary can be seen as an alternative to the construction with auxiliary verb, rather than a derived reduction of it. This proposal aligns with the notion of a link between usage and cognition (Goldberg, Casenhiser and Sethuraman 2004); the usage-based view that, "grammar [is] the cognitive organization of one's experience with language" (Bybee 2006: 711). On this view, language is not held to be structured a priori but instead "apparent structure emerges from the repetition of many local events" (Bybee 2006: 715), such as 'conventionalized word sequences' – what I have been referring to here as 'constructions'. It follows that high frequency construction types are the ones which will first become entrenched.

In the experiments described here, those high frequency constructions – which I propose are entrenched – are likely to be rated as more acceptable and shadowed more accurately than those which are not, because speakers will have more experience of having used or encountered them. I investigate three research questions:

a. To what extent does the zero auxiliary occur, if at all?
b. Does the zero auxiliary occur at different frequencies according to linguistic context (subject type)?
c. Is there any evidence to suggest that the zero auxiliary in any form is cognitively entrenched?
   (c_1) Does experimental evidence from an acceptability judgement task suggest that frequency is a factor in entrenchment?
   (c_2) Does experimental evidence from a continuous shadowing task suggest that frequency is a factor in entrenchment?

First, the corpus study confirms that the zero auxiliary does occur in progressive interrogatives. Furthermore, the corpus frequencies show that, where a subject is supplied, the zero auxiliary is most likely to occur with the second person pronoun – *you* – and least likely to occur with the first person singular pronoun – *I*. These frequency extremes formed the basis of the experiment designs. The results of both an acceptability judgement task and a continuous shadowing task show that frequency is a factor in cognitive entrenchment: (i) participants rated the second person zero auxiliaries as significantly more acceptable than the first person singulars; (ii) in the shadowing task, the participants were more often found to perform a 'fluent restoration' (insertion of auxiliary verb), a repetition error,

hesitation or other disfluency upon hearing the low frequency condition stimuli (i.e., the first person singular).

The corpus study is reported below in §2. This is followed by the experiment studies in §3, including first the acceptability rating task (§3.1) and secondly the continuous shadowing task (§3.2). The paper ends with a discussion of the results, conclusions and directions for future work.

2. Corpus study

For various reasons – chief among them its size, balanced design and availability – the British National Corpus (BNC) was selected as the source for this corpus study. The BNC was constructed in the early 1990s from a broad but balanced range of written and spoken sources. It contains 100 million words, of which 90 million are from written texts and 10 million are transcriptions of spoken language. Only the spoken section of the corpus (sBNC) was used for this study, since written language is more strongly affected by prescriptive rules and therefore the more likely domain in which to find non-standard linguistic forms is speech. Moreover, by definition the progressive aspect and the interrogative are associated more strongly with speech. Indeed, a preliminary survey confirmed that a study of sBNC rather than the written section of the BNC would be more relevant in this case[3], even though the former constitutes just one tenth of the whole corpus.

In the 10 million word sBNC, approximately 6 million words are taken from a more formal setting – business meetings, academic lectures, radio broadcasts and so on – and the remaining 4 million words were recorded by volunteer members of the British public as they went about their daily lives. These volunteers were selected deliberately so that both genders, all age groups and the whole regional and socio-economic spectrum of the United Kingdom would be represented appropriately. Such a balanced design means that the corpus data can, with care, be taken as representa-

3. A test-set of 3000 progressive interrogatives was retrieved from both the written and the spoken sections of the BNC. There were seven times more zero auxiliaries in the spoken test-set than in the written test-set.

tive of the British speech community at that time. The BNC is accessible online at BYU-BNC (Davies 2004).

2.1. Procedure

All progressive interrogatives were retrieved from BYU-BNC. The dataset was manually annotated for subject type, subject person, subject number, subject supplied and realization of the auxiliary verb. In Boolean form, these variables are represented as, 'pronoun: true or false', 'plural: true or false', 'subject supplied: true or false' and 'auxiliary supplied: true or false'. Thus the 'true' value for auxiliary supplied covers both contracted and full forms of the verbs as a group set in contrast to the zero auxiliary. As for subject person, first, second and third person are not scalar but ordinal values and therefore are unsuited to the logistic regression which will be performed on these data. Thus the subject person variable is represented by two subsidiary variables – 'first person: true or false' and 'second person: true or false'. Note that a third variable for 'third person: true or false' would be redundant since the first, second and third persons are not independent of each other. False values for both first and second person variables means the subject must be in third person form. The values for these six variables are listed in Table 1.

Table 1. Annotation variables

Values	Variables					
	Pronoun	1st person	2nd person	Plural	Subject supplied	Auxiliary supplied
0	false (noun)	false (2nd or 3rd)	false (1st or 3rd)	false (singular)	false (zero subject)	false (zero auxiliary)
1	true	true	true	true	true	true (contracted or full)

In Table 2, examples from sBNC are used to exemplify the annotation system:

Table 2. Examples of annotation system

Examples	Variables					
	Pronoun	1st pers	2nd pers	Plural	Subj supplied	Aux supplied
(7) what I'm proposing to do is to (HUD 7)	1	1	0	0	1	1
(8) the Council wasn't doing enough for young people (D95 430)	0	0	0	0	1	1
(9) what you been buying? (KPV 5313)	1	0	1	0	1	0
(10) Ruth and I are trying to get these done (KD0 4657)	0	1	0	1	1	1
(11) going somewhere else? (KBE 5182)	1	0	1	0	0	0

Note that zero subject interrogatives were interpreted as 2nd person singular pronoun subjects, since this is the kind of subject which might be 'restored' to such constructions. The other variables are annotated accordingly, as shown in example (11) in Table 2. Note also that any coordinated noun-plus-pronoun subjects, such as 'Ruth and I' in (10), are annotated '0' for subject type; i.e., as nouns.

2.2. Results

A total of 9950 progressive interrogatives were retrieved from sBNC. Each was annotated as shown in Tables 1 and 2 above. The results are set out in Table 3.

Table 3. Corpus study results

Values	Variables						
	Pronoun	1st pers	2nd pers	3rd pers	Pl	Subj supplied	Aux supplied
0	1033	8620	5025	6519	8326	264	1870
1	8917	1330	4925	3431	1624	9686	8080
total	9950	9950	9950	9950	9950	9950	9950

It is apparent from Table 3 that the zero auxiliary occurs in 18.8% of progressive interrogatives in sBNC. Logistic regression analysis was undertaken to show how strongly the values for each variable predict the zero auxiliary interrogative. The outcome of this analysis is reported in Table 4:

Table 4. Logistic regression analysis

Variable	Predictor coefficient	Sig.
Pronoun	0.171	.371
1st person	0.033	.811
2nd person	2.333	.000
Plural	1.436	.000
Subject supplied	−2.806	.000
Constant	−0.696	.000

The second person, subject number and zero subject coefficients are highly significant in the logistic regression ($p < 0.001$). ($p < 0.001$). The regression analysis shows that subject supplied, plural and second person are the strongest predictors of the zero auxiliary interrogative. Note that the minus value for subject supplied indicates that a zero subject is associated strongly with the zero auxiliary.

In Table 5, the data are presented by construction type. The variables featured in the study are concatenated into various construction types defined by subject form. For example, a clause such as (7) marked as 1-1-0-0-1-1 according to the annotation system described above (Table 1, Table 2) is now labeled a 'first person singular'; 0-0-0-0-1-1 as in (8) is a 'third person singular pronoun'; 1-0-1-0-1-0 as in (9) is a 'second person singular/plural'[4]; and so on.

It is clear from Table 5 that zero and second person subject interrogatives are most likely to occur in zero auxiliary form, while the first person singular, third person singular and third person plural nouns are least likely to occur in zero auxiliary form. The first person plural and third

4. The second person is labeled 'singular/plural' on the grounds that the pronoun form in English – *you* – is ambiguous for number. In conversation, this ambiguity is usually resolved by discourse context, number of participants in the conversation (there may only be two), or speaker gesture. The sBNC transcriptions are devoid of this information, and thus in this analysis number is collapsed for second person subjects.

Table 5. Auxiliary realization in progressive interrogatives by construction type in sBNC

Construction type	Total	Auxiliary supplied		Zero auxiliary (%)
		0 (zero)	1 (contracted or full)	
First person singular	394	3	391	0.8
First person plural	936	137	799	14.6
Second person singular/plural	4925	1332	3593	27.0
Third person singular noun	610	30	580	4.9
Third person singular pronoun	2150	68	2082	3.2
Third person plural noun	154	8	146	5.2
Third person plural pronoun	517	71	446	13.7
Zero subject	264	221	43	83.7

person plural pronoun are intermediate categories. It is this ranking which informs the design of the experiments. In order to keep the tasks a manageable size, only two of these construction types were taken forward and incorporated in the task design – one high and one low frequency zero auxiliary construction. The construction types with minimum and maximum values in the zero auxiliary column were selected: the first person singular progressive interrogative and the second person singular/plural.

In sum, this corpus study has shown that (i) the zero auxiliary is used, contrary to standard assumptions; (ii) it is used more frequently in spoken language than written; (iii) it occurs in 18.8% of progressive interrogatives in sBNC; (iv) it occurs most frequently with zero subject interrogatives; (v) it occurs most frequently with second person interrogatives, and least frequently with first person singular interrogatives – this is the pairing which is taken forward to the experiment section.

3. Experiments

The results of the corpus study reported in the previous section are brought forward and applied to the design of two experiments. As a first attempt, only two of the eight construction types described in Table 5 would be tested in the experimental section of the study. The second person singular/

Table 6. 2 × 2 factorial design for experiments

		Auxiliary realization	
		Full auxiliary	Zero auxiliary
Subject type	1st person singular	what am I doing	what I doing
	2nd person singular/plural	what are you doing	what you doing

plural and first person singular interrogatives were selected as the conditions for these experiments, since these showed the maximal and minimal values for the zero auxiliary. Thus we have a modest 2 × 2 design (Table 6), with scope to include the other construction types – first person plural subject, third person subject, and zero subject zero auxiliaries – in future work.

The two experiments reported in this section are an acceptability judgement task (§3.1) and a continuous shadowing task (§3.2). The first task was carried out using the magnitude estimation method, in which subjects construct their own scale without limits. The second task requires that subjects listen to pre-recorded conversations and repeat what they hear. The key measures in this task are response time, accuracy and alteration – especially, when subjects perform so-called 'fluent restorations', inserting an auxiliary verb even though it had been absent in the stimulus. Note that a different subject group was recruited for each experiment.

The rationale for this combination of tasks is the following: the judgement task requires explicit ratings of zero auxiliaries set against control items which observe grammatical standards; the shadowing task collects implicit judgements, inferred from the manner in which subjects imitate the stimuli. These two tasks are complementary in the way that they elicit deliberate and unaware reactions to the same set of test and control items.

Based on the corpus evidence, the predictions for these experimental tasks are that there will be observable processing differences between the first person singular and second person zero auxiliary interrogatives. The frequency data suggest that the second person zero auxiliary will be more familiar to the human subjects, it being used five hundred times more often than the first person singular zero auxiliary. If there is a link of some sort between an individual's linguistic experience and his cognitive storage, then the prediction is that the second person zero auxiliary will be rated more acceptable and shadowed more fluently and accurately than the first person singular zero auxiliary.

As reported below, the results of these experimental studies align with each other and converge with those of the corpus study.

3.1. Acceptability Judgement task

An acceptability judgement task was used to collect data on the zero auxiliary. Subjects were encouraged to make *naturalness* judgements rather than ones of *correctness.* It was felt that any suggestion of the latter would have subjects trying to recall prescribed rules of grammar rather than the desired behaviour which was simply that they consider their linguistic experience, naturalness and idiomaticity, and the language they would use from day to day. The ratings instrument was 'magnitude estimation' (Bard et al 1996, Cowart 1997). Subjects were asked to give a numerical value for the acceptability of a given stimulus with reference to a benchmark value they have assigned at the outset. This approach avoids many of the pitfalls of fixed scales (such as 1 to 5 or 'very good' to 'very bad') and encourages subjects to think in terms of relative acceptability rather than absolute right and wrong. The method is therefore fitting for analysis of the zero auxiliary – a construction which is absolutely incorrect in terms of the traditional (and taught) rules of grammar but which, as it turns out, is not considered by speakers to be absolutely unacceptable.

3.1.1. Design

Since magnitude estimation involves subjects' independent and implicit creation of ad hoc scales, distinctions among the test items would be enhanced if 'extremes of acceptability' could be created around the test items. In other words, through the inclusion of filler items of deliberately exaggerated unacceptability and of unquestionable acceptability – (hypothetically at first until the results confirm or contradict such categorization) – the first and second person full and zero auxiliary interrogatives may be placed more accurately on the scale of acceptability and at the same time distinguished from each other.

As a consequence, the filler items were not a random assortment of sentences but were instead a controlled collection of subgroups of varying expected acceptability. In addition a certain degree of similarity with the test items was required so that the latter were not identified through their idiosyncrasy. Therefore all fillers were interrogatives also, with a mixture of first, second and third person subjects. The five filler subgroups were: 'fine', 'casual', tense agreement violations, minor word order violations and 'scrambled'. Examples of each are given in Table 7 along with a prediction for their position on an acceptability continuum.

Table 7. Acceptability judgement task test and filler items

Item type	Construction type	Auxiliary	Example	Acceptability
Test	1st person singular	Full	what am I doing	positive
		Zero	what I doing	negative
	2nd person singular/plural	Full	what are you doing	positive
		Zero	what you doing	positive
Filler	Fine	–	how did they get here	positive
	Casual	–	you know what I mean	positive
	Tense agreement violations	–	has she make you a cup of tea	negative
	Minor word order violations	–	where go you on holiday	negative
	Scrambled	–	that sure about you	negative

3.1.2. Materials

There were forty test items in this task, all of which are taken verbatim from the conversational subsection of sBNC. For the sake of comparability, these items were an exact match for the forty test items to be used in the continuous shadowing experiment (§3.2). All forty test items were progressive interrogatives – half with first person singular and half with second person subjects.

