Systematic Exploration of Collocation Profiles

Rainer Perkuhn[1]

> "In a corpus-driven approach the commitment of the linguist is to the **integrity** of the **data as a whole**, and descriptions aim to be **comprehensive** with respect to corpus evidence."
>
> (Tognini-Bonelli, 2001: 84, emphasized by author)

Abstract

The central issue in corpus-driven linguistics is the detection and description of patterns in language usage. The features that constitute the notion of a pattern can be computed to a certain extent by statistical (collocation) methods, but a crucial part of the notion may vary depending on applications and users. Thus, typically, any computed collocation cluster will have to be interpreted hermeneutically. Often it might be captured by a generalized, more abstract pattern. We present a generic process model[2] that supports the recognition, interpretation, and expression of the patterns inside and of the relations between clusters. By this, clusters can be merged virtually according to any notion of a 'pattern', and their relations can be exploited for different applications.

1 Introduction

For any kind of investigation of language usage, a look back into the field of corpus linguistics may help to find which questions were already asked and which ones are already answered. Corpus linguistics contributed to

- the availability of large sets of authentic data,
- the restricted usefulness of small specialized corpora vs. the incentives and the challenge of very large, multi-purpose archives,
- the availability of pre-structuring, high lighting methods (searching, sorting, rendering) resulting in functionality embedded in working environments and considering application concerns.

Of course, these components simply can be added to the linguist's workflow and workplace. Many case studies show that they also offer new possibilities how to work and how to investigate linguistic phenomena (amongst others Hanks, 2004). Tognini-Bonelli pointed out that a new, so-called 'corpus-driven' approach has to follow new principles and strategies (Tognini-Bonelli, 2001). As a consequence, this means that the approach also requires a new methodology and working framework. However, we are not aware of any systematic study contrasting the two ways of work, describing the new workflow, developing a process model,

[1] Research Group for Corpus Linguistics, Institut für Deutsche Sprache, Mannheim, (Institute for the German Language, Mannheim, Germany)
e-mail: perkuhn@ids-mannheim.de

[2] The process model was developed by Cyril Belica and the author.

and clarifying the assumptions, constraints, and consequences of the new approach. We developed a generic process model for data-driven investigations. In the following, we illustrate this model by applying it to a certain database and a family of methods. But, in addition to the usefulness in this scenario, our claim is that this model can be reasonably applied to any data and any structuring method for any application.

The generic process model covers three areas of interest to support the exploration of collocation profiles. First, the results are visualized in an interactive manner so that an interpreting linguist can switch between different views to capture the computed structure in breadth and depth, always in combination with the underlying data. Second, the user can augment the computed structures with working notes and annotations (and additional structural information) expressing anything that might be useful during the interpretation by using an existing or a new annotation ontology. Third, the model supports the preparation of the pre-structured and post-structured data for the application-oriented presentation. Thus, the trace of the language use is documented consistently from the source to the presentation.

Our main concern is the detection of multi-word units and the deployment of patterns for teaching German as foreign language (DaF). For our current project environment, we apply the generic process model to the collocation analysis of a very large corpus of written German (DEREKO, 2007). However, the generic process model can be applied to any corpus and any structuring method and can be customized to any application by providing the respective ontologies.

2 Data and methods

It is (and will be) nearly impossible to make perfect representative corpora available. Of course, the postulation that a corpus should be as large as possible (Church/Mercer, 1993) holds in general and especially for corpus-driven investigations. But an unbalanced composition might distort the analysis. For different purposes it is reasonable to treat the data collections as archives and a corpus as a view on these archives. These views can be explicitly defined by the linguist so that he/she is working with a customized virtual corpus. Query and analysis results can be cross-validated against the dimensions that might have an influence of the linguistic phenomenon under investigation, e.g., text genre, respectively, text type. For the examples we use for illustration, we decided for a very large subset of our corpus archive (texts from DEREKO with approximately 2.2 billion words) cleaned from duplicates (*cf.* Kupietz, 2006) and reduced in over-represented genres.

To characterize the typical usage of a word, we structure its contexts with a family of collocation methods elaborated by Belica (Belica, 1995) that are integrated in our proprietary corpus interface COSMAS II (COSMAS II, 2007). This access service is offered by a partner project free of charge; the only prerequisite is that the users register and commit themselves to non-commercial use. In this environment, any virtual corpus and any configuration of the analysis method can be deployed dynamically for a special research interest. After choosing a corpus, the next step is normally searching for an expression (for – amongst others – a word form, all elements of a paradigm, simple regular expression, in combination with different logical and proximity operators). The collocation analysis allows to define different windows around the hit (left and right context, respecting sentence boundaries or not) and to select values for different parameters: accuracy, granularity, considering function words, using lemmatization, and using auto focus (for obtrusive positional distributions). The dynamic definition of context sizes enables to capture long-distance dependencies and discontinuous constituents. The configuration determines which strings are actually statistically evaluated with respect to a certain level of significance. For the primary partner words, log-likelihood, for the following partner words, mutual information is applied (*cf.* Dunning, 1993). So, any configuration can be understood as a slightly different question to the data yielding a slightly different answer. But apart from which strings were actually included in the computation, our claim is that the detected structures not only can be explained from the statistical point of view but also have an (approximate) counterpart in a functional model of the language representation in our mind.

