It May Be in the Structure, Not the Combinations: Graph Metrics as an Alternative to Statistical Measures in Corpus-Linguistic Research

Anna Shadrova
Humboldt University of Berlin
Berlin, Germany

Abstract

The following contribution summarizes a number of problems associated with a methodological focus on statistical measures in corpus linguistics, specifically in the subfield of lexical and lexicosyntactic analysis: the practical impossibility of collecting sufficient amounts of data to fully account for all of the factors that are known to interact with linguistic output; the reliance on mathematical assumptions that are generally not met by language data; and the epistemological limitations of considering corpora language samples from a presumed superpopulation versus the evolutionary and non-ergodic nature of language. It then proposes graph metrics as a viable alternative to statistical measures. Rather than comparing groups of factors, graph metrics capture relational aspects of the whole dataset, and unlike inferential statistics, they do not project to a presumed external totality or population. Neither do they rely on assumptions of randomness, independence, stationarity, or ergodicity. They thus avoid many of the concessions to the linguistic model that are inherent to probabilistic models of the lexicon, which in turn may result in an epistemologically sounder operationalization and quantification of corpus data. However, the high level of abstraction inherent in graphs, and especially in applications of

In: Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera and Joris van Zundert (eds.): Graph Technologies in the Humanities - Proceedings 2020, published at http://ceur-ws.org.

graph theory, comes with its own caveats. Drawing on the example of Kobalt, a mid-sized corpus of texts written by learners of German, two approaches to the application of graph metrics in corpus-linguistic research are outlined and demonstrated in this paper, including a detailed discussion of the steps necessary for validation and linguistic embedding.

1 Statistics in Quantitative Corpus Linguistics

Quantitative corpus linguistics has been a prolific field of research since the early 1990s. A number of large, general-purpose corpora, such as the British National Corpus (BNC World, 2007) and the German Reference Corpus DeReKo (Leibniz-Institut für Deutsche Sprache, 2019), have been curated and refined for decades, while many more specific corpora (focusing, for example, on spoken, historical, task-based, and/or learner language), are continuously being compiled and developed.

While a wide-range of phenomena are analyzed in linguistics, a particularly prevalent strand of research in corpus linguistics addresses the co-occurrence of words and the formation of meaningful larger units that do not appear to follow higher-order syntactic rules, such as collocations (words that habitually co-occur despite potentially available alternatives, like *strong coffee* vs. *powerful coffee*), idioms (word bundles with elements of meaning that cannot be derived from the words themselves, like *hit the road*), and constructions (semi-syntactic elements that are partially lexically specified and tend to be expandable by analogy, like *He worked his way up*, *She studied her way to the top*, and *They sang their way into our hearts*).

The major methodological focus of this work has been on frequentist statistics over words, lexemes (base word forms, such as *to be* for *is, are, were,* etc.) and relatively coarse syntactic categories (part-of-speech tags, such as *preposition*, or dependency labels, such as *accusative object*) that rely on automatic extraction. The term 'frequentist' here refers to the 'traditional' understanding of statistics, whereby probability is defined as the convergence to the limit of an expected relative frequency in an infinite series of experiments.[1]

[1] Bayesian models have not played a role in corpus-linguistic analysis so far. Bayesian statistics defines probability in a different manner, namely as a range of reasonable belief based on prior experience. This increases the suitability to the modeling of dynamic systems and interacting factors. Whether or not this also allows for an improved quantification of corpus data remains to be seen. Although highly relevant to the methodological development of corpus linguistics, this question will be set aside for the remainder of this paper in or-

For example, if I toss a fair coin an infinite number of times, the relative frequency of it landing on either side will approximate 0.5, which is also the probability value. Corpus linguistics operates within precisely such a framework, relying on various types of quantitative analysis such as productivity measures, which look at word distributions and quantify the openness of slots to accept new members (Zeldes, 2013; Baayen, 2002), and lexical association measures, which investigate conditional probability and transformations thereof (Baayen, 2002; Gries and Stefanowitsch, 2004; Stefanowitsch and Gries, 2003, 2005; Gries, 2013; Evert, 2005; Evert et al., 2017). In recent years, mixed effect modeling has also begun to attract attention as a more sophisticated technique for controlling the many interacting factors in language data (Linck and Cunnings, 2015; Speelman et al., 2018; Gries, 2019).

In the wake of these conceptualizations of corpora as plausibly representative samples of language, a dynamic interplay has developed between amassing larger amounts of data to allow for statistical analysis and an increasing reliance on automatic modes of data extraction and classification. There are, however, a number of issues with viewing language through the lens of stochastics, which can be broken down into three general categories: a) practical – controlling for all interacting factors tends to greatly limit sample size as well as linguistic depth; b) mathematical – the lexicon is likely not a stochastic system; and c) epistemological – corpora frequently do not qualify as samples that allow inferences to a population, especially those that have been compiled in elicited or quasi-experimental settings.

1.1 Practical Problems

A major reason for the reliance of corpus linguistics on surface or surface-near forms lies in the difficulty and high cost of in-depth annotation. Linguistic categories are generally ambiguous, fuzzy-edged, and often only implicitly represented in surface forms.

For example, while the grammatical roles of *subject* and *direct object* in the sentence *The boy broke the vase* are easily extractable – especially in English, where the position in left adjacency to the finite verb is reserved for the grammatical subject – the semantic, pragmatic, textual, and intertextual implications can be much harder to pinpoint. Consider, for example, the sentence *The vase broke*, in which the former direct object to the verb *break* is now the grammatical subject, while the process of breaking still belongs to the same object semantically.[2]

der to allow for a more in-depth discussion of the pitfalls of frequentist approaches and the potential of graph-based modeling as a structural alternative.

[2] This is an example of a so-called *unaccusative* verb. Features and problems of categor-

Even aspects of limited ambiguity, such as the morphological features of a word (for example grammatical gender or types of inflection) or the syntactic structure of a sentence, cannot always be parsed with high accuracy in languages with flexible word order and a high degree of inflection (Seddah et al., 2013). This is particularly true of non-canonical language, i.e. language that does not follow the explicit and implicit rules of written formal language, for example spoken or learner language (Krivanek and Meurers, 2011; Ott and Ziai, 2010; Choi et al., 2015; Bechet et al., 2014).

In a research setting that works with rich morphology and/or syntactic flexibility, untrained, and non-canonical data (as is the case in the analysis of German learner language), automatic classification and parser performance generally do not suffice for research purposes. This is even more so the case for analyses that go beyond surface-near forms, where the accuracy of automatic parsing of semantic or pragmatic information is generally not very high. For example, Morey et al. (2018) compare the performance of various parsers for rhetorical structures, i.e. functional descriptions of discourse blocks such as *example, elaboration,* or *concession*, and show that the most effective parsers only reach accuracy levels of around 50%. This is due to the fact that meaning is not simply encoded in the individual words themselves, but emerges from their combination, composition, and contextual embedding.

Annotation of deeper level information generally requires large amounts of tedious manual or, at best, semi-automatic annotation. Rather than being a simple labeling task, most linguistically interesting annotation is a complex categorization process. It requires the development of guidelines for ambiguous cases, measurements of inter-annotator agreement, and the iterative revision of previous annotations. With this amount of manual input, deep linguistic annotation is generally not feasible for large corpora.

At the same time, a number of recent and ongoing studies (Hirschmann, 2015; Lüdeling et al., 2017; Shadrova, 2020; Lüdeling et al., 2021) have shown that the simultaneous consideration of deeper levels of analysis and several annotation layers in the same study can provide considerable insight into deep linguistic aspects of corpus data. With analyses of this kind, it is generally only possible to assure an adequate quality for small to mid-sized corpora, i.e. several hundred to several thousand texts at most. This potential is further limited by the fact that many corpora require intensive preprocessing in the form of multiple tokenization, alignment of token layers and parallel text, etc. Particularly where spoken or signed data is involved, transcription and normalization are very resource-intensive, and effectively

ization are discussed in depth in Kuno and Takami (2004).

limit the data that is available for analysis to only a few dozen texts. It appears that for lexical analysis, this upper bound often lies *below* the lower bound for the secure application of statistical measures: effect sizes are too small to be convincing, and with the high degree of individual and stratified variation, controlling for all factors splits the data into even smaller subsets, or creates spurious significance (overfitting) through the inclusion of too many factors in the model.

There is thus an inevitable trade-off between the depth of linguistic analysis and corpus size. In other words, large corpora generally underexplore the potentials of linguistic analysis. Statistical measures were introduced into corpus linguistics in order to capture gradual effects and to distinguish between true differences and random fluctuation – but in reality, a strong focus on these metrics greatly limits the analytical capacity of linguistic research. However, as I will argue in the following sections, the problems with such an approach run deeper.

1.2 Mathematical Problems

There are a number of debates centered around statistics in linguistics which deserve to be mentioned here, but cannot be discussed in detail due to spatial constraints:

- There appears to be a general lack of understanding of the underlying distribution of words in corpora. It has been assumed that words are Zipf-distributed, but recent research suggests that this might not be the case, and/or that the Zipf distribution is itself an artifact (Williams et al., 2015; Piantadosi, 2014; Aitchison et al., 2016);
- Treating text as randomly sampled, i.e. as the outcome of a series of random experiments, is a simplification *ad absurdum*. Structured by its very nature, "language is never, ever, ever random" (Kilgarriff, 2005);
- Significance testing is problematic for a number of reasons, but particularly so in language, since essentially all assumptions about the relation between population and sample are mathematically unmet (Schmid and Küchenhoff, 2013; Koplenig, 2017).