The corpus data indicated there would be insufficient instances of the first person singular zero auxiliary interrogative from which to construct a meaningful number of test items (n = 3 in sBNC; Table 5). As a solution, twenty examples of the first person singular *full* auxiliary interrogative were located (n = 387 in sBNC), and then the first person *zero* auxiliary test items were created by removal of the auxiliary verb from those same twenty sentences. This method also controlled for semantic content and so the second person test items were collected in the same way. The full and zero auxiliaries could consequently be compared on a like-for-like basis in terms of context.

Forty texts were extracted from sBNC – twenty featuring a first person singular and twenty featuring a second person progressive interrogative. Each test item would feature in both full and zero auxiliary form. Thus

there were eighty experiment items in total, which were divided into four scripts: 1a, 1b, 2a, 2b. Each contained twenty test items; five first person singular and five second person full auxiliary interrogatives, five first person singular and five second person zero auxiliary interrogatives. Scripts 1a and 2a were equivalents to 1b and 2b except only for the auxiliary verb.

Forty filler items were prepared so as to achieve the 2:1 filler-to-test item ratio demanded by psycholinguistic best practice (Sprouse 2009: 335). Twenty of these were of the type 'fine' (Table 7), another five were 'casual', five were tense agreement violations, five were minor word order violations and five were 'scrambled'. Each of the scripts contained twenty test and forty filler items.

All stimuli were presented as audio files to avoid distancing the construction types from the environment in which they are more likely to be encountered. As Cowart (1997: 64) puts it, spoken language "does not rely on literacy skills of the informant" which are known to be "variable in the general population" and, more importantly, "speakers may have different expectations (or tolerances) or the syntax of written sentences than they do for spoken sentences". There would have been a danger of falsifying subject responses if the stimuli had been presented in written form: subjects would be more likely to refer to standard rules, which are most often based on and taught referring to written language.

The eighty test items (4 scripts $\times$ 20 per script), ten practice items and forty fillers were recorded in advance of testing. The items were read by a male native speaker of Southern British English and recorded digitally in a sound-proofed room. The recordings were made as mono wave (.wav) files at a rate of 44.1 kHz. A Microsoft PowerPoint slideshow was then prepared so that the subjects could run through the task at their own pace and control. First there was an introductory section about the magnitude estimation concept, then came a practice section of ten items, and thirdly the main experimental section.

The explanation of magnitude estimation included a training section in which the subjects were required to use it in judging the length of an assortment of horizontal lines, a task which along with judging loudness and brightness was one of the original successful applications of the scoring method (Bolanowski 1987; Gescheider 1997: 269; Zwislocki 1983). This demonstration was included so that subjects could become familiar with the scoring system and not have to apply an unfamiliar method to the experiment items. The ten practice items then allowed the subjects to transfer the concept to language, and become comfortable and competent

with this skill. The order of presentation of the sixty filler and test items (40 of the former, 20 of the latter) was pseudo-randomized so that there were no two consecutive test items.

For both the practice and main sections, the reference item – the benchmark against which all further items would be measured – was selected as an expected intermediate point on the scale of acceptability, so that subjects would in theory be able to score succeeding items above and below this opening mark. A pilot study on the filler items demonstrated that the 'casual' group would be of intermediate acceptability below the 'fine' and above the other groups. The reference item for the task was therefore chosen from among the 'casual' filler item group.

3.1.3. Procedure

Twenty students from the University of Cambridge were recruited to participate in the study. The age range was 19-35 (mean = 23.25 years, median = 22.5 years) and all students were native speakers of English. Their degree subjects were in various disciplines. None had known hearing difficulty and all received payment for their participation. Each subject was assigned to one of the four scripts – 1a, 1b, 2a, 2b. The experimenter offered a brief spoken summary of the task ahead and subjects were then required to read the introductory section with its more detailed written instructions. Subjects were encouraged to ask questions of the experimenter if at any point they were unsure of the procedure, and an active check was made after the practice section for their understanding of the task before allowing them to proceed to the main experimental section.

The subjects could work through the PowerPoint slideshow at their own pace, and were required to write their acceptability ratings down on a scoring grid. Subjects heard each test and filler item through headphones and after they had first heard and scored the opening reference item, there was an option to replay this item for direct comparison at any point throughout the remainder of the slideshow. The task took approximately twenty minutes to complete.

3.1.4. Results

The main feature of the magnitude estimation scoring system is that the scale – the maximum and minimum score given – is decided by each subject ad hoc. As a consequence the twenty different subjects gave their acceptability judgements on twenty different scales. In order to compare

Table 8. Acceptability judgement task results

Item type	Construction type	Auxiliary	Example	Acceptability (*z*-score)
Test	1st person singular	Full	what am I doing	+0.42
		Zero	what I doing	−0.80
	2nd person singular/plural	Full	what are you doing	+0.63
		Zero	what you doing	+0.13
Filler	Fine	–	how did they get here	+0.52
	Casual	–	you know what I mean	+0.28
	Tense agreement violations	–	has she make you a cup of tea	−0.47
	Minor word order violations	–	where go you on holiday	−0.95
	Scrambled	–	that sure about you	−1.63

ratings, all scales were standardized as *z*-scores[5]. Thus the acceptability judgements are expressed as units of standard deviation, plus or minus, from the mean. Average acceptability scores, category by category, are given in Table 8.

A mixed-measures analysis of variance (ANOVA) of the test (and not filler) items indicates main effects in the data for the factors, subject type ($F(1,16) = 37.7$, $p < 0.001$) and presence of auxiliary ($F(1,16) = 147$, $p < 0.001$). There is also an interaction between the two factors ($F(1,16) = 16.6$, $p < 0.001$). There was, however, no significant between-subjects effect due to the group in which they were placed ($F(3,16) = 1.27$, $p > .05$). That is, the four different scripts were themselves not a factor in the variation found.

The acceptability ranking for the filler items comes out as predicted in Table 7: fine > casual > tense agreement violations > minor word order violations > scrambled. An awareness of this hierarchy allows correct placement of the full and zero auxiliary interrogatives with reference to the filler categories. It is clear from Table 8 that the second person zero

5. The z-score is calculated by the following equation (where x stands for each raw score, μ is the mean, and σ is the standard deviation): $z = (x - \mu) \div \sigma$

Table 9. Acceptability judgement task analysis of variance

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Subject	6.450	1	6.450	37.722	.000
Auxiliary	14.558	1	14.558	146.693	.000
Subject * Auxiliary	2.597	1	2.597	16.627	.001
Error	2.499	16	0.156		

Table 10. Acceptability judgement task pairwise comparisons

Factor a	Factor b	Mean diff.	Std. Error	Sig.
1st person singular	Full vs Zero auxiliary	1.214*	.130	.000
2nd person	Full vs Zero auxiliary	0.493*	.093	.000
Full auxiliary	1st vs 2nd person	−0.208	.143	.165
Zero auxiliary	1st vs 2nd person	−0.928*	.111	.000

auxiliary interrogative is deemed to be more acceptable than the first person singular subject subtype.

Pairwise comparisons reveal statistically significant differences between the full and zero auxiliary whichever the subject type, but only between first and second person subject type for a zero auxiliary realization (Table 10). That is, the first and second person full auxiliary interrogatives are judged to be similarly acceptable but their zero auxiliary equivalents are not. These comparisons demonstrate that the first person singular and second person subjects are not in themselves more or less acceptable than each other; it is only with a zero auxiliary that any difference emerges.

The first person singular subject zero auxiliaries are rated less acceptable than the tense agreement violations but more acceptable than the minor word order violations. The second person subject zero auxiliaries, meanwhile, are placed above the tense agreement violations and below the casual fillers. It is apparent that the first singular and second person subject full auxiliaries group with the filler 'fine' category and therefore can be considered to be a benchmark for comparison with second person zero auxiliaries. The second person zero auxiliaries and casual filler category are close together at a slightly less acceptable level. Nevertheless, all of these were adjudged to be more acceptable than the first person zero

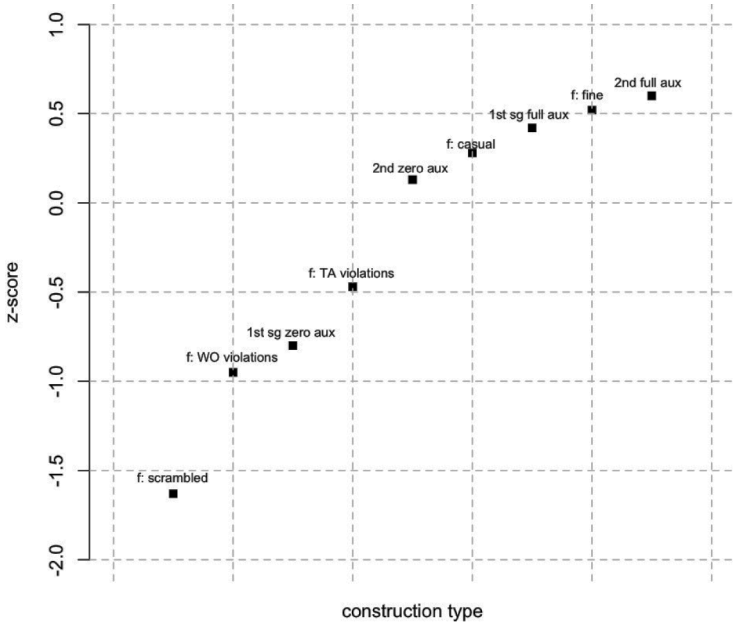


Figure 1. Acceptability judgement task construction type ranking

auxiliaries, the tense agreement violations, minor word order violations and scrambled fillers.

Figure 1 illustrates the data from Table 8 in chart format, with the categories ordered by acceptability rating.

Each construction type is ranked by mean *z*-score, and fillers are labelled with an opening "f:".

3.1.5. Discussion

The above results indicate that the zero auxiliary construction is rated at varying levels of acceptability depending on subject type. A first person singular zero auxiliary is rated much less acceptable than a second person zero auxiliary. This outcome corroborates the corpus evidence, which showed that for progressive interrogatives the second person is the most likely and the first person singular is the least likely to occur in zero auxiliary form (Table 5). The zero auxiliary interrogative is an acceptable construction therefore, but for now this is true only with a second person subject and not with a first person singular subject.

Meanwhile, the second person *full* auxiliary interrogative is rated more acceptable than the first person singular full auxiliary – albeit not to a statistically significant degree – again in parallel with the frequency evidence. This may well be a usage effect whereby the more frequently experienced form is deemed to be more acceptable, even though there is no difference between the two construction types grammatically.

So far, then, the corpus and experimental evidence converge. The results are only an indication, however, in that the acceptability judgement task only included interrogatives, only included the progressive and only included first person singular and second person subjects. Other construction types – declaratives, non-progressives, first person plural, third person and non-pronominal subjects – should be included in a follow-up study before the corpus and experimental evidence may be said to be fully convergent. Nevertheless, this study offers a firm foundation on which to build.

3.2. Continuous Shadowing task

As a complement to the acceptability judgement task a continuous shadowing task was designed as a further test for experimental convergence with or divergence from the corpus data. This paradigm has a long history in the psycholinguistic field (Chistovich 1960; Chistovich, Aliakrinskii, and Abul'ian 1960; Chistovich, Klaas, and Kuzmin 1962; Marslen-Wilson 1973, 1975, 1985). It requires that subjects repeat recorded speech as closely as they can. In this case the input is a dialogue – hence this is 'continuous' rather than word by word 'cued' shadowing. The similarities and differences between the subjects' speech and the original material can offer an insight into cognitive linguistic structure. Here, the results point to the stochastic nature of linguistic knowledge, as error rates correspond with the frequency data.

The key measure in this study is accuracy. In what has become a classic study, Marslen-Wilson and Welsh (1978) distorted certain sounds in a speech recording (for example, *travedy* for *tragedy*) and did not give any warning of the distortions in the input. They found that subjects would repeat the words without distortion approximately half the time, saying *tragedy* not *travedy*. These so-called 'fluent restorations' indicate that people will override what they actually hear with what they expect to hear. Here, fluent restorations would point to expectations that an auxiliary verb is usually supplied in that constructional context. On the other hand, a lack of fluent restorations would suggest that the zero auxiliary is to some extent acceptable in that constructional context.

Other response errors such as hesitation, stutters or incorrect repetitions, will indicate the stimuli with which subjects have the most difficulty in maintaining fluency. Even allowing for straightforward slips of the tongue, such errors might be revealing as to whether people would expect the zero auxiliary in that context, the first person singular or second person subject.

In this study, as in the acceptability judgement task, only the second person and first person singular subject progressive interrogatives feature. The results of the corpus study and the acceptability judgement task reported above predict that fluent restorations and other errors will occur more frequently for first person singular than second person subject zero auxiliaries, since these have been found to be less acceptable (Figure 1) and less frequent (Table 5) than the latter. The outcome of this continuous shadowing experiment is as predicted, with many more fluent restorations and repetition errors in response to first person singular compared to second person singular/plural zero auxiliaries. Moreover, there are several fluent *omissions* of the auxiliary verb in second person full auxiliary interrogatives, whereas there are none for first person singular full auxiliaries. This detail offers further indication that the second person zero auxiliary is cognitively more entrenched thanks to its higher frequency.

