HOLGER KEIBEL / CYRIL BELICA / MARC KUPIETZ / RAINER PERKUHN

Approaching grammar: Detecting, conceptualizing and generalizing paradigmatic variation

Abstract

This paper presents ongoing research which is embedded in an empirical-linguistic research program, set out to devise viable research strategies for developing an explanatory theory of grammar as a psychological and social phenomenon. As this phenomenon cannot be studied directly, the program attempts to approach it indirectly through its correlates in language corpora, which is justified by referring to the core tenets of Emergent Grammar. The guiding principle for identifying such corpus correlates of grammatical regularities is to imitate the psychological processes underlying the emergent nature of these regularities.

While previous work in this program focused on syntagmatic structures, the current paper goes one step further by investigating schematic structures that involve paradigmatic variation. It introduces and explores a general strategy by which corpus correlates of such structures may be uncovered, and it further outlines how these correlates may be used to study the nature of the psychologically real schematic structures.

1. Introduction

Much of what we linguists call grammar concerns schematic structures, i.e., structures that display some kind of paradigmatic variation. While, as an informal concept, the notion of schematic structures is fairly straightforward and intuitive, the precise nature of such structures as an integral part of language is not at all clear. This paper addresses the following two research questions:

1) On what empirical basis is it justified to infer schematic structures?

2) How can these schemas be captured conceptually?

These questions are pursued here as part of an overarching research program which was outlined by Kupietz and Keibel (Keibel / Kupietz 2009, Kupietz / Keibel 2009b) and is summarized in the next section. It is followed by a brief review of some previous empirical work that the current paper is based on (Section 3). The central Section 4 presents methodological and empirical explorations towards question (1) before a tentative approach to question (2) is outlined in the final section.

2. An empirical linguistic research program

The primary goal of the research program described by Kupietz and Keibel (Keibel / Kupietz 2009, Kupietz / Keibel 2009b) is to devise viable research strategies for developing an explanatory theory of grammar.[11] As the only fundamental assumption, this program adopts the general framework of Emergent Grammar (Hopper 1987, 1998), according to which any grammatical regularities are emergent by nature, being constantly influenced and reshaped by language use. These regularities are ascribed a psychological reality in the form of individual speakers' *language routines*, and these routines arise as a continuous result of each speaker's aggregating language experience. Likewise, the grammatical regularities are attributed a social reality which takes the form of *language conventions* in a language community, and these conventions may be characterized informally as the overlap between the individual grammars (i.e., language routines) of most speakers.

As one immediate consequence of their dual reality, grammatical regularities in turn necessarily shape language use: obviously, speakers routinely use their individual language routines, and in order to ensure successful communication, they are likely to use the conventions of the respective language community. If these general assumptions are valid, one would expect to find correlates of any grammatical regularity in an appropriate corpus of authentic language productions, provided that the corpus is sufficiently large and stratified.

The program proposes to adopt a strictly empirical research strategy which is founded on this prediction. Language routines of individual speakers and language conventions in a community cannot be accessed directly, but one may attempt to access and study them through their putative corpus correlates. As authentic corpus data are lexically specific, the best option to do this is by a bottom-up, inductive approach: to start from individual lexical items and to proceed by incrementally deriving increasingly complex and abstract structures around these items. Given the reciprocal dynamics of an emergent grammar that were described above, many – though not all – abstract regularities may have become psychologically real for most speakers along very similar inductive paths. In other words, the bottom-up strategy that we advocate constitutes an attempt to mimic the inductive psychological processes underlying the emergence of grammatical regularities.

[1] Any progress that we make towards this goal is published in a series of talks and papers with the same running title "Approaching grammar".

In a nutshell, observations at the data level are incrementally generalized to more abstract structures, by means of inductive methods that are motivated by psychological facts and premises. It then has to be demonstrated that the resulting generalizations are, in general, corpus correlates of structures that are psychologically and socially real – a priori, they are merely candidates for such correlates. Exploring a large number of such corpus correlates at any level of abstraction prompts researchers to abductively formulate new hypotheses at the theoretical level. Each of these new hypotheses has to be tested empirically, in terms of deduction and falsification.

This complex, iterative strategy is not targeted at studying a specific phenomenon in a quick, direct fashion. Instead, it is an attempt to approach the very notion of *grammatical structure*, and as its focus is on explanation, centrality is given to empirical facts in at least two ways:

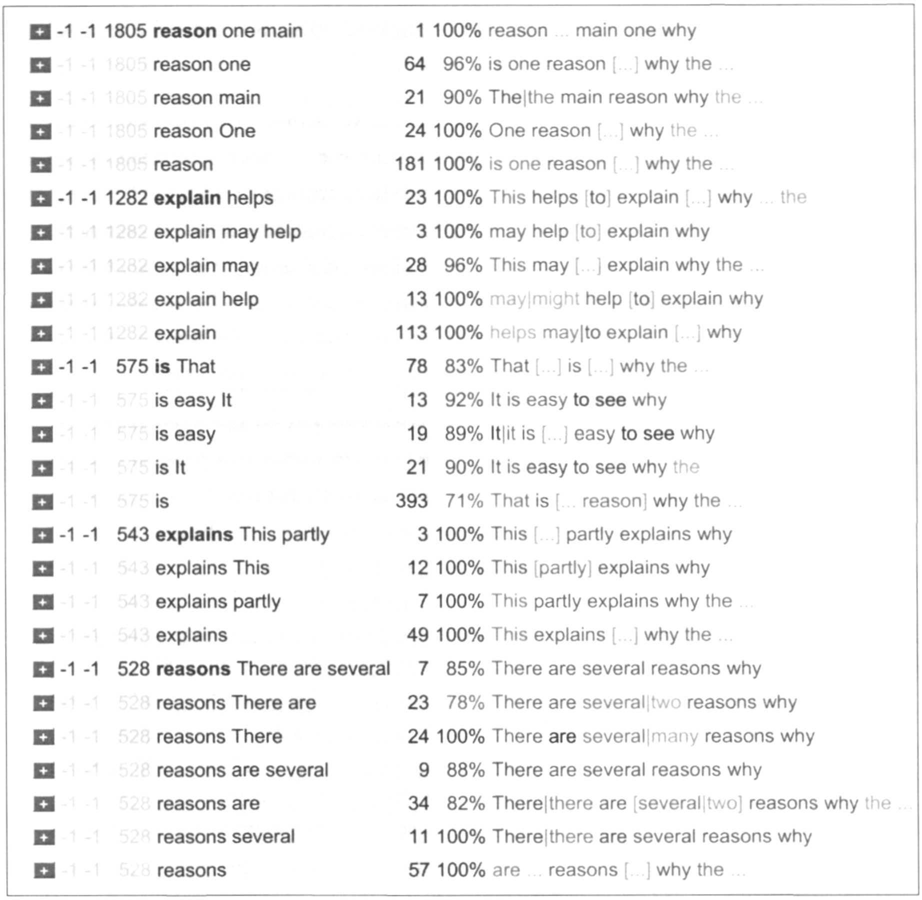
a) induction is driven by empirical observation;

b) any hypotheses derived from the resulting generalizations are tested against empirical data.

3. Previous work

The empirical research reported here crucially builds on previous work (an overview may be found in Keibel / Kupietz / Belica 2008), the primary goal of which was to study the emergent nature of syntagmatic structures that are psychologically and socially real. These structures may be referred to as *psychological collocations* (cf. Hoey 2005). As there is no way to access them directly, we studied them through their putative corpus correlates which take the form of *statistical collocations*. The specific notion of statistical collocation that we used to this end is that of *higher-order collocations* which are detected by processes that are meant to imitate those underlying the psychological collocations. It is a notion that is more flexible than that of n-grams, as the collocates in a higher-order collocation may be non-contiguous, and their order and distances may vary. This concept dates back to 1995 (Belica 1995, Keibel / Belica 2007) and it was rediscovered recently as the very similar, albeit not identical, concept of *concgrams* (Cheng / Greaves / Warren 2006). Due to their positional flexibility, higher-order collocations are sometimes hard to relate to the intuition of competent speakers. Therefore, each higher-order collocation is typically listed together with a *syntagmatic pattern* which summarizes the collocation's most typical word order.