Beyond these already grave concerns, there are two even deeper problems concerning aspects that are intrinsic not only to corpus compilation, but to the nature of language itself.

First, frequentist statistics is rooted in a concept of probability that derives overall probabilities from previous observations – it predicts the future from the past. The bridge between the two is provided by the central limit theorem, which states that in sufficiently large samples, regardless of the underlying distribution, relative frequencies will reach limits and those limits can be idealized to probabilities. This is expressed in the assumption that the overall system will reach expected values if the experiment is repeated *ad infinitum*: if a fair coin is tossed a sufficient number of times, its idealized property of being 0.5 heads and 0.5 tails will become defined over time, even if heads or tails cluster in parts of the series. However, for the central limit theorem to hold true, the underlying system must be stochastic – it must have probabilities – which means it must be *stationary* and *ergodic*, with stationarity referring to the property of a system to have stable and unchanging probabilities. If, for example, the coin is damaged after a few tosses, the outcome may be skewed and the system is no longer stationary. Language is obviously not stationary, since it evolves significantly over time. If this was the whole extent of the problem, then it would suffice to define stationary subsets of language, e.g. corpora spanning only a decade or less – but in fact, there are reasons to believe that stationarity is not reached in thematically diverse corpora even over short periods of time (Piantadosi, 2014). Moreover, there are other language-intrinsic features that create discrepancies to the model, including cognitive, discourse, and inter-speaker dynamics, and, perhaps most problematically, productivity.

Since language is at least in part a cognitive function, it underlies perceptual biases and cognitive influences. One major factor in this is priming. It is well-known that speakers are very easy to prime on all linguistic levels – syntactically (Pickering and Branigan, 1998; Gries, 2005; Loebell and Bock, 2003), lexically (Hoey, 2012; Jones and Estes, 2012), semantically (Lucas, 2000; McNamara, 2005), pragmatically (Bott and Chemla, 2013), phonetically, and/or phonologically (Luce et al., 2000; James and Burke, 2000). This holds true for priming from any source: experimental or incidental cues; self-priming from elements used by the speaker; and interlocutor priming. Since primed activations do not fully persist, but subside over time, they facilitate another feature of natural language that is called *burstiness*. It refers to the frequent re-occurrence of elements in dialogue or small parts of a text within a short period of time, during which their probability to occur leaps to much higher levels, only to swiftly collapse again to a rate that approximates zero. Another contributing factor to burstiness is text structure – certain parts of a story or report mention certain things, which are not picked up again.

Priming also contributes to inter-speaker convergence or alignment, a phenomenon in which the participants in a dialogue adapt to one another in their linguistic expression. Alignment has been shown to exist in contexts as diverse as abstract frames of reference (e.g. up/down vs. north/south in map description tasks, as investigated by Steels and Loetzsch (2008)) and highly specific and unconscious processes such as articulatory movement (Pardo, 2006), underlining the fact that probabilities may increase or decrease in a specific dialogue not by virtue of any features that are intrinsic to the linguistic elements themselves, but rather through the intentional or unintentional influence of the speaker.

A more global factor influencing the frequency of occurrence of linguistic elements is the ebb and flow of discourse interaction, by which debates gain momentum and then die down. This can play out within days, hours, or in the space of a single document, and is inextricably linked to the issue of intention: speakers do not utter certain words to fix imbalances in the probability distribution, but rather consciously choose to speak in specific ways and/or about a specific topic, which in turn increases or decreases the probability for any given word to occur.

Lastly, perhaps the strangest effect is produced by productivity (the process of coining new words incidentally), and creativity (the intentional creation of new words). Nouns, verbs, and adjectives are open classes, meaning new lexemes can be introduced to the system at any time. However, in stochastic terms, this is the equivalent of rolling dice that keep changing the number of their sides – with every new word, the relative frequencies for *all* words change. Far from being a marginal phenomenon, productivity is highly prominent in all spoken and written language, and is intrinsically problematic for the concept of stationarity in language.

The second major issue with respect to the mapping of the linguistic to the statistical model is ergodicity or path-independence. It is also the second necessary condition for the upholding of the central limit theorem. In an ergodic system, no matter which path I take through the system, relative frequencies for all elements can approximate limits and the overall system will reach expected values. For example, by rolling a fair die a sufficient number of times, noting down the values, and then averaging over them, I will find an approximation of the expected value 3.5. No matter which path I take, whether I roll a 6 seven times in a row and then a 1 seven times in a row, or any other combination – with sufficient repetitions, the system will converge. In a non-ergodic system, this is not the case: if I roll the die once and define the final result as "1" for 1, 2, or 3, and "6" for 4, 5, or 6, the expected value is still 3.5. But this value can never actually be reached, as the path taken through the system determines the outcome. There is some research on the mathematics of non-ergodicity in cognitively oriented research fields, including cognitive neuroscience (Medaglia et al., 2011; Franco et al., 2007; Papo, 2013) and social and developmental theory (Molenaar, 2008; Lerner, 2012). For language, however, the problem is rarely discussed, although it is now beginning to attract more scholarly attention (Lowie and Verspoor, 2018; Dębowski, 2018).

Without ergodicity and/or stationarity, the central limit theorem fails. This is not a minor inconvenience, but equivalent to the collapse of the single bridge that spans the space between two tall mountains – without the central limit theorem, mathematically, there is no connection between prior observations and a presumed population. There is no reason to assume that words counted in one corpus hold any precise, probabilistic information about what will be found in another. In a changing system, the past has limited capacity to predict the future. While one may still choose to trust corpus data as empirical evidence, there is little justification to rely on concepts like p-values or effect size to safeguard against chance results.

1.3 Epistemological Problems

The epistemology of statistics over words is unsatisfying for reasons that go beyond mathematical concerns. Suggesting that the production of words can be modeled stochastically, i.e. as a distribution of probabilities, entails at least two philosophical extensions:

1. There is a stochastic system, perhaps in a latent superpopulation[3] (as it is sometimes conceptualized in the social sciences), that produces unchanging and stochastically deterministic output;
2. Linguistic data collection is capable of capturing samples of this output.

Both of these extensions are, in fact, rather strange. The first suggests that while we cannot know *when* certain things will be said, or which words will co-occur; the conversations we have are effectively *pre-determined* in the stochastic system. This implies that for all new word forms that are productively generated, and for all new processes and items in the world that did not exist until recently and are now named (such as *to google* or *Tesla*), the stochastic system *has always held a quantitatively predefined slot*. This idea is utterly alien to linguistics, and is unlikely to be entertained by anyone in the field. On the contrary: usage-based linguistics metaphorically draws on systems theory by suggesting that language is a complex, adaptive (i.e. evolving) system – one that is neither stationary, nor ergodic (Beckner et al., 2009; Ellis, 2016; Holland et al., 2005; Massip-Bonet, 2013; Steels, 2000, among others). In complex adaptive systems, cross-system statistical modeling is pointless,

[3] The concept of superpopulation refers to a model in which a real population is seen as a sample of a latent infinite population. For example, one might think of the distribution of career choices in the population of a country as a sample from a stochastic process. This is used to allow for inferential statistics where data is effectively not reproducible (since there is only one of each country at a given time), but carries the philosophical problem of implying determinism in dynamic and possibly unique systems.

since things keep changing. Parts of the system may of course be compared, and compared quantitatively, but this requires an intelligent way to distinguish between random fluctuation and structural shift.

The second extension touches on the problem of the representativeness of corpus data, and specifically of corpus data that is compiled in what is sometimes called a quasi-experimental design. For example, in studying learner language, one may try to control for certain variables such as topic, task, setting, degree of formality, etc., which typically results in small to midsized corpora like the learner corpora Kobalt (Zinsmeister et al., 2012), Falko (Reznicek et al., 2010; Lüdeling et al., 2008), and the International Corpus of Learner English (Granger et al., 2009), or task-based dialogue corpora such as the Berlin Map Task Corpus (BeMaTaC, Sauer and Lüdeling (2013)). Both offer an excellent way to collect data, which promises to yield crisp and highly relevant results. However, by performing statistical analysis on this data, or more precisely by borrowing methods from inferential statistics, via reliance on the central limit theorem, I do not simply measure what I find in the data, but rather *infer* a presumed superpopulation that works as a stochastic system.

This approach is epistemologically risky in two ways. First, unlike newspaper archives or historical sources, it does not involve naturally occurring data that has been collected from existing language – it works with language that has been intentionally created and that may never have existed without the researcher's intervention. This means that the data collection created the presumed superpopulation from which certainty or effect size is then derived, rendering the argument circular. Second, if the amount of data turns out to be too small for certain types of analysis (as it typically does for collocations, see section 1.2), one may be tempted to collect more data of the same kind. Doing so is not the same as taking samples, however. Rather, it means expanding language in a way in which it probably would not have developed on its own. Continuing with this method until data sizes of two higher orders of magnitude are reached – for example by collecting tens of thousands of texts written by learners of German instead of a few hundred – would: a) skew the proportions of all documented data of learner German significantly; b) interfere in the writing development of hundreds of thousands of learners of German; and c) create a population, rather than sampling from one.