3.2.1. Design & Materials

The continuous shadowing task was constructed according to the previously described 2×2 factorial design (Table 6), consistent with the acceptability judgement task. The script design and test item set were copied from the acceptability judgement task (§3.1). Thus there were four scripts of twenty test items each: 1a, 1b, 2a, 2b. In this task, additionally, the test items were preceded and followed by sections of the actual dialogue they occurred in, since for best effect continuous shadowing – in that it should be prolonged – requires streams of speech and not just isolated sentences. A by-product of this decision was that the need to include filler-to-test items at a ratio of 2:1 or more would be fulfilled with ease through the preparation of those test items. The fillers outnumber the test items by a ratio of 11:1.

The target constructions were located, extracted with surrounding text and then recreated as dialogue for the task. Four native speakers of British English (two male, two female) were recruited to recreate the eighty experiment scripts as spoken dialogue. Recordings were made on computer in a sound proofed room. To control against extraneous effects which might

arise from differences in the surrounding material, the exact same recording of the filler items was used for both versions of the text. It only remained for the full or zero auxiliary construction to be 'spliced' in at the appropriate point in the sound file. Note that the speakers were required to read the full and zero auxiliary variants of the same question in an intonationally consistent way.

The recordings were made as mono wave (.wav) files at a rate of 44.1 kHz. The experiment scripts were then assembled as RealPlayer playlists according to the four versions of the task. The order of the scripts within the playlists was pseudo-randomized so that there were no two consecutive instances of the same category of test item. The end product was four playlists of approximately fifteen minutes length.

3.2.2. Procedure

Another twenty University of Cambridge students were recruited to participate in the study. The age range was 18–35 (mean = 22 years, median = 21 years), all students were native speakers of English, and they came from a range of disciplines. None had known hearing difficulty and all received payment for their participation. Each subject was assigned to one of the four playlists – 1a, 1b, 2a, 2b – and then the experimenter would offer a brief spoken summary of the task ahead.

It was emphasized to the subjects that they should try to strike a balance between speed and accuracy of response, but also that they should not attempt to go back and correct any deviations they might feel they had made from the original recording. Most importantly, the subjects were instructed not to wait until the end of a phrase before they began speaking, but to do so as soon as possible.

The subjects were then required to wear headphones and begin the playlists. The opening item was a recorded explanation and demonstration of the continuous shadowing technique, followed by five practice items. At the end of every item there was a ten second pause. The playlist proceeded automatically.

The subjects' responses were recorded on computer by way of a cardioid microphone, DAT machine, mixer and external sound card. The DAT machine acted as an amplifier for the microphone. The mixer ensured that subject response (henceforth the output, adopting a subject-centric perspective) was recorded in one channel and the original material (the input) was re-recorded in the other channel. The sound card acted as the line-in for recordings to be made directly and digitally onto computer.

By separately recording input and output in the same sound file, reaction time and restoration analysis could be undertaken comparatively between the two. Reaction time is here referred to as 'latency' and is measured as the time delay between onset of the stimulus and onset of subject response. Each subject's latencies to test items were measured using Praat software (Boersma and Weenink 2009), and all recordings were listened to in order to ascertain fluent restoration and other error frequencies.

3.2.3. Results

Mean latencies for the four test conditions are reported in Table 11. With errors excluded, the overall mean latency was 966 ms. The slight latency differences among conditions are not statistically significant, as shown by regression analysis (Table 12) and pairwise comparison.

Analysis of variance shows that there was no significant between-subject effect of which playlist the subjects were assigned to ($F(1,3) = 0.55$, $p > .05$). Instead the crucial means of analysis is not quantitative but in fact qualitative: errors and fluent restorations are key.

In the course of testing, four hundred subject responses to experiment items were recorded (20 subjects $\times$ 20 items per playlist), one hundred to each condition. Fluent restorations and other errors were encountered on

Table 11. Continuous shadowing task latencies

Subject	Auxiliary	Mean latency (ms)
1st person singular	Full	891
	Zero	1002
2nd person	Full	981
	Zero	990

Table 12. Continuous shadowing task regression analysis coefficents

	Predictor coefficient	Sig.
Subject	.033	.610
Auxiliary	.054	.398
Constant	.836	.000

a total of 112 occasions out of these four hundred responses. Transcribed examples of each are given below:

(12) Recorded stimulus: I going out?
(from sBNC_conv, KDE 105)
Subject response: Am I going out?

(13) Recorded stimulus: Right, what I having for dinner?
(from sBNC_conv, KDB 1126)
Subject response: Right, what am I having for dinner?

(14) Recorded stimulus: The green knight said, I having some chocolate?
(from sBNC_conv, KDW 3658)
Subject response: The green knight said, I'm having some chocolate

(15) Recorded stimulus: What I getting, the spuds?
(from sBNC_conv, KD8 61)
Subject response: but... what I... get [sp?]

The number of fluent restorations (defined as production of the auxiliary verb without apparent interruption of speech flow) and other errors (defined as production of wrong word, incomplete attempt to shadow the phrase or no attempt to shadow phrase) for each condition is given in Table 13.

For the zero auxiliary categories there was a 30% rate of fluent restoration for the first person subject in contrast to the 17% rate for the second person subject. The occurrence of other errors was recorded at a rate of 46% for the first person zero auxiliaries, whereas the rate for the three other categories was in the range 4–7%. Since the dependent variable is a count variable, we used poisson regression to analyze the predictive power of subject person and auxiliary realization for the occurrence of other

Table 13. Continuous shadowing results

Subject	Auxiliary	Fluent restorations	Fluent omissions	Other errors
1st person singular	Full	–	0/100	4/100
	Zero	30/100	–	46/100
2nd person	Full	–	4/100	4/100
	Zero	17/100	–	7/100

Table 14. Continuous shadowing poisson regression parameter estimates

Parameter	B	Std. Error	Wald Chi-Square	df	Sig.
Intercept	2.257	.3056	54.565	1	.000
person1	1.514	.3330	20.671	1	.000
person2	0[a]	.	.	.	.
aux	−1.891	.3793	24.851	1	.000
Scale	1				

[a] Set to zero because this parameter is redundant.

errors. Person is recoded as two Boolean variables – person 1 and person 2 – because first and second person are not points on a scale, strictly speaking.

Table 14 shows that the first person subject and zero auxiliary are significant predictors of other errors in the continuous shadowing task. Table 14 confirms that it is the first person zero auxiliary errors and not the second person errors which are the cause of this significance.

On the whole, the errors involved omission of the subject and auxiliary altogether (for example, *coming over to see you?* instead of *I coming over to see you?*), replacement of *I* with a negated declarative opening – *why I'm not*, or replacement of *I* with another subject pronoun *she*, *we* or *you*. Moreover, these errors were characterized by hesitation, stuttered syllables and retakes: all typical signals of confusion.

In the event, as well as these errors and fluent restorations, a number of fluent omissions were observed – as in (16):

(16) Recorded stimulus: What are you going to do?
(from sBNC_conv, KB3 491)
Subject response: What you going to do?

On four occasions, the subject shadowed a full auxiliary interrogative in the input by reproducing a zero auxiliary interrogative. Tellingly, all four of these fluent omissions were in response to second person and not first person interrogatives.

3.2.4. Discussion

The results of this task indicate that the first person subject zero auxiliary incurs many more 'trip-ups' (errors of repetition other than fluent restorations) than its second person counterpart. Meanwhile, the assumption –

whether conscious or unconscious – that the construction features an auxiliary verb (the fluent restorations) was a response to both zero auxiliary subtypes, but a significantly more common one to the first person singular. This outcome is consistent with the corpus and acceptability judgement data: processing of the first person singular zero auxiliary – already shown to be of relatively low frequency and low acceptability – is markedly different from that of the second person zero auxiliary and the full auxiliary forms.

In this task the second person zero auxiliary was processed in an approximately similar way to the first and second person full auxiliary interrogatives. There were a number of fluent restorations of the auxiliary for the second person zero auxiliary items, but at a significantly lower rate than that for the first person condition. Moreover, when compared with the number of fluent omissions of the auxiliary from the second person full auxiliary items, it seems as though the two types of auxiliary realization are being used interchangeably. The fact that no such fluent omission in the first person singular full auxiliary condition lends further credibility to this conclusion. The number of other errors was not significantly different between the first person full auxiliary, second person full auxiliary and second person zero auxiliary conditions. As a group however, these three differed significantly from the first person zero auxiliary.

Once again, experimental evidence on the zero auxiliary suggests that frequency is a factor in entrenchment. In this continuous shadowing task, clear differences were found in the subjects' response to the (low frequency) first person singular and (high frequency) second person subject zero auxiliary. These differences again suggest that the second person zero auxiliary construction is processed in a similar manner to established constructions such as the first singular and second person full auxiliaries. The first person singular zero auxiliary, on the other hand – the construction least frequently found in sBNC – is processed with some difficulty (other errors) or an avoidance strategy (fluent restoration).

It is an unfortunate but oft-encountered consequence of running a linguistic study on a university campus that the subjects are mostly students, and therefore mostly of a certain age and social background (social inasmuch as it relates to education). It is difficult to hypothesise in which direction this might skew the results. Non-standard linguistic features are traditionally associated with younger generations and therefore younger subjects might judge the zero auxiliaries to be more acceptable than older subjects might and shadow them more fluently. The opposite responses might be expected of educated speakers since they are more likely to have been exposed to

prescriptive grammar training. Further investigation is needed into the sociolinguistics of zero auxiliary use. To do so, subjects must be recruited from a range of generations and educational backgrounds.

4. General discussion

To conclude, we return to the research questions introduced above (§1):

a. To what extent does the zero auxiliary occur, if at all?
b. Does the zero auxiliary occur at different frequencies according to various subject types?
c. Is there any evidence to suggest that the zero auxiliary in any form is cognitively entrenched?
   (c_1) Does experimental evidence from an acceptability judgement task suggest that frequency is a factor in entrenchment?
   (c_2) Does experimental evidence from a continuous shadowing task suggest that frequency is a factor in entrenchment?

In response to these questions, the corpus study showed that the zero auxiliary occurs in 18.8% of all progressive interrogatives in sBNC. This proportion was found to vary according to subject type, ranging from the first person singular at 0.8% to the zero subject at 83.7%. The next most frequent subject type was the second person subject, at 27%. However, this was by far the most numerous zero auxiliary construction with 1332 occurrences (the next most frequent was the zero subject with 221; Table 5). Since it was frequency and not relative frequency under investigation, the second person subject was selected as the high frequency condition for the experiment studies, set against the first person singular as the low frequency condition.

The experiment results offer evidence that frequency is a factor in entrenchment. The high frequency second person interrogative is processed in ways which are approximately similar to established constructions such as the full auxiliary interrogative. Its appearance among test stimuli can still induce performance errors and is still judged to be at a lower level of acceptability than the full auxiliaries, but nevertheless it is differentiated through the data collected here from the low frequency first person singular interrogative. In this light the second person zero auxiliary interrogative is seen as closer to the established full auxiliary constructions, being significantly more acceptable and inducing significantly fewer errors in experimental tests of processing than the first person singular zero auxiliary. In

other words, there is evidence that the second person zero auxiliary is cognitively entrenched and the first person singular zero auxiliary is not.

In the first experiment – the acceptability judgement task with magnitude estimation – second person zero auxiliaries were rated at a similar level to the 'casual' filler category. The first person singular zero auxiliaries were rated at a lower level of acceptability, comparable to the tense-agreement and word order violation fillers. The judgement of acceptability for the second person zero auxiliary condition (and 'casual' fillers) was not quite at the same high level as the 'fine' filler category, that which was matched by the established first person singular and second person full auxiliary items.

In the second experiment, the continuous shadowing study, the straightforward imitative nature of the task meant that the quantitative measure – the latencies between test item onset and onset of the subject's response – did not distinguish between the full and zero auxiliary conditions to any statistically significant degree. Instead, the measure which set the second person zero auxiliaries apart from the first person singular zero auxiliaries was that of fluent restoration and other error rates. Performance for the first person singular zero auxiliaries was significantly more error prone than that for the second person zero auxiliaries. This result suggests that the latter condition was much less problematic for subjects to process and re-produce than the former condition.

The converging evidence presented in this paper suggests that the second person zero auxiliary interrogative, through exposure, has become cognitively entrenched whereas the first person singular zero auxiliary has not. That it should be the second person at the forefront of this development fits with its status as the most frequent interrogative subject. These conclusions tie in with usage-based theories of grammar which posit a link between linguistic experience and cognitive structure. In future, zero auxiliaries with other subject types might be tested in psycholinguistic research, to verify whether the corpus-experimental correspondence found for first person singular and second person zero auxiliaries would be repeated. Additionally, investigation of the inter-relationship of further variables with the zero auxiliary would be desirable: both semantic – in terms of any meaning nuances between full and zero auxiliaries – and sociological – in terms of prestige, covert or otherwise. Finally, since all subjects for both tasks were aged 18–35, it is also necessary to recruit subjects of other age groups in any future study, so as to understand the sociolinguistics of the zero auxiliary more fully.