To illustrate these concepts, Figure 2 shows some of the most cohesive higher-order collocations for "*why*" which were derived from a fairly small web-based corpus composed of written English (2.5 million words). Each line corresponds to one higher-order collocation, the collocates are listed in the central column while the right-most column gives the predominant syntagmatic pattern.



⊞	-1	-1	1805	**reason** one main	1	100%	reason ... main one why
⊞	-1	-1	1805	reason one	64	96%	is one reason [...] why the ...
⊞	-1	-1	1805	reason main	21	90%	The|the main reason why the ...
⊞	-1	-1	1805	reason One	24	100%	One reason [...] why the ...
⊞	-1	-1	1805	reason	181	100%	is one reason [...] why the ...
⊞	-1	-1	1282	**explain** helps	23	100%	This helps [to] explain [...] why ... the
⊞	-1	-1	1282	explain may help	3	100%	may help [to] explain why
⊞	-1	-1	1282	explain may	28	96%	This may [...] explain why the ...
⊞	-1	-1	1282	explain help	13	100%	may|might help [to] explain why
⊞	-1	-1	1282	explain	113	100%	helps may|to explain [...] why
⊞	-1	-1	575	**is** That	78	83%	That [...] is [...] why the ...
⊞	-1	-1	575	is easy It	13	92%	It is easy **to see** why
⊞	-1	-1	575	is easy	19	89%	It|it is [...] easy **to see** why
⊞	-1	-1	575	is It	21	90%	It is easy to see why the
⊞	-1	-1	575	is	393	71%	That is [... reason] why the ...
⊞	-1	-1	543	**explains** This partly	3	100%	This [...] partly explains why
⊞	-1	-1	543	explains This	12	100%	This [partly] explains why
⊞	-1	-1	543	explains partly	7	100%	This partly explains why the ...
⊞	-1	-1	543	explains	49	100%	This explains [...] why the ...
⊞	-1	-1	528	**reasons** There are several	7	85%	There are several reasons why
⊞	-1	-1	528	reasons There are	23	78%	There are several|two reasons why
⊞	-1	-1	528	reasons There	24	100%	There are several|many reasons why
⊞	-1	-1	528	reasons are several	9	88%	There are several reasons why
⊞	-1	-1	528	reasons are	34	82%	There|there are [several|two] reasons why the ...
⊞	-1	-1	528	reasons several	11	100%	There|there are several reasons why
⊞	-1	-1	528	reasons	57	100%	are ... reasons [...] why the ...

Figure 2: Collocation profile of "*why*" (only top portion shown)

The *collocation profile* of a given node word is defined as the full spectrum of higher-order collocations around this node word, together with the dominant syntagmatic patterns and some related characteristics. Again, as an illustration, Figure 2 is the top portion of the collocation profile of "*why*". For many higher-level research questions it is useful to have fast access to large numbers of such collocation profiles, and this is also the case for the present study. To this end, we take advantage of the collocation database CCDB (Belica 2001-2007, Keibel / Belica 2007) which currently provides collocation profiles for more than 220 000 node words. These profiles are based on the virtual corpus CCDB2007 with approximately 2.2 billion text words (Institut für Deutsche Sprache 2007b) which was composed as a subset of the German Reference Corpus DeReKo (Institut für Deutsche Sprache 2007a, Kupietz / Keibel 2009a, Kupietz et al. 2010).

It should be pointed out that node words in these collocation profiles are lemmas, whereas their collocates are word forms. We believe that this different treatment of node words and collocates imposes minimal assumptions at the psychological level. On the one hand, the fuzzy denotational and connotational structure associated with the entire paradigm of a word – irrespective of its inflectional properties – seems to be, in general, ontologically more primitive than any of its specific grammatical forms (cf. Belica et al. 2010). The units under investigation should therefore be lemmas. On the other hand, it is only the grammatically expressed word forms that can be, in general, directly observed in language use. Thus, the collocates – as observed properties of the node words – should be assessed at the level of word forms.

4. Detecting schematic structures

For the next inductive step our primary goal was to study the emergent nature of schematic structures that are psychologically and socially real. These structures may be characterized as syntagmatic structures involving paradigmatic variation. The research interest thus is on schemas as emergent psychological phenomena, but just as for psychological collocations in the previous section, there is no way to access them directly. We therefore attempted to study them through their putative corpus correlates. However, unlike the case of collocations, it is not at all clear what these correlates are and how they may be uncovered.

We are not the first ones interested in inducing syntagmatic-paradigmatic structures from corpora and there is a growing body of useful concepts and approaches in the corpus-linguistic literature, including *collocational frameworks* (Renouf / Sinclair 1991), *phrase frames* (Stubbs 2004, Fletcher 2003), *Pattern Grammar* (Hunston / Francis 2000), *local grammar patterns* (Mason

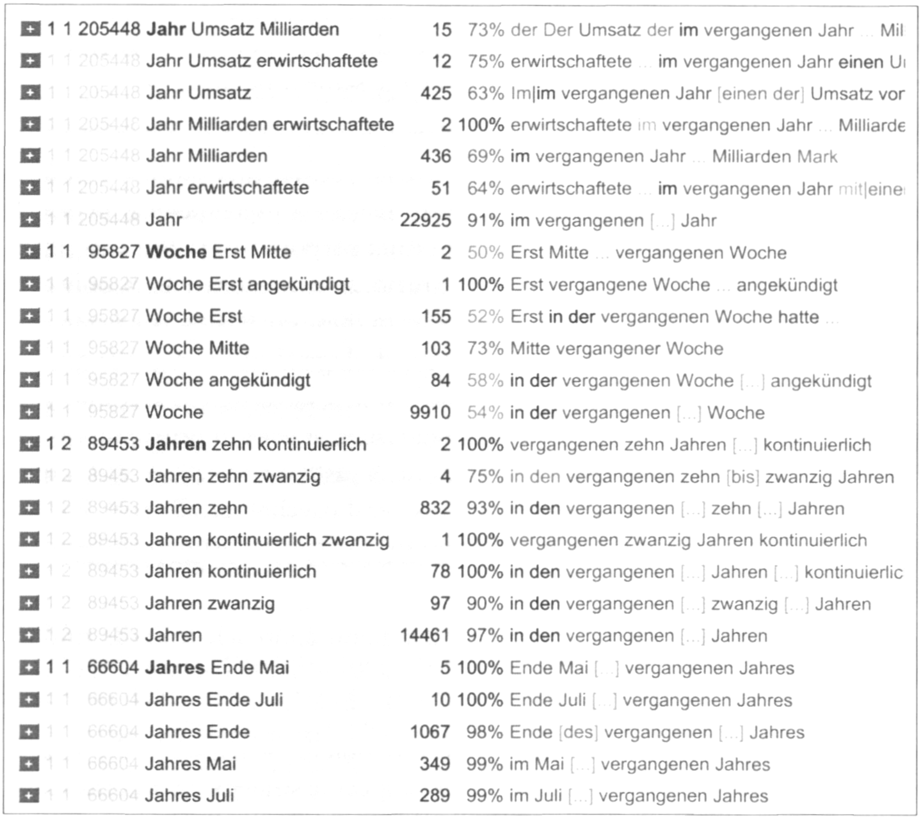
2004), and *formulaic frames* (Biber 2009). We propose a different approach that is motivated by the following rationale. As schemas are syntagmatic-paradigmatic structures, they can be thought of as being instantiated by syntagmatic structures, viz. collocations. The general idea therefore is that schematic structures may in turn be uncovered as abstractions across collocations. This idea is, again, an attempt to imitate the psychological processes underlying the emergence of schematic structures because it is likely that many schematic structures have become psychologically real for most speakers as abstractions across psychological collocations, and that the way they are constantly reshaped in speakers' minds is also driven by the same influence.

Before further outlining the idea, we first need to refine our terminology. When we talk of *schematic structures* or simply of *schemas* without further explication, we refer to structures that are real in a psychological or social sense. As stated before, what can be found from corpora – or in this case, from corpus-induced statistical collocations – are not schemas but, at best, only correlates of schemas. In the following we refer to these correlates as *schema corpus correlates* (short: *SCCs*). However, the psychological and social status of the structures that a given approach induces from collocations is not known a priori. Therefore, until this status has been tested, at least in principle, we call these induced structures *SCC candidates*. It is justified to talk of SCCs only when the induced SCC candidates are in general psychologically real. For the remainder of this paper, it is thus important to strictly distinguish between schemas, SCCs, and SCC candidates.

