Stefan Th. Gries, Marlies Jansegers, and Viola G. Miglio

Quantitative methods for corpus-based contrastive linguistics

Abstract: The present paper makes a methodological contribution to the field of corpus-based contrastive linguistics. Contrary to the large majority of studies in contrastive linguistics that are mainly based on observed (relative) frequencies of (translation) data and are essentially monofactorial in nature, our study leverages more complex contrastive data that do justice to the complexity and multifactorial nature of cross-linguistic phenomena. Specifically, we focus on four challenging notions for the study of cross-linguistic near-synonymy: polysemy, degree of sense distinctiveness, prototypicality and identification of discriminatory variables. Each of these phenomena is tackled by means of a variety of statistical analyses based on two different kinds of input data that offer different kinds of resolutions on the data: (i) annotated concordance data and (ii) Behavorial Profile vectors. In an attempt to add to the toolbox of contrastive linguistics, we pay special attention to visualization techniques for cross-linguistic (dis)similarities such as hierarchical agglomerative cluster analysis, fuzzy clustering, and network analysis. These statistical methods will be illustrated on the basis of a case of cross-linguistic near-synonymy, namely the verb *sentir(e)* in Romance Languages.

Keywords: cross-linguistic near-synonymy, Behavioral Profile, data visualization (fuzzy clustering, network analysis), Romance perception verbs

1 Introduction

1.1 General introduction

Over the last few decades, linguistics has experienced a strong empirical and quantitative turn towards both experimental and observational, esp. corpus, data. Much of corpus linguistics was originally centered on monolingual corpora, but

Stefan Th. Gries, Department of Linguistics, University of California, Santa Barbara, United States of America; Justus Liebig University Giessen, Germany
Marlies Jansegers, Erasmus University College Brussels, Zespenningenstraat, Brussels
Viola G. Miglio, Department of Spanish and Portuguese, University of California, Santa Barbara, United States of America

https://doi.org/10.1515/9783110682588-003

over time corpus methods also became more widespread in contrastive-linguistic studies. However, although much of the revival of Contrastive Linguistics in the 1990s is due to its meeting with corpus linguistics, cross-fertilization between both disciplines is still rather limited as there are two main challenges that have not yet been fully addressed, namely (i) an empirical assessment of the nature of the data which are commonly used in cross-linguistic studies (namely translation data vs. comparable data), and (ii) the development of advanced methods and statistical techniques suitably adapted to the methodological challenges that are raised by contrastive research questions. Contrary to the other contributions in this volume, which largely focus on the nature of the data, the present paper focuses on this second challenge and focuses on making a methodological contribution to the field of contrastive linguistics (even though it should go without saying that improved methodology also has huge implications for what is possible in the areas of theory development and testing).

Even anno 2015, Gast (2015: 6) states that "the methodological branch of corpus-based contrastive linguistics is still tender", an inconvenient truth that becomes particularly evident when considering for example the specific field of contrastive semantics. Indeed, a closer look at the recent bibliography in contrastive corpus-based semantics shows that, with the notable exception of studies such as Levshina (2016), many analyses are based exclusively on mere frequency counts of translation equivalents (among others Viberg 1999, 2002, 2005; Altenberg 2002; Schmied 2008). Other studies make use of comparable corpora instead of translations or a combination of both, but are again largely based on mere (relative) frequencies (among others Enghels and Jansegers 2013; Comer and Enghels 2016; Rozumko 2016; Lansari 2017; Molino 2017).

With the objective of making the methodological branch of corpus linguistics less tender, the present paper is both programmatic and methodological in nature in that we aim to showcase the use of different statistical methods that can be applied to contrastive corpus-based semantics, which will be illustrated on the basis of a data set on cross-linguistic near-synonymy. Specifically, we are following up on Enghels and Jansegers (2013), a study of the semantics of the cognate verbs *SENTIR(E)* in the three Romance languages (French, Spanish, and Italian) combining parallel and comparable corpora.[1] The two main findings of this study were the following:

1 The definition of these kinds of data and the difference between parallel and comparable corpora has been discussed elsewhere in this volume, see especially the papers from De Baets et al. and Viberg.

(i) It showed that the *tertium comparationis* at its most basic level can be defined as "general physical perception without any modality of perception being specified", as exemplified by the translation equivalents in (1):

(1) a. *Harry **sentit** la chaleur se répandre autour de lui comme s'il venait de plonger dans un bain tiède*. (French)
b. *Harry **sintió** que el calor lo cubría como si estuviera metido en un baño caliente*. (Spanish)
c. *Harry **sentì** il calore inondarlo come se si fosse immerso in un bagno caldo*. (Italian)
'Harry felt the warmth wash over him as though he'd sunk into a hot bath.' (Harry Potter and the Philosopher's stone)

In other words, this translation equivalence shows that *SENTIR(E)* has been defined as a general physical perception verb in all three languages and it is this classification that constitutes the *tertium comparationis* at its most basic level. Therefore, *tertium comparationis* or "common ground" of comparison (Altenberg and Granger 2002: 15) for this study does not only refer to formal identity but also this basic semantic similarity between the three verbs.

(ii) However, apart from this small common core of perfect lexical correspondence, there seem to be some important language specific features: French *sentir* most dominantly covers the field of cognitive (but often intuitive) perception (see (2)). Italian seems to be the language where *sentire* most clearly belongs to the category of perception verbs, referring in the vast majority of the cases to auditory perception (see (3)). Spanish, on the other hand, has strongly developed the emotional sense of the verb and related to this, refers to the emotional meaning "regret, deplore" in a unique way (see (4)):

(2) a. *Il l'avait **senti** plus **qu'entendu**: quelque chose ou quelqu'un se trouvait dans l'espace étroit entre le muret et le garage de la maison devant laquelle il s'était arrêté*. (French)
b. *Más que **oírlo**, lo **intuyó**: había alguien detrás de él, en el estrecho hueco que se abría entre el garaje y la valla*. (Spanish)
c. *Lo **avvertiva**, più che **sentirlo** con le orecchie: c'era qualcuno o qualcosa lì nello stretto passaggio tra il garage e la staccionata alle sue spalle*. (Italian)
'He had **sensed** rather than **heard** it: someone or something was standing in the narrow gap between the garage and the fence behind him.' (Harry Potter and the prisoner of Azkaban)

(3) a. *Elle **entendit** soudain battre son propre cœur. Ma famille?* (French)
b. *De pronto Sophie se **oía** los latidos de su corazón. ¿Mi familia?* (Spanish)
c. *Sophie aveva **sentito** che il cuore accelerava i battiti. La mia famiglia?* (Italian)
'Sophie suddenly could **hear** her own heart. My family?' (Da Vinci Code)

(4) a. ***Je suis désolée**, Potter, reprit-elle, mais c'est mon dernier mot.* (French)
b. ***Lo siento**, Potter; pero es mi última palabra.* (Spanish)
c. ***Mi dispiace**, Potter, ma è la mia ultima parola.* (Italian)
'**I'm sorry**, Potter, but that's my final word.' (Harry Potter and the prisoner of Azkaban)

In the present study, we take these observations as a starting point but we would like to make several suggestions for how it can be extended, both from a methodological and a more qualitative perspective:

– While the study by Enghels and Jansegers (2013) mainly addresses the issue of the comparability / compatibility between translation and comparable corpus data, it is based on observed (relative) frequencies, and is essentially monofactorial in nature. Our study, by contrast, focuses on the methodological challenge for the field of Contrastive Linguistics. It leverages more complex contrastive data derived from Behavioral Profiles (BPs) that are based on the similarities of vectors in order to explore the question of how this degree of cross-linguistic near-synonymy can be operationalized and investigated on an empirical and quantitative basis. That is, how can we compare multifactoriality behind this case of near-synonymy between sister languages? In an attempt to add to the toolbox of contrastive linguistics, we also extend this method for better visualization of cross-linguistic differences.
– On a more qualitative level, the study by Enghels and Jansegers focuses largely on the semantics of the verbs, and adopts moreover a coarse-grained perspective by focusing on three general semantic categories such as physical perception, emotional perception and cognitive perception. Since the BP method starts from the distributional hypothesis, namely the idea that differences in function/meaning are reflected in differences in distribution, we performed a very fine-grained manual annotation of dozens of features that include not only semantic, but also morphological, syntactic, and other characteristics. In this way, we hope to answer the question to what extent do these semantic differences correlate with syntactic diverging patterns.

In what follows, we will briefly describe the outline and of this chapter as well as the kinds of phenomena we discuss.

1.2 Overview of the present paper

As mentioned above, this paper is intended to be two things: (i) programmatic in nature and (ii) methodological. Specifically, we wish to discuss how a variety of research questions that are common in contrastive linguistics (with a special emphasis on semantic questions) can be studied on the basis of corpus data and their differently sophisticated statistical analyses. Given constraints of space, the proposed methodologies can only be exemplified briefly, which makes it even more necessary than generally to structure this overview well. Two issues need to be covered in particular: the range of phenomena we will cover and the kinds of input data whose statistical analysis will be discussed.

