Keibel, Holger/ Belica, Cyril (2007): CCDB: A Corpus-Linguistic Research and Development Workbench.
Proceedings of the 4th Corpus Linguistics conference, Birmingham.

CCDB: A Corpus-Linguistic Research & Development Workbench

Holger Keibel and Cyril Belica
Institute for the German Language (IDS)
Mannheim, Germany
{keibel, belica}@ids-mannheim.de

1 Introduction

Within a strictly corpus-driven paradigm, an in-depth profiling of many linguistic phenomena requires fast access to massive amounts of data derived from very large corpora. This poster presentation describes an empirical baseline framework – the CCDB – established for this purpose in 2001 at the Institute for the German Language (IDS) in Mannheim. We use the CCDB for the study, development, and evaluation of methods for the data-driven exploration and modelling of language use. The CCDB can be accessed through a public web interface at the URL `http://corpora.ids-mannheim.de/ccdb/`.

The paper is structured as follows: We first describe the kind of data that the framework provides. Then, we briefly discuss the notion of similarity of collocation profiles. Finally, we give examples of specific CCDB-based methods that we have been recently working on.

2 Baseline Data

As its core, the CCDB framework provides a large set of empirical data that were established from a 2.2 billion subset of the Mannheim German Reference Corpus (DEREKO) by performing a variety of collocation analyses on the entire stemmed corpus vocabulary. We used an extended iterative collocation algorithm to extract *higher-order collocations*. These are potentially non-contiguous n-tuples (in contrast to n-grams) and may occur with varying relative positions (unlike positional n-grams). By default, the algorithm looks for collocates of any order within a context window (around the focal word) whose size is determined dynamically. The algorithm we used is since 1995 part of the proprietary *Corpus Search, Management and Analysis System* (COSMAS) that is run at the IDS and operates on essentially the same corpus archive. Each single collocation analysis in the CCDB could, thus, in principle be reconstructed on-line by any COSMAS user.

2.1 Collocation Profiles

The CCDB comprises several alternative baseline data sets that were generated in separate analysis runs, each one with different settings for the extended collocation algorithm. For each run, the CCDB presents the resulting (higher-order) collocations around a given word as a hierarchical cluster tree, visualising the increasingly fine-grained collocation structure. The cluster nodes at the top level (i.e., on the first branch in the tree) pertain to the word pairs defined by the focal word together with one of its primary collocates. Their subclusters pertain to the word triples consisting of the parent cluster together with one of its secondary collocates, and so forth.

The resulting collocation clusters together constitute the word's *collocation profile* (the top portion of one such profile is shown in Figure 1). Currently, the CCDB comprises more than 220,000 nontrivial collocation profiles per analysis run, up to 250 top-level cluster nodes per profile, and up to 100,000 concordance lines per collocation cluster. The collocation profiles can be browsed via the focal word or searched via the collocates that appear anywhere in these profiles.

Figure 1. Annotated snippet of a CCDB web page showing the basic view of the collocation profile of *machen*.

2.2 Syntagmatic Patterns

Inspecting the concordance lines within a given collocation cluster gives rise to *syntagmatic patterns* that summarise the collocation's predominant usages in a sort of *wild card* expression, preserving the collocates' word order and, to improve the legibility, with inserted filler words that were observed to occur in this position at a certain rate.

The notion of syntagmatic pattern is illustrated in the following English examples.

(1) is **not** [...] at **all**

(2) … **should** do|be made **as** [simple|soon as] **possible** but not …

The CCDB currently lists each collocation cluster together with the single most typical syntagmatic pattern (cf. the rightmost column in Figure 1).

Syntagmatic patterns can be thought of as illustrations of the respective collocation, and they facilitate relating the collocation cluster both to the data and to the intuition of competent speakers. In our own corpus-driven studies, they serve as the starting point for the qualitative interpretation of collocations.

3 Examples of Ongoing Research

Our main objective is to conduct research on methodologies for discovering structures in the high-dimensional similarity space that is spanned by the collocation profiles and eventually to help understand the role of these structures in the theory of language use. Some of these efforts are illustrated below.

3.1 Pairwise Similarity of Collocation Profiles

Although a large number of similarity measures have been proposed and evaluated in the literature, it remains an open research question and a nontrivial challenge how to devise a distance metrics that operationalises a plausible notion of similarity for comparing objects as complex as the collocation profiles in the CCDB.

Folgende verwandte Kookkurrenzprofile zu Hindi wurden gefunden

Chinesisch
Englisch
Spanisch
Türkisch
Urdu
Portugiesisch
Japanisch
Arabisch
Italienisch
Landessprache
Polnisch
Französisch
Muttersprache
Griechisch
Hebräisch
Ungarisch
Amtssprache
Tschechisch
Russisch
Niederländisch
Rumänisch
Sprache
Schwedisch
Kroatisch
Albanisch
Slowenisch
Serbokroatisch
Dänisch
Umgangssprache
Koreanisch

mehr ...