With these concepts we can formulate a general approach towards finding schemas which consists of three stages. The subject of the first stage is to manually explore collocations for traces of paradigmatic variation in order to obtain some inspiration as to where and how to look for schemas. In the second stage, these observations are used to devise a specific strategy for automatically inducing SCC candidates from collocations. Once such a strategy has been formulated, the general psychological reality of the SCC candidates it induces has to be tested, and these tests constitute the third stage. It is unlikely that the inductive strategy formulated upon the first attempt will be sophisticated enough to detect genuine SCCs, so these tests will prompt one to go back to stage 2 and to revise this strategy until one arrives at a setup that is believed to generally detect genuine SCCs. In the remainder of this section we explore and discuss this three-stage approach for written German, based on the CCDB and the same virtual corpus CCDB2007 as before.

4.1 Stage 1: Exploring collocations for traces of paradigmatic variation

To address the first stage of this strategy, a large number of collocation profiles have to be inspected. For instance, consider the profile of the German adjective and past participle *vergangen* (English: *last*, *past*, *elapsed*), the top portion of which is shown in Figure 3.[2]



⊞	1	1	205448	**Jahr** Umsatz Milliarden	15	73%	der Der Umsatz der **im** vergangenen Jahr ... Mil
⊞	1	1	205448	Jahr Umsatz erwirtschaftete	12	75%	erwirtschaftete ... **im** vergangenen Jahr **einen** U
⊞	1	1	205448	Jahr Umsatz	425	63%	Im|**im** vergangenen Jahr [einen der] Umsatz vor
⊞	1	1	205448	Jahr Milliarden erwirtschaftete	2	100%	erwirtschaftete im vergangenen Jahr ... Milliarde
⊞	1	1	205448	Jahr Milliarden	436	69%	**im** vergangenen Jahr ... Milliarden Mark
⊞	1	1	205448	Jahr erwirtschaftete	51	64%	erwirtschaftete ... **im** vergangenen Jahr mit|eine
⊞	1	1	205448	Jahr	22925	91%	im vergangenen [...] Jahr
⊞	1	1	95827	**Woche** Erst Mitte	2	50%	Erst Mitte ... vergangenen Woche
⊞	1	1	95827	Woche Erst angekündigt	1	100%	Erst vergangene Woche ... angekündigt
⊞	1	1	95827	Woche Erst	155	52%	Erst **in der** vergangenen Woche **hatte** ...
⊞	1	1	95827	Woche Mitte	103	73%	Mitte vergangener Woche
⊞	1	1	95827	Woche angekündigt	84	58%	**in der** vergangenen Woche [...] angekündigt
⊞	1	1	95827	Woche	9910	54%	**in der** vergangenen [...] Woche
⊞	1	2	89453	**Jahren** zehn kontinuierlich	2	100%	vergangenen zehn Jahren [...] kontinuierlich
⊞	1	2	89453	Jahren zehn zwanzig	4	75%	in den vergangenen zehn [bis] zwanzig Jahren
⊞	1	2	89453	Jahren zehn	832	93%	**in den** vergangenen [...] zehn [...] Jahren
⊞	1	2	89453	Jahren kontinuierlich zwanzig	1	100%	vergangenen zwanzig Jahren kontinuierlich
⊞	1	2	89453	Jahren kontinuierlich	78	100%	**in den** vergangenen [...] Jahren [...] kontinuierlic
⊞	1	2	89453	Jahren zwanzig	97	90%	**in den** vergangenen [...] zwanzig [...] Jahren
⊞	1	2	89453	Jahren	14461	97%	**in den** vergangenen [...] Jahren
⊞	1	1	66604	**Jahres** Ende Mai	5	100%	Ende Mai [...] vergangenen Jahres
⊞	1	1	66604	Jahres Ende Juli	10	100%	Ende Juli [...] vergangenen Jahres
⊞	1	1	66604	Jahres Ende	1067	98%	Ende [des] vergangenen [...] Jahres
⊞	1	1	66604	Jahres Mai	349	99%	im Mai [...] vergangenen Jahres
⊞	1	1	66604	Jahres Juli	289	99%	im Juli [...] vergangenen Jahres

Figure 3: Collocation profile of *vergangen* (only top portion shown)

By scanning this collocation profile, one may uncover traces of paradigmatic variation around this word *vergangen*. However, doing so for a profile that is represented as a simple list is not very efficient, and, more importantly, a lot of paradigmatic variation may be missed in this way. The evidence for some paradigmatic structure is often scattered across the profile which in turn is gener-

[2] The full profile may be inspected at: http://corpora.ids-mannheim.de/ccdb/.

ally rather large. A more systematic approach is needed for this exploratory stage, and ideally this involves the possibility of progressively recording any evidence for paradigmatic variation as it is encountered. To this end, we took advantage of the collocation explorer VICOMTE (Perkuhn 2007).

Figure 4 shows the same collocation profile of *vergangen* in VICOMTE's default visualization. The node word is displayed in the center and its primary collocates are given on the inner-most circle around the node word, sorted by decreasing cohesion (*Jahr*, *Woche*, *Jahren*, ...). In order to keep the visualization simple, only five primary collocates are displayed at full size while the others appear miniaturized. VICOMTE offers several interactive ways of inspecting all regions of the collocation profile at normal size (e.g., mousing over the respective boxes or rotating the entire tree diagram). Secondary collocates appear attached to the corresponding primary collocates, and ternary collocates are in turn attached to their corresponding secondary collocate, and so forth. Like this, any higher-order collocation is represented by a unique radial path connecting the node with collocates on the various circles.

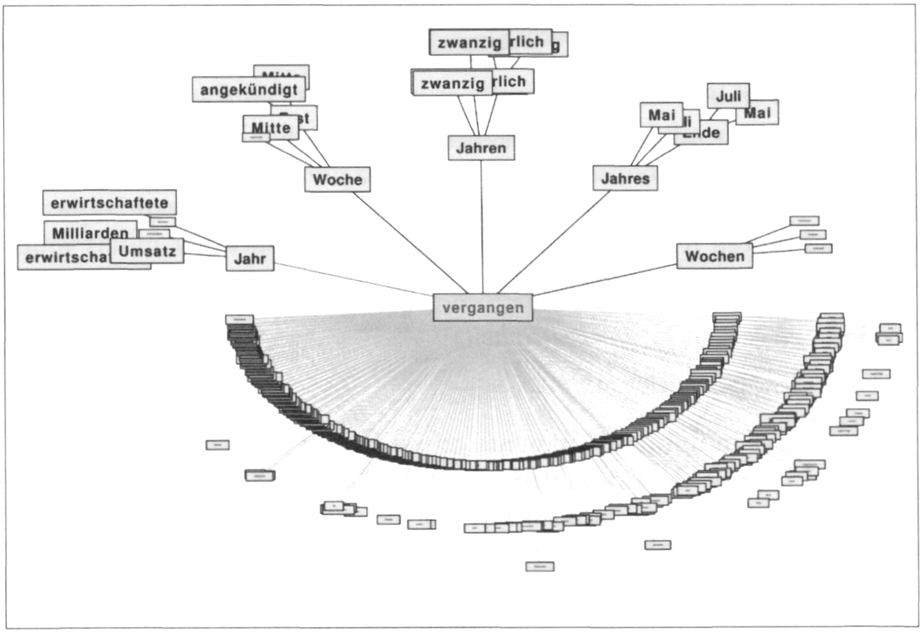


Figure 4: Collocation profile of *vergangen* (VICOMTE visualization)

To keep the explorations simple, we only look at primary collocates (in the visualization on the inner circle). Scanning through this single profile, we encounter a lot of evidence for schematic structures around this node word. For example, many of the primary collocates of *vergangen* refer to larger units of time such as *Jahr*, *Jahrzehnt*, *Jahrhundert*, *Monat*, *Woche*, *Periode*, etc. (English: *year*, *decade*, *century*, *month*, *week*, *period*, etc.). All words of this group found in the profile are highlighted in Figure 5.