With regard to the former – the phenomena – we will focus on the following concepts, each of which will be briefly addressed in a separate section below:

- *degree of sense distinctiveness*: How many different senses of an expression are there in each language separately and how do these senses relate to each other within and across languages?
- *polysemy*: To the extent that senses can be delineated/operationalized, which senses are there and how do they differ especially across languages?
- *prototypicality*: To what degree are prototypical meanings of cognate words similar or different across languages? Is it possible to identify one cross-linguistic prototype?
- *identification of discriminatory variables*: What are the (morphosyntactic and semantic) variables correlating with a specific sense that most strongly discriminate between languages?

With regard to the latter – the input data – there are two kinds of data we will consider, since they offer different levels of resolution and of usefulness for further analysis. In particular, we will focus on the following kinds of data:

- *Annotated concordance data*: where the input will consist of, typically, a spreadsheet kind of structure in which each row represents one line of a concordance output (i.e., one match) and in which each column represents one variable with regard to which the match has been annotated (for what follows, such variables will also be referred to as ID tags, see Atkins 1987) and the different values that each variable/ID tag can assume will be referred to as ID Tag levels; for example, each subject of a verb could be annotated for the variable/ID tag *subject animacy* using one of, say, four, ID tag levels

(e.g. *human*, *animate*, *concrete inanimate*, *abstract*). This format is commonly referred to as the case-by-variable format (e.g. Maindonald and Braun 2010 or Fox and Weisberg 2011) or "the long format".

– *Behavioral Profile vectors* (based on annotated concordance data): this format is based on percentages. Behavioral Profiles is a statistical method to analyze semantic and syntactic aspects of corpus/concordance data with regard to semantic questions such as (near) synonymy, polysemy, and others. It was developed by Gries (2006) and Divjak (2006). If one created the above kind of annotated concordance data for – say – a set *x* of near-synonymous verbs in one language, then Behavioral Profile (BP) vectors are generated from it by computing for each of the *x* verbs, the percentage that each ID tag level makes up each ID tag. This is the technical way for saying something statistically quite easy: It means that, to use the above example of *subject animacy*, for each verb, we compute how many instances in % of the subjects are *human*, are *animate*, are *concrete inanimate*, and are *abstract*; these percentages will add up to 1 (100%), and we do the same for each verb and for each other ID tag. That way, each verb's overall behavior will be characterized by a concatenation of ID tag percentages (each adding up to 1), which can then be analyzed in various ways; for applications, see Divjak and Gries (2009); Gries (2010a); Gries and Otani (2010).

In the next section, we discuss the data we use in this paper to exemplify our analyses, first the annotated concordance data (Section 1.3.1), then the BP vectors (Section 1.3.2).

1.3 The current data

1.3.1 The annotated concordance data

In order to study *SENTIR(E)* from a cross-linguistic perspective, we compiled a comparable corpus consisting of authentic texts in each language that match as far as possible in terms of text type, subject matter and communicative function (Altenberg and Granger 2002: 8), but are not translations of each other. From this corpus, 1,500 occurrences of the verb *sentir(e)* were retrieved – 500 per language – half of which were drawn from literature (fiction) and the other half from press texts.[2] From these comparable data we generated and annotated

2 The availability of representative corpora differs considerably from one language to another. The Spanish database CREA contains both fiction and journalistic data, but for French the literary

pseudo-randomly sampled concordance lines of *SENTIR(E)* in all three languages for a large variety of morphosyntactic and semantic properties, called ID tags (Atkins 1987). A wide range of objectively verifiable (observable) parameters were distinguished according to four general levels of analysis, that is (i) the properties of the verb itself, (ii) the argument structure of the verb, (iii) the characteristics of other adjuncts, and (iv) discourse phenomena. Table 1 presents an example of such ID tags and their levels:

AU: We have highlighted the cross references for author/editor's check which will be removed in the next stage. Please check and confirm.

Table 1: Examples of ID tags and their levels.

General level	Type of ID tag	ID tag	ID tag level
Verb	morphosyntactic properties	Tense	present, past, future, infinitive
		Person	1, 2, 3
		number	singular, plural
	semantic properties	semantic category	general physical, specific physical, emotional, cognitive, ambiguous
		fine-grained sense (40)	emotional experience, to hear, general physical experience, to realize, to consider/judge, to intuit, tactile experience, to regret, … (=70%)
Argument structure	properties of subject form	lexical S	with S, without S
	properties of object form	lexical DO	with DO, without DO
	semantics of DO	referent DO	person, concrete entity, abstract entity, situation, ambiguous
Adjunct	properties of adverbial adjuncts	presence of adverbial adjunct	w/ adverbial adjunct, w/out adverbial adjunct
		form of adverbial adjunct	adverb, prepositional phrase, nominal phrase, etc.
Discourse	scope	predicational autonomy	no, yes

database FRANTEXT was complemented by data retrieved from the newspaper Le Monde. The Italian journalistic database *Il Corriere della Sera* (CdS) was supplemented with data drawn from two novels: *La luna di carta* (A. Camilleri) and *L'intreccio di universi paralleli* (A. Lo Gatto).

As an essential part of the analysis, the sense annotation merits some additional comments. As indicated in Table 1, the semantic analysis of the verb itself was done in two different resolutions. First, we resorted to a very fine-grained annotation of the different possible senses that were minimally different. Second, this fine-grained analysis then led to a more coarse-grained classification into four general semantic categories, namely (i) general physical perception, (ii) specific modality of physical perception, (iii) emotional perception and (iv) cognitive perception. This was done manually and mainly on the basis of the Romance comparative study of *SENTIR(E)* by Enghels and Jansegers (2013) where a lexicographic analysis was complemented with the results of a parallel corpus, based on translation data.[3]

The output of this first step then is a spreadsheet with one row for every concordance match of *SENTIR(E)*, some columns describing the language and maybe corpus of each match, and minimally one additional column for every ID tag that has been annotated, as exemplified in Table 2.

Table 2: Snippet of a concordance spreadsheet with annotation.

Preceding	Match	Subsequent	X	Y	Z	...
a b c	**sentir**	d e f	k	l	m	...
o p q	**sentir**	r s t	w	x	y	...
...	...	...	...	...	...	...

1.3.2 The BP vectors

After the retrieval and manual annotation of all the occurrences, we converted these data into a co-occurrence percentage table that provides the relative frequency of co-occurrence of each sense of the verb *sentir* (in the columns) with each ID tag level (in the rows). This procedure was performed with Gries's (2010b)

3 The consulted Spanish dictionaries are: the *Diccionario de la Lengua Española* (DRAE), the *Diccionario de Uso del Español* (DUE), the *Diccionario del Español Actual* (DEA) and the *Gran Diccionario de la Lengua Española* (GDLE) for the synchronic data. For French, the lexicographic study is based on *Le Nouveau Petit Robert: Dictionnaire alphabétique et analogique de la langue française* and for Italian the *Grande dizionario Italiano dell'uso*. The translation corpus (approx. 2,5 million words) contains source texts written in a non-Romance language and their translations in Spanish, French, and Italian. Ideally, all of the annotation could have been double-checked by additional annotators, a practice not yet very widespread in corpus linguistics.

BehavioralProfiles 1.01 script using the R statistical software package. As exemplified in Table 3, the percentages of ID tag levels add up to 1 within each ID tag so that each column represents a set of co-occurrence percentages for one sense of the verb. It is precisely these vector of co-occurrence percentages – i.e. 0.3, 0.35, 0.01, 0.34, 0.18, 0.82, ... for "experience: physical perception" – that are called "Behavioral Profiles".

Table 3: Examples of BP vectors.

ID tag	ID tag level	experience: physical perception	experience: emotional perception	auditory perception	consider, judge	...
tense	present	0.30	0.36	0.29	0.55	...
	past	0.35	0.40	0.53	0.30	...
	future	0.01	0.01	0.00	0.02	...
	infinitive	0.34	0.23	0.18	0.13	...
lexical S	with S	0.18	0.41	0.24	0.41	...
	without S	0.82	0.59	0.76	0.59	...
...	...	...	...	...	...	...

This BP method has proven useful for the analysis of different phenomena in lexical semantics such as near-synonymy (Divjak and Gries 2006; Divjak 2010), antonymy (Gries and Otani 2010) and polysemy (Gries 2006; Berez and Gries 2009; Jansegers et al. 2015) and has recently also been successfully applied to diachronic data (Jansegers and Gries to appear). However, we will make and exemplify two suggestions for how it can be extended. First, while most existing BP studies focus mainly on monolingual corpora, we will apply the BP approach to contrastive linguistic research questions in lexical semantics. Second, whereas most BP studies used hierarchical agglomerative cluster analysis (HAC) as their main exploratory tool, we will also pay special attention to other visualization techniques for cross-linguistic (dis)similarities.