For our own experimental platform CCDB (Cooccurrence Database[3], CCDB, 2007), we applied two approved configurations of the methods to the above mentioned large, cleaned, re-balanced corpus. We stored the results of the analysis of more than 220.000 words, their collocation profiles, in a database. Thus, we have immediate access to the descriptions of the contextual characteristics of these words and make them subject to further investigations. One thread aims at the definition of an algorithm and a metric for the computation of the similarity between and further comparison of collocation profiles. In this report, we concentrate on a second thread that targets at creating an environment for user guidance and support for the study of the local substructures.

For our process model, we defined an interface to the CCDB and to COSMAS II, but we use primarily data from the CCDB to complement the insights of the other thread.

3 Methodology and process model

Our claim is that – even if this or any other setting is nearly perfect – it will be always necessary to offer support and guidance in working with the empirical data and especially in interpreting the analyses results. Of course, depending on their question, the linguists could and should configure their environment, but we cannot anticipate any linguistic question and reduce finding the answer to one step. Working with corpora is a complex process that needs to be structured.

By providing large data archives, the data preparation can be reduced to virtual composition. Good training and guidance helps in applying the methods as close as possible to the needs. But then, still the researchers need a way to work further with the pre-structured empirical data.

3.1 Viewing of automatically prepared structures

For different reasons, the application of the methods does not apparently yield the answer to the linguist's question: The language itself changes, and the methods cannot anticipate the different interests of different linguists that also can change during time. So, to avoid that important results are filtered out in advance, the methods normally prefer recall to precision. The results contain more and finer information than necessary. It is the task of the linguists to try to understand as much as possible from what the structure offers. They have to decide to what extent, respectively, how they have to consider the substructures when they commit themselves "to the integrity of the data as a whole" and when their "descriptions aim to be comprehensive with respect to data evidence" (*cf.* Tognini-Bonelli, 2001, p. 84). If the data set is large enough for making reasonable judgements on language usage, the query and analyses results are very complex, so complex, that they cannot be handled with simple means. We suggest a presentation that shows the complete resulting structure with the possibility of interactive browsing. The limitations of the space due to the screen size require

[3] The German term 'Kollokation' has a narrower meaning than the English term 'collocation'. So, to avoid the naming conflict in the field of studies of the German language, we introduced the term 'cooccurrence' and use it in combination with the analysis and the database.

a special technique to show the complete structure while displaying at least a part clearly and readable – the focus and context technique (Lamping et al., 1995).