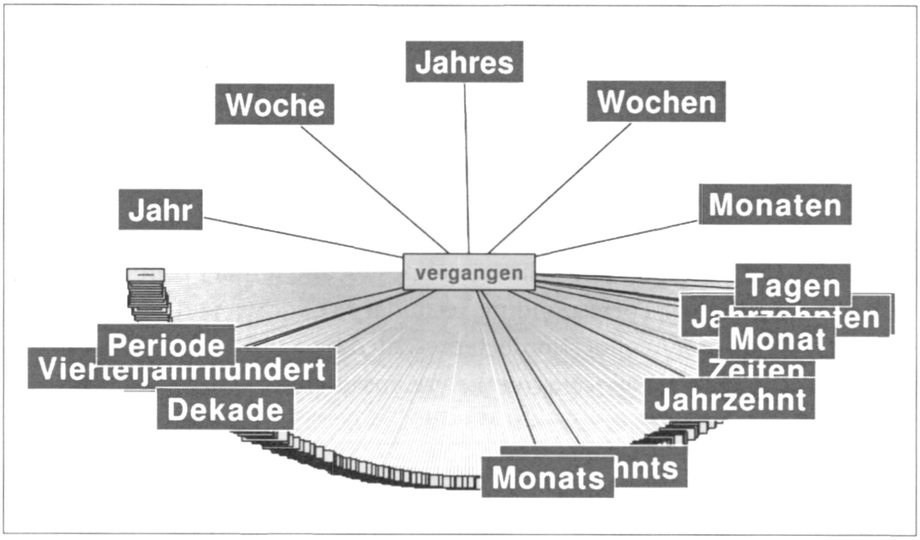


Figure 5: Profile of *vergangen*: primary collocates referring to units of time are highlighted

Based on our competence as speakers of German, we believe that this group of collocates indeed relates to a schematic structure in speakers' minds which might be summarized as follows:

(1) *vergangen*$_{+inflection}$ *<Zeiteinheit>*
last / past <unit of time>

The collocates in this description are generalized to a *paradigmatic class* (represented here by the placeholder variable *<Zeiteinheit>*). Importantly, this description is merely an intuitive label for the putative schema (and its paradigmatic class), but not necessarily the schema itself. An adequate representation of the full schema is likely to be more complex, e.g., involving relations between its components and so forth. Given the goals of this research, it seems advisable not to make any a-priori assumptions about the representation of schematic structures. Therefore, a label as in (1) is to be understood only as an intuitive shorthand for a schema, SCC or SCC candidate.

It should also be stressed that the collocations of *vergangen* with the group of collocates highlighted in Figure 5 constitute good evidence for a possible schema not because there is a nice descriptive label for it but because there appears to be some more general structure underlying these collocations in speakers' minds. In other words, there is a competence-based response in speakers who are subsequently exposed to collocations such as *vergangene Woche* and *vergangenes Jahr* which triggers schematic entities in their implicit language knowledge. It does not matter whether or not speakers are able to make this knowledge explicit – what matters is the response itself.

In many cases, therefore, it is difficult to capture the essence of a paradigmatic class of collocates by a concise label such as <*Zeiteinheit*>. Therefore, to be able to describe all putative schemas, SCCs and SCC candidates in the same universal way, we use non-interpretative labels such as (2).

(2) *vergangen*$_{+inflection}$ {*Jahr, Jahrzehnt, Jahrhundert, Monat, Woche, Periode, ...*}
last / past {year, decade, century, month, week, period, ...}

To record this particular evidence, we restructure and annotate the VICOMTE representation such that the collocates referring to units of time are grouped together (Figure 6). In this fashion, we continue to scan the remaining profile and to record any indications of paradigmatic variation by grouping together the respective collocates to a putative paradigmatic class.

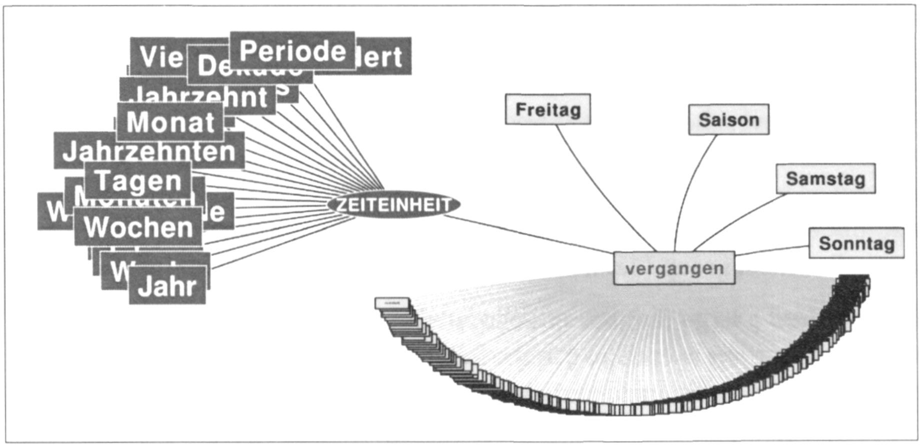


Figure 6: Annotated profile of *vergangen*

In our explorations for the specific profile of *vergangen* we identified a large number of likely paradigmatic classes of collocates. Some examples are the groups listed in (3).

(3a) *Montag, Dienstag, ..., Samstag, Sonnabend, Sonntag*
Monday, Tuesday, ..., Saturday, Sunday

(b) *zwei, fünf, sieben, zwölf, fünfzehn, anderthalb, ...*
two, five, seven, twelve, fifteen, one and a half, ...

(c) *Gewinn, Verlust, Umsatz, ...*
profit, loss, revenue, ...

(d) *gestiegen, gewachsen, zugelegt, zurückgegangen, gesunken, ...*
increased, gained, decreased, dropped, ...

(e) *deutlich, erheblich, drastisch, kräftig, stark, ...*
considerably, substantially, drastically, strongly, ...

Some of these putative paradigmatic classes in conjunction with the node word *vergangen* do not readily relate to any intuitive structures. However, inspecting the range of underlying concordances of the individual collocations does prompt respective structures in competent speakers, albeit this response may be weaker than for the earlier example (2). For instance, collocations underlying the paradigmatic classes (3d) and (3e) are often instantiated in sentence fragments such as

(4) *... in den vergangenen Jahren deutlich gestiegen ...*
... over the past few years considerably increased ...
... increased considerably over the past few years ...

and likewise for any collocate of group (3d) in the position of *gestiegen*, and any collocate of group (3e) instead of *deutlich*.

We explored a large number of collocation profiles in the same fashion, and these explorations overall lead to the following general observations. First, any evidence for paradigmatic variation that we observed for a fixed node word involved a group of collocates that are semantically fairly similar, where similarity is assessed in terms of intuitive speaker judgments. This refers to a non-categorial but rather associative psychological notion of semantic similarity (cf. Belica et al. 2010). Second, the collocates in each such paradigmatic class tended to belong to the same lexical category. Third, where applicable, the collocates in each such paradigmatic class were often observed to share morphosyntactic features. For instance, in example (3d), all collocates grouped together were past participles in their basic form (i.e., not inflected as adjectives). Fourth, the collocations underlying each such paradigmatic class display very similar positional preferences. That is, the different collocates tend to occur in (nearly) the same position relative to the node word.

To sum up, the explorations so far suggest that simple two-word collocations for a fixed node word may be good candidates for relating to an underlying schematic structure if the respective collocates are similar in terms of their associative semantics and their positional preferences relative to the node word (observations 1 and 4). In particular, they suggest that the paradigmatic classes underlying the schemas of a node word are no language-general word classes, but rather specific to this node word, if not specific to the individual schemas (very much as in Construction Grammar, especially in the approach by Croft 2001). Note, however, that, due to our methodological framework, the collocates' agreement with respect to lexical categories and morphosyntactic features (i.e., observations 2 and 3) does not qualify as additional criterion for detecting schemas. Instead, these two observations may be construed as epiphenomena of the other two observations (cf. 4.2.1).