1.3.3 Final preliminary comments

It should be mentioned that we are not particularly concerned with how this kind of annotation was arrived at. We understand that there is no tried and true mechanistic way of distinguishing between different senses of a polysemous lexeme in general and that any such sense discrimination will need to consider not only

the immediate linguistic context of the sentence it appears in, but possibly also the pragmatic context of use (Rozovskaya and Girju 2009). Much like the lexicographic work of sense identification, the annotation leading to BPs is usually an iterative process, where for instance, ID tags are modified, corrected or their number extended as different contexts of usage in the corpus come to light. In other words, we are not considering the question of sense identification/discrimination as theoretically or methodologically unambiguously resolved, just as tractable for practical purposes (again as in lexicographic work)[4] or there would be much fewer problems with lexical semantics and lexical relations within the context of machine learning, where different senses need to be extracted automatically, rather than manually coded (Romeo et al. 2013). In coding the ID tags for this work, we are first of all building on decades of traditional semantic, cognitive-linguistic, and psycholinguistic research attesting to the fact that it is possible to distinguish between senses and meanings of polysemous terms, and we adopted a pragmatic view that linguistically trained coders, who are also speakers of the languages at hand, would be able to disambiguate the senses through the perusal of the term's linguistic context. Secondly, our use of a concordance avoids looking at the term out of context or hand-picking terms occurring in a limited syntactic context or with predetermined senses. We accepted instead the full gamut of natural language usage and its complexity, as found in the corpora we used, and we also allowed for senses being annotated as *ambiguous/unidentifiable*. Thirdly, the methods outlined in this study can actually help in perfecting some of these fine-grained distinctions in meaning, which is what BPs were originally developed for (Gries 2006). Finally, results can of course be made even more robust and replicable by implementing any method requiring inter-annotator agreement.

2 Polysemy and senses' differences/distinctiveness

Assuming one has data of the above kind, the question of the most important and common senses there are across languages and which ones are language-

4 We realize how much this sounds like a cop-out, but such situations abound in linguistics in many domains other than semantic as well; after all, it is not like scholars would agree on the syntactic analysis of constructions, the morphological status of affixes, the status of certain morphemes or words in child language acquisition data, etc. In all these disciplines, researchers adopt solutions that are not perfect but feasible enough for certain analytical or practical purposes.

specific can sometimes be relatively straightforward to answer: The simplest way is cross-tabulation of the annotated concordance data and visualization (which could be followed by significance testing (χ^2 or G^2), if one's data meet the assumption of independence of data points).

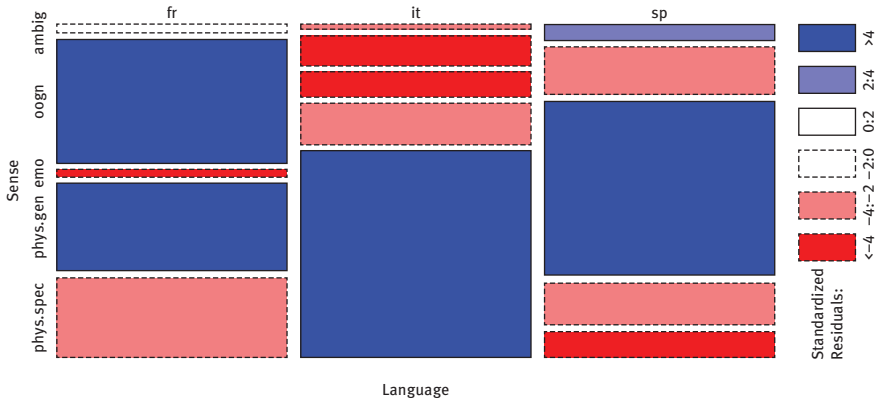
This shows minimally that the three languages differ significantly with regard to which senses *SENTIR(E)* expresses (in the coarse resolution of just five senses): In French, the cognitive and the physical.general senses are more frequent than expected, in Italian, the physical.specific sense is, and in Spanish the emotional one is. Also, in French, the emotional sense is very rare. Obviously, this can be done with more fine-grained sense classifications: Adopting the more fine-grained classification discussed above, Cramer's *V* increases to 0.69.

In addition, data such as Table 4 also permit us to compare how similar the different languages are in their sense frequencies. One way to do so would involve a hierarchical agglomerative cluster analysis (see Gries 2013: Section 5.6, Moisl 2015), where the language are clustered on the basis of how similar the senses' frequencies are to each other; the result of such an analysis (based on Euclidean distances and the "complete" amalgamation method) is shown in the left panel of Figure 2. Another way to do so would be a correspondence analysis (see Glynn 2010, Desagulier 2017: Section 10.4–10.5).

Table 4: Cross-tabulation ($G^2 = 852.3$, $df = 8$, $p < 10^{-100}$, Cramer's $V = 0.54$).

	French		Italian		Spanish	Totals
ambiguous	13	(2.6%)	7	(1.4%)	26	(5.2%)	46
cognitive	203	(40.6%)	49	(9.8%)	77	(15.4%)	329
emotional	11	(2.2%)	40	(8%)	288	(57.6%)	339
physical.general	144	(28.8%)	68	(13.6%)	69	(13.8%)	281
physical.specific	129	(25.8%)	336	(67.2%)	40	(8%)	505
Totals	500		500		500		1,500

The left panel shows that the sense frequencies in French and Italian are much more similar to each other than they are to Spanish. The right panel shows that, too: The three languages are clearly separated along the *x*-axis, with French and Italian being close together and far apart from Spanish; moreover, French and Italian are associated more with cognitive and physical senses, whereas Spanish is more closely associated with the emotional sense; notice also how, in a nicely intuitive way, the ambiguous uses occupy a central position in the plot. In this case, both powerful tools do not offer much beyond the simpler analyses of



AU: Please mention Figure 1 in the text.

Figure 1: Mosaic plot of Table 4.

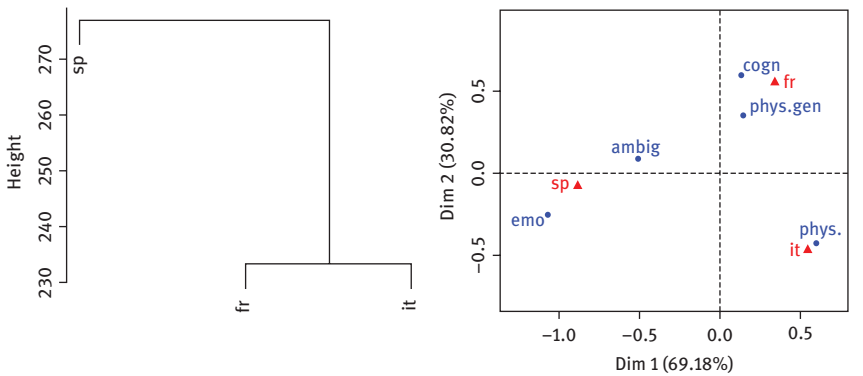


Figure 2: Further analytical plots of Table 4 (left: hierarchical cluster analysis, right: correspondence analysis).

the mosaic plot above, but that is why these simple data exemplify the kinds of attainable outcomes well. With more complex multivariate data, cluster and correspondence analysis have of course more to offer.

Different analytical possibilities arise when we change the resolution, which we can do in two ways. First, we can switch from the annotated concordance data to the BP vectors; second, we can create a new variable that combines – for each line – its language (i.e. French, Italian, and Spanish) with its sense (in either a coarse or a fine-grained resolution). This can be used to determine which (groups of) senses behave alike across (which) languages. Let us first briefly discuss the result of a cluster analysis of the combination of languages with coarse-grained

senses based on the BP vectors. Before we show the results, it is instructive to consider the range of results one might get:

- one theoretical extreme is that the dendrogram would group together all senses (i.e. their BP vectors) within each language, therefore, we would get three clusters (essentially as in the left panel of Figure 2);
- another theoretical extreme is that the dendrogram would group together all senses (i.e. their BP vectors) across languages, therefore, we would get four clusters (because we are leaving out the ambiguous cases now);
- a complete mess, either because there is no discernible structure in the data or there is, but it makes no sense either way.

The actual results are now shown in Figure 3 below and they are remarkably clear.

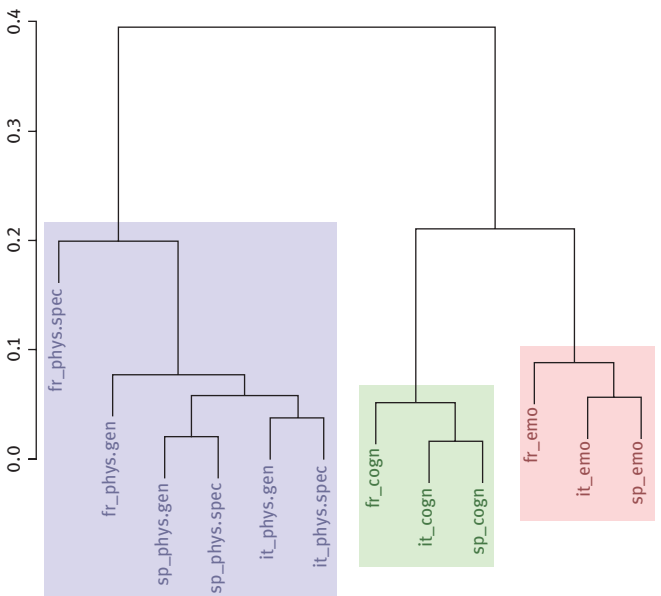


Figure 3: Dendrogram from a hierarchical cluster analysis of combinations of languages and coarse senses.