Folgende verwandte Kookkurrenzprofile zu Charakteristikum wurden gefunden

Merkmal
Eigenheit
Eigenschaft
Eigenart
Ausprägung
Charakteristik
Anliegen
Element
Besonderheit
Charaktereigenschaft
Ausformung
Stilelement
Kriterium
Charakterzug
Stilmittel
Parameter
Charakter
Eigentümlichkeit
Ereignis
Spielart
Attribut
Auswahlkriterium
Qualitätsmerkmal
Gemeinsamkeit
Herausbildung
Vorzug
Aspekt
Argument
Ausdrucksmittel
Manko

mehr ...

Figure 2. Browser view of the similar collocation profiles lists for *Hindi* and *Charakteristikum*.

We currently examine the applicability of various similarity measures to the comparison of collocation profiles. Note, however, that semantic proximity of words is only one component of usage-based similarity between collocation profiles; others pertain to paradigmatics, syntagmatics, terminology, idiomaticity, etc. We focus on the task of determining how the fine-grained information about the usage patterns of each word manifested by the hierarchical structure of collocation profiles can be best captured and modelled. For epistemic goals, the explanatory power of such modelling is crucial but still poorly understood. An example of web output from the current version of the Similar Collocation Profiles method is shown in Figure 2.

3.2 Discovering Cross-Profile Similarity Structures

We are currently pursuing two complementary research directions that involve a set of words whose collocation profiles are most similar to the profile of a given focal word. Both of them aim at discovering and visualising crucial aspects of the similarity structure among these words, breaking down the complex constellation in the high-dimensional vector space to a two-dimensional representation that can be directly accessed and inspected visually.

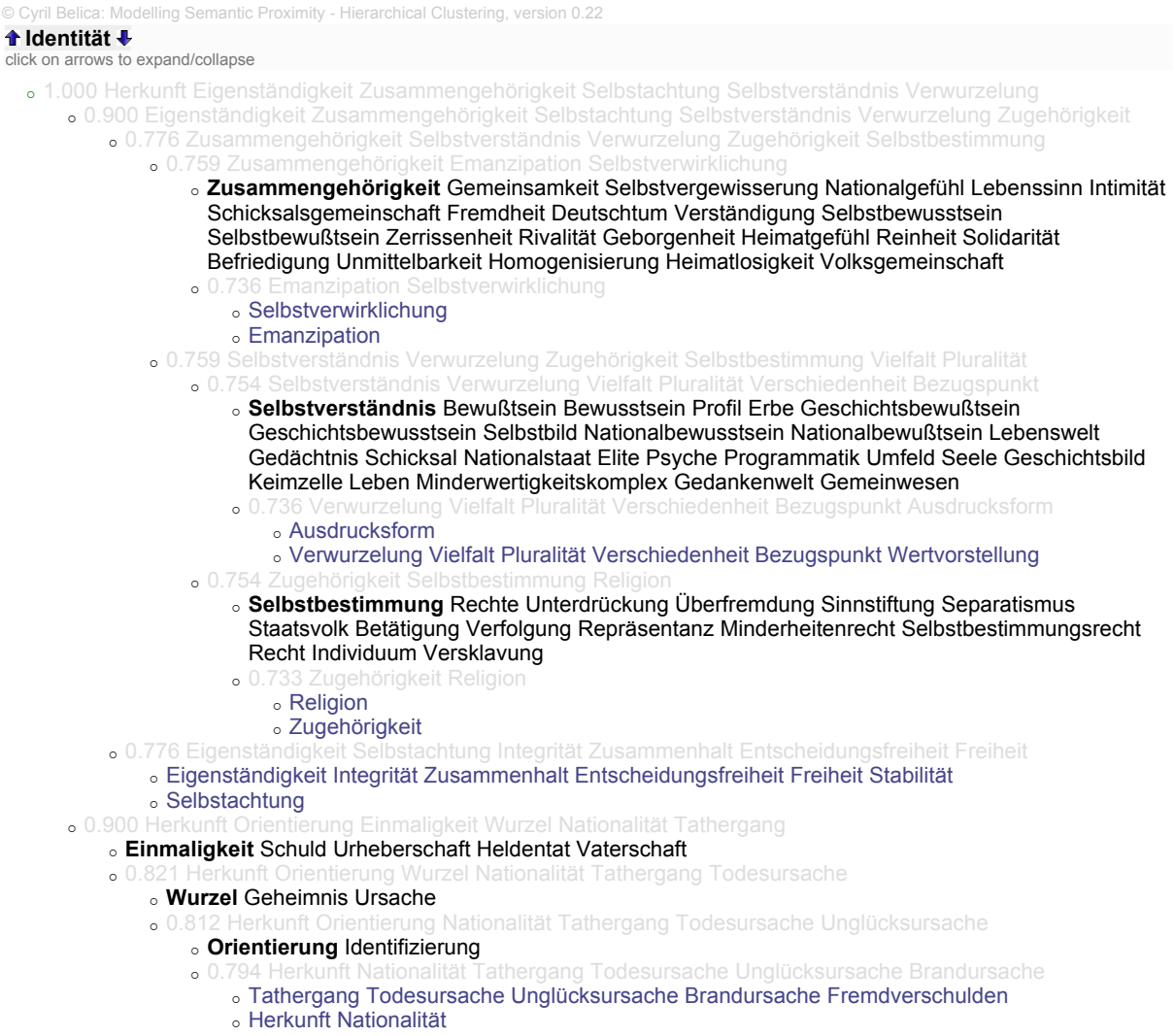

Figure 3. Partially expanded result of hierarchical cluster analysis of collocation profiles related to *Identität*.