Consider for a moment the following analogy: if a naturalist went into the woods to count black and blue birds belonging to a particular species, the birds would have existed without their involvement. Provided that the proper methodology is chosen and applied well, reliable insights into the features of this population are to be expected. However, if the naturalist began breeding the birds in order to gain control over data collection, and accidentally bred green birds never before observed in the wild, no amount of new data would allow them to infer from their laboratory population to the original population, as the existence of these birds has permanently changed the entire species. In other words, in an attempt to create representative amounts of data (birds), our naturalist will have created irresolvable path-dependency in their research. The population that could have been inferred to from the initial sample in the wild is now no longer representative of the whole species, and thus of little to no use for comparison or significance testing.

At first glance, this may seem like a far-fetched and purely theoretical problem, but the reality of learner language is that most of it is *not* spoken in the context of the argumentative essays that are typically found in learner corpora. Instead, it is found in: a) immigration offices and areas with high density immigrant populations; b) middle schools; and c) tourist destinations – none of which are prone to eliciting essay-length argumentative texts on controversial topics in a formal written setting. This means that the compilation of learner corpora as they exist today already interferes in learner language and influences projections of a latent superpopulation.

This is not to say that there are no quantifiable patterns or measurable differences between language cohorts. However, it is important to map the demands of the mathematical model to the subject matter, and it appears as though frequentist statistical interpretations of lexical distributions offer a rather unlinguistic perspective. The same is not necessarily true of other types of linguistic distributions. More abstract syntactic features like word order, cases, part-of-speech distributions, determiner systems, and so on, are much more stable historically, and converge quickly even in data of limited size. They also exhibit lower variance than word distributions and are greatly influenced by changes in subcategories, as well as in any of the complementary categories. For example, phonological changes may influence the obligatoriness of articles, thereby also influencing the case system, as is the case in diachronic language change. Using different words and inventing new ones, on the other hand, does not have the same effect. It is therefore quite possible that probabilistic analysis of syntactic aspects in corpora is mathematically and conceptually valid. For largely lexically oriented analysis, this does not appear to be the case.

It should be noted here that criticizing the application of statistical methods on lexical aspects of corpora is not merely a matter of methodological taste or belief. If words do not have probabilities, probability-based measures are meaningless in the same way that it is meaningless to measure the temperature of a philosophical debate (even if it is heated) or the width of the history of Europe (even if it is wide-ranged). Validity matters in scholarly research, and that includes both the internal validity of the mathematical model, and the mapping of mathematical to subject-specific concepts and empirical observations.

2 Graph-Based Modeling in Linguistics

With an abundance of problems around statistical measures in corpus-based lexical and lexicosyntactic analysis, and with the practical consideration that reliable measures of smaller data are necessary, graph metrics would appear to possess intrinsic appeal to corpus linguistics. Yet they are practically unused. In fact, graph theory in general is underrepresented in linguistics, despite some early work centered mainly around Markov chain modeling (Goodman, 1961; Jelinek et al., 1975; Brainerd and Chang, 1982). While computational linguistics does include graph-based models, they are largely implemented in engineering ways (i.e. for underlying databases and algorithms) rather than for the analysis of language itself. In core linguistic research, graphs are mainly used for the visualization of complete analyses, most notably in syntax trees, ontologies, and taxonomies. However, graphs can also be measured in different ways.

2.1 Example 1: Modularity

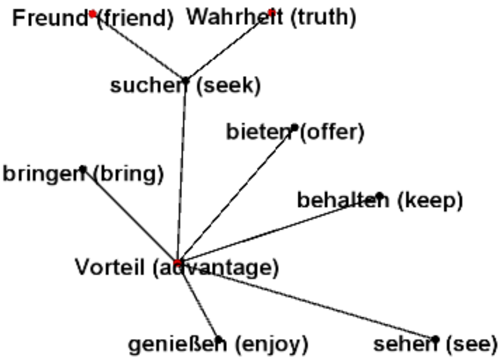

Figure 1: Sample graph for lexical co-occurrence as filtered through syntax in Kobalt

The first example shows the modeling of lexical information and its measurement through the metric of Louvain modularity (Blondel et al., 2008). The data used for this analysis is taken from the Kobalt corpus (Zinsmeister et al., 2012), which consists of 151 topic-controlled argumentative essays written by learners of German from China and Belarus, and 20 control texts written by German native speakers. From these texts, graphs based on corrected syntactic dependency parses were extracted. Verb lexemes and noun, adjective, and adverb lexemes that depend on those verbs as arguments (different kinds of objects, subjects, and predicates) were modeled as nodes, and edges were modeled from the existing dependence (co-occurrence in the respective syntactic environment of the verb). The full analysis is discussed in detail in Shadrova (2020) and includes other layers of linguistic analysis, as well as an in-depth internal validation through a range of sampling techniques and hyperparameter settings. Figure 1 provides an example of a small subgraph. The full graph of the Kobalt native speaker subcorpus is visualized in Figure 2.[4]

In conducting this analysis, the main research question was whether graphs of this type show different degrees of structuredness for different stages of foreign language acquisition in learners, as well as between learners and native speakers. The underlying concept of coselectional constraint or idiomaticity describes the tendency of native speakers to constrain their choice in word combinations to a relatively small set out of a potentially large combinatorial space. For example, the verb chosen for describing the action of cleaning teeth in English is *brush*, whereas the German equivalent is *putzen* ('to clean'). While *clean teeth* would be understood and is semantically acceptable, the combination is highly unlikely to occur in a text written by an English native speaker (unless it was describing the act of having one's teeth professionally cleaned by a dentist). This particular example would typically be learned early in both unguided learning and in language classes. However, there are vast amounts of subtle constraints of this kind, and they are known to be very difficult to master even at an advanced stage of second language acquisition (Howarth, 1998; Pawley and Syder, 1983; Paquot and Granger, 2012, and many others).

The central hypotheses of the present study are as follows:

1. Graphs are more structured in more vs. less advanced learners;
2. Graphs are more structured in native speakers than learners;
3. Learners show a u-shaped learning development in their coselectional structuredness.[5]

[4] Since these graphs are generally to large and detailed to be legible in print, further visualizations, alongside the data and scripts for analysis, are available in a separate Zenodo repository (doi: 10.5281/zenodo.3584091).

[5] The linguistic background of this is discussed in detail in Shadrova (2020). Very briefly put, this hypothesis is rooted in the theoretical premise that learners typically undergo a process of randomization of structures. First, they learn in chunks during early acquisition (everyone learns more or less the same in early language classes), and then they acquire and

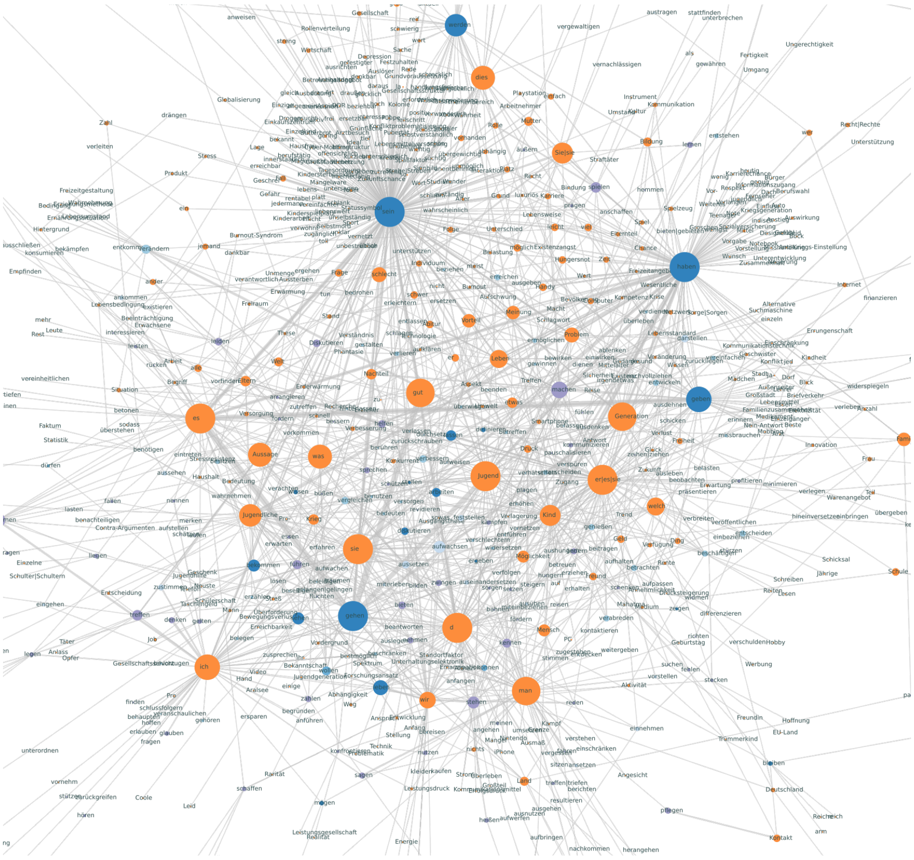

Figure 2: Full graph of lexical co-occurrence (verb-argument structures) for native speakers in Kobalt. Graph serves as conceptual demonstration only, scalable images are available via Zenodo: `10.5281/zenodo.3584091`).