References

Arppe, Antti and Juhani Järvikivi
2007 Every method counts – combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2): 131–160.
Bard, Ellen Gurman, Dan Robertson and Antonella Sorace
1996 Magnitude estimation of linguistic acceptability. *Language* 72(1): 32–68.
Boersma, Paul and David Weenink
2009 Praat: doing phonetics by computer (Version 5.1.05).
Bolanowski, Stanley J., Jr.
1987 Contourless stimuli produce binocular brightness summation. *Vision Research* 27(11): 1943–1951.
Bybee, Joan L.
2006 From usage to grammar: the mind's response to repetition. *Language* 82(4): 711–733.
Chistovich, Ludmilla
1960 Classification of rapidly repeated speech sounds. *Akusticheskii Zhurnal* 6: 392–398.
Chistovich, Ludmilla, V.V. Aliakrinskii and V.A. Abulian
1960 Time delays in speech repetition. *Voprosy Psikhologii* 1: 114–119.
Chistovich, Ludmilla, Yu. A. Klaas and Yu. I. Kuz'min
1962 The process of speech sound discrimination. *Voprosy Psikhologii* 6: 26–39.
Cowart, Wayne
1997 *Experimental syntax: Applying objective methods to sentence judgments*. London: Sage.
Davies, Mark
2004 BYU-BNC: The British National Corpus. URL: http://corpus.byu.edu/bnc
Divjak, Dagmar S. and Stefan Th. Gries
2009 Converging and diverging evidence: corpora and other (cognitive) phenomena? Workshop at Corpus Linguistics 2009, University of Liverpool.
Ellis, Nick C. and Rita Simpson-Vlach
2009 Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics and education. *Corpus Linguistics and Linguistic Theory* 5(1): 61–78.
Gescheider, George A.
1997 *Psychophysics: the fundamentals*. 3rd edn. Hove: Psychology Press.
Goldberg, Adele E., Devin Casenhiser and Nitya Sethuraman
2004 Learning argument structure generalizations. *Cognitive Linguistics* 15(3): 289–316.

Gramley, Stephan and Kurt-Michael Pätzold
2004 *A Survey of Modern English.* 2nd edn. London: Routledge.
Greenbaum, Sidney
1991 *An introduction to English grammar.* London: Longman.
Gries, Stefan Th. and Gaëtanelle Gilquin
2009 Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1): 1–26.
Gries, Stefan Th., Beate Hampe and Doris Schönefeld
2005 Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4): 635–676.
Gries, Stefan Th., Beate Hampe and Doris Schönefeld
2010 Converging evidence II: more on the association of verbs and constructions. In: Newman, John and Sally Rice (eds.), *Experimental and empirical methods in the study of conceptual structure, discourse and language*, 59–72. Stanford, CA: CSLI.
Hoffmann, Thomas
2006 Corpora and introspection as corroborating evidence: the case of preposition placement in English relative clauses. *Corpus Linguistics and Linguistic Theory* 2(2): 165–195.
Jespersen, Otto
1933 *Essentials of English Grammar*. London: George Allen and Unwin.
Kepser, Stephan and Marga Reis (eds.)
2005 *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*. Berlin: Mouton de Gruyter.
Labov, William
1969 Contraction, Deletion, and Inherent Variability of the English Copula. *Language* 45(4): 715–62.
Leech, Geoffrey
1987 *Meaning and the English verb*. 2nd ed. London: Longman.
Marslen-Wilson, William
1973 Linguistic structure and speech shadowing at very short latencies. *Nature* 244: 522–523.
Marslen-Wilson, William
1975 Sentence perception as an interactive parallel process. *Science* 189: 226–228.
Marslen-Wilson, William
1985 Speech shadowing and speech comprehension. *Speech Communication* 4: 55–73.
Marslen-Wilson, William and Alan Welsh
1978 Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* 10: 29–63.
de Mönnink, Inge
1997 Using corpus and experimental data: a multimethod approach. In: Ljung, Magnus (ed.), *Corpus-based Studies in English: Papers*

from the seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17), Stockholm May 15–19, 1996. Amsterdam: Rodopi.

Nordquist, Dawn
2004 Comparing elicited data and corpora. In: Barlow, Michael and Suzanne Kemmer (eds.), *Usage-Based Models of Language*. Stanford: CSLI.

Nordquist, Dawn
2009 Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory* 5(1): 105–130.

Rickford, John R.
1998 The Creole Origins of African-American Vernacular English: evidence from Copula Absence. In: Mufwene, Salikoko S., John R. Rickford, Guy Bailey and John Baugh (eds.), *African-American English: Structure, History, and Use*. London: Routledge.

Rizzi, Luigi
1993 Some Notes on Linguistic Theory and Language Development: The Case of Root Infinitives. *Language Acquisition* 3(4): 371–393.

Roland, Douglas and Dan Jurafsky
2002 Verb sense and verb subcategorization probabilities. In: Stevenson, Suzanne and Paola Merlo (eds.), *The Lexical Basis of Sentence Processing: formal, computational and experimental issues*. Amsterdam: John Benjamins.

Sprouse, Jon
2009 Revisiting Satiation. *Linguistic Inquiry* 40(2): 329–341.

Theakston, Anna L., Elena V. Lieven, Julian Pine and Caroline Rowland
2005 The acquisition of auxiliary syntax: BE and HAVE. *Cognitive Linguistics* 16(1): 247–277.

Wells, John
1982 *Accents of English: volume I, an introduction*. Cambridge: Cambridge University Press.

Wulff, Stefanie
2009 Converging evidence from corpus and experimental data to capture idiomaticity. *Corpus Linguistics and Linguistic Theory* 5(1): 131–159.

Zwislocki, Jozef. J.
1983 Absolute and other scales: Question of validity. *Perception and Psychophysics* 33: 593–594.

Primary source

The British National Corpus, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

The predictive value of word-level perplexity in human sentence processing: A case study on fixed adjective-preposition constructions in Dutch

Maria Mos, Antal van den Bosch and Peter Berck

Words are not distributed randomly; they tend to co-occur in more or less predictable patterns. In this chapter, one particular pattern, the Fixed Adjective Preposition construction (FAP), is investigated. We focus on six specific adjective-preposition sequences in Dutch, and contrast two contexts (varying the finite verb in the utterance) and two interpretations (i.e. as a unit or as a coincidental sequence). In an experimental copy task in which participants can revisit and switch to the sentence they are asked to copy, we log at which points subjects do this. The switching data provide insight in the question to what extent FAPs are a unit in human processing. In addition, we relate our data to the word-level perplexities generated by a stochastic n-gram model, to analyze whether the model is sensitive to FAPs. On the basis of a correlation analysis of human task behaviour (reflecting human processing) and stochastic likelihood measures (reflecting local co-occurrence statistics) we discuss what aspects coincide and where the two differ. Based on our observation that the stochastic model explains more than 25% of the variance found in the human data, we argue that it is reasonable to assume that to the extent of the correlation, (co-) occurrence frequencies are behind the language processing mechanisms that people use.

1. Introduction

Words are not distributed randomly; they tend to co-occur in more or less predictable patterns. Some sequences are highly fixed: given the sequence *they lived happily ever* . . . most speakers of English would assume that the next word is *after*, based on their previous exposure to this expression. Because such word combinations are conventional, they must be part of speakers' linguistic repertoires. Not all collocations are as fixed as *happily ever after*. In his analysis of the resultative construction, Boas (2003, 2008) points out that, when the verb *drive* occurs with a resultative phrase, this phrase nearly exclusively denotes a state of mental instability. *Crazy* is the

most frequent instantiation, but other words and phrases also occur (e.g. *insane*, *up the wall*). Speakers must have this as part of their representation of the verb *to drive* in the resultative construction, Boas convincingly argues, or else the distributional data cannot be accounted for. We do not know to what extent speakers process such combinations as one unit. One the one hand, they consist of more than one meaning-bearing element, while on the other hand they are unit-like in that they have a clear joint meaning.

In this contribution, we focus on a specific construction: the Fixed Adjective Preposition construction (FAP). Section 2 contains a description of the formal and semantic characteristics of this construction, using the Construction Grammar framework. We examine human sentence processing data for the FAP-construction and its sentential context, and look at the probability of the sequence and its influence on the probability of other elements in an utterance with a measure of word-level perplexity. By then comparing these data, we can analyze to what extent the likelihood of a FAP sequence influences sentence processing.

In language use, it is difficult to determine if a multi-word sequence is really one unit, although sometimes there are indications, e.g. when there is phonological reduction in speech. Bybee and Scheibman (1999), for instance, observe that *do not* is most likely to be reduced to *don't* in its most frequent context, i.e. when preceded by *I* and followed by *know*, *think*, *have to* or *want*. Advances in technology in the past decade have made it possible to follow people's gaze in reading, offering a tool for investigating 'units' in language processing. An example of such research is Schilperoord and Cozijn (2010) who use an eye-tracking experiment to investigate anaphor resolution. They found reading times were longer for antecedents and anaphoric information if the antecedent was part of a longer fixed expression (e.g. … *by the skin of his teeth*$_i$. *They*$_i$ *have to be brushed twice every day*), than if the antecedent was not part of a fixed expression. This indicates that elements inside the fixed expression are less available for anaphor resolution, which in turn may be interpreted as evidence for the unit-status of such expressions in reading. Besides measuring eye movements in reading, however, it is hard to tap into processing directly without interfering with it. We introduce an experimental technique attempting to do just that: investigating the units people divide an utterance into when they memorize it for reproduction.

In the experimental task described in Section 3 we focus on six specific adjective-preposition sequences in Dutch, and contrast two contexts (varying the finite verb in the utterance) and two interpretations (i.e. as a unit

or as a coincidental sequence). The task and test items are outlined in sections 3.1.4 and 3.1.5. One likely indicator of what is used as a unit is co-occurrence patterns in attested speech. For that reason, the processing data from the experiment are compared to a likelihood measure for observing the next word, computed by a memory-based language model trained on a corpus (cf. Section 3.1.3). This corpus only contains letter sequences and word boundaries. No Part of Speech tags or constituent structures are added to the input the model receives, on which to base its measures. Thus, the model can be classified as a knowledge-poor stochastic language model as typically used in speech recognition systems (Jelinek, 1998).

The final section of this contribution discusses the relation between such measures and human language processing. On the basis of a correlation analysis of task results (reflecting processing) and likelihood measures (reflecting co-occurrence patterns) we attempt to point out what aspects coincide and where the two differ; finding the latter would indicate that humans do something else (or more) than what a simple stochastic model does. Based on the correlations we find, we argue that it is reasonable to assume that to the extent of the correlation, (co-)occurrence frequencies are behind the language processing mechanisms that people use.

2. The Fixed Adjective-Preposition construction (FAP)

2.1. The FAP-construction: Form

In this contribution we look at a specific type of conventional pairs of words: the fixed adjective-preposition construction in Dutch. An example of such a combination is *trots op* 'proud of'. At first glance, this seems to be a two-word fixed expression, but a closer look reveals that the adjective tends to be preceded by a subject and a verb (see below) and the preposition is followed by a nominal constituent, with which it forms a prepositional phrase. Example (1) is a prototypical instance of the pattern.

(1) *de boer is* ***trots op*** *zijn auto*
the farmer is **proud of** his car (fn001204.26)[1]
structure: NP V_{fin} [proud [of [his car]$_{NP}$]$_{PP}$]$_{AP}$

1. All Dutch examples are taken from the Spoken Dutch Corpus (CGN). This is a 10-million word corpus of contemporary Dutch, as spoken in the Netherlands and in Flanders. The codes between brackets identify utterances in this corpus. See http://lands.let.ru.nl/cgn/ehome.htm for more information.

The fixed elements of this construction are the adjective and the preposition. This combination is conventional: the selection of the preposition is not semantically transparent (in fact, the literal translation of *trots op* is 'proud on', instead of 'proud of'). There are quite a number of these conventional pairings in Dutch (we provide more examples in Section 3.1.2). Because they are purely conventional and fixed, it must be assumed that speakers of Dutch have these patterns stored in their linguistic repertoire or constructicon. In addition to these two lexically specific elements, the construction also consists of further underspecified elements. An underspecified element of a construction is an element that is not expressed with the same exact words in each instantiation of a construction. One of the underspecified elements is obligatory: the nominal constituent that combines with the preposition to form a prepositional phrase. Two other underspecified elements are discussed in Section 2.1.2. They collocate strongly with the FAP-construction: while these elements are not obligatory, they do occur in most instantiations.

2.1.1. Underspecified element 1: The nominal constituent

The lexical content and internal structure of the noun phrase are not specified for the construction; it may take different forms. These range from anaphoric pronominal expressions, as in Example (2), to referential lexical NPs with a noun (+ optional modifiers, cf. (3) and (4)), and even full clauses.

(2) *Jonas wou dat niet dus die was* ***boos op ons***
Jonas wanted that not so that was **angry at us**
'Jonas didn't want that, so he was angry at us' (fv901185.238)

(3) *Titia is voortdurend* ***boos op haar vader***
Titia is continually **angry at her father** (fn001240.20)

(4) *die is heel erg* ***boos op z'n uh Deense opponent*** *omdat*
he is very much **angry at his uhm Danish opponent** because
ie zomaar gaat liggen
he simply lies down. (fn007444.108)

Depending on its sentential context, the prepositional and/or nominal phrase may precede the adjective,with the NP reduced to the placeholder *er* or *daar* 'there', as in (5).

(5) *maar zoals wij uh dat ingericht hebben nou* ***daar*** *was*
But like we uhm that arranged have well **there** was
iedereen ***jaloers op***
everyone **jealous of**.
'but the way we arranged that, well everyone was jealous of it.'
(fn008213.286)

2.1.2. Underspecified element 2: The verbal constituent and the subject

All example sentences so far have a copula in them. The adjective in the FAP-construction is used predicatively, and the utterance expresses that a property is ascribed to the subject. Most instantiations of the FAP-construction take this form. This is not the case when the A in the FAP occurs adverbially, rather than as an adjective, which is something the majority of Dutch adjectives can do. In Dutch, adjectives used attributively occur before the noun. The pre-nominal adjectival phrase is usually short, as it is in English. Adjectives can be stacked (*I saw a tall, dark, handsome man*) or modified (*he was an extremely good-looking guy*), but, in English and Dutch alike, they cannot contain a modifier in the form of a prepositional phrase (**I saw the jealous of his colleague professor*).