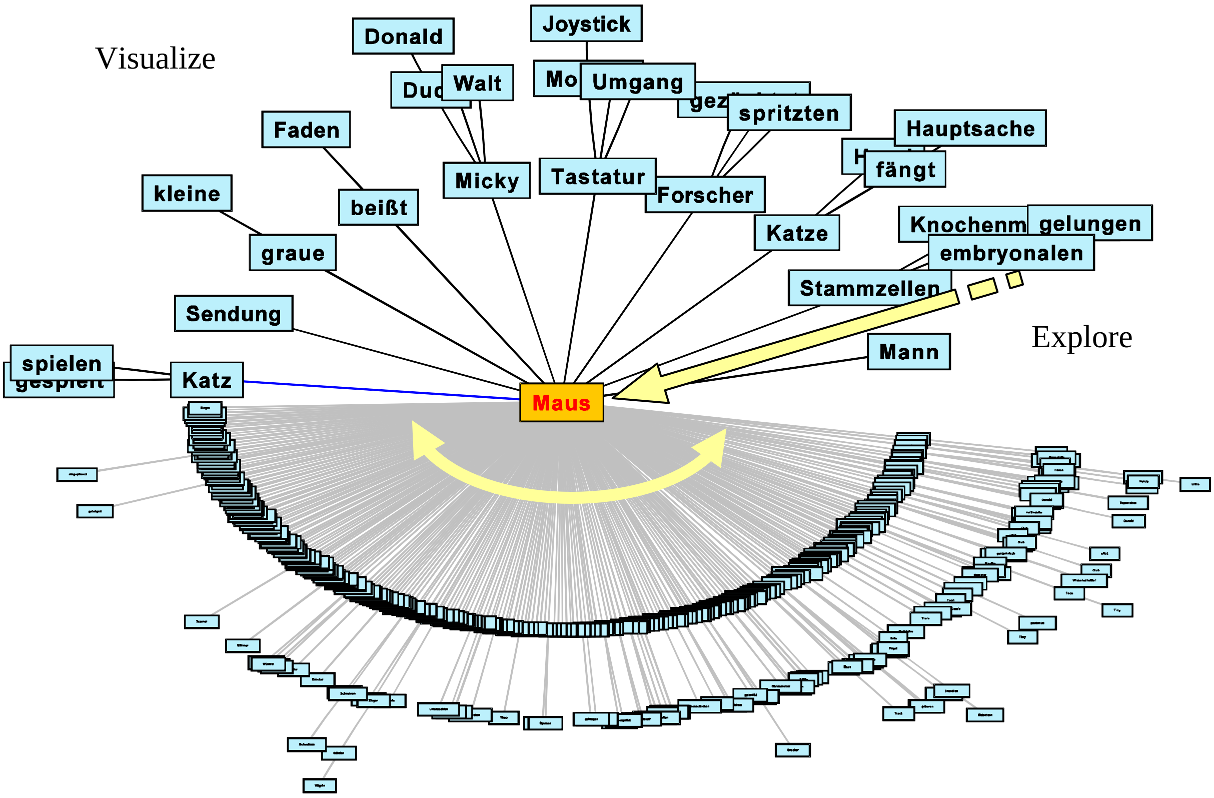


Figure 1: Visualization of a computed structure of the word 'Maus' (*engl.* 'mouse') (arrows indicate ways of interaction).

The centred rectangle is named with the word that is the subject of the investigation. The other rectangles are arranged on several circles around this centre. Each rectangle stands for a computed cluster containing the text fragments and its quantitative data.

last partner tok.	interval	Llr	freq.	ratio	syntagmatic pattern
"Faden"	[-2,-2]	822	66	100%	Da|da beißt die Maus keinen Faden ab[4]

kwics

```
Z98  zur Entschuldigung.  "Da beißt die Maus keinen Faden ab."
Z98  facto eine Kürzung, da beißt keine Maus den Faden ab. Als Privatdozent muß
Z01                 Nein, da beißt die Maus keinen Faden ab: Wenn beide Partner
B00 lung: Von solchem "Köder" beißt die Maus keinen Faden ab.
B01 Roland Koch hat sich - da beißt die Maus keinen Faden ab - um die politische
B01 unktur eingetrübt hat. Da beißt die Maus keinen Faden ab.
T92                       "Da beißt die Maus keinen Faden ab", sprach der Richte
T93 in Ordnung gewesen" - "da beißt die Maus keinen Faden ab". Die Umzugskosten
T94 st sind verschieden, da beißt keine Maus den Faden ab. Der banale Satz "Über
T95 Na, vorbei is' vorbei, da beißt die Maus keinen Faden ab!
T96 creamin' Jay Hawkins - da beißt die Maus keinen Faden ab. Die berühmte Schau
```

[4] *engl.* "It's Lombard Street to a China orange."

```
T96 n ist die halbe Miete, da beißt die Maus kein' Faden ab. Wohl dem, der dazu
T96 getan wird. Indes, und da beißt die Maus keinen Faden ab, im Zeitalter der "
T97  hat, ist verwerflich. Da beißt die Maus keinen Faden ab. Aber es kommt auch
T97 e "Ich sag' dir, die Alte beißt der Maus keinen Faden ab" - nur albernes "Ha
T97 orleben geben."König: "Da beißt die Maus keinen Faden ab."
T97 en gefälscht habe, "daran beißt die Maus keinen Faden ab", so Salditt. Aber
 …                                         …
```

Figure 2: Content of a cluster (quantitative data and text fragments).

The inner circle lists the primary clusters, initially sorted according to the statistical measure of cohesion. The outer circles add the n-ary (sub-)clusters. The primary clusters especially ten in the upper hemisphere are privileged over the others. They are assigned more space so that they can be displayed bigger. If the substructure around them is so dense that they would mostly overlap with other rectangles, all other rectangles have to be displayed very tiny.

The distances between the circles shrink from the centre to the peripheral circles. This is an effect of a three-dimensional hyperbolic projection of equal distances to enable more space for the centre.

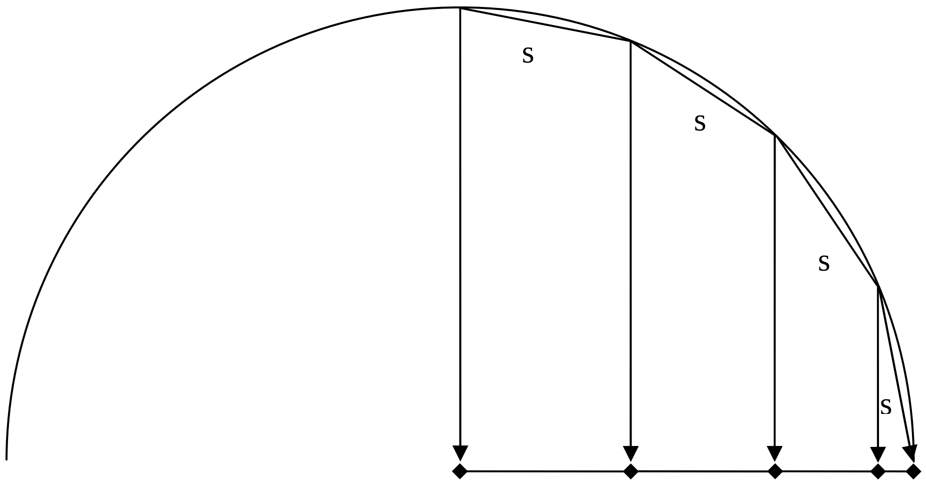


Figure 3: Hyperbolic Projection.