The measure used to capture these effects is Louvain modularity (Blondel et al., 2008), one of a number of community detection algorithms[6] deriving a value in the range of [-1,1] for the modularity of a graph. A more modular graph contains strongly structured (i.e. interconnected) communities that are less connected to the rest of the graph. A less modular graph contains

recombine more vocabulary to succeed in more complex communicative situations. At this second phase, their language cannot yet be fully restricted in ways aligned with the target language for lack of experience and influence of the learners' native language. This phase of randomization is followed by semantic differentiation and lexical and semantic specialization (people learn more specific language over time). This decrease in accuracy at intermediate stages commonly described as a u-shaped learning trajectory, and can be found wherever a learning process is guided by both rules and exemplar-based exceptions.

[6] It appears that Louvain modularity prefers certain community sizes and constellations over others, raising the question of whether other algorithms may be better suited for corpus-linguistic analysis. This issue has not been considered in the analysis presented and remains to be addressed in future research.

many nodes that are more randomly connected to other parts of the graph. A graph of negative modularity contains fewer edges between nodes than would be expected by chance, which does not appear to be the case in lexical graphs generally. Importantly, modularity is not an artifact of graph size. As exemplified in Figure 3, graphs of the same size in terms of nodes and edges can be more or less structured, and thus possess higher or lower modularity.

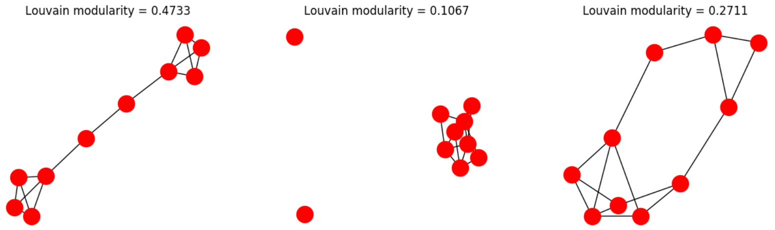
Louvain modularity = 0.4733
Louvain modularity = 0.1067
Louvain modularity = 0.2711

Figure 3: Louvain modularity in different graphs of the same size (10 nodes, 15 edges)

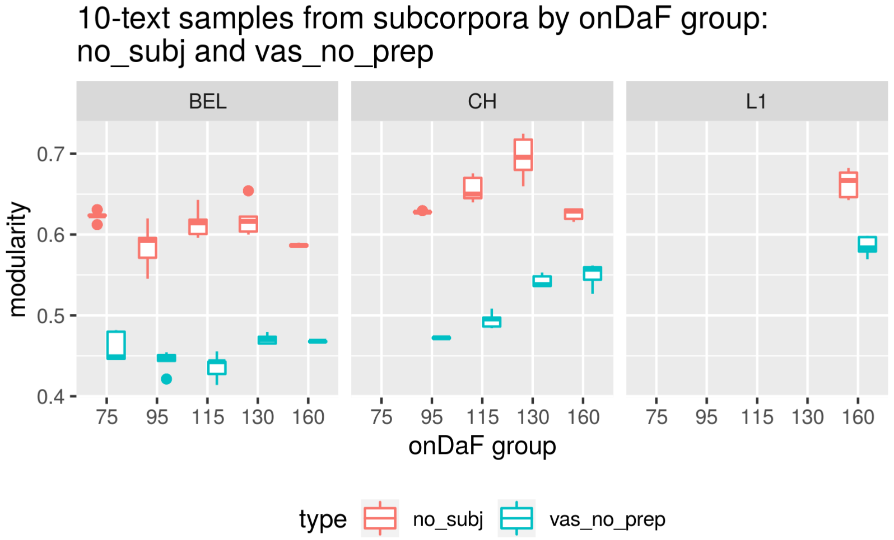

Figure 4: Louvain modularity for graphs of verbs and argument lexemes with (vas_no_prep) and without (no_subj) subjects. Subcorpora were sampled five times. Samples partially overlap. The onDaF test is a standardized c-test for German language assessment (Eckes, 2010).

An analysis of a range of Kobalt subcorpora, each represented by five samples in Figure 4, shows that hypotheses 1 and 2 are met by the data, and hypothesis 3 is met in the Belarusian, but not in the Chinese subcorpus. A sliding window analysis of the same data is presented in Figure 5. Windows

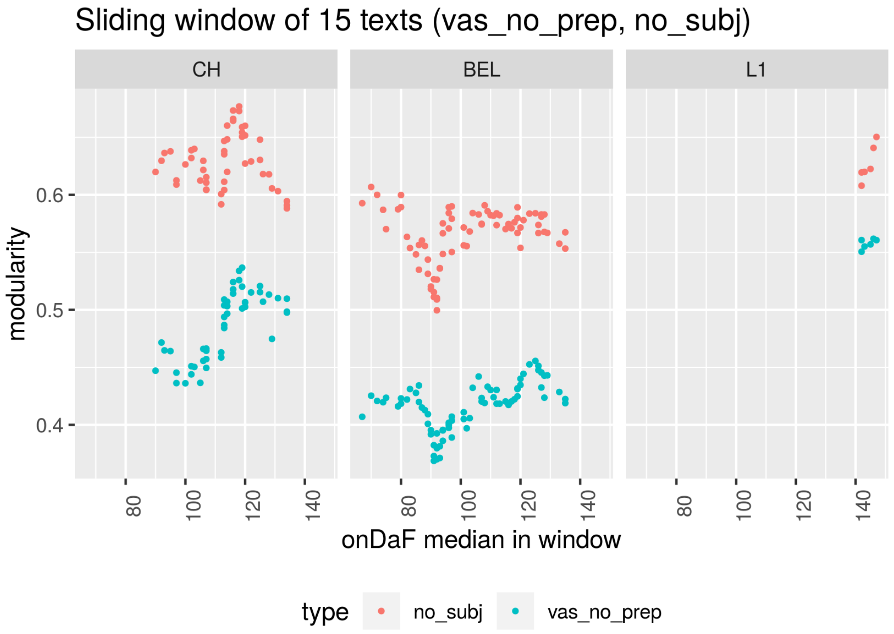

Figure 5: Sliding window analysis of Kobalt: Louvain modularity for graphs of verbs and argument lexemes with (vas_no_prep) and without (no_subj) subjects, computed for windows (subcorpora arranged by onDaF scores, neighboring windows overlap in all but two texts). The onDaF test is a standardized c-test test for German language assessment (Eckes, 2010).

of 15 texts were used for each data point. Window 1 is created from texts 1-15 arranged by test scores, window 2 from texts 2-16 and so on. This analysis shows that, beyond the grouped samples, a clear trajectory emerges over test scores, which suggests that test scores and graph structures indeed correlate in meaningful ways. If they did not, the trajectory would be expected to be much more erratic.

For comparison, a statistical analysis of the same data based on lexical association measures was also performed. However, despite cohort homogeneity and the limitation to an identical task/topic, there are very few similar or identical collocational pairs across subcorpora. Absolute occurrences quickly dwindle into very low numbers (below 10), which means that a high and low rate of co-occurrence would falsely be identified from the same order of magnitude. While 6 is of course three times as many as 2, the absolute difference is not very telling in terms of the underlying difference in coselectional association. In addition, the combinatorial power of lexical material as it occurs even in small corpora is counter-intuitively large and poses a hindrance to statistical interpretation. In fact, the number of verb and accusative object lexemes as they are used in a small subcorpus of Kobalt

(21 texts, 148 unique verbs, 304 unique nouns in accusative object position) results in a potential of 44,992 combinations. Even if one were to assume that only 10% of those are semantically possible, and that only 10% of the total 726 realizations of verbs and adjective objects in the subcorpus are freely combined in the first place, this results in 72 draws from 449 elements. The combinatorial potential of this is higher by several orders of magnitude than the estimated number of atoms in the universe ($3.352 \cdot 10^{89}$ vs. 10^{78} – 10^{82}). Constructing a scalar measure of coselectional constraint statistically or stochastically would require a definition of thresholds for relative frequencies at which the occurrence of word pairs would be considered within or outside the range of expectation. However, drawing *any* specific combination from this vast combinatorial space is extremely unlikely. Overall, results from the statistical analysis remain inconclusive and difficult to interpret. This is discussed in detail in Shadrova (2020, chapter 4).

In contrast, the graph-based analysis shows clear trajectories and effects in line with hypotheses 1 and 2, as well as a linguistically meaningful divergence from hypothesis 3.

2.2 Example 2: Lexicosyntactic Graphs and Grammar as a Graph

The second example is not yet quantified and serves to demonstrate the underlying modeling problem – a suggestion as to its quantification appears at the end of this section.

In the previous analysis, syntax is involved as a hierarchical lexical filter: only those words that occur in certain syntactic slots of the verb are considered in the structural analysis. Unlike positional models of collocation, which consider words in a window of *n* tokens around the target word, a syntactic model of this kind is able to effectively filter for the relevant collocations in languages with flexible word order such as German. However, in the theoretical framework of usage-based linguistics that many corpus linguists ascribe to, a strict division of lexicon and syntax is not usually presumed. Instead, a number of models suggest the existence of a continuum, partial overlap, and/or interdependence of lexical and syntactic elements. In fact, some strands of grammar theory are largely focused on the very interface of lexicon and syntax – these include construction grammar, which is entirely based on the idea of the inseparability of lexicon and syntax (Goldberg, 2005, 2013; Sag et al., 2012; Croft, 2001; Boas, 2013), and modern approaches to valency and dependency grammar, according to which grammar is generated from specified features of lexical items (Ágel and Fischer, 2010; Herbst, 2014; Faulhaber, 2011).[7]

[7] The latter is also partially true of some merging theories such as head-driven phrase

While the idea of the syntax-lexicon continuum is appealing at first glance, it quickly runs into conceptual problems and is difficult to operationalize. For one thing, it requires clarity over whether the compared elements constitute categories/variables or exemplars. Since many words are exemplars (like *on, after, if*), but many are also categories (like *to be* with its paradigm *are, were, is,* etc.), this is not trivial. Second, syntax and higher-order functions structure text in such a way that once a category is decided, many others fall into place. At the same time, there is plenty of space for inter- and intra-individual variation.