The absence of FAP-instantiations in attributive positions means that this construction typically co-occurs with one of a very short list of verbs, namely those that are found in copula constructions. Of these, *zijn* 'to be' is by far the most frequent; it occurs in all examples given so far (ex. 1–5). Other copula verbs with the FAP-construction are rare, but they can be found (6).

(6) *je* ***zou*** *toch* ***jaloers op*** *die beesten* ***worden***
You would still **jealous of** those animals **become**
'It would make you jealous of those animals' (fn001227.46)

In terms of likelihood, the occurrence of a FAP-instantiation is a much more reliable cue for the co-occurrence of a copula verb than vice versa: copula constructions are a lot more frequent and the complement constituent may take many different forms, with both adjectival (*she is very intelligent*) and nominal constituents (*she is a doctor*). The verb links the property expressed in the adjective to a subject; the copula construction means that this link consists of ascribing the property to the subject. Since the verb is a strong collocate of the FAP-sequence, this will be part of the experimental design (see Section 3.1.4 below). As there do not seem to be any clear distributional patterns for specific subjects, other than that

subjects are typically references to humans, that underspecified element of the construction will not be varied systematically in the test items.

In sum, the fixed adjective-preposition construction consists of two fixed elements: an adjective and a preposition, which are conventionally paired. The underspecified elements are the subject, the verb and the nominal complement to the adjective. Although there are a large variety of subjects, the verb that occurs with this construction is usually a copula verb. The nominal element can take different forms. We now turn to a description of the meaning of the FAP-construction.

2.2. The FAP-construction: Meaning

The adjectives that occur as part of a FAP-construction also occur outside of these, carrying largely the same meaning. A direct comparison of a FAP instance and an utterance with the same adjective but no prepositional phrase allows for a first approximation of the construction's meaning:

(7) *hij is **jaloers op** jullie mooie huis*
He is **jealous of** your beautiful house (fn001175.116)

(8) *en als je met de postbode stond te praten was ik echt*
and when you with the postman stood to talk was I really
*heel erg **jaloers***
very much **jealous**
'and when you were talking with the postman I was really very jealous' (fn001041.21)

In both utterances there is a person to whom the property of an emotion is ascribed – jealousy. The prepositional phrase *op jullie mooie huis* 'of your beautiful house' in (7) is a lexically specific reference to the object the emotion is aimed at, the cause of this emotion. In (8) there is no prepositional phrase. The cause of the jealousy, however, is clear from the context. A corpus search of *jaloers* in the Spoken Dutch Corpus (118 occurrences) reveals that the object of the jealousy is not always so explicitly expressed. In those cases where *jaloers* co-occurs with *op* (39 tokens, sometimes with intervening sequences of up to 5 words), this prepositional phrase always refers to the object or person that causes this emotion, i.e. who or what the jealousy is aimed at.

The case of *jaloers op* is not unique. In many FAPs the prepositional phrase contains a reference to the cause and/or recipient of the emotion that the adjective describes (e.g. *blij met* 'happy with', *bang voor* 'afraid

of', *verbaasd over* 'surprised about' etc.). While not all adjectives that occur in FAPs refer to emotions, very often the prepositional phrase refers to a cause, as in *allergisch voor* 'allergic to', *kwijt aan* 'lose to', *ziek van* 'sick of' etc., as illustrated in Examples (9) and (10).

(9) *dus hij is een week van zijn vakantie* ***kwijt aan*** *stage?*
So he is a week of his holiday **lost on** internship?
'So he will lose a week of his holidays to doing an internship?' (fv400194.22)

(10) *ze ging wel vroeg naar bed want ze was* ***ziek van*** *'t hete eten of zo*
She went sure early to bed because she was **sick of** the spicy food or something
'She did go to bed early, because the spicy food or something had made her sick'. (fn000384.169)

2.3. The FAP-construction: Overall analysis

The general pattern that seems to underlie all the examples reviewed so far, is visualized in Figure 1. For reasons of space, the subject and the verb are left out. The adjective expresses a property, and the noun phrase

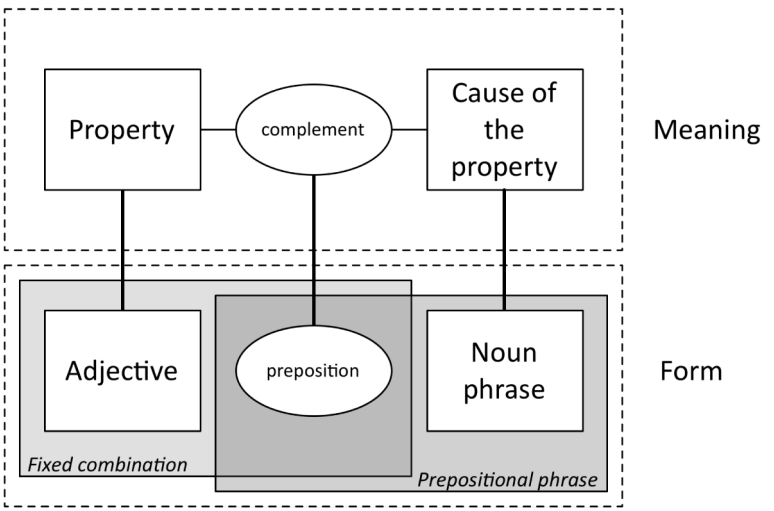


Figure 1. Visual representation of the FAP-construction

is the cause of that property. They are linked in a complement relation. The lexical expression of this relation is the preposition.[2]

The fact that it is possible to formulate a form-meaning pattern that the different FAPs all seem to follow, does not mean that this general schema is a level of representation that is psycholinguistically real, i.e. that speakers of the language have this in their constructicons. Here, we will not attempt to find out whether the latter is the case. Instead we focus on the degree to which specific FAPs are used as a unit in human sentence processing. Even when both fixed elements of the FAP construction occur in the same utterance, they may still not trigger the FAP interpretation, as is the case in (11) – although these examples are very rare.

(11) *meester Sjoerd gaat **trots** **op** zijn stoel zitten*
Teacher Sjoerd goes **proud** **on** his chair sit
'Mister Sjoerd proudly sits down on his chair' (fn001281.6)

The cause of a property is not something that a speaker always needs to express. For that reason it is not surprising that the adjectives in FAPs frequently occur without the prepositional phrase. The fact that adjectives referring to emotions seem to take up a large part of the distribution of FAPs may be explained by the construction's semantics: emotions are caused by something or someone, and this is a relatively salient aspect of this type of adjectives, compared to many other semantic groups of adjectives (e.g. colors, adjectives related to size etc.). For adjectives referring to emotions, when this property is ascribed to someone, the cause of that emotion must be retrievable in the discourse context. It can be expressed lexically with a FAP, but this is not necessary: sometimes an earlier mention suffices, or the cause could be left unexpressed on purpose.

The relative salience of the cause of an emotion entails that in a significant number of instances it will be useful information to express lexically. An in-depth analysis of the distribution of FAPs, for instance with the help of a behavioral profile (e.g. Divjak & Gries, 2008), could shed more light on the specifics of this construction, but is outside of the scope of the present contribution.

2. It is possible to express both the property and the cause in other ways, e.g. in two separate clauses (*he scored a goal. His mother was very proud*) or with a nominalization (*his goal made his mother very proud*).

3. Experiment

Determining what units people use when they produce language is difficult, because the unit boundaries are not visible: speech is continuous. Earlier research has shown that pauses in speech mainly occur at boundaries between constituents (e.g. Hawkins, 1971, Goldman-Eisler, 1972, Grosjean & Deschamps, 1975). This indicates that these boundaries are real, but does not tell us anything about the points where pauses are absent. Recent research into the effects of global and local text structure on pauses in speech shows that most speakers use pauses to indicate text structure to listeners (cf. Den Ouden, Noordman & Terken, 2009). This means that pauses in speech serve a communicative purpose, yet they do not provide conclusive evidence about the processing units in speech production. An alternative to the analysis of speech is to ask participants to divide sentences into units or ask them how strongly consecutive words are related, as Levelt (1969) did. The strength-of-relationship measure thus obtained strongly resembles the constituent structure. The problem or shortcoming of this technique, however, is that participants have to rely on explicit knowledge. It is an off-line task and therefore does not measure language processing. Participants who have been taught in school to analyze sentences in terms of constituents, may use this to perform the task, regardless of whether this actually conforms to the units they use themselves. Unfortunately, it is impossible to determine with certainty whether this strategy is employed (participants may even do this subconsciously, such that asking them about their strategies may not solve this problem).

Griffiths (1986) introduced an inventive design that sought to overcome the vagueness of pauses in speech as a measure, while maintaining the online character of the task. He asked participants to copy sentences in writing. They saw a sentence, which they then had to copy on a sheet of paper. The sentence was not in sight of the copy sheet. Participants were told that they could look back at the original sentence whenever they were unsure about how to continue. Each time they did this, it was registered at what point they were in the copying process (i.e. after which word they had to look at the sentence again). Much like the pauses, the look-back points are an indication of a unit boundary, although not looking back does not mean that there is no unit boundary: participants can remember more than one unit at a time. Since Griffiths' introduction of the method, technological advances have made it much easier to execute such an experiment: straightforward software programs can create log files in which switches are registered (cf. Ehrismann, 2009, for a similar experiment).

This experimental design is not a speech production task. In Gilquin and Gries' classification of kinds of linguistic data in terms of naturalness (Gilquin & Gries 2009: 5) this task rates rather low. All experimentally elicited data rank in the lower half, and experiments involving participants to do something they would not normally do with units they do not usually interact with come at the bottom of the range. Arguably, memorizing long sentences verbatim with the aim of reconstructing them is not a 'natural' task. We return to this issue in the closing section.

The task requires participants to store (a part of) a sentence in working memory for the duration of the copying process. Because participants have to reconstruct the exact sentence – paraphrasing is not allowed – they will have to store the specific sequences the sentence is made up from. This provides us with the advantage that we can see into what parts the participants break up the test sentences, including the FAPs, while they perform the task, thus giving us an insight in their sentence processing. The switch behavior between the original sentence and the copy screen indicates unit boundaries.

We use this experimental design to answer the following research questions:

1. Are FAPs a unit in human processing?
   a. Is the switching behavior different for sentences in which the adjective-preposition sequence is coincidental (no semantic link between the adjective and the prepositional phrase) than for sentences in which the prepositional phrase expresses the cause of the property (i.e. has a "FAP interpretation")?
   b. Is the switching behavior influenced by the identity of the verb that precedes the coincidental construction or the FAP?
2. Does a stochastic word probability measure recognize FAPs as one unit?
   a. Does the metric distinguish between coincidental sequences and FAP constructions?
   b. Is the metric influenced by the identity of the verb that precedes the coincidental construction or the FAP?
3. Is the word perplexity measure relevant in a predictive sense to aspects of human sentence processing?

In order to answer these questions, we designed an experiment in which participants were asked to reconstruct sentences that they had just seen. These sentences each contained an adjective-preposition sequence. By varying the context, we were able to determine the influence of the presumed unit status of the sequence (research question 1a and 2a) and of the verb

hypothesized to be associated with it (question 1b and 2b). The participants' data are compared to a word probability measure to answer the third research question. The relevant details of the experimental set-up are explained in Section 3.1, followed by the results (Section 3.2). The final part of this chapter discusses the findings in light of the converging and diverging evidence they present.

3.1. Experimental design

3.1.1. Participants

The participants were 35 children in sixth grade ('groep acht'), mean age 12;5 years. They came from two primary schools in Tilburg, a city in the south of the Netherlands. All children participated on a voluntary basis, with consent given by their parents. The experiment was part of a larger research project (Mos, 2010) focusing on knowledge and processing of complex lexical items. An attempt to replicate the experiment with adult participants failed: the task proved too easy for them, resulting in too few switches to perform statistical analyses.

3.1.2. Item selection

In order to select frequent adjective-preposition pairs, we first made an inventory of all combinations that were listed in the *Prisma woordenboek voorzetsels* (Reinsma & Hus, 1999). This dictionary lists combinations of a preposition and another word for over 5,000 different lexical items, and contains 472 different FAPs. Since our primary interest is in frequent combinations, we restricted this list to items where the adjective occurs at least 100 times in the Spoken Dutch Corpus, and the combination is found at least 10 times as a continuous sequence. 75 combinations met this requirement. For the experimental task we selected six combinations that allowed for the construal of test sentences in which the prepositional phrase could express either the 'cause' of the adjective ("FAP interpretation") or a separate location/prepositional phrase connected to the verb

Some adjectives occur with more than one preposition. In total 365 different adjectives are listed. When the dictionary contains two entries for the same adjective, in many cases there is a clear semantic difference between the two types of extensions, e.g. with *blij met* and *blij voor*, 'happy with' and 'happy for', respectively (cf. Examples (12) and (13)). While both types of referents can be construed as 'causes' for the emotion, the prepositional phrase introduced with *met* refers to the thing (physical object

or achievement/result) that causes happiness, and the *voor* phrase contain a reference to a person whose presumed happiness causes vicarious happiness in the speaker.

(12) *laten we dat vooropstellen we zijn natuurlijk wel* ***blij*** ***met***
let us that first_put we are of_course indeed **happy** **with**
onze vrouwen
our wives
'we must stress that we are of course happy with our wives' (fn000377.289)

(13) *ik ben* ***blij*** ***voor*** *de prins dat hij eindelijk een mooie*
I am **happy** **for** the prince that he finally a beautiful
jonge vrouw heeft kunnen vinden
young woman has been_able_to find (fv600215.13)

With other fixed combinations, this is less the case, e.g. *aardig tegen* and *aardig voor*, 'nice towards' and 'nice to'. Both prepositional phrases express the person kindness is directed towards (see (14) and (15)).