To capture the complete analysis in full depth and breadth, the linguist can

- search for clusters according to different criteria,
- examine each cluster with the underlying data and the quantitative measures,
- focus on clusters of the outer circles,
- change the selection of the upper hemisphere by turning the star,
- change the ordering by sorting the primary clusters according to different criteria (cohesion, frequency, alphabetically, reverse alphabetically, and by position),
- view several analyses in parallel,
- compare and contrast two juxtaposed analyses.

As a kind of memory for interesting clusters, the rendering of the rectangles can be toggled (switched from tiny to big and vice versa) or added to a selection so that they are rendered high lighted. Both mark-ups remain visible even if the rectangle moves to the lower hemisphere.

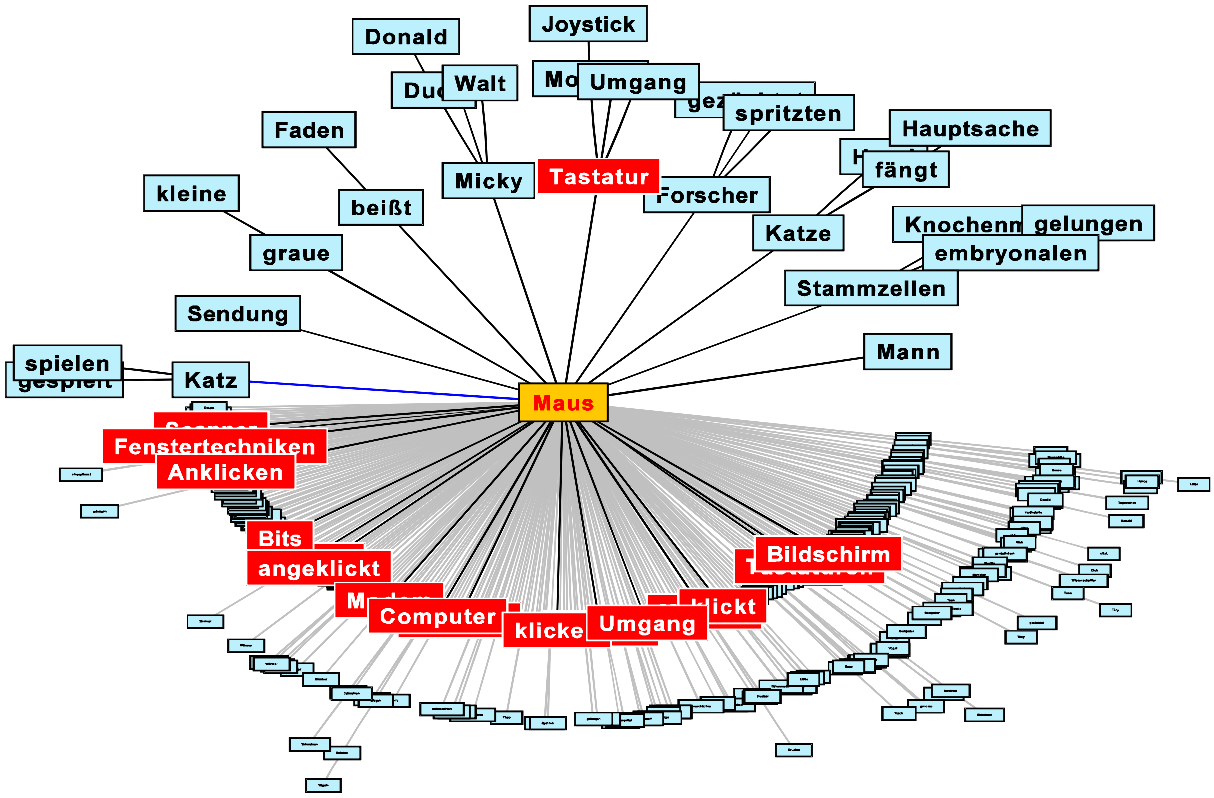


Figure 4: High lighted, selected clusters that document the use of the word 'Maus' (*engl.* 'mouse') in computer contexts.

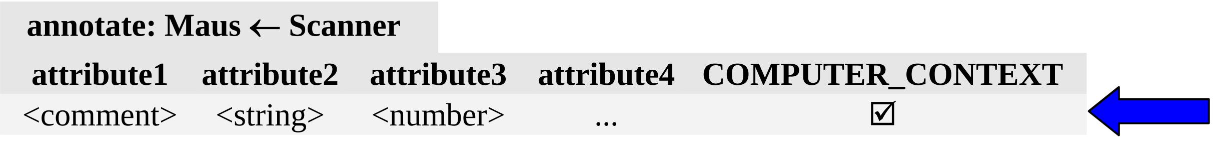
All of these facilities deal with the first aspect of how the process model supports the perception and understanding of the computed structure. The linguists can inspect single clusters with their data and quantitative data, they can repeat this step for several clusters inside one or across different structures to understand their similarity or relatedness, and they can compare and contrast complete structures.