So far, corpus studies have tended to look at the coselectional patterns of individual lexemes or types of lexemes, or individual verb-argument structures or classes thereof. For example, Dux (2016) inspects the argument structure patterns of selected verbs from the semantic fields of *changing* and *stealing*, whereas Faulhaber (2011) compares the argument structure patterns of 88 different verbs. Both conclude that semantics alone cannot predict syntactic patterns with a high level of accuracy, and that lexical idiosyncrasies need to be taken into consideration. A similar observation is made by Zeldes (2013) concerning the productivity of verbs in their argument selection. The unifying thread in these studies is that grammar is informed by the preferences or selections of individual lexical items – something that is not considered possible in more traditional approaches to syntax, where lexical items are merged into syntactic patterns in a way in which syntax imposes rules on the lexicon, but not vice versa.

Technically, the studies mentioned do not model syntax and lexicon in the same space – rather, they discuss the combinatorics of two sets of elements, where the sets are selected by semantic or syntactic rules, similar to the filtering function of syntax in the previous analysis.

The mapping of two sets of elements, however, can be summarized and made explicit as one system in a graph. An example is provided in Figures 6 and 7. In these graphs, lexemes and syntactic functions such as grammatical subject (SUBJ) and direct or accusative object (OBJA) are modeled as nodes, and the syntactic dependency between all syntactic functions is encoded explicitly. At the same time, forces of association between lexemes and syntactic functions are also represented in the graph. Thus, a connection between what would be rows of a table of factor combinations are made explicit in the graph structure. The visualization is produced by a so-called force algorithm, a physics simulation in which higher frequency of occur-

structure grammar (HPSG, Pollard and Sag (1994)). However, syntax still has major independence in these schools of thought, and lexical items can only select, but not generate, syntactic rules. For a critical discussion of the various approaches to this issue and an attempt to unify them, see Müller (2013) and Müller and Wechsler (2014).

rence of a lexeme in a syntactic slot, or of one syntactic slot depending on another, pull the two respective nodes into proximity.

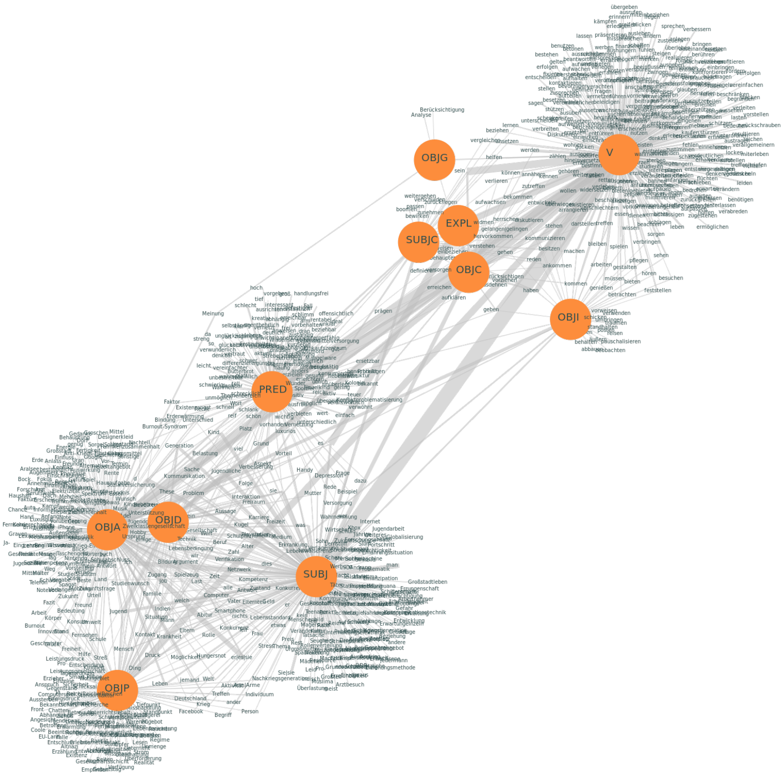

Figure 6: Lexicosyntactic graph of the native speaker subcorpus in Kobalt. Accusative and dative objects (OBJA and OBJD) are pulled into proximity by shared lexemes, subjects are more distant.

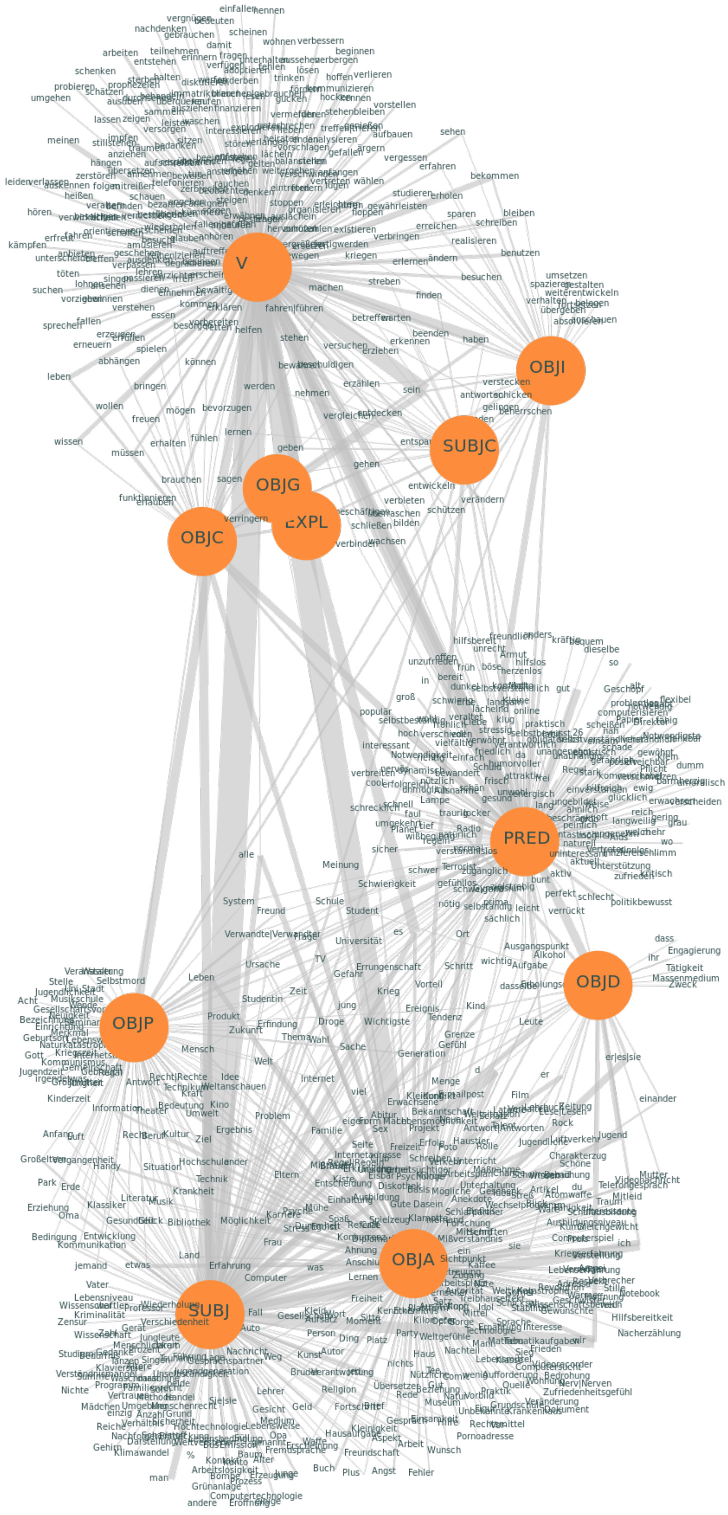

Figure 7: Lexicosyntactic graph of the intermediate Belarusian subcorpus in Kobalt. Subjects (SUBJ) and accusative objects (OBJA) are pulled into proximity by shared lexemes, other object types float further away.

A closer look at the two figures reveals that the graph for intermediate Belarusian learners of German in Figure 7 shows much more proximal positions (i.e. shared lexemes) for subjects and accusative objects compared to the native speaker graph in Figure 6, while objects appear to cluster in a more coherent group in the native speaker graph. From a linguistic perspective, this is interesting in two ways.

First, as has already been noted, syntactic category distributions tend to converge quickly in corpus data, and in many cases do not differ much between different cohorts. Despite some minor pattern deviations, when the same data is analyzed for distributions of syntactic categories, it does not yield any substantial differences according to native language group. However, through the combination of syntactic categories and lexical items, substantial differences in the graphs of the two cohorts do emerge. This suggests that graph structure is capable of capturing linguistically meaningful effects that are *not encoded* in the individual co-occurrences of lexemes in the syntactic structures themselves, but only in their *interrelation*: it appears that lexicosyntax lives in the cross-systemic structure, and not in the combinations of words.