4.2 Stage 2: Inducing SCC candidates

With respect to automatically inducing SCC candidates, the results of the exploratory stage 1 so far are instructive in several ways. First of all, they suggest that the task may be simplified by splitting it into two subtasks (cf. Figure 7). The first subtask is to identify for a given node word – on the basis of its collocation profile – the paradigmatic classes that appear to be relevant for its schemas. Once this is done, the resulting paradigmatic classes may be used to derive from the same corpus a range of SCC candidates for this node word which constitutes the second subtask.

In the following, we briefly describe a possible way of accomplishing these two tasks. The specific operationalizations given below are only of secondary importance because our epistemic goal is to imitate and model the inductive processes underlying the emergent nature of schematic structures in language. We approach this goal here by mimicking the competence-based procedures in 4.1.

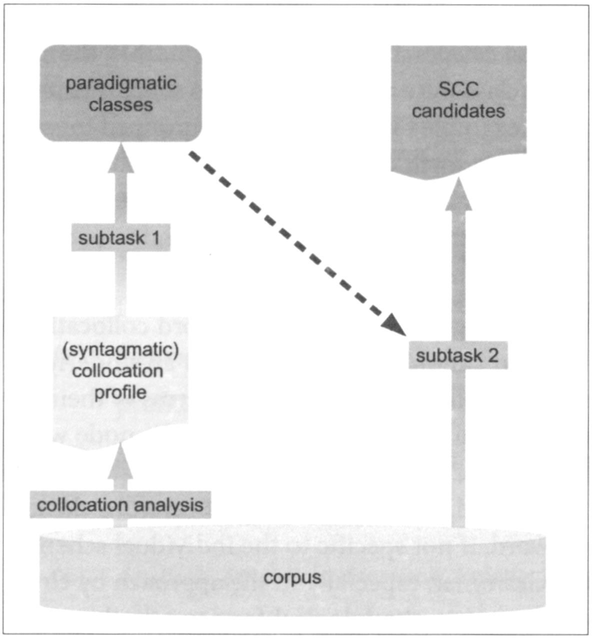


Figure 7: Two subtasks for inducing SCC candidates

4.2.1 Subtask 1: Inducing paradigmatic classes for a node word

The basic idea for accomplishing subtask 1 is to group together all primary collocates of a given node word that are sufficiently "similar", and to use a notion of similarity that imitates the four intuitive similarity criteria observed in Subsection 4.1. However, it turns out that the second and third of these observations cannot be exploited here for at least two reasons. First, they involve preexisting linguistic categories (lexical categories and morphosyntactic features) and therefore unnecessary theoretical assumptions which should be avoided given the explanatory objectives of this research program (cf. Section 2). Second, even if linguistic categories were admitted in this program, available taggers and parsers are usually not reliable enough, especially for the less frequent phenomena. This imperfect reliability would be much less problematic if it constituted evenly distributed statistical noise in the classifications – but as it involves systematic errors, the scientifically sound use of taggers and parsers generally requires time-consuming manual intervention (cf. Belica et al. in this volume) which would be too costly for the kind of research pursued here.

By contrast, the two remaining observations of Subsection 4.1 – associative-semantic and positional similarity – are fully valid concepts in this program. Both are meaningful constitutive criteria constraining the systematic search for realistic paradigmatic classes. While this is immediately obvious for the criterion of associative-semantic similarity, the following example is intended to demonstrate it also for positional similarity. Among the primary collocates of the node word *Haar* (English: hair), there are many color adjectives (in various inflected forms), including the German counterparts of *red*, *white*, *snow white*, *black*, *blonde*, *salt-and-pepper*, *brunette*, etc. Surprisingly, however, *Haar* also collocates with different forms of *blau* (English: blue), which is an unlikely hair color – at least not likely enough for the word to be traceable as a significant collocate of *Haar*. Closer inspection of the underlying concordances reveals that this collocation is mainly due to instances of the phrase (5a), or some variant of it, where *blau* is not used to refer to an attribute of *Haar*. An inductive reasoning guided by associative-semantic similarity alone would probably face difficulties to distinguish *blau* (in its various forms) from those color adjectives that are in fact used significantly as attributes of *Haar*, as in example (5b). Incorporating information about the typical word position of the collocate (relative to the node word) constrains associative-semantic induction and helps to induce appropriate paradigmatic classes.

(5a) *blonde Haare und blaue Augen*
blonde hair and blue eyes

(5b) *mit roten Haaren*
with red hair

(5c) *ihr Haar ist rot* [*gefärbt*]
her hair is [dyed] red

Note that this information also helps to distinguish predicative uses (5c) from attributive uses (5b) of the same color adjective, provided that they are different word forms (which is generally the case in German). This points to the more general observation that the positional preferences of a collocate, relative to its node word, often correlate with lexical categories and morphosyntactic features. In other words, although lexical classes and morphosyntactic features were excluded from our theoretical assumptions, the underlying type of information seems to remain available in this approach.

In the appendix we briefly describe how we operationalized the two notions of associative-semantic and positional similarity between collocates, and how both may be integrated into a single similarity measure which can be thought of as quantifying the overall *paradigmatic similarity* between any two collocates.[3]

In order to complete subtask 1, one additionally needs a way to group paradigmatically similar collocates into paradigmatic classes. Of course, it is easy to find for each individual collocate *x* the set of other collocates that are paradigmatically most similar to *x*, but it is a nontrivial optimization problem to partition the set of collocates into classes such that all collocates simultaneously are sufficiently *happy* with the other words that they are grouped with. The human mind is extremely proficient at this kind of optimization problem and performs it all the time, but it is not at all clear how this human skill could best be imitated. *Self-organizing methods* such as *hierarchical cluster analysis* are probably a good starting point. When applied to the present situation, such methods produce tree diagrams (so-called *dendrograms*) which represent the global similarity structure of all collocates. For instance, clustering the primary collocates of the node word *vergangen* produces the following dendrogram (the specific cluster analysis method is irrelevant here).

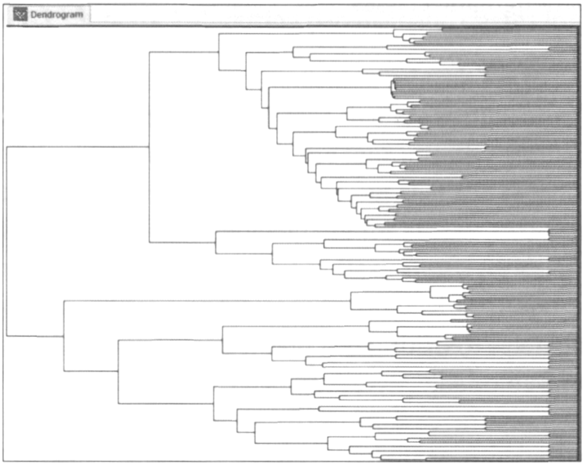


Figure 8: Dendrogram for the collocates of *vergangen* (collocates not shown)

[3] For the purposes of this study, nothing crucial hinges on these particular operationalizations – in fact, there may be more appropriate ones – they are merely tentative proposals and only serve to provide a proof of concept for the general research strategy.

Each line to the right represents one collocate (the collocates themselves are not shown to avoid overcrowding the figure), and the tree structure (from right to left) visualizes how similar collocates are successively merged to increasingly large clusters. The more to the right they are merged, the more similar they are.

One straightforward way of deriving (hypotheses on) paradigmatic classes from such a dendrogram would be to cut the tree at a given similarity level and delete the structure to the left of this level. The remaining clusters of collocates would then be interpreted as the relevant paradigmatic classes: the higher (i.e., further to the left) the cut-off level, the greater and more general the derived classes. For the above dendrogram, an intermediate cut-off level yielded, among others, the following putative paradigmatic classes for *vergangen* which contain names of weekdays (6a), months (6b), cardinal numerals (6c), past participles expressing decrease or increase (6d), and intensifying degree adverbs (6e) respectively (square brackets give each collocate's positional preference relative to *vergangen*, cf. Appendix). These examples correspond closely to the competence-based paradigmatic classes (cf. 4.1). Note that the collocates in these examples are listed in the order in which they appear in the dendrogram such that the class-internal similarity structure is still partly reflected in these lists.