3.2 Assessing relevance

From the statistical point of view, every cluster is legitimised since the accuracy of the computed cohesion is above a level of significance. But from the linguist's point of view, it may vary whether a cluster is relevant or not. Of course, there might have been noise in the data resulting in mysterious clusters. Nevertheless, it might be dangerous to introduce a wastebasket for all non-relevant clusters, but they might be hidden for representation. The non-relevance is already important information for a more appropriate customization of the analysis method, and, possibly, they become relevant later.

Some clusters can be very similar, so similar that it is not reasonable to distinguish between them, or they contain different instantiations of the same pattern or phenomenon. Only part of this can be treated as universal subsumptions; most of this will depend on the target application. Anyway, the linguist can add his 'opinion', his ideas and his comments and anything that comes into his mind at parts of the structure. As long as the modifications are monotonic with respect to the initial computed structure (no delete, no duplicate), it is still only an intermediate layer to access the data as a whole (in the sense of Tognini-Bonelli, 2001).

The linguist can declare and use different kinds of comments and other attributes. The attribute can be typed (strings, numbers, truth values, enumerations), each type allowing a simple and controlled way of assigning a value to the attribute.



annotate: Maus ← Scanner

attribute1	attribute2	attribute3	attribute4	COMPUTER_CONTEXT
<comment>	<string>	<number>	...	☑

Figure 5: Annotating a cluster.

The comments can be displayed synoptically. The self-declared attributes can be used as the built-in attributes for searching and sorting. With the annotations, the linguist keeps a record of his insights and his epistemic assessment of the computed clusters and structures. The annotated values of the attributes are stored persistently together with the structure; therefore, they can be used across sessions and users. As a consequence, the attributes can also be used to trace the status, respectively, progress of the work, to schedule the working plan, to exchange intermediate results with partners, and to coordinate collaborative work.

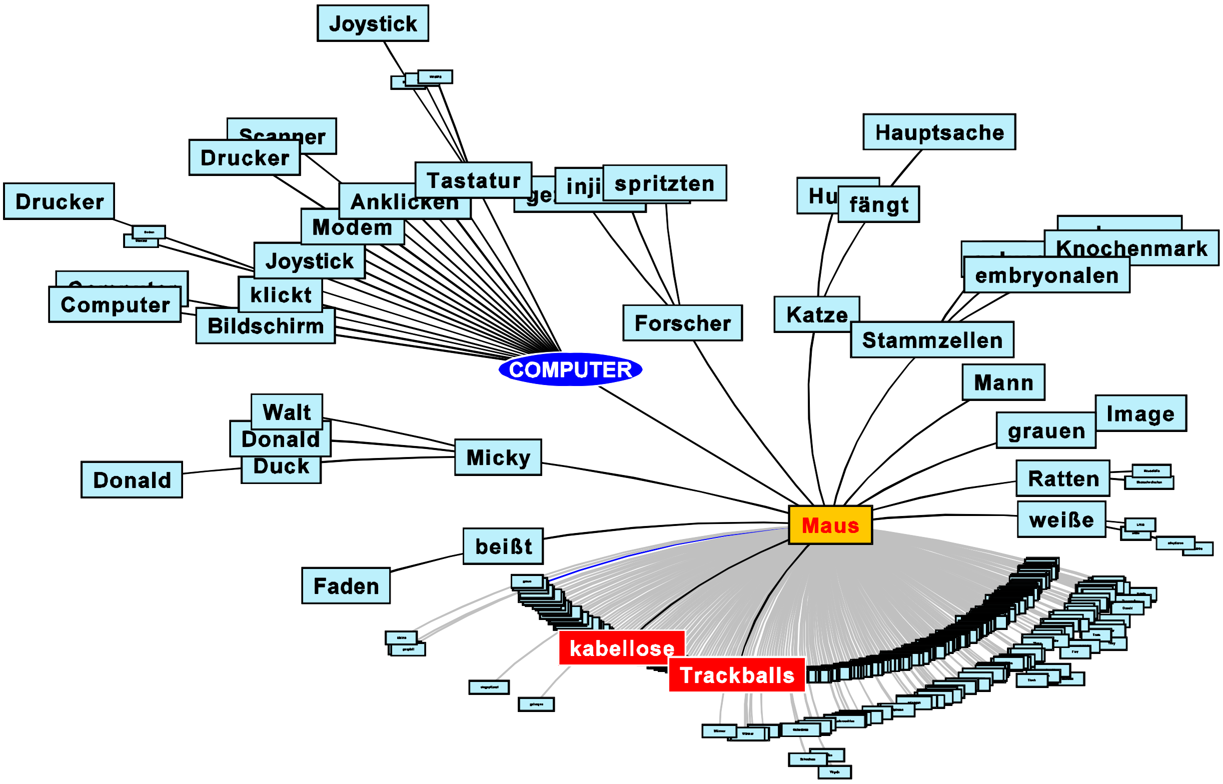


Figure 6: Augmented structure with structural element for a group of computer contexts with two further candidates for the group high lighted

The clusters that match a search query are high lighted as shown in Figure 4. The high lighting represents a selection. In addition to the combination with other queries via the logical operators (‘and’, ‘or’, ‘not’), selections can be modified by hand either by adding clusters to or by removing clusters from selections. Thus, groups of clusters can be developed somehow like the answer of a series of questions (e.g., “select all clusters that are relevant from a certain perspective above a certain level of confidence *and* that are *not* already marked as special cases”). If the linguist wants to keep the information about the composition of a group, e.g., because he/she considers it a relevant category, he/she can express this graphically or virtually. Graphically means that the structure is modified: A new item for the category is introduced, and the members of the group become structurally dependent on this item.