Second, the specific differences between the two graphs correspond to higher-level linguistic concepts: a higher interchangeability of subject vs. object lexemes may indicate passivization, or it might point towards differences in anaphoricity (i.e. the types and structure of referentiality to previously introduced discourse referents). This is particularly interesting because Russian and Belarusian are partially pro-drop languages, meaning that subjects can be left out or go unrealized in many contexts, which raises the plausibility for different encodings of the relation between subjects and objects in the minds of native speakers of those languages.

Figure 8, which shows the same type of graph for intermediate Chinese learners of German in Kobalt, underscores this point. Here, subjects and predicates are more closely aligned, which may be rooted in typological aspects of Mandarin Chinese; a theme-rheme language in which a sentence topic is presented and then commented upon, and which can thus be plausibly mapped to a subject-copula-predicate construction in German. However, since these analyses are post-hoc, and the graphs involved are opportunistically derived from previously analyzed data, more research is needed to verify their genuine usefulness in the study of lexicosyntax.

This includes the need for quantification. Even if the visualization suggests certain effects, these can only be verified through a comparison with other graphs from the same as well as other cohorts. Assessing differences of this kind visually can be tedious and unreliable. Furthermore, a quantification would allow for an analysis of variance (i.e. more or less similar graphs of the same cohort), while a visual assessment can only extract obviously diverging patterns.

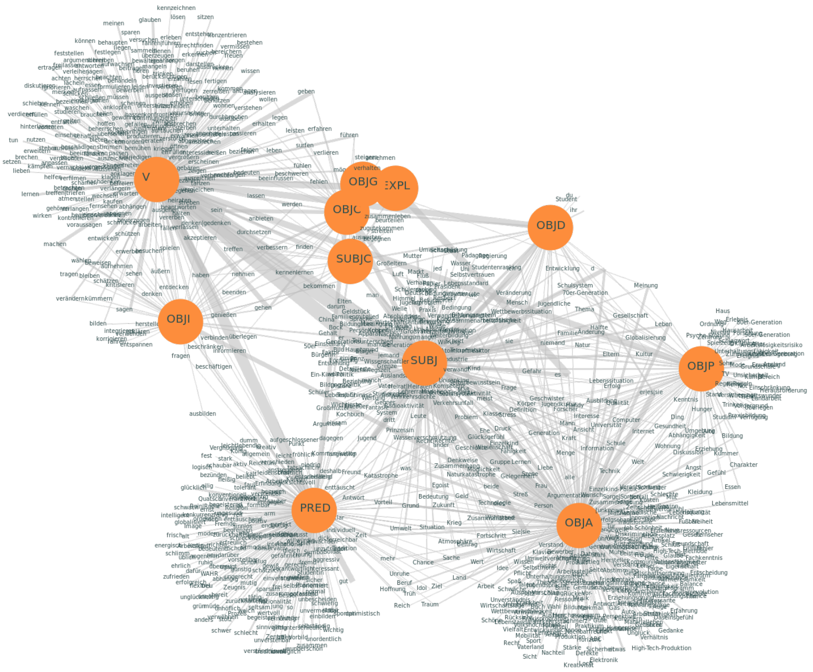

Figure 8: Lexicosyntactic graph of the intermediate Chinese subcorpus in Kobalt. Subjects (SUBJ) and predicates (PRED) are located more centrally in the graph. Object types appear more separate than in the other graphs.

The visualization encodes structural information that is also present quantitatively in the graph itself, and may be used for external quantification. One possible way to approach this would be through non-negative matrix factorization (NMF, Gillis (2014)), which is used in information extraction techniques such as topic modeling (de Paulo Faleiros and de Andrade Lopes, 2016; Chen et al., 2019; Kuang et al., 2015). NMF is a dimensionality reduction algorithm that takes a matrix and deterministically reduces it to a unique vector. Vectors can then be compared via cosine distance, as is done in certain applications of computational linguistics such as word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Ethayarajh, 2019, among others).

While conceptually and computationally well-implemented, the central question arising here is what should be encoded in the matrix to ensure linguistic validity. One approach would be to span a matrix for all nodes times all nodes (i.e. both lexemes and syntactic functions). But since words can only occur in syntactic slots, this would leave most cells empty, perhaps overestimating similarity. The other would be to abstract from the concrete lexemes and span the matrix for syntactic nodes only, entering only the number of shared lexemes in each field. However, this implies a double dimensionality reduction and leaves out not only the exemplar information, but also structural information that is relevant to the graph visualization and relates to linguistically meaningful aspects of the graph structure, such as the issue of how node clusters are distributed in relation to other node clusters with which they do not share information.

This goes to show that computing just any graph metric will not suffice for the development of a linguistically interpretable method of graph measurement. As for all methodological development, in-depth validation, replication, and extension to new data are of crucial importance - but perhaps more so is the embedding into the theoretical frameworks of linguistic study, and the triangulation with results obtained through other methods.

3 Conceptual Validation and Future Research

Graph-based modeling could bring new perspectives into quantitative corpus linguistics. Its major advantages over frequentist statistical modeling are that it: (a) captures relational information directly, without the internal triangulation of various measures; and b) does not infer to concepts such as stationary and ergodic stochastic systems or a stable and existing superpopulation. However, it is going to take extensive further research to fully assess the potential of graph metrics. This is particularly true since unlike in mathematical graph theory or network analysis, the words of a language are both representative of linguistic categories and relevantly individual. That is to say, each item differs in important aspects from all other items, which means that, generally speaking, a full abstraction over all words will not be well-aligned with the linguistic model. Detailed theoretical modeling and empirical conceptual validation are therefore necessary to ensure that graph metrics do indeed measure linguistic aspects and not spurious information of a hyperstructure that is only marginally important in linguistic terms.

An example of such an effect is the measurement of the in- and out-degree of lexemes in corpora, a measurement that consistently detects so-called *small-world-effects*. These are graph structures in which some nodes have many in- or outgoing edges, while most others have only few, and have been measured consistently through a range of published research across many corpora (Choudhury and Mukherjee, 2009; Ferrer i Cancho and Solé, 2001; Wachs-Lopes and Rodrigues, 2016, among others). While one might be tempted to view this as a graph-based discovery, the fact of the matter is that quantitative linguistics has always known that lexemes follow a Zipf, long-tailed, or power-law distribution, also called a distribution with large numbers of rare events (LNRE). In large corpora, a small number of words are very frequent – mostly functional words like prepositions or pronouns – and typically up to half of the lexemes are so-called *hapax legomena* or *dis legomena*, words that occur only once or twice in the corpus. Arranged in a graph, this would trivially imply degrees of one or two for most words, leaving an impression of high centrality or connectivity of individual lexemes. However, to corpus linguistics, this is merely a visualization of an already well-established fact and not a structural discovery. In order to avoid the dead-ends of epiphenomenal observations, a clear theoretical foundation as well as a structured approach to the validation of graph-based research in linguistics is crucial.

With the development of sufficient computing power, graph-based analysis has become more advanced and central not only in traditional STEM subjects, but also in the humanities and social sciences. Accordingly, there is an ongoing development of new metrics. Some of these metrics are derived from graph theory directly, which means that they abstract more strongly from individual nodes and focus on the more abstract properties of node and edge classes. Isographs, i.e. structurally identical (sub-)graphs that may differ significantly in their node contents, are a particularly relevant issue in this context. Their existence may or may not have implications for the underlying theoretical model, and should be considered in any application of graph metrics to a subject-specific research question.

In order to make the most efficient use of graph metrics in linguistics and other fields of study, it is important to consider implications of this kind not only mathematically, but also epistemologically by asking questions like the following:

1. Is the theoretical model well-represented by the mathematical model?
2. Are there aspects to the mathematical model that may interfere with subject-specific theoretical underpinnings and/or interpretations of the data?

Much more validation and research is required to identify concerns in this regard, including: the application of measures to more data; the detailed modeling and mapping of theoretical to mathematical concepts; the calibration and triangulation of metrics; and the development of evaluative frameworks for all of the above.

4 Conclusion

Graph metrics are very new to the field of corpus linguistics, especially in core-lingustic (rather than psycholinguistic or computational) research. The analyses discussed in this paper suggest that graph-based modeling and quantification may provide an alternative to operationalizations in the more traditionally applied framework of frequentist statistics. They may even provide solutions to some of the common practical problems of corpus linguistics, since by encoding information at a higher density, graphs allow for the quantification of small data. By abstracting from individual entities such as lexemes they also allow for structural assessment without the triangulation and comparison of a large number of words or word pairs, which implies fewer artifacts from text length, a problem that is notorious with all corpus research.

With advances in computing power and the ease provided by graph infrastructures such as neo4j, the corpus search engine graphANNIS (Krause, 2019), the community API in Python, the GUI-based graph analysis program Gephi (Bastian et al., 2009), and the igraph package in R (Csardi and Nepusz, 2006), graph metrics are becoming more usable outside of traditional computationally oriented subjects. However, this should not tempt linguists and other researchers to blindly compute graph metrics on data without further consideration of the underlying model. Much more research is required to reliably map the concepts of graph theory and network analysis in a way that is fully compatible with linguistic concepts, and to ensure that the application of graph metrics does not produce undesirable effects like the ones associated with the use of frequentist statistics in corpus-based studies of the lexicon and lexicosyntax. The most immediate desideratum for future research into the graph-based analysis of corpus data thus lies in the fields of theoretical and quantitative modeling, validation, and replication.