It makes sense to recognize three clusters as highlighted (based on average silhouette widths), and they are of the second theoretical kind: the obtained clusters point to the fact that senses pattern together *across languages* rather than patterning together *within the same language*. The pink cluster indicates that the emotional uses behave more similarly across the three languages than within each one of them, as it contains all and only all emotional senses; another cluster

contains all cognitive senses across the three languages, and the biggest one contains all physical senses with one "outlier sense" (French physical.specific). Follow-up analyses of such dendrograms (along the lines of Divjak and Gries 2006) can then help determine which of the annotated ID tags and their levels drive this particular clustering outcome.

A similar analysis of the fine-grained senses returns many more and more diverse clusters, but it still offers a result that groups senses together more than languages. For instance, all and only all emotional.experience senses are together in one cluster, as are all consider.judge senses, whereas the cognitive.realize, cognitive.think, and cognitive.intuit senses are also together in the same cluster.

Obviously, alternative cluster-analytical approaches are conceivable – for instance, these data can also be analyzed with more cognitively plausible fuzzy clustering approaches, which allow for graded cluster memberships of the clustered elements, towards senses within one language only.[5] For example, Figure 4 is one possible visualization of a fuzzy clustering of the BP vectors of the fine-grained senses in Italian (with 4 desired clusters); this clustering is quite fuzzy (normalized Dunn coefficient = 0.25), but the membership values clearly support, among others, a fairly robust cognitive cluster (red, on the left), a fairly robust cluster of multiple physical.specific senses (green, foreground), and one of physical.general_experience (turquoise, in the center).

A final analytical example involves the use of network analysis as discussed in Ellis et al. (2013), where senses and their interrelations are plotted as nodes/vertices and connecting links/edges respectively in an undirected network graph. In the present case and just to exemplify the method, we built a network of the French senses observed with SENTIR, where

- vertices and their sizes represent fine-grained senses and their frequencies in the French data;
- edges and their thickness represent the similarity of the BP vectors of all pairs of senses whose similarity (Euclidean distance) was greater than the 40% quantile of all pairwise similarities (this was done to avoid having to plot even edges that reflect low degrees of sense similarity; the cut-off point of 40% is arbitrary and was chosen here on the basis of visual inspection);

5 We are considering these cognitive more plausible for the simple reason that they allow for graded category membership and prototypicality in a way that is extremely compatible with the kind of cognitive-linguistic or usage-based approach we are adopting here as well.

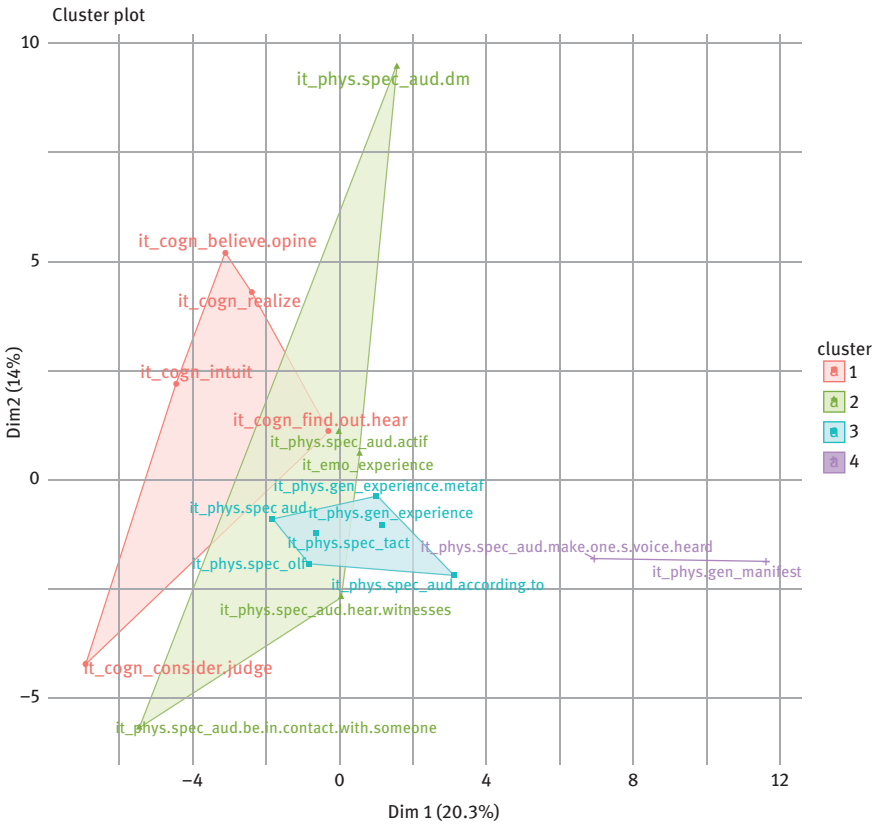


Figure 4: Dendrogram from a fuzzy cluster analysis of the Italian fine-grained senses.

– the vertices' colors represent the three "communities" of senses identified by a multi-level modularity optimization algorithm for finding community structure based in the pairwise similarities mentioned above.[6]

The network algorithm finds three communities whose elements are differently strongly related to each other and which are represented in Figure 5: (i) a red community consisting of all cognitive senses as well as emotional.experience, (ii) a green community consisting of all physical.general senses and one physical.

6 Modularity in graph theory is treated as a quality measure of the amount and "cleanliness" of a cluster structure in a network. Much like in cluster analysis, it refers to the notion of clusters in a network exhibiting (i) high internal connectivity/similarity but (ii) low connectivity/similarity to other clusters.

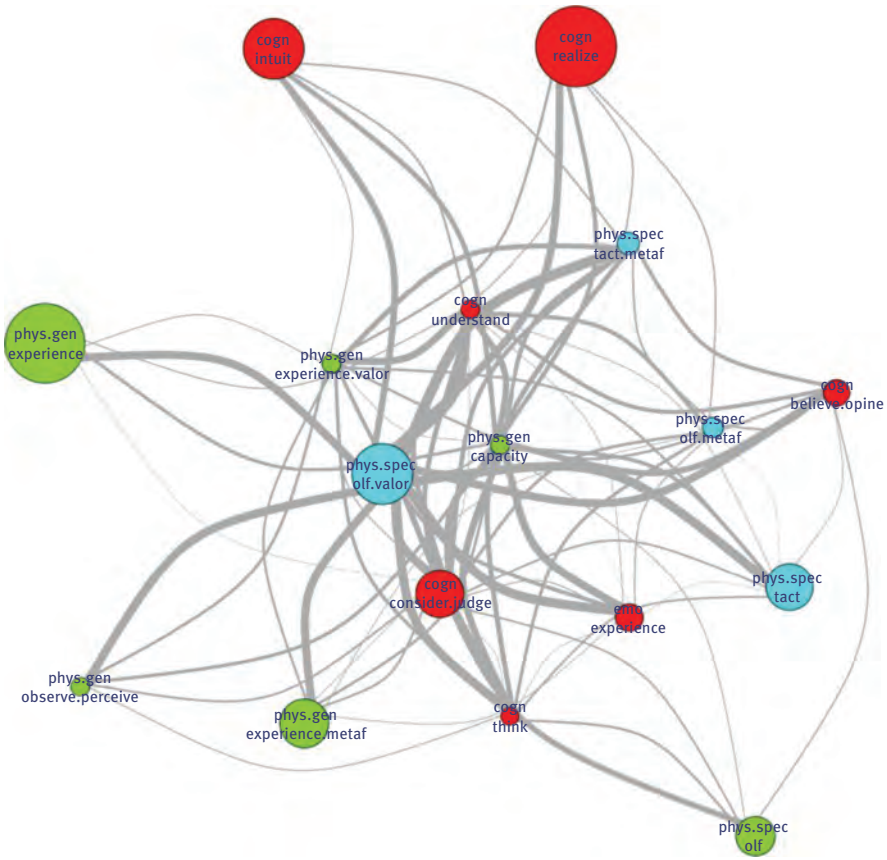


Figure 5: Semantic network analysis plot of the fine-grained French senses of *SENTIR*.

specific sense (bottom right), and (iii) a light blue community consisting of all remaining physical.specific senses. In this case, the result is quite clear and the bottom-up and multivariate method lends strong support to both the method per se, the sense annotation, and, most importantly for contrastive linguists from a cognitive perspective, a grouping of senses that is compatible with cognitive-linguistic theorizing, e.g. the clear distinctions of more mental (cognitive and emotional) senses on the one hand and more physical ones on the other.

We now turn to the questions of how to identify prototypical configurations of ID tag levels and prototypes as well as how to identify which ID tag levels are most discriminatory within and across languages.

3 Prototypicality, markedness, and identifying discriminatory/predictive variables

The question of identifying prototypical senses of *SENTIR(E)* is one that is best approached by, first determining a likely candidate for "the prototype" in each language (or, alternatively, a candidate set), and then compare those prototypes across languages. Gries (2006) discusses a variety of ways in which prototypes of the verb *to run* may be explored. Given the nature of the concept "prototype" itself, it is not surprising that there are few, if any, individual necessary and jointly sufficient conditions/diagnostics; there are, however, different ways to approach the issue.