The graphical correspondence of a group is very convenient because it immediately shows the categorical information but it is restricted to strict taxonomies without cross-classification. In these cases or as an alternative to the structure modification, virtual groups may be used. Similarly, a new item for the group is created, but no further structural modification is performed. Solely, the information of the composition of the group is assigned to the item. Thus, problems with cross-classification do not matter. The information can be used later to reconstruct the selection independent of how many groups a cluster belongs to.
The linguists will hardly be able to modify the structure straightforward to express their findings. They might change their opinion about the relevance of what they found or about the way they express their insights, or they just want to be more accurate to express the relations between the clusters. Of course, the structures can be revised and refined. The manually introduced elements can be renamed, and like the structural modifications, they can be rearranged.

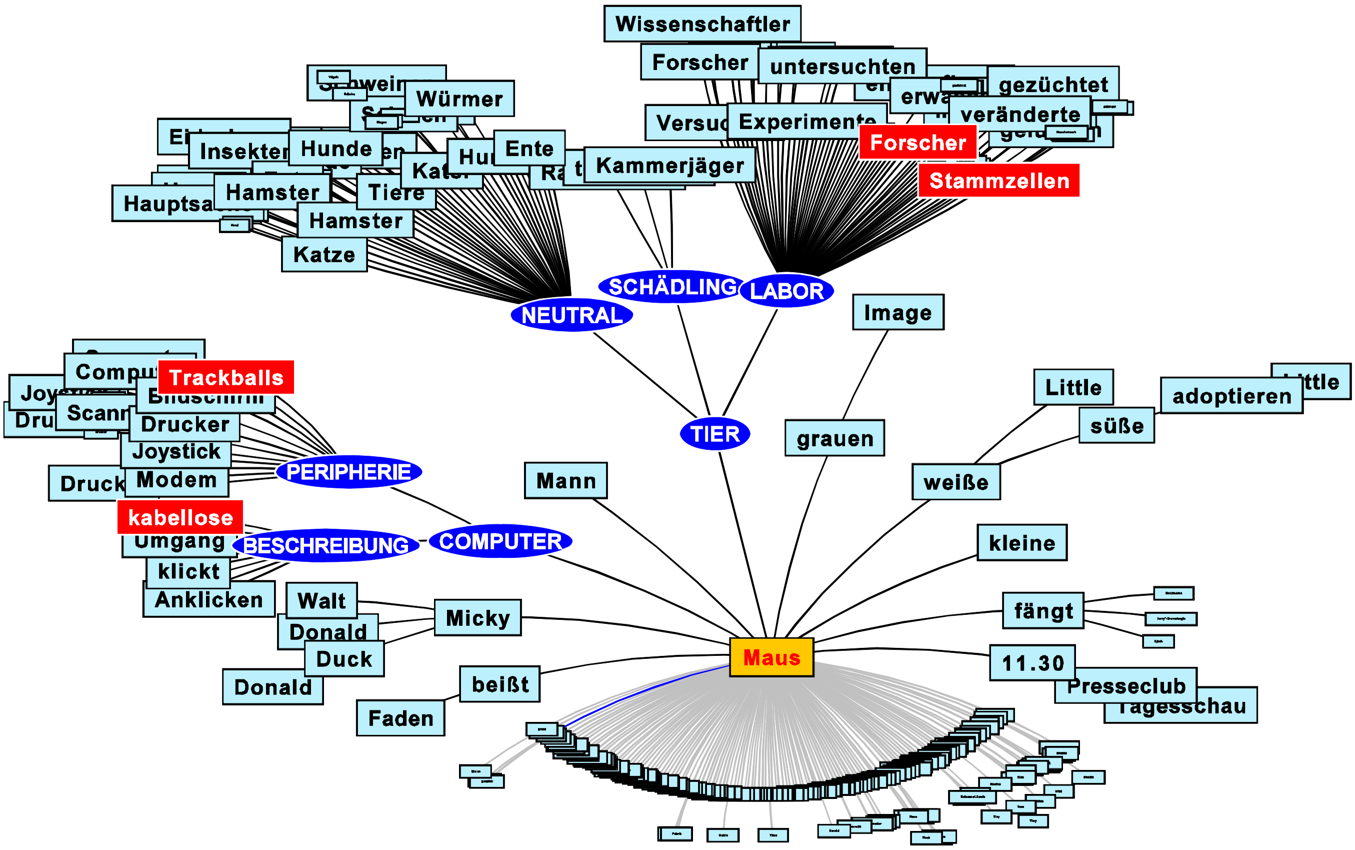


Figure 7: Augmented structure after revisions and refinements with structural elements for groups of computer and animal contexts and substructures.