5 A Note on Software

Kobalt was preprocessed with TreeTagger[8] and Malt Parser (Nivre et al., 2006) based on Foth et al. (2006)'s dependency grammar of German, which was slightly adjusted for the purposes of this analysis. Further details can be found in Shadrova (2020). For graph extraction, analysis, and visualization, R (R Core Team, 2015) and RStudio (RStudio Team, 2015) with packages reshape2 (Wickham, 2007), dplyr (Wickham et al., 2018), jsonlite (Ooms, 2014), and ggplot2 (Wickham, 2016) were employed. Graphs have been visualized with D3.js (Bostock et al., 2011), Python matplotlib (Hunter, 2007), and networkX (Hagberg et al., 2008). Modularity was computed with Python's community API developed by Thomas Aynaud (`https://python-louvain.readthedocs.io/en/latest/api.html`). Figure 1 was created with Gephi (Bastian et al., 2009).

[8] `https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

Acknowledgements

I would like to thank Joris van Zundert for his very kind and constructive comments on a previous version of this essay, which have greatly helped to improve its focus and clarity. All remaining errors are of course my own.

References

Ágel, V. and Fischer, K. (2010). 50 Jahre Valenztheorie und Dependenzgrammatik. *Zeitschrift für germanistische Linguistik*, 38(2):249–290.

Aitchison, L., Corradi, N., and Latham, P. E. (2016). Zipf's Law Arises Naturally When There Are Underlying, Unobserved Variables. *PLoS computational biology*, 12(12).

Baayen, R. H. (2002). *Word Frequency Distributions*, volume 18. Springer, Dordrecht, Netherlands.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Bechet, F., Nasr, A., and Favre, B. (2014). Adapting Dependency Parsing to Spontaneous Speech for Open Domain Spoken Language Understanding. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., et al. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning*, 59:1–26, DOI: 10.1111/j.1467-9922.2009.00533.x.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, DOI: 10.1088/1742-5468/2008/10/p10008.

BNC World (2007). The British National Corpus. http://www.natcorp.ox.ac.uk/.

Boas, H. C. (2013). *Cognitive Construction Grammar*. Oxford University Press, DOI: 10.1093/oxfordhb/9780195396683.013.0013.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3 Data-driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, DOI: 10.1109/TVCG.2011.185.

Bott, L. and Chemla, E. (2013). Pragmatic Priming and the Search for Alternatives. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Brainerd, B. and Chang, S. M. (1982). Number of Occurrences in Two-state Markov Chains, with an Application in Linguistics. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 225–231.

Chen, Y., Zhang, H., Liu, R., Ye, Z., and Lin, J. (2019). Experimental Explorations on Short Text Topic Mining between Lda and Nmf Based Schemes. *Knowledge-Based Systems*, 163:1–13, DOI: 10.1016/j.knosys.2018.08.011.

Choi, J. D., Tetreault, J., and Stent, A. (2015). It Depends: Dependency Parser Comparison Using a Web-based Evaluation Tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396.

Choudhury, M. and Mukherjee, A. (2009). The Structure and Dynamics of Linguistic Networks. In *Dynamics on and of Complex Networks*, pages 145–166. Birkhäuser, Boston, MA.

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, DOI: 10.1093/acprof:oso/9780198299554.001.0001.

Csardi, G. and Nepusz, T. (2006). The igraph Software Package for Complex Network Research. *InterJournal*, Complex Systems:1695, http://igraph.org.

de Paulo Faleiros, T. and de Andrade Lopes, A. (2016). On the Equivalence between Algorithms for Non-negative Matrix Factorization and Latent Dirichlet Allocation. In *ESANN*.

Dębowski, Ł. (2018). Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited. *Entropy*, 20(2):85, DOI: 10.3390/e20020085.

Dux, R. J. (2016). *A Usage-based Approach to Verb Classes in English and German*. PhD thesis, https://tdl-ir.tdl.org/handle/2152/ETD-UT-2011-05-3114.

Eckes, T. (2010). Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung. In Grotjahn, R., editor, *C-Test: Contributions from Current Research*, volume 18 of *Language Testing and Evaluation*, pages 125–192.

Ellis, N. C. (2016). Salience, Cognition, Language Complexity, and Complex Adaptive Systems. *Studies in Second Language Acquisition*, 38(2):341–351.

Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *arXiv*, page 1909.00512, https://arxiv.org/abs/1909.00512.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Evert, S., Uhrig, P., Bartsch, S., and Proisl, T. (2017). E-view-affilation–a Large-scale Evaluation Study of Association Measures for Collocation Identification. *Proceedings of eLex 2017–Electronic lexicography in the 21st century: Lexicography from Scratch*, pages 531–549.

Faulhaber, S. (2011). *Verb Valency Patterns: A Challenge for Semantics-based Accounts*, volume 71. De Gruyter.

Ferrer i Cancho, R. and Solé, R. V. (2001). The Small World of Human Language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265.

Foth, K., By, T., and Menzel, W. (2006). Guiding a Constraint Dependency Parser with Supertags. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*. Association for Computational Linguistics, DOI: 10.3115/1220175.1220212.

Franco, L., Rolls, E. T., Aggelopoulos, N. C., and Jerez, J. M. (2007). Neuronal Selectivity, Population Sparseness, and Ergodicity in the Inferior Temporal Visual Cortex. *Biological cybernetics*, 96(6):547–560.

Gillis, N. (2014). The Why and How of Nonnegative Matrix Factorization. *Regularization, optimization, kernels, and support vector machines*, 12(257):257–291.

Goldberg, A. E. (2005). *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, DOI: 10.1093/acprof:oso/9780199268511.001.0001.

Goldberg, A. E. (2013). *Constructionist Approaches*. Oxford University Press, DOI: 10.1093/oxfordhb/9780195396683.013.0002.

Goodman, N. (1961). *Graphs for Linguistics*. American Mathematical Society, DOI: 10.1090/psapm/012/9982.

Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). *International Corpus of Learner English*. Presses universitaires de Louvain, Louvain-la-Neuve.

Gries, S. T. (2005). Syntactic Priming: A Corpus-based Approach. *Journal of psycholinguistic research*, 34(4):365–399.

Gries, S. T. (2013). 50-something Years of Work on Collocations. *International Journal of Corpus Linguistics*, 18(1):137–166.

Gries, S. T. (2019). 15 Years of Collostructions. *International Journal of Corpus Linguistics*, 24(3):385–412.

Gries, S. T. and Stefanowitsch, A. (2004). Extending Collostructional Analysis: A Corpus-based Perspective Onalternations'. *International journal of corpus linguistics*, 9(1):97–129.

Hagberg, A., Swart, P., and Chult, D. S. (2008). Exploring Network Structure, Dynamics, and Function Using Networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11–15. https://www.osti.gov/biblio/960616.

Herbst, T. (2014). The Valency Approach to Argument Structure Constructions. In *Constructions Collocations Patterns*, number 282 in Trends in Linguistics. Studies and Monographs [TiLSM], pages 167–216. De Gruyter, Mouton, DOI: 10.1515/9783110356854.167.

Hirschmann, H. (2015). *Modifikatoren im Deutschen: Ihre Klassifizierung und varietätenspezifische Verwendung*. Stauffenburg, Tübingen.

Hoey, M. (2012). *Lexical Priming: A New Theory of Words and Language*. Routledge, London, DOI: 10.4324/9780203327630.

Holland, J. H., Gong, T., Minett, J., Ke, J., et al. (2005). Language Acquisition as a Complex Adaptive System. *Language acquisition, change and emergence*, pages 411–435.

Howarth, P. (1998). Phraseology and Second Language Proficiency. *Applied Linguistics*, 19(1):24–44, DOI: 10.1093/applin/19.1.24.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, DOI: 10.1109/MCSE.2007.55.

James, L. E. and Burke, D. M. (2000). Phonological Priming Effects on Word Retrieval and Tip-of-the-Tongue Experiences in Young and Older Adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6):1378–1391, DOI: 10.1037/0278-7393.26.6.1378.

Jelinek, F., Bahl, L., and Mercer, R. (1975). Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech. *IEEE Transactions on Information Theory*, 21(3):250–256, DOI: 10.1109/tit.1975.1055384.

Jones, L. L. and Estes, Z. (2012). Lexical Priming: Associative, Semantic, and Thematic Influences on Word Recognition. In Adelman, J. S., editor, *Visual Word Recognition: Meaning and Context, Individuals and Development*, volume 2 of *Current Issues in the Psychology of Language*, pages 44–72. Psychology Press, London, DOI: 10.4324/9780203106976.

Kilgarriff, A. (2005). Language Is Never, Ever, Ever, Random. *Corpus Linguistics and Linguistic Theory*, 1(2):263–276, DOI: 10.1515/cllt.2005.1.2.263.

Koplenig, A. (2017). Against Statistical Significance Testing in Corpus Linguistics. *Corpus Linguistics and Linguistic Theory*, 15:321–346, DOI: 10.1515/cllt-2016-0036.

Krause, T. (2019). *ANNIS: A Graph-based Query System for Deeply Annotated Text Corpora*. PhD Thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, DOI: 10.18452/19659.