One is obviously frequency, which is arguably at least somewhat related to prototypicality and can easily be obtained from the annotated data. According to this criterion the prototype for Italian *SENTIR(E)* would be physical.specific (specifically, from the fine-grained data, the sense physical.auditory), whereas for Spanish it would be emotional.experience (especially including the grammaticalization of the construction into the fixed form *lo siento*).[7] For French, however, we immediately recognize the problem of granularity: Table 4 suggests that the cognitive sense (of cognitive.realize) is most frequent, but it is also obvious that the two physical senses *together* would outnumber this one. This might be a case of a radial category with multiple centroids (in the same way that, to use a well-known example, the word *game* might have multiple local prototypes, e.g. one for games involving sports-like physical activity like "playing catch", one for games involving no physical aspects but mental acuity such as card games like Poker, ...).

7 The connection between the Spanish emotional meaning of *sentir* and its prototypicality is due to the historical evolution of this sense in the language. While "being affected by something already exists in the Latin meaning of the verb" (Verbeke 2011: 21), Verbeke also shows that the verb in Spanish evolves from denoting physical sensation ('to feel cold'), extends to feeling emotions ('to feel joy, anger, sadness'), among which both dictionaries (for instance Covarrubias in 1611) and corpora start numbering a few examples of feelings of regret or dissatisfaction already in the 15th century. By the 18th century the *Diccionario de Autoridades* marks one of the senses of *sentir* as 'to feel anguish or sorrow' (p. 23) and soon competes with *lamentar* 'to regret'. *Sentir* has become the go-to generic perception verb by the 20th century with 83% of modern uses, among them many with negative and emotional perceptions and 9% more exclusively in the 'to regret' sense according to Verbeke (2011: 47). In this last sense, the subjectivized verb has also undergone grammaticalization along a morphosyntactic cline producing many instances of the fixed form *lo siento* in the 20th century corpus. Its literal meaning is 'I regret it', nowadays used as an interjection with the simple meaning of 'sorry!'. As such it significantly increases the frequency of the emotional sense of *sentir* in modern texts, and contributes to the prototypicality of the emotional sense in Spanish.

Another way of approaching prototypicality is based on the notion of cue validity. Much research on prototypes now argues that prototypes are an abstract entity combining the properties with the highest cue validity for the category in question, where cue validity is essentially the conditional probability *p*(category membership | property); for instance the property 'having feathers' has a high cue validity for the category BIRD (because most birds have feathers and most non-birds don't) whereas 'having eyes' does not have a high cue validity for BIRD because while most if not all birds have eyes, most other animals do too.

This simple definition is instructive in how it points to the possibility of exploring prototypicality on the basis of classifiers and similar techniques such as regression models, (linear/quadratic) discriminant analysis, classification trees or random forests, and many others. This is because these techniques can all do two things: they can identify which ID tags and levels have the highest degree of predictive power per sense (per language) and they can compute a predicted probability of a sense (per language) for each case in the data. How does that relate to prototypicality? It does along the lines argued first by Gries (2003a, 2003b), who used the probabilities with which a binary constructional choice was predicted to identify the most prototypical instances in the data: the highest predicted probabilities for a constructional choice reflected that these instances combined many features that raised the (conditional) probability for that constructional choice, making them (close to) prototypical. The same logic can be applied here: one can run a classifier on either the senses (across all languages) or the language-sense combinations and then, if the classifier does a good job,

- use measures of variable importance to determine which predictors are most important for predictions;
- determine for each level of the dependent variable, which cases yield the highest and correctly predicted probabilities to abstract away to a prototype.

To briefly exemplify this kind of analysis, we used random forests to try and predict from all annotated ID tags and their levels a variable that consisted of the language and the coarse-grained senses; in other words, the dependent variable had levels such as "fr_emo", "fr_cog", "fr_phys.gen", etc. Random forests are an extension of simple classification (and regression) trees. Classification (and regression) trees are a partitioning approach that consists of successively splitting the data into two groups based on predictors (here ID tags) such that the split maximizes the classification accuracy regarding the dependent variable. This process is recursive, i.e. repeated until no further split would increase the classification accuracy enough anymore. Random forests in turn add two layers of randomness to the analysis, which help (i) recognizing the impact of variables or their combinations that a normal classification tree might not register and (ii) protecting against overfitting.

On the one hand, the algorithm constructs many different trees (we used 500), each of which is fitted to a different bootstrapped sample of the full data. On the other hand, each split in each tree could choose from only a randomly-chosen subset of predictors (we set that parameter to five predictors). The overall result is then based on amalgamating all 500 trees that have been generated by identifying the majority vote of the forest's predictions for each cases.[8]

The baseline for such a classifier is typically computed as the highest probability of any level of the dependent variable, which here is 0.224 (for the most frequent sense of ita_phys.spec). We then ran a random forest (using all default settings of the function party::cforest in R, see Hothorn et al. 2006) on the data and obtained a very good prediction accuracy of 0.656, i.e. nearly three times as good as, and significantly different from, the baseline. The most important ID tags (as determined by variable importance plots) for this excellent result were the semantic role and form of the subject as well as the referent and the form of the direct object; in fact, those four ID tags alone already yield a prediction accuracy of 0.648. We then finally looked at the combinations of ID tag levels for each language-sense combination that were most frequent, had the highest predicted probability, and were correctly predicted, which yielded, among others, the following prototypes:

– French cognitive: a pronominal experiencer SUBJ and a clausal DO referring to a situation/event, which is very similar to the Spanish cognitive: a non-lexical SUBJ (since Spanish is pro-drop, the subject is typically not expressed by a pronoun or a NP) with the same kind of DO. The Italian cognitive uses were hardly ever predicted correctly by the classifier.

  (5) *Le militantisme était devenu une contrainte. Je* ***sentais*** *que le monde était plus complexe que nos discours* (French, Le Monde, 1998).
  'Activism had become a constraint. I realized that the world was more complex than our speeches.'

– Italian physical.specific: a non-lexical (since Italian is also pro-drop, the subject is typically not expressed by a pronoun or a NP) perpt (perceptor, i.e. an entity that experiences *physical* perception, visual, auditory, tactile etc.) with a concrete-entity DO NP (6) or infinitive; the corresponding French sense has a stimulus NP as a subject and no DO. The corresponding Spanish sense was hardly ever predicted correctly.

8 Note that random forests do not require the same kind of training vs. test sampling procedure because the predictions that the algorithm returns are OOB (out-of-bag) predictions, i.e. predictions made not for the data points on which a tree was trained, but the ones held out.

(6) *Ho* ***sentito*** *un boato – racconta Aurora Falcone – è poi sono stata catapultata sulla strada.* [Italian, CdS, 2010]
'I heard an explosion – says Aurora Falcone – and then I was catapulted on the road.'

Thus, random forests or any other classifier that returns predicted probabilities can help identify both concrete examples in the data as well as abstract combinations of features with high cue validities that correspond to what in cognitive linguistic approaches are prototypes.

The next method to be briefly mentioned is that of association rules, a much more exploratory and extremely granular machine learning method that looks at potentially quite large data sets of categorical variables. This method is also applied to the annotated concordance data. Association rules are essentially just conditional sentences, consisting of

- an *if*-clause or antecedent, which can contain more than one condition (up to a user-defined number, we used 4); in association-rules terminology, this is referred to as "the left-hand side" (LHS);
- a main clause or consequent, which contains one resultant condition; in association-rules terminology, this is referred to as "the right-hand side" (RHS).

An example of a rule in the present context (using the coarse-grained senses) would be "if Language = "French" and if FormOfDO = "clause" (LHS), then Sense = "cognitive (RHS)". If an analyst wishes to apply this method to a data set (such as the 1,500 concordance lines times 26 ID tag columns of the present data), (s)he usually specifies three parameters that serve to put a cap on the number of such rules that are generated:

- a parameter called *support*: the proportion of data points that contains all conditions/items in the rule (i.e. both LHS and RHS). In the above case, the 1,500 data points contain 106 cases of French uses with the sense "cognitive" where the DO is a clause, i.e. support = 106/1,500 ≈ 0.071. Support is used to state the minimum number of cases to which a rule must apply for it to be returned;
- a parameter called *confidence*: the proportion of times the rule is correct. In the above case, there are 16 additional cases of French cases with a clausal DO that do *not* come with the sense "cognitive", which means the rule is right 106/122 ≈ 0.869 of the time;
- a parameter called *maxlen*, which specifies the number of elements in the rule or, since the length of the RHS is set to 1, the number of conditions usable in the LHS. In the above example, the length of the rule is of course 3.

We applied this approach to our data (with min. support = 0.05, min. confidence = 0.6, maxlen = 5) and obtained approximately 1,5 million rules. However, to see which senses are most different between languages, this number was then reduced to only those rules that featured Language in the LHS and Sense in the RHS, which returned 10,3K rules. Obviously, these cannot all be studied so analysts have a wide range of options to narrow down which rules to study. These options include

- specific statistics that quantify the "noteworthiness" of each rule (examples include statistics such as *lift*, *hyper-lift*, *hyper-confidence*, and just about any other association measures that can be applied to 2×2 tables, see Hahsler and Hornik 2007; Hahsler et al. 2008). Lift is a measure reflecting how much observed co-occurrence differs from expected co-occurrence; hyper-lift is a more robust variant of that statistic.
- common-sense and phenomenon-specific considerations such as the diverging syntax-semantics interfaces across languages, here, being particularly interested in rules, whose LHS differ only by language and whose RHS differ only by sense (which means that they predict different senses).