Finally, the linguists have to decide which part of the information they want to use to describe the detected phenomena. Actually, they have to preserve a mapping between the gathered structure and the structure of their target representation. Given such a mapping, the last step can be just an export of required packages of information.

3.3 Consistent epistemic annotation

Actually, when the linguists think about categories, and relate all the items to each other, and try to explain the (pre-structured) data, they are looking for a model that is compatible to data evidence. ‘Looking for’ can be interpreted as very different efforts: Of course, it can mean that the linguists want to find an existing model, one that is derived from an established linguistic theory and that offers a perfect blueprint for the observed data. But in most cases, phenomena exist that no model can sufficiently capture. Normally, the linguists need to adapt a model (and revise the underlying theory), or they have to decide how to cope with the heretical usage. Eventually, they will brand them as defective, or they coerce them to a nearly fitting category. In both cases, they evade the question what the data really wants to tell.

The other reading of ‘looking for’ can be understood as ‘building a model’: The linguists have to consider possibly new categories; these may be partially borrowed from known model substructures. They can postpone the labelling of a category and its intensional definition and just use it as an extensional container of their observations. After they have gained enough experience in this way of work, they can think about labels and formal definitions. In the meantime, they are just preparing a kind of ‘ontology’. The term is used in the framework less in the sense of Gruber (Gruber, 1993) but in the sense of Guarino (Guarino, 1997, Guarino, 1998): The linguists should make the relevant properties explicit, not necessarily formally, but approximately to an extent that they can achieve a mutual understanding of their models.

The process model supports the management of such simple kinds of ontologies. Up to now, they are more or less restricted to taxonomies without inferential knowledge. The linguists can build an ontology from scratch or they can use and adapt an existing one. Accordingly, the optional part of the workflow of creating or changing an ontology shifts the intention the linguists pursue. They can build up a new model. They can customize an existing one, or they can just describe the data with existing models. Since, up to now, no corpus-obliged model exists, the process model should primarily be used for gaining experience in this field. Later on, these models can be used as other models, too, e.g., for contrasting each other. Presumably, layered models will come into play, combinations of upper models (describing more or less application-independent categories), and application-specific models.

4 Using the process model for ...

> “You shall know a word by the company it keeps.”
> (Firth, 1968:179)

In many applications and domains that want to offer information on words and their relations in the language, the crucial first step is to understand the word and to get an overview of how the word is used. In the following, we sketch briefly three domains where the process model is already in experimental use and its prototypical implementation is released (since spring 2004), or its introduction is planned, respectively, in preparation.

4.1 … Description of multi-word units

A very simple use case is to match the collocation profiles with only a few established categories, e.g., idioms, and to add comments how the patterns are used and to explain their metaphorical, compositional, or non-compositional meaning. This way of editing prepares structures that can easily be assembled to specialised dictionaries, e.g., idiom or collocation dictionaries.

4.2 … Modern lexicography

Due to limited space in printed editions, traditional dictionary compilation has to break down the richness of a ‘meaning’ of a word to only a few readings, senses, facets, or nuances. Electronic editions allow reflection of nearly all usage aspects that are recognized in the collocation profiles. We would not endorse any theory where neither each discourse (Gale et al., 1992) nor each collocation (Yarowsky, 1993) determines a sense. The notion of a ‘sense’ can be understood as the desire for distinctive information – but only with respect to certain application in mind (*cf.* Kilgarriff, 1996). The relevance of the distinctiveness is completely different, e.g., for a translation task or for didactic purposes for foreign language learning on a nearly expert competence level. The process model is designed exactly for this demand of dynamic views on the data evidence. Each abstract, universal, or application-specific category over collocations can be treated as a sense. The empirical foundations for different target domains are the result of different interpretations.

4.3 … Learning German as foreign language (DaF)

The potential of corpus linguistics for advanced foreign language learning is already divined – but cannot be appraised to full extent. Data-driven learning, self-learning from context, and classroom concordances (*cf.* Hadley, 2002, Braun et al., 2006) are established, but the benefits of collocation analysis are only loosely approved. Our framework can afford the premises to introduce collocation cluster networks as self-learning environments.

To detect the patterns in the language, a learner needs sufficient input. Interpreted collocation structures can embody condensed information for German as a foreign language, collected examples of the same pattern. In many situations, understanding is postponed although most parts of the pattern are known until enough examples can help resolving the ‘missing item’. Instead of waiting over a long time for repeating events, clusters are computed collections of these events. Of course, it is somehow different whether an uncertainty will be resolved in a delayed, passive, unconscious process or on demand in an active and conscious process. But to offer didactically prepared cluster structures would enable learning from context par excellence because we do not only present one context but the complete collection of all contexts containing the same pattern.

Even if it is a slight difference whether the missing item is resolved in a natural environment unconsciously or in a training scenario on demand, we are planning to develop a cooccurrence network adequately modelled for this purpose combined with adequate presentation techniques.