(6a) *Freitagabend, Samstagabend, Freitag, Dienstag, Montag, Donnerstag, Mittwoch, Sonntag, Samstag, Sonnabend* (all: [1;1])

(b) *Mai, Oktober, Juli, März, November, September, August, Juni, April, Februar* (all: [-1;1])

(c) *zwei, drei, vier, fünf, sechs, neun, sieben, zehn, zwölf, fünfzig, fünfzehn, zwanzig, halben, zweieinhalb, anderthalb, eineinhalb* (all: [1;1])

(d) *gesunken, gestiegen, angestiegen, zurückgegangen, zugenommen, gewachsen, zugelegt* (all: [2;5])

(e) *kontinuierlich, stetig, dramatisch, drastisch, deutlich, erheblich, kräftig, stark* (all: [2;3])

We derived paradigmatic classes for a range of other node words in the same fashion and found the results to be highly plausible in almost all cases. In short, although the particular operationalizations we chose for these explorations are in part rather provisional, the explorations indicate that the proposed strategy for subtask 1 is a good starting point for identifying potential paradigmatic classes that may be relevant in the schemas of a given node word.

An open issue is how to determine the *optimal* – i.e., psychologically most adequate – cut-off level which has direct consequences on the size and degree of abstraction of the resulting paradigmatic classes. Moreover, the optimal cut-off level is most likely not constant for all branches of the tree.

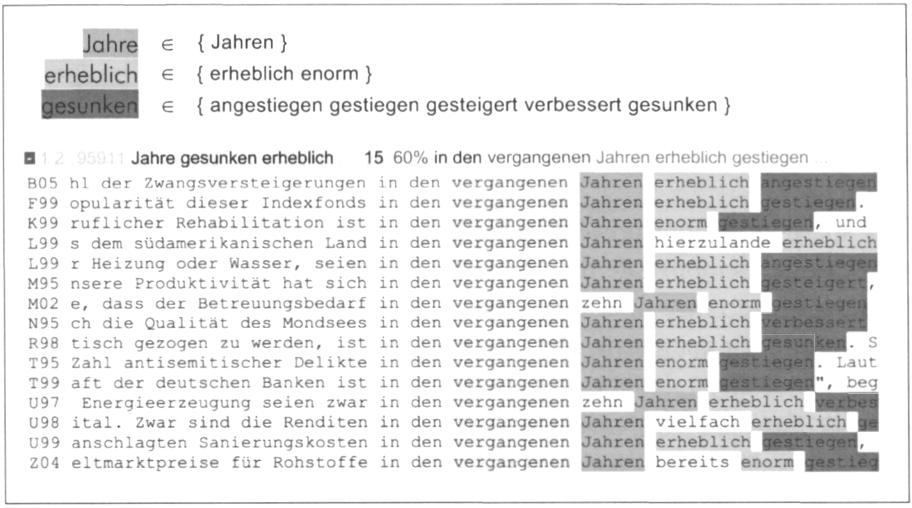
Importantly, the general strategy does not presuppose any predefined language-general classes but empirically derives classes that are potentially node-specific, and this is in line with the general observation in Subsection 4.1. In other words, this strategy supports the possibility that, for different node words, the same collocate word may belong to very different paradigmatic classes. However, as is illustrated in 4.2.2, the realistic classes are not only node-specific but sometimes even schema-specific – the same collocate may generalize to a different paradigmatic class in different SCC candidates around the same node word. To extend the strategy for subtask 1 to also capture this possibility, one would have to allow for a collocate to be a member of multiple classes.

4.2.2 Subtask 2: Deriving SCC candidates for a node word

In this subsection we propose a way by which the paradigmatic classes identified for a given node word by the first subtask may be used to derive SCC candidates for the same node word. The underlying corpus should be the same as the one from which paradigmatic classes were derived – in our explorations, this was again the virtual corpus CCDB2007.

Fortunately, this second subtask may be addressed in a fairly straightforward way, by exploiting an existing methodology that is already well-established within this research program. The basic idea is that SCCs are much like statistical collocations, except that at least one of their collocates is not a specific word but a whole paradigmatic class of words. Therefore, SCC candidates may be detected by re-using the same collocation algorithm as in Section 3, but this time, the paradigmatic classes (identified in subtask 1) are treated as potential collocates, as if the members of each such class were the same word. Like this, the node word may potentially be found to collocate with the classes and any other words that do not belong to these classes. The result is a new type of collocation profile which may be called *syntagmatic-paradigmatic collocation profile*. Some collocations in this profile involve one or multiple paradigmatic classes – these collocations are the SCC candidates proper – while others are entirely lexically specific and are simply higher-order collocations that were found before.

We explored this general strategy again for a range of node words. Figure 9 shows one of the SCC candidates we found for the node word *vergangen*, together with a fraction of the underlying concordances. This SCC candidate is a syntagmatic-paradigmatic collocation consisting of the node word itself, a lexically specific collocate *Jahren* (English: years), a second collocate which is a paradigmatic class {erheblich, enorm} (English: substantially, enormously), and as a third collocate the larger class {*angestiegen*, *gestiegen*, *gesteigert*, *verbessert*, *gesunken*} (English: increased, raised, improved, dropped).



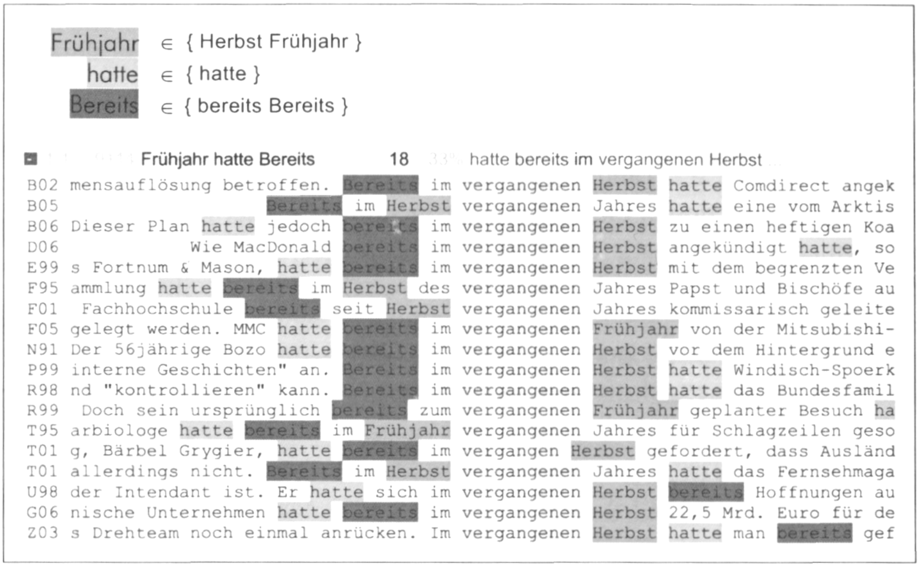
```
     Jahre  ∈  { Jahren }
 erheblich  ∈  { erheblich enorm }
  gesunken  ∈  { angestiegen gestiegen gesteigert verbessert gesunken }

■ 1.2 .95911 Jahre gesunken erheblich   15  60% in den vergangenen Jahren erheblich gestiegen ...
B05 hl der Zwangsversteigerungen in den vergangenen Jahren erheblich angestiegen
F99 opularität dieser Indexfonds in den vergangenen Jahren erheblich gestiegen.
K99 ruflicher Rehabilitation ist in den vergangenen Jahren enorm gestiegen, und
L99 s dem südamerikanischen Land in den vergangenen Jahren hierzulande erheblich
L99 r Heizung oder Wasser, seien in den vergangenen Jahren erheblich angestiegen
M95 nsere Produktivität hat sich in den vergangenen Jahren erheblich gesteigert,
M02 e, dass der Betreuungsbedarf in den vergangenen zehn Jahren enorm gestiegen
N95 ch die Qualität des Mondsees in den vergangenen Jahren erheblich verbessert
R98 tisch gezogen zu werden, ist in den vergangenen Jahren erheblich gesunken. S
T95 Zahl antisemitischer Delikte in den vergangenen Jahren enorm gestiegen. Laut
T99 aft der deutschen Banken ist in den vergangenen Jahren enorm gestiegen", beg
U97  Energieerzeugung seien zwar in den vergangenen zehn Jahren erheblich verbes
U98 ital. Zwar sind die Renditen in den vergangenen Jahren vielfach erheblich ge
U99 anschlagten Sanierungskosten in den vergangenen Jahren erheblich gestiegen,
Z04 eltmarktpreise für Rohstoffe in den vergangenen Jahren bereits enorm gestieg
```

Figure 9: One SCC candidate for the lemma *vergangen* and some underlying concordances

As can be seen in the figure, the word order of the collocates is highly predictable in this example. In other cases, we observed a greater positional variability, as in the following example (Figure 10).