(14) *ja één keer en toen was ie best* ***aardig*** ***tegen***
yes one time and then was he kind_of **nice** **to**
mij
me (fn006990.20)

(15) *maar iedereen is altijd heel erg* ***aardig*** ***voor*** *mannen in Japan*
but everyone is always very much **nice** **to** men in Japan,
hè
right (fn007565.1380)

The distribution of these two combinations may be influenced by various factors, including region (i.e. one form is prevalent in a certain part of the country), genre and others. None of these combinations were selected as test items for the experimental task.

3.1.3. Corpus for the memory-based language model

The corpus on which the memory based language model is trained is a combination of two newspaper corpora: the Twente news corpus[3] and the

3. For more information on the Twente news corpus, see: http://inter-actief.cs.utwente.nl/~druid/TwNC/TwNC-main.html

ILK newspaper corpus[4]. The former contains newspaper articles, teletext subtitles and internet news articles. The latter contains data from several regional Dutch newspapers. The first 10 million lines of the corpus, containing 48.207.625 tokens, were taken to train the language model. Tokens which occurred five times or less were replaced by a special token representing low frequency words. We are aware that this reference corpus is not a close match to the type of input children are exposed to, which we would prefer. There is, however, no evidence that the FAP-construction is especially sensitive to genre difference. On a more pragmatic note, the Twente and ILK corpora are readily available, while a large corpus reliably reflecting the participants' input (both written and spoken) is not.

3.1.4. Test items

The task consisted of 24 sentences, each containing one of the six selected adjective-preposition sequences. For each pair, we determined the most frequently co-occurring verb in the Corpus of Spoken Dutch: *zijn* 'to be' in five cases and *doen* 'to do' for one pair. A sentence was created with a subject, this verb, the fixed combination, and an appropriate nominal phrase as the complement of the preposition, referring to the object the emotion named by the adjective is aimed at (see example (16)). These sentences thus contain the frequent verb + adjective + preposition combination, with the prepositional phrase related to the adjective ("FAP interpretation").

(16) *Al in april* ***was*** *Esra* ***enthousiast over*** *de vakantie naar*
Already in April **was** Esra **enthousiastic about** the vacation to
haar familie in het buitenland (TYPE A sentence)
her family in the foreign_country

In order to allow us to find out how the participants divide this sequence into smaller units, we placed the target sequence near the middle of the test sentence. Had the sequence been placed at the beginning, results would be clouded as the first words of a sentence are easily remembered. Positioning the target sequence too close to the end of the sentence would also create problems, because by that time it is easy to reconstruct the remainder of the sentence, as explained below when we detail the task procedure.

Subsequently we created sentences with exactly the same subject, adjective and prepositional phrase sequence, but in which the prepositional

4. A description of the ILK corpus can be found at http://ilk.uvt.nl/ilkcorpus

phrase does *not* relate to the adjective. These sentences were created both with the frequent verb (example (17), TYPE B sentences) and with a verb that makes a different interpretation for the prepositional phrase, e.g. a location or topic, more likely[5] (example (18), TYPE C sentences).

(17) *Voor de pauze* ***was*** *Esra* ***enthousiast*** ***over*** *de vakantie*
Before the break **was** Esra **enthousiastically** **about** the vacation
aan het kletsen met haar vriendin (TYPE B sentence)[6]
on the chatting with her friend

(18) *Na het weekend* ***begon*** *Esra* ***enthousiast*** ***over*** *de vakantie*
After the weekend **began** Esra **enthousiastically** **about** the vacation
te vertellen aan haar hele klas (TYPE C sentence)
to tell to her whole class

The 2 × 2 design is then completed by generating sentences with a non-frequently co-occurring verb and an FAP interpretation (example (19), TYPE D). Table 1 provides an overview of the sentence types.

(19) *Lang voor vertrek* ***begon*** *Esra* ***enthousiast*** ***over*** *de*
Long before departure **began** Esra **enthusiastic** **about** the
vakantie alvast haar tas in te pakken (TYPE D sentence)
vacation already her bag in to pack

Table 1. Types of test sentences

FAP interpretation	Frequent verb	
	+	−
+	Type A (example 16)	Type D (example 19)
−	Type B (example 17)	Type C (example 18)

5. For some speakers, and depending on the intonational contours, a FAP-interpretation may be available as well. This is particularly true for Type B and Type D sentences, which contain improbable combinations of a frequent verb + coincidental sequence interpretation or an infrequent verb + FAP-interpretation. Although we attempted to construe the sentences in such a way that the intended interpretation was most likely, readers who are speakers of Dutch may feel that for at least some of these sentences another interpretation is at least available if not preferable.
6. Note that the English gloss for 'enthousiast' in this sentence contains an adverbial suffix. In Dutch the same form can be used both adjectivally and adverbially.

In addition to these requirements, for each adjective – preposition pair the sentences were constructed in such a way that the four types contained an equal number of words and differed by no more than two letters in total length. This was done in order to minimize the influence of these factors, so that any difference in test behavior could be reliably attributed to sentence type. Table 2 summarizes these data for the six word pairs tested.

Table 2. Characteristics of test sentences

Pair	Translation	Frequent verb	Non-frequent verb	Word count	Letter count
Boos op	Angry at	Was	Stond *stood*	13	59–61
Enthousiast over	Enthusiastic about	Was	Begon *began*	15	70–71
Geïnteresseerd in	Interested in	Was	Stond *stood*	16	76–78
Jaloers op	Jealous of	Was	Begon *began*	13	63–64
Voorzichtig met[7]	Careful with	Deed *did*	Liep *walked*	15	67–68
Trots op	Proud of	Was	Stond *stood*	15	78–79

Each test sentence had a similar grammatical structure, which is reproduced in Figure 2 below.

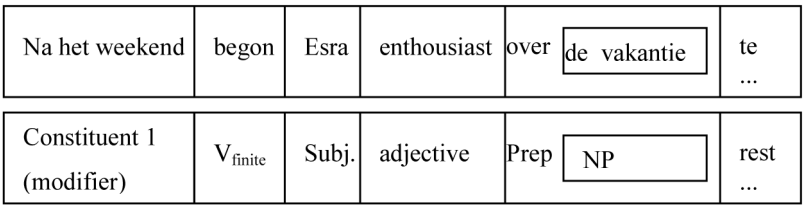


Figure 2. Grammatical structure of test sentences

7. One anonymous reviewer remarked that *voorzichtig met* 'careful with' is used adverbially in each of the four variants since it is combined with *doen* 'to do', rather than *zijn* 'to be'. In our opinion, *voorzichtig doen met* and *voorzichtig zijn met* are near synonyms. The verb *doen* is semantically light and functions rather like a copula verb. The experimental results for the test sentences with this FAP did not differ significantly from the other sentences.

In sum, the test items varied on three points: six fixed pairs of adjectives and prepositions, two finite verbs per pair, and two types of relation between the adjective and the prepositional phrase. The underspecified element 'verb' is thus a variable that is manipulated. The other two underspecified elements, the subject and the noun phrase, remain the same for the different variations of each FAP sequence. A complete list of test sentences can be found in Appendix 1.

3.1.5. Procedure

The experiment took place in the children's schools, in their computer rooms. Each of the participating schools has a dedicated class room equipped with enough computers to let all pupils in one class do the experiment simultaneously. The children were told that they would participate in an experiment designed to find out what kinds of sentences are difficult or easy to remember, to reduce test anxiety. Each child worked at an individual computer.

After starting up the program, the children saw a brief introduction, outlining what they had to do. A short text explained that they would see a sentence, which they were to read. They then had to press the space bar, which replaced the sentence with a new screen. On this screen, they saw a number of words in the top half, and an empty bar at the bottom (see screen shot in Figure 3). The task was then to drag the words down to the bar in the right order to form the sentence they had just read. Only if the correct word was dragged down, it would stay there. Other words would pop back up, i.e. they could only reconstruct the sentence from left to right and had to start at the beginning. If at any point in the sentence they forgot how it continued, they could return to the original sentence by pressing the space bar again.

After completing a sentence, a message appeared saying 'press Enter to continue with the next sentence' in Dutch. All children completed the whole task, copying 24 sentences, with the order randomized for each participant. Before starting with the first test sentence, they had to do a practice sentence first. The researcher was present in the computer room and answered questions about the procedure when necessary.

On average children spent nearly 20 minutes on the task, but some took up to half an hour. Some children got easily distracted, in spite of verbal admonitions by the researcher to try and be as quick as they could. They complained that the task seemed 'endless'. Because the order of the sentences was different for each participant, we assume that this has not compromised our data. All switch data were logged online and later retrieved for analysis.

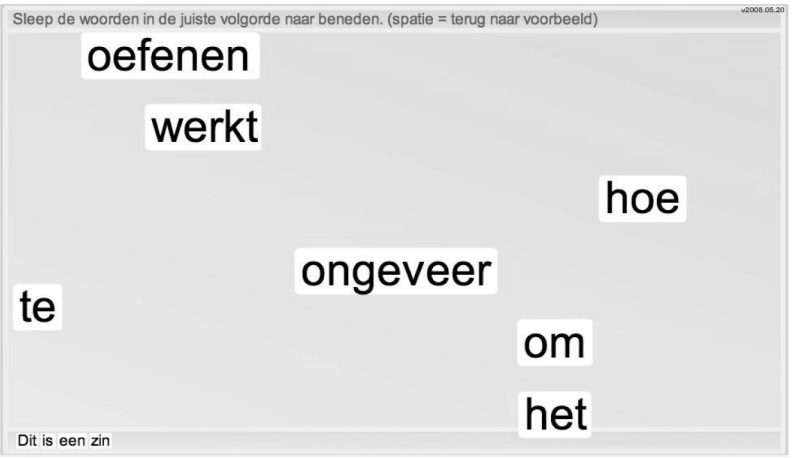


Figure 3. Screen shot copy task (practice sentence). The sentence at the top of the screen reads 'drag the words in the right order down (space bar = return to example)". At the moment this screen shot was taken, a participant had already dragged the first four words of the sentencs 'dit is een zin' *this is a sentence* down in the green bar

3.1.6. Variables

Item-based variables

The test sentences vary with regard to the adjective-preposition pair (six different pairs), the finite verb (two per pair) and the type of relation between adjective and preposition (two per pair). The switches that the participants made between the sentence and the reconstruction screen were all coded for their 'position': at which point in the reconstruction process did a participant go back to the original sentence. This position was defined with regard to the last word reconstructed. If a participant correctly copied *lang voor vertrek*, the first three words of the test sentence given as example (19), and then switched before continuing with *begon*, this was registered as a switch after *vertrek*.

Processing variables

The software program made specifically for this experiment logged for each word how it was handled. The log files therefore show at what points in the sentences each participant switched. Since people will only store a limited number of units in working memory at one time, they will switch

when that storage has run out. Each switch is therefore a sign of a boundary between two units, but the absence of a switch does not indicate that the child is dealing with only one unit. The sum of all switches was determined for each word boundary ('*n* switches adjective', for example, for the word boundary following the target adjective).

Likelihood variables

We trained a computational memory-based language model (Van den Bosch & Berck, 2009) on the aforementioned newspaper text corpora. The memory-based language model predicts, based on a context of n consecutive words to the left, a distribution of possible following words. The computational model can be likened to standard stochastic models that employ backoff smoothing, but without additional smoothing (Zavrel & Daelemans, 1997). Hence, if the model finds a matching context in memory that points to a single possible following word, the model predicts this word with a probability of 1.0. If there is a mismatch between the current local context of the n preceding words and the contexts in memory, the model backs up iteratively to find a match in the preceding $n-1$ words, producing estimates that do include more than a single possible word, with their probabilities adding up to 1.0. We set $n = 3$, yielding a 4-gram memory-based model that was subsequently applied to the 24 sentences to establish word-level perplexities. This means that the model assigns a value of word perplexity to all of the words in the 24 test sentences, based on its predictions given the three preceding words.

For each word, we take the negative base-2 logarithm of the probability assigned by the model to the word that actually occurs as the next word. This measure is typically referred to as the word-level logprob (Jelinek, 1998). The metric is strongly related to word-level perplexity, another often-used metric in statistical language modeling to express the degree of surprise of a language model to observe a word given an earlier sequence of words: word-level perplexity is 2^{logprob}. In the remainder of the text, we perform tests on the logprob measure, but occasionally refer to this metric as the "perplexity" measure.

3.2. Results

3.2.1. Descriptives

The 35 children who participated in the experiment and each copied 24 sentences switched a total of 1,794 times: 2.14 switches on average per

sentence. Each switch (back and forth between the sentence and the reconstruction screen) was coded for position. In other words, for each switch we know after which word it was made. In order to be able to sum switches over different test sentences, the positions were defined in terms of the word's function in the sentence (e.g. 'subject'). Figure 4 visualizes the switch behavior summed over all participants for one sentence in the form of a dendrogram. The switch behavior shown here is representative for the other sentences. The dendrogram illustrates which word sequences are more 'unit-like'. Sequences were iteratively combined starting with the word boundary that caused the fewest switches (in this particular example the last three words *een volle trein*). At each iteration, the sequences are linked that required the least amount of switches. For the sentence shown here, the word boundary between *veranderingen* and *stond* is the most frequent switch point (in this case, 15 of the 35 participants switched).