5 Related work and conclusion

We cannot mention every corpus workbench. Besides the corpus interface of the IDS, COSMAS II, we just want to refer exemplarily to the IMS corpus workbench (Christ, 1994), but neither of these tools supports a process model as we described in this report. Application-oriented environments, e.g., WASP (Kilgarriff/Tugwell, 2001) are deeply influenced by the traditional process models of their domains. The work of others (Elliott et al., 2001, Magnusson/Vanharanta, 2003, Powers/Pfitzner, 2003) overlaps only partially with our concerns. We are not aware of any related work that first starts with a conceptual model, a methodology, for corpus-obliged investigations and, therefore, designs or chooses the functionality and techniques to build a guidance and support framework.

A nice game is derived from the tool: You can guess a hidden word by exploring its collocation profile. Taken seriously, it is more than a game because it simulates a situation where the linguists do not know anything about a word, just all the aspects how it is used. They almost have to learn the word from scratch. They have initially no assumption of how many senses a word has, and they do not need to decide to which sense an aspect belongs. We presented the game (and, of course, the complete working environment, too) on different occasions. It was fascinating that almost every word could be guessed very quickly from its collocation profile. Our conclusion is that the profiles contain enough information for a sufficient description of the word, and our claim is that an application requires a special view on the profile. With our process model and our framework, we provide the means to create adequate collocation networks.

References

Belica, Cyril (1995): *Collocation Analysis and Clustering.* Corpus Analysis Module. Institut für Deutsche Sprache, Mannheim, http://corpora.ids-mannheim.de/, released 1995.

Braun, Sabine, Kurt Kohn, and Joybrato Mukherjee (eds) (2006): *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods.* (English Corpus Linguistics 3, ed. Thomas Kohnen and Joybrato Mukherjee) Frankfurt am Main: Peter Lang.

Christ, Oli (1994): A modular and flexible architecture for an integrated corpus query system. *COMPLEX '94,* Budapest.

Church, Kenneth W., and Robert L. Mercer (1993): Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics,* 19, 1, 1–24.

CCDB (2007): *Cooccurrence Database*, http://corpora.ids-mannheim.de/ccdb/, visited 29 June 2007.

COSMAS II (2007): *Corpus Search, Management and Analysis System*, http://www.ids-mannheim.de/cosmas2/, visited 29 June 2007.

DEREKO (2007): *Deutsches Referenzkorpus* (= German Reference Corpus), http://www.ids-mannheim.de/kl/projekte/korpora/, visited 29 June 2007.

Dunning, Ted (1993): Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19, 1, 61–74.

Elliott, John, Eric Atwell, and Bill Whyte (2001): Visualisation of Long Distance Grammatical Collocation Patterns in Language. *Proceedings of the 5th International Conference on Information Visualisation*, 297–302.

Firth, John R. (1968): A Synopsis of Linguistic Theory 1930–1955. In: Studies in Linguistic Analysis. Philological Society. Oxford, 1957. Reprinted in Palmer, F. (ed.): *Selected Papers of J. R. Firth*. Harlow: Longman, 168-205.

Gale, William A., Kenneth W. Church, and David Yarowsky (1992): One Sense Per Discourse. *Proceedings of the 4th DARPA Speech and Natural Language Workshop,* February 23-26, New York: Harriman, 233–37.

Gruber, Tom R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5, 2, 199–220.

Guarino, Nicola (1997): Understanding, building, and using ontologies. *International Journal of Human-Computer Studies,* 46, 293–310.

Guarino, Nicola (1998): Formal ontology and information systems. In: N. Guarino (ed.): *Formal Ontology in Information Systems*, Amsterdam: IOS Press, 3–15.

Hadley, Gregory (2002): Sensing the Winds of Change: An Introduction to Data-Driven Learning. *RELC Journal*, 33, 2, 99–124.

Hanks, Patrick (2004): The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography*, 17, 3, 245–74.

Kilgarriff, Adam (1996): Word senses are not bona fide objects: implications for cognitive science, formal semantics, NLP. *Proceedings of 5th Conference on the Cognitive Science of Natural Language Processsing.* Dublin, 193–200.

Kilgarriff, Adam, and David Tugwell (2001): WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation. *Proceedings of MT Summit VIII.* Santiago de Compostela, Spain, 187–90.

Kupietz, Marc (2006): *Near-Duplicate Detection in the IDS Corpora of Written German*. Tech. Rep. KT-2006-01. Institut für Deutsche Sprache. ftp://ftp.ids-mannheim.de/kt/ids-kt-2006-01.pdf.

Lamping, John, Ramana Rao, and Peter Pirolli (1995): A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems,* Denver, May 1995, 401-408.

Magnusson, Camilla, and Hannu Vanharanta (2003): Visualizing Sequences of Texts Using Collocational Networks, Machine Learning and Data Mining in Pattern Recognition, *Proceedings of MLDM,* Berlin: Springer, 276–83.

Powers, David W., and Darius Pfitzner (2003). The Magic Science of Visualization. *Proceedings of the Joint International Conference on Cognitive Science,* University of New South Wales, 529–34.

Tognini-Bonelli, Elena (2001): *Corpus Linguistics at Work.* Studies in Corpus Linguistics 6. Amsterdam: John Benjamins Publishing Company.

Yarowsky, David (1993): One Sense Per Collocation. *Proceedings of ARPA Human Language Technology Workshop,* March 21-24, Princeton, New Jersey, 266–71.