Figure 6 shows two plots that would help analysts analyze the data. Both panels plot all 10.3K association rules on the basis of their support (*x*-axis) and their hyper-lift (*y*-axis), with the point size indicating the confidence. The left panel uses RGB coloring to indicate the language to which the rule applies and it is immediately obvious that the rules for Italian are characterized by much less hyper-lift than those of the other languages; the median for Spanish is highest, followed by French, followed by the much smaller Italian. The right panel uses RGB coloring to indicate the coarse-grained sense which the rule involves, and here it is clear that the cognitive sense is characterized by the highest hyper-lift, compared to lower values for emotional, followed by much lower values for the two physical senses.

Sorting all rules by their LHSs and/or by the number of rules in which a certain LHS is embedded is a path towards a more detailed analysis. For instance, we find the following kinds of differences between the languages:

- SEMANTIC_ROLE_S=perpt is correlated with physical.general senses in French and Spanish, but with physical.specific in Italian (see (6) above, esp. when no other adjuncts and complements are present or when the subject is animate/human):

  (7) *Mariana se convirtió en una muchacha de aspecto lánguido, con la sonrisa triste de las personas que padecen sin* ***sentir*** *dolor en el cuerpo* (Spanish, CREA, 1996).
  'Mariana became a languid-looking girl, with the sad smile of those people who suffer without feeling pain in their body.'

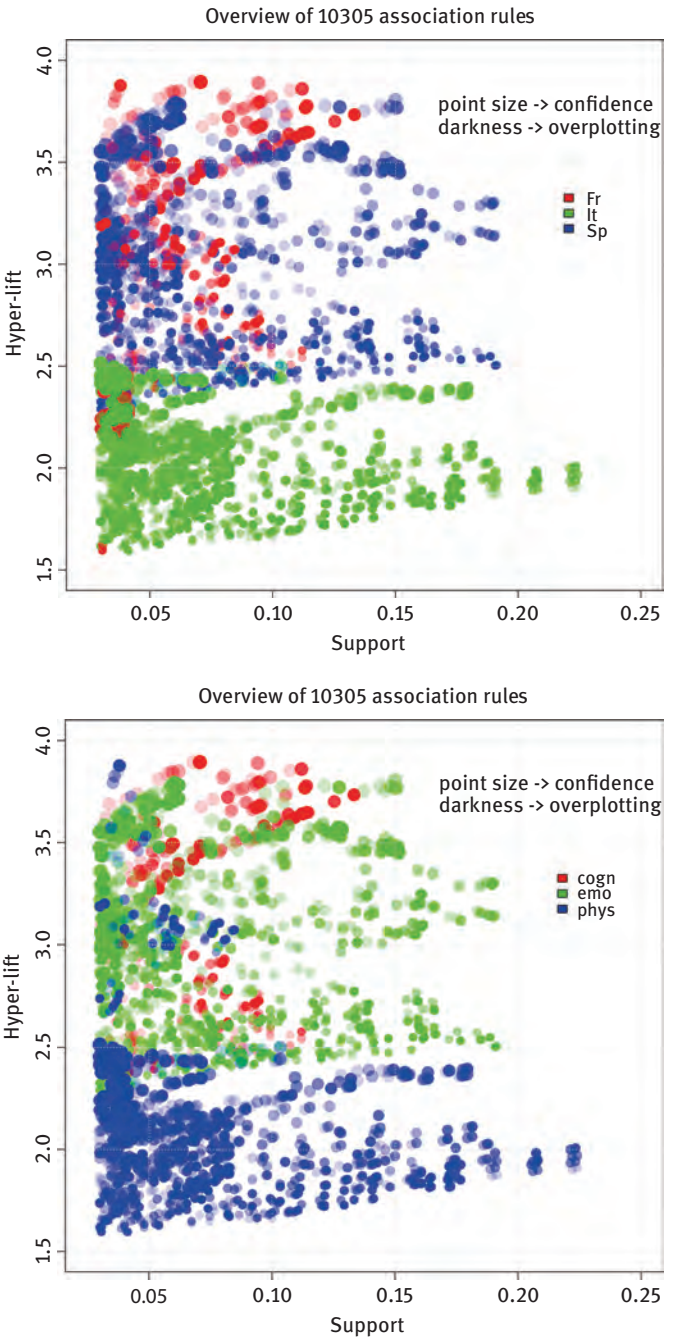


Figure 6: Overview of association rules results.

– infinitive DOs are correlated with physical.general senses in French (see (8)), but with physical.specific in Italian (see (9)):

(8) *Je commençais à suer, à **sentir** sourdre la sueur sous mes aisselles* (French, Frantext, 2006).
'I began to sweat, to feel the sweat well up under my armpits.'

(9) *Come **sentire** squillare un cellulare in sala o vedere un abbigliamento non consono al teatro. Il messaggio è lanciato.* (Italian, CdS, 2017)
'Like hearing a cell phone ring at the cinema or see inappropriate attire at the theater. The message has been sent.'

– BASIC_AS=abs (esp. with no additional adjuncts or complements) are correlated with physical.specific senses in French and Italian, but not in Spanish; etc. These are cases of the absolute use of the verb, without explicit DO. For example, French *sentir* often appears in a copulative construction, expressing a certain valorization of the olfactory process:

(10) *Il n'aimait pas son odeur, ça **sent** le poisson pourri, il ne pouvait pas le faire*. (French, Frantext, 2006)
'He did not like her smell; [lit.] it smells like rotten fish, he could not do it.'

While the technique is highly exploratory, it can help reveal much probabilistic structure in the data, and interactive visualization tools (see Hahsler 2017), which cannot be shown in a printed paper, can serve to highlight patterns in the data that would otherwise remain invisible to the naked eye just studying concordance lines.

Moving on to the BP vectors, another criterion can be derived from markedness considerations, leading to the assumption that the prototypical sense should be (among) the formally least constrained senses. For BP data, this criterion could lead to the question of which senses have the smallest numbers of zeros in their BP vectors, i.e. which senses are attested with the largest variety of ID tags. For the present data, this leads to

– for French: cognitive.consider/judge and physical.general_experience;
– for Italian: physical.specific_auditory and physical.general_experience;
– for Spanish emotional.experience and physical.general_experience.

In other words and maybe unsurprisingly given *SENTIR(E)*'s "general meaning", physical.general_experience is always part of the least restrained senses, but

then the languages differ in terms of the other least restrained sense. Virtually the same results are obtained from using a more advanced approach, namely by computing, for each language separately, how much the ID tag level percentages with a specific sense differ from the same ID tag level percentages with *all* senses with the Kullback-Leibler divergence (see Cover and Thomas 2006: 19–20), a directional measure that quantifies how much one probability distribution differs from another. Then one adds up how much each sense's ID tag level distribution is different from those of all senses because the least marked sense(s) should exhibit the smallest difference(s). We obtain the same results as with the simpler approach – the only difference is that this approach returns cognitive.intuit for French rather than cognitive.consider/judge; everything else stays the same. This leads to two interesting findings: First, all three proposed criteria largely converge in each language, which is reassuring. Second, that in turn makes it less straightforward to want to postulate any prototype more specific than physical.general_experience, since all three languages share that meaning component, but not the other.[9]

Finally, possibly the simplest analysis using BP vectors that is still insightful is to determine what the differences are, if any, within a sense (e.g. of physical.general_experience) between the languages by computing pairwise differences between the BP vectors of – say – French and Italian (because they are in one cluster in Figure 2) or of Italian and Spanish (to see what might be behind their big difference in Figure 2). These comparisons show that the differences between French and Italian are mostly form-related: most larger ID tag differences involve morphosyntactic ID tags – the main semantic differences are that the French DOs of SENTIR(E) are much more often situation/events (see (7) above) and much less often have no DO than in Italian, and that the semantic role of the SUBJ is much more often a perceptor (perpt, i.e. an entity that experiences *physical* perception, visual, auditory, tactile etc.) than an experiencer (exp, i.e. an entity that experiences *mental* perception). The differences between Italian and Spanish are various and both morphosyntactic and semantic in nature. With regard to the former, Italian has many more 1st person uses and many fewer 3rd person uses than Spanish; Italian has many more cases without a DO referent or with a concrete one (see (11)), but many fewer abstract-entity DOs and situations/events than Spanish, for instance. Such results, as visualized in Figure 7, can be pointers for subsequent study.