Krivanek, J. and Meurers, D. (2011). Comparing Rule-based and Data-driven Dependency Parsing of Learner Language. In Gerdes, K., Hajičová, E., and Wanner, L., editors, *Computational Dependency Theory*, pages 207–225. IOS press.

Kuang, D., Choo, J., and Park, H. (2015). Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering. In *Partitional Clustering Algorithms*, pages 215–243. Springer, Cham.

Kuno, S. and Takami, K.-i. (2004). *Functional Constraints in Grammar: On the Unergative–unaccusative Distinction*. John Benjamins, https://www.jbe-platform.com/content/books/9789027295217.

Leibniz-Institut für Deutsche Sprache (2019). Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2019-I (Release vom 18.03.2019). http://www.ids-mannheim.de/DeReKo.

Lerner, R. M. (2012). Developmental Science: Past, Present, and Future. *International Journal of Developmental Science*, 6(1-2):29–36, DOI: 10.3233/DEV-2012-12102.

Linck, J. A. and Cunnings, I. (2015). The Utility and Application of Mixed-Effects Models in Second Language Research. *Language Learning*, 65(S1):185–207, DOI: 10.1111/lang.12117.

Loebell, H. and Bock, K. (2003). Structural Priming across Languages. *Linguistics*, 41(5):791–824, DOI: 10.1515/ling.2003.026.

Lowie, W. M. and Verspoor, M. H. (2018). Individual Differences and the Ergodicity Problem. *Language Learning*, 69:184–206, DOI: 10.1111/lang.12324.

Lucas, M. (2000). Semantic Priming without Association: A Meta-Analytic Review. *Psychonomic Bulletin & Review*, 7(4):618–630.

Luce, P. A., Goldinger, S. D., Auer, E. T., and Vitevitch, M. S. (2000). Phonetic Priming, Neighborhood Activation, and PARSYN. *Perception &psychophysics*, 62(3):615–625, DOI: 10.3758/bf03212113.

Lüdeling, A., Hirschmann, H., and Shadrova, A. (2017). Linguistic Models, Acquisition Theories, and Learner Corpora: Morphological Productivity in SLA Research Exemplified by Complex Verbs in German. *Language Learning*, 67(S1):96–129, DOI: 10.1111/lang.12231.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K., and Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 45(2):67, DOI: 10.37307/j.2198-2430.2008.02.02.

Lüdeling, A., Hirschmann, H., Shadrova, A., and Wan, S. (2021). *Deutsch in Europa: Sprachpolitisch, grammatisch, methodisch*, chapter Tiefe Analyse von Lernerkorpora, pages 235–284. De Gruyter, DOI: doi:10.1515/9783110731514-013.

Massip-Bonet, À. (2013). *Language as a Complex Adaptive System: Towards an Integrative Linguistics*, pages 35–60. Springer, Berlin/Heidelberg, DOI: 10.1007/978-3-642-32817-6_4.

McNamara, T. P. (2005). *Semantic Priming: Perspectives from Memory and Word Recognition*. Psychology Press, DOI: 10.4324/9780203338001.

Medaglia, J. D., Ramanathan, D. M., Venkatesan, U. M., and Hillary, F. G. (2011). The Challenge of Non-ergodicity in Network Neuroscience. *Network: Computation in Neural Systems*, 22(1-4):148–153, DOI: 10.3109/09638237.2011.639604.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in neural information processing systems*, NIPS'13, pages 3111–3119, Red Hook, NY. Curran Associates Inc.

Molenaar, P. C. M. (2008). On the Implications of the Classical Ergodic Theorems: Analysis of Developmental Processes Has to Focus on Intra-individual Variation. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 50(1):60–69, DOI: 10.1002/dev.20262.

Morey, M., Muller, P., and Asher, N. (2018). A Dependency Perspective on Rst Discourse Parsing and Evaluation. *Computational Linguistics*, 44(2):197–235, DOI: 10.1162/coli_a_00314.

Müller, S. (2013). Unifying Everything. *Language*, 89(4):920–950, https://hpsg.hu-berlin.de/~stefan/Pub/unifying-everything.html.

Müller, S. and Wechsler, S. (2014). Lexical Approaches to Argument Structure. *Theoretical Linguistics*, 40(1-2):1–76, DOI: 10.1515/tl-2014-0001.

Nivre, J., Hall, J., and Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2006/pdf/162_pdf.pdf.

Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. https://arxiv.org/abs/1403.2805.

Ott, N. and Ziai, R. (2010). Evaluating Dependency Parsing Performance on German Learner Language. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, volume 9, pages 175–186. NEALT Tartu.

Papo, D. (2013). Why Should Cognitive Neuroscientists Study the Brain's Resting State? *Frontiers in Human Neuroscience*, 7:45, DOI: 10.3389/fnhum.2013.00045.

Paquot, M. and Granger, S. (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, 32:130–149, DOI: 10.1017/S0267190512000098.

Pardo, J. S. (2006). On Phonetic Convergence during Conversational Interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393, DOI: 10.1121/1.2178720.

Pawley, A. and Syder, F. H. (1983). Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In Richards, J. C. and Schmidt, R. W., editors, *Language and Communication*, Language and Communication, pages 191–226. Routledge, London, DOI: 10.4324/9781315836027.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, DOI: 10.3115/v1/d14-1162.

Piantadosi, S. T. (2014). Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, DOI: 10.3758/s13423-014-0585-6.

Pickering, M. J. and Branigan, H. P. (1998). The Representation of Verbs: Evidence from Syntactic Priming in Language Production. *Journal of Memory and language*, 39(4):633–651, DOI: 10.1006/jmla.1998.2592.

Pollard, C. and Sag, I. A. (1994). *Head-driven Phrase Structure Grammar*. University of Chicago Press.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, https://www.R-project.org/.

Reznicek, M., Walter, M., Schmidt, K., Lüdeling, A., et al. (2010). *Das Falko-Handbuch: Korpusaufbau und Annotationen*. Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin.

RStudio Team (2015). *Rstudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, http://www.rstudio.com/.

Sag, I. A., Boas, H. C., and Kay, P. (2012). Introducing Sign-Based Construction Grammar. In Boas, H. C. and Sag, I. A., editors, *Sign-Based Construction Grammar*, pages 1–30. CSLI Publications.

Sauer, S. and Lüdeling, A. (2013). BeMaTaC: A Flexible Multilayer Spoken Dialogue Corpus for Contrastive SLA Analyses. In *ICAME*, volume 34, pages 46–47.

Schmid, H.-J. and Küchenhoff, H. (2013). Collostructional Analysis and Other Ways of Measuring Lexicogrammatical Attraction: Theoretical Premises, Practical Problems and Cognitive Underpinnings. *Cognitive Linguistics*, 24(3):531–577, DOI: `10.1515/cog-2013-0018`.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., et al. (2013). Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, WA. Association for Computational Linguistics, `https://aclanthology.org/W13-4917`.

Shadrova, A. (2020). *Measuring Coselectional Constraint in Learner Corpora: A Graph-based Approach*. Dissertation, Humboldt-Universität zu Berlin.

Speelman, D., Heylen, K., and Geeraerts, D. (2018). *Mixed-Effects Regression Models in Linguistics*. Springer, Cham.

Steels, L. (2000). Language As a Complex Adaptive System. In *Parallel Problem Solving from Nature PPSN VI*, pages 17–26, Berlin/Heidelberg. Springer, DOI: `10.1007/3-540-45356-3_2`.

Steels, L. and Loetzsch, M. (2008). Perspective alignment in spatial language. *arXiv*, (0605012), `https://arxiv.org/abs/cs/0605012`.

Stefanowitsch, A. and Gries, S. T. (2003). Collostructions: Investigating the Interaction of Words and Constructions. *International Journal of Corpus Linguistics*, 8(2):209–243, DOI: `10.1075/ijcl.8.2.03ste`.

Stefanowitsch, A. and Gries, S. T. (2005). Covarying Collexemes. *Corpus linguistics and linguistic theory*, 1(1):1–43.

Wachs-Lopes, G. A. and Rodrigues, P. S. (2016). Analyzing Natural Human Language from the Point of View of Dynamic of a Complex Network. *Expert Systems with Applications*, 45:8–22, DOI: `10.1016/j.eswa.2015.09.020`.

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21(12):1–20, `http://www.jstatsoft.org/v21/i12/`.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, NY, `http://ggplot2.org`.

Wickham, H., François, R., Henry, L., and Müller, K. (2018). *dplyr: A Grammar of Data Manipulation*, `https://CRAN.R-project.org/package=dplyr`. R package version 0.7.6.

Williams, J. R., Lessard, P. R., Desu, S., Clark, E. M., et al. (2015). Zipf's Law Holds for Phrases, Not Words. *Scientific Reports*, 5(1):12209, DOI: `10.1038/srep12209`.

Zeldes, A. (2013). Productive Argument Selection: Is Lexical Semantics Enough? *Corpus Linguistics and Linguistic Theory*, 9(2):263–291, DOI: `10.1515/cllt-2013-0006`.

Zinsmeister, H., Reznicek, M., Brede, J. R., Rosén, C., et al. (2012). Das Wissenschaftliche Netzwerk „Kobalt-DaF“. *Zeitschrift für germanistische Linguistik*, 40(3):457–458, DOI: `10.1515/zgl-2012-0030`.