These figures are no full-scale representations of the identified SCC candidates. They only list the relevant paradigmatic classes and some underlying concordances which makes it easier for analyzers to refer to and talk about the SCC candidates, and to relate them to their language competence, but this kind of description does not fully capture the essence of the SCC candidate (cf. 4.1). Developing an adequate representation will be an important direction for future research (cf. Section 5).



```
Frühjahr  ∈ { Herbst Frühjahr }
   hatte  ∈ { hatte }
 Bereits  ∈ { bereits Bereits }

■ Frühjahr hatte Bereits        18   33%  hatte bereits im vergangenen Herbst
B02 mensauflösung betroffen. Bereits im vergangenen Herbst hatte Comdirect angek
B05                   Bereits im Herbst vergangenen Jahres hatte eine vom Arktis
B06 Dieser Plan hatte jedoch bereits im vergangenen Herbst zu einen heftigen Koa
D06            Wie MacDonald bereits im vergangenen Herbst angekündigt hatte, so
E99 s Fortnum & Mason, hatte bereits im vergangenen Herbst mit dem begrenzten Ve
F95 ammlung hatte bereits im Herbst des vergangenen Jahres Papst und Bischöfe au
F01  Fachhochschule bereits seit Herbst vergangenen Jahres kommissarisch geleite
F05 gelegt werden. MMC hatte bereits im vergangenen Frühjahr von der Mitsubishi-
N91 Der 56jährige Bozo hatte bereits im vergangenen Herbst vor dem Hintergrund e
P99 interne Geschichten" an. Bereits im vergangenen Herbst hatte Windisch-Spoerk
R98 nd "kontrollieren" kann. Bereits im vergangenen Herbst hatte das Bundesfamil
R99  Doch sein ursprünglich bereits zum vergangenen Frühjahr geplanter Besuch ha
T95 arbiologe hatte bereits im Frühjahr vergangenen Jahres für Schlagzeilen geso
T01 g, Bärbel Grygier, hatte bereits im vergangen Herbst gefordert, dass Ausländ
T01 allerdings nicht. Bereits im Herbst vergangenen Jahres hatte das Fernsehmaga
U98 der Intendant ist. Er hatte sich im vergangenen Herbst bereits Hoffnungen au
G06 nische Unternehmen hatte bereits im vergangenen Herbst 22,5 Mrd. Euro für de
Z03 s Drehteam noch einmal anrücken. Im vergangenen Herbst hatte man bereits gef
```

Figure 10: Another SCC candidate for the lemma *vergangen* and some underlying concordances

Only space limitations prevent us from presenting more examples which together would demonstrate that the general strategy for subtask 2 indeed addresses a broad range of SCC candidates that vary in manifold respects, e.g., concerning their complexity (i.e., their number of collocates), their degree of abstraction (i.e., the size of their paradigmatic classes), or the distance between the collocates. All of these aspects suggest that SCC candidates may potentially delve deeply into what is commonly perceived as grammatical (rather than lexical) structure.

In interpreting these results, it is important to keep in mind that SCC candidates are derived from information that is entirely contained in the corpus. No external information is involved in the strategies for subtasks 1 and 2, especially no language competence (except for exploration and evaluation purposes). The relevant information is implicitly present in the corpus, distributed across many usage events, and the approach that we propose here is meant to uncover this hidden information by exclusively employing techniques that plausibly imitate the psychological processes underlying the acquisition and continuous emergence of schemas in language.

As a final remark on subtask 2, it is worthwhile pointing out that re-using collocation analysis for deriving SCC candidates is not only convenient but also a way of avoiding over-generalization. For instance, suppose the result of subtask 1 included the following two paradigmatic classes for *vergangen*:

(7a) *Montag, Dienstag, ...*
Monday, Tuesday, ...

(b) *Woche, Monat(s), Jahr(es), ...*
week, month, year, …

Given these classes, a naive solution to subtask 2 might infer from higher-order collocations (or more precisely: from syntagmatic patterns) of the form (8a) that a label such as (8b) constitutes a good SCC candidate.

(8a) *Montag vergangener Woche*
Monday last week

(b) {*Montag, Dienstag, ...*} *vergangene(r|n)* {*Woche, Monats, Jahres, ...*}
{Monday, Tuesday, ...} last {week, month, year, ...}

(c) {*Montag, Dienstag, ...*} *vergangener Woche*
{Monday, Tuesday, ...} last week

However, such an SCC would obviously be too general to correspond to anything in speakers' minds for it would be very surprising for someone to talk about "Tuesday last month" etc. A more realistic SCC candidate would be (8c). The collocate *Woche* most likely does generalize to a paradigmatic class like (7b) in some collocations of *vergangen*, but at the same time, it may have idiosyncratic properties in other collocations that are not shared by other members of (7b). More generally, in different collocations around the same node word, a collocate may instantiate different classes – including the primitive class consisting just of the collocate itself.

This observation emphasizes the necessity of subtask 2. The paradigmatic classes obtained from subtask 1 alone do not reveal much about the SCC candidates around a node word – each of these classes is likely to play a role in *some* SCC candidate, but one still needs to determine the specific SCC candidates in which they actually do, and in particular, the SCC candidates in which different classes combine.

Our provisional implementation of subtask 2 does not avoid over-generalized SCC candidates directly. However, over-generalizations are likely to receive a low cohesion score (i.e., a low statistical significance) – and this score tends to be lower for a greater degree of over-generalization. Better treatment of the danger of over-generalization would be to work with an extended notion of paradigmatic classes that may overlap (i.e., partly include the same collocates; cf. 4.2.1) and to determine for any conflicting SCC candidates which of them would be an optimal generalization.

4.3 Stage 3: Evaluating the psychological reality of SCC candidates

In the previous subsection we described a general approach for automatically inducing SCC candidates from corpora, which involved an abstract strategy and a sequence of technical modeling decisions (e.g., choice of a corpus, similarity measures, a particular clustering algorithm, cut-off levels). Our competence-based evaluations sufficed to provide a general proof of concept, but given the ultimate goals of this line of research (cf. Section 2), a simple "looks good to me" evaluation is certainly not enough. Before SCC candidates can be used to indirectly study the real schemas that are entrenched in individual speakers and in a language community, their status as true SCCs must be established in a conclusive way.

What is needed is a systematic and rigorous evaluation of the derived SCC candidates in terms of appropriate psychological studies. SCC candidates that do not correlate to anything that is psychologically real may point either to systematic shortcomings of the general strategy, or to the deficiencies of its specific technical implementation, which in both cases might prove vital for critical revision and further improvements.

5. Future prospects

A corpus-based detection of genuine SCCs would enable us linguists to study through these SCCs the emergent schemas – i.e., syntagmatic-paradigmatic structures which are socially and psychologically real. Of particular interest might be questions like the following:

1) How do schemas operate in language processing?

2) What would an appropriate cognitive conceptualization look like?

While ultimately, such questions will necessarily involve, again, psychological investigations, it is possible to use SCCs to generate hypotheses about the nature of these structures. To this end, a good strategy would involve the following three steps. First, closely inspect a large number of corpus-derived SCCs and attempt to characterize them individually. Second, by abstracting across many SCCs, try to identify more general characteristics of this type of structure. Third, formulate these meta-descriptions as specific hypotheses (at the theoretical level) about the real schemas in language. This third step constitutes *abductive reasoning*: the inductive steps described in Sections 3 and 4 generalize from specific observations in the corpus data to more abstract structures, but all these structures still only have the status of descriptions, none of them reaches the theoretical level. To do that, abduction – in the sense of an "inference to the best explanation" (Harman 1965) – is required (cf. Figure 1).