9 See Lester (2018) for a similar approach using the Kullback-Leibler divergence to explore prototypicality.

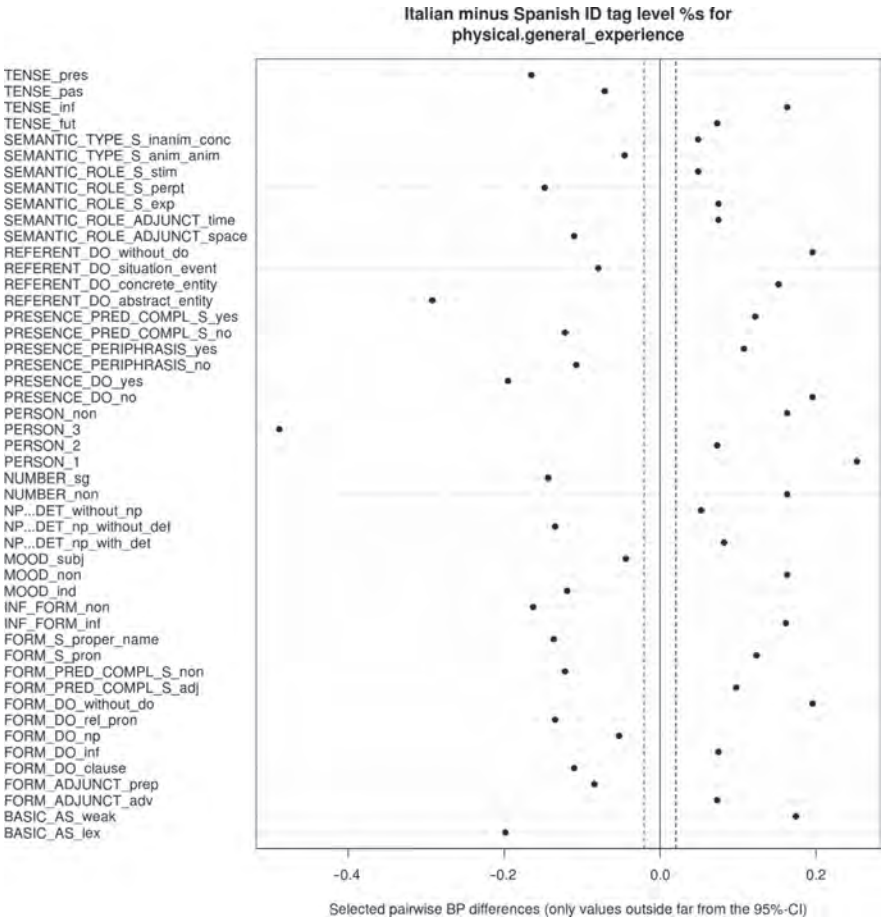


Figure 7: Differences between Italian and Spanish: positive and negative values reflect over- and underrepresentation in Italian (relative to Spanish) respectively.

(11) *Ora **sento** l'adrenalina, sono in grado di fare cose che il 90 per cento delle persone non si sogna nemmeno.* (Italian, CdS, 2010)
'Now I feel the adrenaline, I can do things that 90% of people do not even dream of.'

4 Discussion and concluding remarks

4.1 Interim summary

In the previous sections, we discussed how a variety of research questions common in contrastive linguistics can be studied on the basis of corpus data and advanced statistical techniques. Specifically, we focused on four (highly) interrelated challenging notions for the study of cross-linguistic near-synonymy, namely: polysemy and degree of sense distinctiveness (Section 2) and prototypicality and identification of discriminatory variables (Section 3). Each of these phenomena was tackled by means of a variety of statistical analyses based on two different kinds of input data that offer different kinds of resolutions on the data: (i) annotated concordance data and (ii) BP vectors.

First, the simplest way to determine the most important and common senses across and within languages, is cross-tabulation of the annotated concordance data. This cross-tabulation can then be visualized in a variety of ways such as a mosaic plot, hierarchical cluster analysis and correspondence analysis. These visualizations, in turn, allow for comparing the precise extent to which the languages differ in their senses' frequencies. Although the analysis based on the raw data shows what the most important senses across and within languages are, it is not the most appropriate way to tackle the question of sense distinctiveness. That is, the question of how many different senses there are in each language and how these senses relate to each other can be addressed better by shifting the resolution from the annotated data to the BP vectors. For example, one might want to do a kind of broad clustering covering all languages and all possible senses in order to determine which senses behave alike across which languages. However, a significant downside of a hierarchical agglomerative cluster analysis is that it implies forced binary splits of the data, which will typically not be a cognitively realistic representation of a phenomenon like near-synonymy. Therefore, in order to visualize the senses and their interrelations in a more faithful way, alternative cluster-analytical approaches can be used such as fuzzy clustering and network analysis.

The identification of the prototypical sense can be done on the basis of a variety of criteria. Three ways of approaching prototypicality on the basis of the annotated concordance data were discussed here: (i) frequency, (ii) cue validity, and (iii) association rules. The notion of cue validity allows exploring prototypicality on the basis of classifiers such as random forests. Changing again the resolution from the raw, annotated data to the BP vectors adds other analytical possibilities. Another way of handling prototypicality is based on markedness

considerations: More prototypical elements are taken to be less formally constrained and thus could appear in a wider variety of (formally and/or lexically defined) contexts. Using BP vectors, this can be done by taking into account the smallest numbers of zeros in the BP vectors or by computing how much the ID tag level percentages with a specific sense differ from the same ID tag level percentages with all senses. The latter can be done with the Kullback-Leibler divergence. Finally, we also illustrated how one can compute pairwise differences between the BP vectors in order to determine and visualize which variables are responsible for the differences between the same senses in different languages.

4.2 Implications

The above overview of different advanced methods and statistical techniques for addressing contrastive linguistic research questions leads to several substantial implications for contrastive linguistic and lexical semantic studies. Although recent research has gradually moved away from comparing individual dictionary definitions or using a philological approach to text analysis, towards the use of corpora with examples from actual natural language use, the methodology applied has by and large impeded substantial advances in the field. The problems this chapter aimed at raising, if not solving, are essentially of a dual nature: one aspect of the question concerns the more basic nature of data visualization, and the other, a more theoretical one, concerns the impossibility of finding the existing structure in larger data sets without the help of different statistical approaches. Without them it is impossible to highlight the connections between forms and functions, between different senses of the same word, diverging evolutions of the same etymon in sister languages, or different translations of a term in parallel corpora.

Visualization is a fundamental tool for exploratory analyses, and yet even accurate and detailed analyses in the contrastive linguistic and lexical typology tradition, often use tables comparing raw data or percentages to describe the frequencies of senses of a lexeme or near-synonymous terms (verbs of emotion, mental verbs etc.), or side-by-side comparison of constructions used in the parallel corpora of translated texts (see for instance Viberg's (2008) contrastive analysis of Swedish verbs of perception as an example: p. 129, 132 for tables, and 131, 133 for side-by-side comparisons). Side-by-side examples may be useful to elucidate members of a specific category, but tables of raw data or percentages can never allow the analyst (or the reader) to construct a mental overview of the results: we simply cannot analyze large amounts of data without statistics and we miss the generalizations obtained by graphing them in colors and patterns that highlight the most relevant variables causing some specific distribution.

The more theoretical point is related to the need for improved data analysis. The large majority of studies in contrastive linguistics are mainly based on observed (relative) frequencies of (translation) data and are essentially monofactorial in nature. However, most linguistic problems are intrinsically multifactorial, as is the case of near-synonymy between sister languages analyzed in this chapter. We have shown that different statistical analyses can provide more or less granularity (the sense frequencies from the mosaic plot in Figure 1 *vs*. BP vectors or cluster analysis based on BP vectors in Figure 3 or fuzzy clustering in Figure 4).

More specifically, this chapter also presents some improvements with regard to previous applications of the BP approach. First, while most existing BP studies mainly focus on monolingual corpora (Divjak and Gries 2009 being an exception), our study presents an application of the approach to contrastive linguistic data. Second, whereas most BP studies use hierarchical agglomerative cluster analysis as their main exploratory tool, we paid special attention to other visualization techniques for cross-linguistic (dis)similarities such as network analysis and fuzzy clustering. While it is true that Behavioral Profiles require a lot of largely manual annotation and are still exploratory in nature, what we gain is a very high level of analytical detail, which allows for a wide range of exploratory possibilities of the data. It not only facilitates comparability within and across languages, but also allows comparing specific senses within and across languages both in general and with regard to their structural manifestations.

The use of advanced statistical techniques also has implications on a more qualitative, theoretical level. For example, the application of methods such as network analysis and fuzzy clustering offers usage-based evidence for cognitive linguistic theorizing concerning polysemous networks: As mentioned above, HAC results can arguably overemphasize discreteness and mutual exclusivity of (elements within) meaning clusters, whereas the use of fuzzy clustering exemplified here allows for a clearer identification of graded cluster memberships of the clustered elements in the semantic space both within and across languages. From a cross-linguistic perspective, then, the present paper offers some powerful tools for the analysis and visualization of cross-linguistic (dis)similarities. As illustrated by the big cluster analysis in Figure 3, the BP method allows for comparing multiple languages not on the basis of their mere senses' frequencies, but on the basis of a very fine-grained annotation that includes semantic, morphological, syntactic, and other characteristics shared across languages, thus uncovering the source of the cross-linguistic (dis)similarity.

Finally, the proposed analyses also highlight features in the data that would otherwise remain concealed, such as language-specific structural reflexes of grammaticalized/ constructionalized senses (see e.g. Hilpert 2013; Traugott and

Trousale 2013). A clear example is the extreme position of the "it_phys.spec_aud.dm" sense (graphed in purple) in the Italian data visualized in the dendrogram from the fuzzy cluster analysis of the Italian fine-grained senses (Figure 4). This sense underlines the different behavior of a discourse marker derived from the verb *sentire* in Italian, which does not exist in the corresponding verbs of its sister languages French or Spanish. In other words, the proposed methods in this paper are an excellent way to display diverging grammaticalization/constructionalization patterns in cognate languages, confirming Viberg's (1999) conclusion that grammaticalization can drive cognates apart semantically.