The dendrogram arising from the aggregated switches is remarkably consistent with the constituent structure of the sentences: the three prepositional phrases in the sentence *door de veranderingen*, *op haar school* and

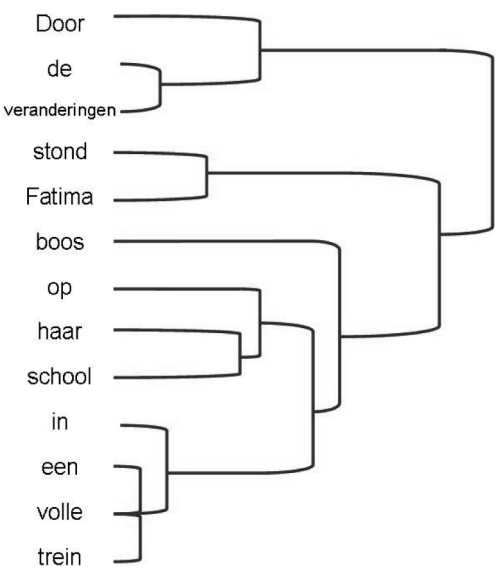


Figure 4. Visual representation of switch behavior (sentence contains *boos op* 'angry at', a FAP interpretation and a non-frequent verb)
Door de veranderingen stond Fatima boos op haar school in een volle trein
Because_of the changes stood Fatima angry at her school in a full train

in een volle trein each are linked together before they are integrated in the rest of the sentence. The noun phrase inside the prepositional constituent is also visible in the switch data. The sequence finite verb – subject is a relatively strong unit as well. All of these are semantic as well as structural units. The first constituent is linked to the rest of the utterance at the last iteration. This is a recurrent pattern for all test sentences: at the end of the three- or four-word sequence that constitutes the first phrase, many of the participants switch. These patterns are significant: the number of switches between the last word of the first constituent and the finite verb, and between the finite verb and the subject are significantly different (mean number of switches at boundary first constituent = 12.2 (sd 2.8) and after the finite verb = 3.1 (1.6), $t(23) = 12.75$, $p < .001$).

Finally, the number of switches tapers off near the end. This is most likely due to the research design: the participants had to select the next word from the remaining words in the sentence, all visible on the screen. For the last couple of words, there were only a few left to choose from.

The same kind of dendrogram can be construed on the basis of the probability measure, linking sequences that are progressively less likely to follow each other. Figure 5 contains a dendrogram for the same sentence as Figure 4, this time using the probability measure to construct it.

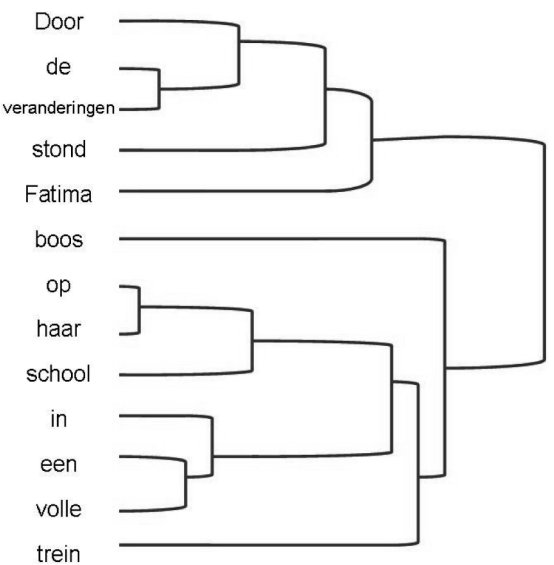


Figure 5. Visual representation of logprob (sentence contains *boos op* 'angry at', a FAP interpretation and a non-frequent verb)

The two figures are rather similar: the probability measure too results in a constituent-like structure for prepositional phrases. Note that within the prepositional phrase, the structure is different for *op haar school*. In terms of probability, the preposition – determiner sequence is more of a unit than the determiner – noun sequence. This is a recurrent pattern: the switch data tend to cluster a noun phrase within a prepositional phrase, whereas the probability data often lead to a cluster of preposition – determiner. A second difference is the absence of the sequence finite verb – subject as a unit in the probability-based dendrogram. Again, this is a recurrent finding for many of the sentences. We will return to these differences and discuss them in terms of human processing and the role of local co-occurrence probabilities therein in the last section of this chapter. Finally, unlike the human data, the probability measure does not profit from a reduction in possible candidates towards the end of the sentence, as the stochastic language model has no access to the diminishing list of possible continuations that the human subjects have. We already suggested that this effect in the switch behavior is due to the experimental design. This possible explanation is corroborated by the absence of this effect in the probability data.

3.2.2. Statistical analyses

a. Research question 1: Human switch data

For each adjective-preposition sequence, there were test sentences with and without a frequent verb and with and without a semantic link between the prepositional phrase and the adjective, a FAP interpretation. In a direct comparison of switch behavior for all four types of sentences (recall the 2×2 design), an analysis of variance was carried out with verb and interpretation as factors. The total number of switches per participant per sentence does not differ significantly depending on either the verb ($F(1,20) = 1.34$, p = n.s.), the interpretation ($F(1,20) = 0.15$. p = n.s.) or the interaction between the two factors ($F(1,20) = 2.18$, p = n.s.). The mean number of switches per sentence type is given in Table 3.

Since there is no significant difference in total number of switches per sentence type, we may assume that overall they did not differ significantly in processing difficulty. There is, however an overall difference between the six FAPs: the total number of switches is larger for sentences with *geïnteresseerd in* than *boos op* and *jaloers op* ($F(5,18) = 3.83$, $p < .05$, post-hoc Bonferroni tests show significant differences only between *geïnteresseerd in* and *boos op* and *jaloers op*). This difference is likely to be due to

Table 3. Mean total of switches per sentence, for each sentence type

Sentence type	Verb	Interpretation	Nr of switches	SD
A	Frequent	FAP	66.50	19.29
B	Frequent	Locative/other	74.50	16.87
C	Infrequent	Locative/other	72.17	10.32
D	Infrequent	FAP	85.83	23.02

Note: the maximum nr. of switches for each sentence is 35* (N words-1)

the difference in sentence length: when the number of words per sentence is entered as a covariate, the difference between the FAP-pairs is no longer significant ($F(4,18) = 0.92$, p = n.s.).

To determine whether the verb made any difference to the switch behavior, sentences with a frequent verb (frequent in co-occurrence with the FAP) and those with an infrequent verb were contrasted for the different word boundaries: statistical analyses were conducted to see if the number of switches differed at each point in the sentences. The only boundary where switch behavior was significantly different was at the 'subject' position. Fewer switches were made after the subject when the verb was a collocate of the adjective-preposition sequence: the mean number of switches was 10.33 (sd 2.64) for sentences with a collocate verb, and 14.00 (2.95) for a non-collocate verb ($F(1,20) = 11.08$, $p < .01$, $r^2 = .425$). This is an effect of the verb on the amount of switches made at a later point in the sentence: in each sentence, the verb came directly before the subject. Participants switched fewer times after the subject if the preceding verb was a collocate of the sequence. The effect of the verb on the number of switches is very likely caused by differences in frequency. Unfortunately, these data do not allow us to distinguish between an effect of the verb's overall frequency and the frequency of the collocation: in both respects, the +frequent verbs occur more often. In order to be able to determine which of these two aspects of frequency (general frequency of the verb or contextual frequency within the construction) is most relevant, a set of test items with more variation in verbs used is necessary. For the constructions tested here, one single verb accounts for the vast majority of tokens, which is the reason frequency was operationalized as a binary rather than continuous variable.

The effect of the interpretation (FAP interpretation or coincidental sequence) was not significant at this word boundary ($F(1,20) = 1.85$,

p = n.s.) or at any other word boundary. The interaction between the factors verb and interpretation is also not significant.

The 2 × 2 design for the test sentences entails that some of the sentences contained a verb likely to co-occur with the adjective – preposition sequence in a FAP interpretation, but where the prepositional phrase turned out to serve a different purpose (type B sentences). Likewise, other test sentences include a verb that does not co-occur often with the FAP, but was still followed by it (type D sentences). Such sentences may lead to garden-path effects: the first part seems to lead to one interpretation, but as the sentence continues, it becomes clear that a different interpretation is the correct one. To check whether this caused a difference in switch behavior, we contrasted type A and C sentences (no garden path effect likely) with type B and D sentences (garden path likely). There was indeed a significant effect: the likely garden path sentences required more switches at the word boundary after the first word following the prepositional phrase, that is once the whole FAP construction, including the full prepositional phrase, had been copied (mean nr. of switches for non-garden path sentences at this word boundary = 1.75 (1.54) and for potential garden path sentences = 4.75 (3.11), $t(22) = 2.99$, $p < .05$). This is a clear indication of a garden path effect: the sentences are processed with equal effort until *after* the entire target sequence. It is at this point that participants had to reanalyze the structure of the sentences.

In sum, we find a small significant effect of the verb in the test sentences: a frequent verb reduces the amount of switches needed after the subject in comparison with a verb that did not occur frequently with the adjective-preposition sequence. The distinction between FAP interpretation and coincidental sequences did not directly lead to differences in switch behavior. This suggests that co-occurrence patterns are influential for the temporary storage of the sentence by the participants, an observation that is confirmed by the effects of the likely garden path sentences.

b. Research question 2: Probability measure

For each word in the 24 test sentences, the model determined the probability of that word occurring. The measure used to reflect this probability is the aforementioned logprob measure. Analogous to the sum of switches for the copy task data, this results in a numerical value for each word boundary. Values for the logprob measure are higher when the likelihood is lower: a completely expected word with no competitors scores closest to zero, while strongly unexpected words receive a high value. The statistical tests that were used to identify any significant differences between groups of sentences and/or position, were also applied to the logprob measure.

Unlike the switch data, the logprob data were influenced by one variable only: the choice of verb. Sentences with a frequent verb have a perplexity score that is closer to zero for the subject than sentences with an infrequent verb (note that the subject is the first word after the verb, ($F(1,20) = 39.13$, $p < .001$). This indicates that the subject is more expected after the frequent verb than after the less frequent verb. Note again, that the infrequent verb is especially infrequent in combination with the FAP construction, and not infrequent in and of itself.

Similarly to the switch data, there are no significant differences in word perplexity for sentences with or without FAP interpretation ($F(1,20 = 0.17$, $p =$ n.s.). Unlike the switch data, any visible effects of garden path sentences are absent.

c. Research question 3: Comparing switch data and the probability measure

The memory-based language model provides us with a probability measure for each word, and the switch data summed over all participants provide a numerical indication of processing units. These two measures turn out to correlate quite strongly: Pearson's correlations are significant ($p < .05$) for 19 out of 24 sentences (r ranging from .518 to .841), with correlations for a further 4 sentences stronger than .40. The remaining sentence has a correlation of .249. We may remind ourselves that the probability measure is based on the preceding three words only. This means that for the first word overall frequency is the only guideline. Near the end of each sentence, the switch data go down a lot, due to the experimental design (see above), whereas there is no such help in the selection of the next word for the model. Given these restrictions, the correlations are reasonably strong, and indicate that the participants were more likely to remember the next word when this word was a probable word.

4. Conclusion and discussion

The experimental task that the participants performed provided evidence about the psycholinguistic reality of FAP sequences and the FAP construction.[8]

8. Note that this is not the same as saying that the representation in Figure 1 is psycholinguistically real: that figure abstracts over all FAP sequences in one highly general knowledge representation, whereas we tested six *specific* FAPs.

Sentences that contained a verb that frequently co-occurs with the FAP sequence were apparently easier to remember (i.e. fewer switches after the subject were observed). Moreover, if the sentence did contain a collocate verb with the construction, but no FAP interpretation or vice versa, this led to more switches later in the sentence, a finding which can be interpreted as a garden path effect.

Clearly, the FAP sequence is not word-like for all children and all sentences: there are sometimes switches between the adjective and the preposition. The results do suggest, however, that the sequence is processed more as a unit, when it is preceded by a verbal collocate. In our opinion, this result is most compatible with a view of 'units' as a gradient phenomenon: sequences are more or less unit-like, and this is influenced by distributional patterns, i.e. co-occurrence frequency, and by constituency. Note that with the current experimental design we can never be certain that a sequence is in fact analyzed as a whole. When there is no switch, the sequence may or may not be stored as a whole. It can still be the case that it is remembered as two units that are stored together in short-term memory (just like any other sequence of words). The fact that switch behavior is not random, though, makes this a less likely interpretation.

Our probability measure, reflecting the probability of the next word given the three preceding ones, was also significantly influenced by the main verb. In contrast to the switch data, however, a garden path effect was not found. This lack of effects can be attributed to the fact that the stochastic model bases its estimates on a local window of three words, meaning that it is oblivious to dependencies spanning beyond this width. Standard stochastic language models are known to be empirically bounded by a local window of three or four neighboring words (Jelinek, 1998), beyond which observations become too sparse, and estimates too unreliable. In addition to its limitations in locality, the probability measure is also based on overall frequencies and co-occurrence patterns alone: effects of (recent) context such as activation and decay are not part of the model. In human processing, this is akin to only having a long-term memory. Recency and priming effects are not captured by the present model. There is ongoing work on more flexible language models (e.g. Guthrie et al. 2006) that could prove useful here – this is a departure point for future research.

Yet, stating that the human switch data and the probability measure are unrelated would be an underestimation. For most of the sentences, the two measures correlated quite strongly in spite of limitations in the comparability (switches taper off near the end of the sentence, probability

is quite low for any first word of a sentence). A more detailed comparison of the two measures reveals, in addition to the similarities (see also the two dendrograms in Figures 4 and 5), two points of difference. These differences concern prepositional phrases and the finite verb – subject sequence.

The 24 test sentences contain a total of 71 prepositional phrases (PPs), each consisting of a preposition, a determiner (sometimes an adjective) and a noun. In the switch data, for only 14 PPs the number of switches between the preposition and the determiner was lower than that between the determiner and the noun. The stochastic measure showed a lower value for the preposition-determiner sequence than for the determiner-noun sequence 28 times. This seems to indicate that the noun phrase within the PP was sometimes more unit-like for the participants than would be expected on the basis of the stochastic measure (see Figure 6 below for the comparison).