To provide some guidelines with respect to the first step, a good starting point for characterizing a given SCC would be to inspect each of its classes relative to the SCC – and not just relative to its node word. This inspection could initially proceed along paradigmatic and syntagmatic lines. For the paradigmatic inspection, one could first attempt to describe commonalities between the class members observed in instances of this SCC. Based on these insights, one could then choose a name for the class which facilitates metadiscourse. Crucially, however, this name only constitutes a convention and is not to be confused with the class itself. An additional approach would be to attempt to generalize the class beyond observation and test its predictive power for unseen events. This may lead to insights into the dynamic nature and productivity of the SCC. After many SCCs have been inspected in this fashion, one further exciting question would be whether similar node words tend to have similar paradigmatic classes.

For the syntagmatic inspection of an SCC, one could start by attempting to characterize the relation between the SCC and each of its classes. Initially, this process should not involve pre-existing categories such as colligation, semantic preference (Sinclair 1998), or more traditional categories such as phrasal categories, subcategorization frames, thematic roles, etc. Eventually, after having inspected a broad range of SCCs in this way, the process might lead to the confirmation or modification of existing relational categories, or to the introduction of new categories, if inevitable.

Appendix

For our explorations, in order to operationalize the notions of associative-semantic and positional similarity between collocates, we took advantage of information already provided by the CCDB (Belica 2001-2007, Keibel / Belica 2007). With respect to the collocates' positional similarity, we used autofocus information that is available in the CCDB profiles for each primary collocate: the *positional focus* of a given collocate is the context window around the node word in which this collocation is most *cohesive* (i.e., statistically most significant, but not necessarily most frequent). In other words, it is a measure for the surface positions (relative to the node word) that the collocate occurs in most typically. For example, a collocate with the positional focus [-1;3] is likely to occur anywhere between one word to the left of the node word and three words to its right. Given this information, we defined the *positional similarity* between any two collocates of the same node word as the similarity between their positional foci.

The measure we used for quantifying the similarity between positional foci guarantees that two (nearly) identical positional foci are deemed the more similar the smaller they are because a smaller focus is more specific and thus conveys more information. For instance, a collocate that was assigned the largest possible focus – in the current online version of the CCDB this is the context window [-5;5] – essentially exhibits no positional preferences at all.

The operationalization of *associative-semantic similarity* between any two collocates *x* and *y* of a fixed node word is slightly more complex. To this end, we used the collocation profiles of the collocates, thus treating *x* and *y* themselves as node words.[4] The similarity between these two profiles then quantifies the degree to which *x* and *y* are used in similar ways. Formally, this similarity was assessed in terms of a measure that has proven to implement a plausible notion of similarity which is most sensitive to semantic and pragmatic factors, but also to other aspects of usage similarity between words (e.g., Belica et al. 2010).

Finally, to obtain a measure of the overall *paradigmatic similarity* between any two collocates, we combined the measures of their semantic and positional similarities (e.g., by means of multiplication). Future research should seek to match the operationalizations described in this appendix with available psychological evidence, and revise them if necessary.

[4] As node words are lemmas and collocates are word forms, we first had to lemmatize the collocates.

References

Belica, Cyril (1995): Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode. Mannheim: Institut für Deutsche Sprache. Internet: http://corpora.ids-mannheim.de (last visited: 11 / 2010).

Belica, Cyril (2001-2007): Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Mannheim: Institut für Deutsche Sprache. Internet: http://corpora.ids-mannheim.de/ccdb/ (last visited: 11 / 2010).

Belica, Cyril / Keibel, Holger / Kupietz, Marc / Perkuhn, Rainer (2010): An empiricist's view of the ontology of lexical-semantic relations. In: Storjohann, Petra (ed.): Lexical-semantic relations: Theoretical and practical perspectives. Amsterdam et al.: Benjamins, 115-144.

Biber, Douglas (2009): A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. In: International Journal of Corpus Linguistics, 14, 3: 275-311.

Cheng, Winnie / Greaves, Chris / Warren, Martin (2006): From n-gram to skipgram to concgram. In: International Journal of Corpus Linguistics 11, 4: 411-433.

Croft, William (2001): Radical construction grammar: Syntactic theory in typological perspective. Oxford: Oxford University Press.

Fletcher, William H. (2003): Phrases in English (PIE). Internet: http://pie.usna.edu (last visited: 11 / 2010).

Harman, Gilbert (1965): The inference to the best explanation. In: The Philosophical Review 74, 1: 88-95.

Hoey, Michael (2005): Lexical priming: A new theory of words and language. London: Routledge.

Hopper, Paul J. (1987): Emergent grammar. In: Berkeley Linguistics Society 13: 139-157.

Hopper, Paul J. (1998): Emergent grammar. In: Tomasello, Michael (ed.): The New Psychology of Language: Cognitive and functional approaches to language structure. Mahwah, NJ: Erlbaum, 155-175.

Hunston, Susan / Francis, Gill (2000): Pattern Grammar: A corpus-driven approach to the lexical grammar of English. (= Studies in corpus linguistics 4). Amsterdam et al.: Benjamins.

Institut für Deutsche Sprache (2007a): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2007-I (Release vom 31.01.2007). Mannheim: Institut für Deutsche Sprache. Internet: www.ids-mannheim.de/kl/projekte/korpora/archiv.html (last visited: 11 / 2010).

Institut für Deutsche Sprache (2007b): Virtual corpus "CCDB2007" composed from the German Reference Corpus (Institut für Deutsche Sprache 2007a).

Keibel, Holger / Belica, Cyril (2007): CCDB: A corpus-linguistic research and development workbench. Proceedings of the 4th Corpus Linguistics Conference, Birmingham. Internet: http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf (last visited: 11 / 2010).

Keibel, Holger / Kupietz, Marc (2009): Approaching grammar: Towards an empirical linguistic research programme. In: Minegishi / Kawaguchi (eds.), 61-76. Internet: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/061-076.pdf (last visited: 11 / 2010).

Keibel, Holger / Kupietz, Marc / Belica, Cyril (2008): Approaching grammar: Inferring operational constituents of language use from large corpora. In: Šticha, František / Fried, Mirjam (eds.): Grammar & Corpora 2007: Selected contributions from the conference Grammar and Corpora, Sept. 25-27, 2007, Liblice, Czech Republic. Prague: Academia, 235-242.

Kupietz, Marc / Belica, Cyril / Keibel, Holger / Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the seventh conference on International Language Resources and Evaluation (LREC'10), 1848-1854. Internet: www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf (last visited: 11 / 2010).

Kupietz, Marc / Keibel, Holger (2009a): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi / Kawaguchi (eds.), 53-59. Internet: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf (last visited: 11 / 2010).

Kupietz, Marc / Keibel, Holger (2009b): Gebrauchsbasierte Grammatik: Statistische Regelhaftigkeit. In: Konopka, Marek / Strecker, Bruno (eds.): Deutsche Grammatik – Regeln, Normen, Sprachgebrauch. Jahrbuch des Instituts für Deutsche Sprache 2008. Berlin/New York: de Gruyter, 33-50.

Mason, Oliver (2004): Automatic processing of local grammar patterns. Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, January 6-7, 2004, University of Birmingham. Birmingham: University of Birmingham Press, 166-171.

Minegishi, Makoto / Kawaguchi, Yuji (eds.) (2009): Working Papers in Corpus-based Linguistics and Language Education, Vol. 3. Tokyo: Tokyo University of Foreign Studies (TUFS). Internet: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/index.pdf (last visited: 11 / 2010).

Perkuhn, Rainer (2007): Systematic exploration of collocation profiles. Proceedings of the 4th Corpus Linguistics Conference, Birmingham. Internet: http://corpus.bham.ac.uk/corplingproceedings07/paper/132_Paper.pdf (last visited: 11 / 2010).

Renouf, Antoinette / Sinclair, John M. (1991): Collocational frameworks in English. In: Aijmer, Karin / Altenberg, Bengt (eds.): English corpus linguistics: Studies in honour of Jan Svartvik. London: Longman, 128-143.

Sinclair, John (1998): The lexical item. In: Weigand, Edda (ed.): Contrastive lexical semantics. Amsterdam et al.: Benjamins, 1-24.

Stubbs, Michael (2004): On very frequent phrases in English: Distributions, functions and structures. Plenary given at ICAME 25 (25th anniversary meeting of the International Computer Archive for Modern and Medieval English), Verona, Italy, May 19-23, 2004.