4.3 Where to go from here

Considering the possibilities for further analysis mentioned at the end of Section 3 above, it would be interesting to compute pairwise differences between the BP vectors of French and Italian to see why they are in one cluster in Figure 2, or those of Italian and Spanish to see what might cause the big difference between the two languages in the same figure. These comparisons show that the differences between French and Italian are mostly form-related, whereas there are several morphosyntactic and semantic differences separating Italian and Spanish. A detailed study looking at these differences could uncover their causes, and supply a more thorough linguistic analysis of the data that this chapter, because of its methodological nature, did not provide.

The techniques suggested in this chapter have wide potential applications both for lexical semantic analyses within and across languages, and potentially also for the diachronic evolution of the senses of polysemous terms, possibly revealing phenomena of subjectification and grammaticalization. In this sense, further work on new or previously published data applying these statistical methods is bound to uncover extremely interesting tendencies and generalizations.

References

Altenberg, Bengt. 2002. Causative constructions in English and Swedish A corpus-based contrastive study. In Bengt Altenberg & Sylviane Granger (eds.), Recent trends in cross-linguistic lexical studies, 97–116. Amsterdam/Philadelphia: John Benjamins.

Altenberg, Bengt & Sylviane Granger. 2002. Recent trends in cross-linguistic lexical studies. In Bengt Altenberg & Sylviane Granger (eds.), *Lexis in contrast: corpus-based approaches*, 3–48. Amsterdam & Philadelphia: John Benjamins.

Atkins, Beryl. T. Sue. 1987. Semantic ID tags: Corpus evidence for dictionary senses. In *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary*, 17–36.

Berez, Andrea L. & Stefan Th. Gries. 2009. In defense of corpus-based methods: a behavioral profile analysis of polysemous *get* in English. In Steven Moran, Darren S. Tanner, & Michael Scanlon (eds.), *Proceedings of the 24th Northwest Linguistics Conference*, 157–166. Seattle, WA: Department of Linguistics.

Comer, Marie & Renata Enghels. 2016. La polisemia de los verbos de colocación. Descripción sincrónica y evolución diacrónica de los cuasi-sinónimos *poner/meter* y *poser/mettre*. *Revue Romane* 51 (1). 70–94.

Cover, Thomas M. & Joy A. Thomas. 2006. *Elements of information theory*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Desagulier, Guillaume. 2017. *Corpus linguistics and statistics with R: introduction to quantitative methods in linguistics*. Berlin & New York: Springer.

Divjak, Dagmar. 2006. Ways of intending: delinieating and structuring near synonyms. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, 19–56. Berlin & New York: Mouton de Gruyter.

Divjak, Dagmar. 2010. *Structuring the lexicon. A clustered model for near-synonymy*. Berlin & New York: Mouton de Gruyter.

Divjak, Dagmar & Stefan Th. Gries. 2006. Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2 (1). 23–60.

Divjak, Dagmar & Stefan Th. Gries. 2009. Corpus-based cognitive semantics: A contrastive study of phasal verbs in English and Russian. In Katarzyna Dziwirek & Barbara Lewandowska-Tomaszczyk (eds.), *Studies in cognitive corpus linguistics*, 273–296. Frankfurt am Main: Peter Lang.

Ellis, Nick C., Matthew B. O'Donnell & Ute Römer. 2013. Usage-based language: Investigating the latent structures that underpin Acquisition. *Language Learning* 63 (Suppl 1). 25–51.

Enghels, Renata & Marlies Jansegers. 2013. On the cross-linguistic equivalence of *sentir(e)* in Romance languages: a contrastive study in semantics. *Linguistics* 51 (5). 957–991.

Fox, John & Sanford Weisberg 2011. *An R companion to applied regression*. 2nd ed. Thousand Oaks, CA & London: Sage.

Gast, Volker. 2015. On the use of translation corpora in contrastive linguistics. A case study of impersonalization in English and German. *Languages in Contrast* 15 (1). 4–33.

Glynn, Dylan. 2010. Correspondence analysis: exploring data and identifying patterns. In Dylan Glynn & Kerstin Fischer (eds.), *Corpus methods for semantics: quantitative methods in polysemy and synonymy*, 443–485. Amsterdam & Philadelphia: John Benjamins.

Gries, Stefan Th. 2003a. *Multifactorial analysis in corpus linguistics: a study of Particle Placement*. London & New York: Continuum Press.

Gries, Stefan Th. 2003b. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1. 1–27.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, 57–99. Berlin & New York: Mouton de Gruyter.

Gries, Stefan Th. 2010a. Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon* 5 (3). 323–346.

Gries, Stefan Th. 2010b. BehavioralProfiles 1.01. A program for R 2.7.1 and higher.

Gries, Stefan Th. 2013. *Statistics for linguistics with R: a practical introduction*. Berlin & Boston: De Gruyter Mouton.
Gries, Stefan Th. & Naoki Otani. 2010. Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34. 121–150.
Hahsler, Michael. 2017. arulesViz: Visualizing association rules with R. *R Journal* 9 (2). 163–175.
Hahsler, Michael, Christian Buchta, & Kurt Hornik. 2008. Selective association rule generation. *Computational Statistics* 23 (2). 303–315.
Hahsler, Michael & Kurt Hornik. 2007. New probabilistic interest measures for association rules. *Intelligent Data Analysis* 11 (5). 437–455.
Hilpert, Martin. 2013. *Constructional change in English: Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press.
Hothorn, Torsten, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro, & Mark Van Der Laan. 2006. Survival Ensembles. *Biostatistics* 7 (3). 355–373.
Jansegers, Marlies, Clara Vanderschueren, & Renata Enghels. 2015. The polysemy of the Spanish verb *sentir*: a Behavioral Profile analysis *Cognitive Linguistics* 26 (3). 381–421.
Jansegers, Marlies & Stefan Th. Gries (to appear). Towards a dynamic behavioral profile: a diachronic study of polysemous *sentir* in Spanish. *Corpus Linguistics and Linguistic Theory*.
Lansari, Laure. 2017. *I was going to say / j'allais dire* as discourse markers in contemporary English and French. *Languages in Contrast* 17 (2). 205–228.
Lester, Nicholas A. 2018. The syntactic bits of nouns: How prior syntactic distributions affect comprehension, production, and acquisition. Unpublished Ph.D. dissertation, University of California, Santa Barbara.
Levshina, Natalia. 2016. Verbs of letting in Germanic and Romance languages. A quantitative investigation based on a parallel corpus of film subtitles. *Languages in Contrast* 16 (1). 84–117.
Maindonald, John & W. John Braun. 2010. *Data analysis and graphics using R: an example-based approach*. 3rd ed. Cambridge: Cambridge University Press.
Molino, Alessandra. 2017. A contrastive analysis of reporting clauses in comparable and translated academic texts in English and Italian. *Languages in Contrast* 17 (1). 18–42.
Moisl, Hermann. 2015. *Cluster analysis for corpus linguistics*. Berlin, Munich, & Boston: Mouton de Gruyter.
Romeo, Lauren, Sara Mendes & Núria Bel. 2013. Using qualia information to identify lexical semantic classes in an unsupervised clustering task. *Proceedings of COLING 2012: Posters*, 1029–1038.
Rozovskaya, Alla & Roxana Girju. 2009. Identifying semantic relations in context: near-misses and overlaps. *International Conference RANLP 2009 Recent Advances in Natural Language Processing, Borovets, Bulgaria*, 381–387.
Rozumko, Agata. 2016. Adverbs of certainty in a cross-linguistic and cross-cultural perspective English-Polish. *Languages in Contrast* 16(2). 239–263.
Schmied, Josef. 2008. Contrastive Corpus Studies. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An international handbook*, 1140–1159. Berlin & New York: Mouton de Gruyter.
Traugott, Elizabeth C. & Graeme Trousdale. 2013. *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.
Verbeke, Charlotte. 2011. *Sentir: ¿un verbo de percepción o un verbo de emoción?* Ghent: Ghent University MA thesis.
Viberg, Åke. 1999. The polysemous cognates Swedish *gå* and English *go*: universal and language-specific characteristics. *Languages in Contrast* 2 (1). 87–113.

Viberg, Åke. 2002. Polysemy and disambiguation cues across languages. The case of Swedish *få* and English *get*. In Bengt Altenberg and Sylviane Granger (eds.), *Lexis in contrast: corpus-based approaches*, 119–150. Amsterdam & Philadelphia: John Benjamins.

Viberg, Åke. 2005. The lexical typological profile of Swedish mental verbs. *Languages in Contrast* 5 (1). 121–157.

Viberg, Åke. 2008. Swedish verbs of perception from a typological and contrastive perspective. In María de los Ángeles Gómez González, J. Lachlan Mackenzie, & Elsa M. González Álvarez (eds.). *Languages and cultures in contrast and comparison*, 123–172. Amsterdam & Philadelphia: John Benjamins.