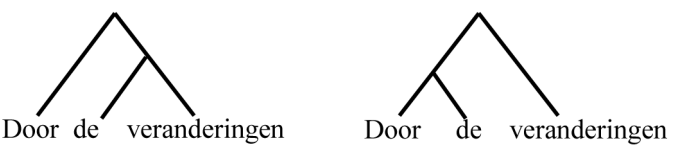


Figure 6. Clustering an NP within a PP (left) vs. a DET + PREP sequence within a PP (right)

The sequence of two relatively frequent closed-class items (preposition + determiner) is often associated with a high probability. This can be taken as a weakness of the stochastic measure, but at the same time reflects the intrinsic ambiguity between regarding the preposition + determiner pair or the determiner + noun pair as the primary unit; it could be argued that these two readings may exist in parallel. This would be another reason to expect gradient aggregated switching behavior rather than all-or-none.

The second point of divergence concerns the sequence of two open class words: the verb and the subject. In the test sentences, the subject was always a proper noun. Although this is a theoretically infinite set, the word boundary between the verb and the subject was never a more frequent switch point than either the word boundary before or after this sequence in the switch data. In other words, for all 24 test sentences, more participants switched before the verb and after the subject than between these two words. The perplexity measure does not reflect this completely: in one third of the sentences (8 times) the logprob value for either the preceding

or the following word is closer to zero. It can be expected that there are no clear predictions at this point of the sentence from the memory-based model. For the participants in our experiment, however, the sequence was very much part of one unit in memory.

The copy task proved to be an informative task; the switch data show clear effects of conceptual units and of co-occurrence patterns. An attempt to replicate the experiment with adult participants, however, failed: the task proved so simple that adult participants hardly ever switched (less than 1.5 switches per sentence). Apparently, having to remember sentences of the type that was used in this experiment does not tax their memory enough to cause switches. Increasing the task demands is one possible way to make this experiment more difficult for adults. Ehrismann (2009) did this by adding a secondary task: participants had to add a second sentence to each utterance they had copied. Other options include adding distracting sound or limiting view time to the original sentence.

In order to distinguish between a 'simple' frequency effect of the verb, regardless of the remainder of the sentence, and a co-occurrence frequency effect based in the FAP-construction, a follow-up experiment should also include test sentences with simple predicative structure (Type A sentences), but a non-collocate copula verb.[9] While the frequent verbs may have facilitated processing locally (effects are found at the subject position), differences in the total amount of switches per sentences type are not significant, indicating that this effect does not extend to the sentence level and that the possibly higher syntactic complexity of Type B and Type D sentences does not lead to more switches overall.

The broader question behind the comparison between experimental data and a measure that is based in frequency and co-occurrence patterns is to what extent human sentence processing reflects these distributional patterns. Given the caveats that were discussed earlier, we feel that the correlations are encouraging: the simple stochastic model explains more than 25% of the variance.

The copy task requires participants to briefly remember parts of utterances. It is therefore likely that their working memory capacity influences performance on this task. In this chapter, we were interested in finding out more about the unit-status of FAP sequences and the influence of the (preceding) verb and the interpretation on these units. While including a WM-measure in a follow-up design would certainly be interesting, work-

9. We are grateful to the audience at the Fifth Corpus Linguistics Conference in Liverpool (July 2009) and an anonymous reviewer for these suggestions.

ing memory span tasks measure the amount of units (digits, tones, words etc.) someone can briefly story. They do not provide information about the nature of the units that are stored. A sequence of words will be more likely to be memorized in one attempt by someone with a large WM-capacity, but it is not more likely that this sequence is stored as one unit.

Earlier in this contribution, we indicated that our experimental task ranks rather low on the 'naturalness scale' that Gilquin and Gries (2009) introduced. At the 2009 Corpus Linguistics conference in Liverpool the editors of this book organized a workshop entitled 'converging and diverging evidence'. The participants at this workshop each compared corpus-based frequency measures to human language use. Interestingly, tasks that scored relatively high on the naturalness scale, such as defining different word senses or recognizing neologisms (Littlemore & MacArthur, 2009; Svanlund, 2009), correlated less strongly with frequency measures than the more unnatural task of a word class identification task also presented at this workshop (Teddiman, 2009), which is also not normal language use.

In this case, 'naturalness' may not be the most appropriate scale to place these tasks on. The word class identification task and the copy task we report on here have in common that they are online tasks: they measure language processing as it happens. In word definition tasks, on the other hand, people have time to reflect on their answers, or at least do not have to instantly decide or act. We suggest that it is this difference that is reflected in a higher (online tasks) or lower (off-line tasks) correlation with frequency measures. Time allows people to not only rely on frequency, but to take other factors (possibly pragmatics, larger textual and co-textual contexts, and meta-linguistic knowledge etc.) into account.

Seen on the gradient scale of time allotted to test subjects, from immediate to indefinite, our task is not among the most immediate online tasks: participants read the sentence first before they started the copying process. We believe that this is the core reason we find no significant correlation of our simple stochastic metric and the presence or absence of the FAP construction – while at the same time we do observe a quite strong overall correlation. Arguably, the room that our task grants to participants to think about the sentence causes divergence of their behavior with respect to the simple stochastic model. In sum, we feel that the results of our experimental task provide converging evidence for the relevance of frequency in human processing. At the same time, however, the data show that people take into account more aspects of structure and meaning than can be captured by the current simple stochastic model.

Acknowledgements

Discussions with Joost Schilperoord, Ad Backus and Anne Vermeer have greatly helped our thinking about FAPs. The software program for this experiment was written by Michiel van der Ros. For more information about this program, please contact the first author.

Appendix 1: test sentences

Sentences are grouped by FAP. For each adjective-preposition pair, there are four test sentences:

A: frequent verb, FAP interpretation
B: frequent verb, FAP is a coincidental sequence
C: infrequent verb, FAP is a coincidental sequence
D: infrequent verb, FAP interpretation

Boos op, *angry at*

A. Vanwege die beslissing was Fatima boos op haar school en de nieuwe leraar
'Because of that decision, Fatima was angry at her school and the new teacher'
B. Ondanks de toestemming was Fatima boos op haar school brood aan het eten
'In spite of the permission, Fatima was angrily eating a sandwich at school'
C. Om te protesteren stond Fatima boos op haar school folders uit te delen
'In order to protest, Fatima angrily handed out flyers at her school'
D. Door de veranderingen stond Fatima boos op haar school in een volle trein
'Because of the changes, Fatima, angry at her school, was standing in a full train'

Enthousiast over, *enthusiastic about*

A. Al in april was Esra enthousiast over de vakantie naar haar familie in het buitenland
'Already in April, Esra was enthusiastic about the vacation to her family outside the country'

B. Voor de pauze was Esra enthousiast over de vakantie aan het kletsen met haar vriendin
'Before the break, Esra was chatting enthusiastically about the vacation to her friend'
C. Na het weekend begon Esra enthousiast over de vakantie te vertellen aan haar hele klas
'After the weekend, Esra enthusiastically started to tell her whole class about the vacation'
D. Lang voor vertrek begon Esra enthousiast over de vakantie alvast haar tas in te pakken
'Long before departure, Esra, enthusiastic about the vacation, started to pack her bag'

Geïnteresseerd in, *interested in*

A. Vanwege zijn loopneus was Tim geïnteresseerd in een folder van de huisarts over griep en snot
'Because of his runny nose, Tim was interested in a brochure from the doctor about the flue and mucus'
B. Met zware hoofdpijn was Tim geïnteresseerd in een folder van de huisarts tips aan het lezen
'With a bad headache, Tim was reading a brochure from the doctor with interest'
C. Met hoge koorts stond Tim geïnteresseerd in een folder van de huisarts te lezen over ziektes
'With a high fever, Tim stood reading with interest a brochure from the doctor about diseases'
D. Met erge buikpijn stond Tim geïnteresseerd in een folder van de huisarts te wachten op hulp
'With severe stomachache, Tim, interested in a brochure from the doctor, stood waiting for help'

Jaloers op, *jealous of*

A. Zonder enige twijfel was Sophie jaloers op de tas van haar nichtjes uit Breda
'Without any doubt, Sophie was jealous of her nieces' bag from Breda'
B. Zonder te aarzelen was Sophie jaloers op de tas van haar buurvrouw gaan staan
'Without hesitating, Sophie had jealously stepped on her neighbor's bag'

C. Zonder enige aanleiding begon Sophie jaloers op de tas van haar zus te spugen
'Without any reason, Sophie started to spit jealously on her sister's bag'
D. Zonder goede reden begon Sophie jaloers op de tas van haar vriendin te worden
'Without a good reason, Sophie started to get jealous of her friend's bag'

Voorzichtig met, *careful with*

A. Na dat ongelukje deed Fleur voorzichtig met de vaas die enorm veel geld gekost had
'After that accident, Fleur was careful with the vase that had cost an enormous amount of money'
B. Na het schoonmaken deed Fleur voorzichtig met de vaas in haar handen een stap opzij
'After cleaning, Fleur carefully stepped aside with the vase in her hands'
C. Na het uitpakken liep Fleur voorzichtig met de vaas naar de vensterbank in de kamer
'After unpacking, Fleur carefully walked with the vase to the windowsill in the room'
D. Na het verhuizen liep Fleur voorzichtig met de vaas maar druk pratend naar de gang
'After moving, Fleur walked to the hallway, careful with the vase, but talking busily'

Trots op

A. Net zoals zijn broer was Daan trots op de carnavalswagen waaraan ze maandenlang hadden gewerkt
'Just like his brother, Daan was proud of the carnival float they had been working on for months'
B. Met zijn blauwe pruik was Daan trots op de carnavalswagen geklommen voor hij eindelijk vertrok
'With his blue wig, Daan had proudly climbed onto the carnival float before it finally took off'

C. Ondanks de koude regen stond Daan trots op de carnavalswagen te zwaaien naar iedereen onderweg
'In spite of the cold rain, Daan stood waiving proudly on the carnival float to everyone on the way'
D. In zijn nieuwe kostuum stond Daan trots op de carnavalswagen tussen de mensenmassa te wachten
'In his new costume, Daan, proud of the carnival float, stood waiting between the crowd'

References

Boas, H. C.
2003 *A Constructional Approach to Resultatives*. Chicago: University of Chicago Press.

Boas, H. C.
2008 Determining the structure of lexical entries and grammatical constructions in Construction Grammar. *Annual Review of Cognitive Linguistics*, 6, 113–144.

Bybee, J. and Scheibman, J.
1999 The effect of usage on degrees of constituency: the reduction of *don't* in English. *Linguistics*, 37, 575–596.

Divjak, D. and Gries, S. Th.
2008 Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon*, 3, 188–213.

Ehrismann, M.
2009 *Eén enkele eenheid? De wijze waarop gefixeerde expressies in het lexicon zijn opgeslagen*..[One single unit? The way in which fixed expressions are stored in the lexicon] MA Thesis, Tilburg University, The Netherlands.

G. and Gries, S. Th.
2009 Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5, 1–25.

Goldman-Eisler, F.
1972 Pauses, clauses, sentences. *Language and Speech*, 15, 103–113.

Griffiths, P.
1986 Constituent structure in text-copying. *York Papers in Linguistics*, 12, 75–116.

Grosjean, F. and Deschamps, A.
1975 Analyse contrastive des variables temporelles de l'anglais et du français. *Phonetica*, 31, 144–184.

Guthrie, D., Allison, B., Liu, W., Guthrie, L. and Wilks, Y.
2006 A closer look at skip-gram modeling. Paper presented at the Fifth International Conference on Language Resources and Evaluation (LREC-2006) in Genoa, Italy.

Hawkins, P. R.
1971 The syntactic location of hesitation pauses. *Language and Speech*, 14, 277–288.
Jelinek, F.
1998 *Statistical Methods for Speech Recognition*. Cambridge, MA: The MIT Press.
Levelt, W. J. M.
1969 The scaling of syntactic relatedness: A new method in psycholinguistic research. *Psychonomic Science*, 17, 351–352.
Littlemore, J. and MacArthur, F.
2009 *Figurative extensions of word meaning: How do corpus data and intuition match up?* Paper presented at the Fifth Corpus Linguistics Conference in Liverpool, UK.
Mos, M.
2010 *Complex Lexical Items*. Ph.D. Thesis, Tilburg University / LOT.
Ouden, H. Den, Noordman, L. and Terken, J.
2009 Prosodic realizations of global and local structure and rhetorical relations in read aloud news reports. *Speech Communication*, 51, 116–129.
Reinsma, R. and Hus, W. J. B.
2005 *Prisma Woordenboek Voorzetsels*. Utrecht: Het spectrum.
Schilperoord, J. and Cozijn, R.
2010 The representation and processing of fixed and compositional expressions. In: M.Everaert, T. Lentz, H. de Mulder, Ø. Nilsen, and A. Zondervan (eds.), *The linguistics enterprise: From knowledge of language to knowledge in linguistics* (pp. 291–314). Amsterdam: John Benjamins.
Svanlund, J.
2009 *Frequency, familiarity, and conventionalization*. Paper presented at the Fifth Corpus Linguistics Conference in Liverpool, UK.
Teddiman, L.
2009 *Conversion and the lexicon: Comparing evidence from corpora and experimentation*. Paper presented at the Fifth Corpus Linguistics Conference in Liverpool, UK.
Van den Bosch, A. and Berck, P.
2009 Memory-based machine translation and language modeling. *The Prague Bulletin of Mathematical Linguistics*, 91, 17–26.
Zavrel, J., and Daelemans, W.
1997 Memory-based learning: Similarity for smoothing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, July 1997, 436–443.

Index