The first method ("Modelling Semantic Proximity", cf. Figure 3) subjects these words to a *hierarchical cluster analysis*, whereas the second method ("Topographic Profile", not shown here) uses *self-organizing maps* (SOM; Kohonen, 1984) to arrange them on a two-dimensional grid such that proximity on the grid reflects similarity between the words' collocation profiles. Both methods lead to interesting insights into a word's semantic structure and other usage properties, especially for cases where these properties are particularly complex.

3.3 Contrasting Near-Synonyms*

In this method, we use self-organizing maps to contrast the usage properties of near-synonyms. Given two focal words, we first retrieve a set of words whose collocation profiles are most similar to that of either near-synonym. We then attempt to arrange the retrieved words on a two-dimensional grid according to the entangled pairwise similarity relations between their collocation profiles, as in 3.2 above. The cells on the grid are then colour-marked to reflect whether and to what degree they are related to either member of the given pair of near-synonyms (cf. Figure 4). The distribution of colours generally provides a reliable idea how similar the two near-synonyms really are with respect to their usage properties; moreover, it points to the particular usage aspects that the two words do not have in common.

Einsamkeit/Zweisamkeit

Heimatlosigkeit Vergänglichkeit existentiell existenziell Mühsal beklemmen Tragik Endlichkeit	Alleinsein Stille Weite Erhabenheit Vergeblichkeit Verlorenheit unendlich Unendlichkeit	Gemeinsamkeit Abgeschiedenheit Geborgenheit Zusammengehörigkeit Vertrautheit Harmonie danach Glückseligkeit	Beschaulichkeit Abwechslung Nachtruhe Privatheit Geselligkeit Idylle Muße Behaglichkeit	Zusammensein Lebensabend sorgenfrei vergönnt Naturerlebnis vergönnen harmonisch Atmosphäre
Verlassenheit Ausweglosigkeit Ohnmacht Zerrissenheit Sinnlosigkeit Trostlosigkeit Auflehnung Gefühl	Bitterkeit Melancholie Trauer Lebenslust verspüren Verzweiflung Wehmut Schwermut	Sehnsucht Zärtlichkeit Intimität Liebe Zartheit Empfindsamkeit herzzerreißend Innerlichkeit	Wonne Innigkeit Erlösung immerwährend Gegenwelt Ganzheit Heldentum Seligkeit	schönen Kindheit unvergesslich unvergeßlich Leben Moment schön Augenblick
Langeweile Leere Resignation Ratlosigkeit Verbitterung Tristesse Niedergeschlagenheit Frust	Wut Selbstmitleid Scham Selbsthaß Selbsthass abgrundtief Aggression Haß	unerfüllt grenzenlos unstillbar Leidenschaft Verliebtheit Verlangen Zuneigung hergerissen	unerfüllbar Erotik Selbstverwirklichung pubertär Träumerei voyeuristisch Körperlichkeit erotisch	innig nachhängen anrührend Zwiesprache Liebesfilm intim Kindheitserinnerung sehnsuchtsvoll
Hoffnungslosigkeit Hilflosigkeit Sprachlosigkeit Orientierungslosigkeit Apathie Frustration Erstarrung Gleichgültigkeit	Erniedrigung Schmerz Angst Schuldgefühl Existenzangst Selbstzweifel Demütigung Müdigkeit	Suff Selbstzerstörung Gier Todessehnsucht Eifersucht Abenteuerlust übersteigert übersteigern	Freundschaft platonisch Feindschaft Sexualität Hassliebe Haßliebe Treue entfremdet	Männerfreundschaft Umarmung Zwiegespräch Herzblatt Küssen Liebesszene Liebesleben Flirt
Vereinsamung Isolation Vereinzelung Erschöpfung Armut Entwurzelung Überforderung Verwahrlosung	Teufelskreis ausweglos Not Qual Depression Elend Entbehrung Leid		Familienleben Ehe Kleinfamilie ehelich außerehelich ehelichen Flitterwochen Ehejahr	Liebesbeziehung Liaison Verliebte unzertrennlich schenken Liebesspiel turteln umarmen

Figure 4. Topographic profile of the near-synonym pair *Einsamkeit* and *Zweisamkeit*.

* Joint research with Marie Vachková and Marek Schmidt, Charles University, Prague.

4 Future Directions

All the methods described above are currently being evaluated and refined. Additionally, our ongoing research focuses on measuring and exploring the similarity between syntagmatic patterns, and on the analysis of frequency distributions of the data with respect to text-external evidence such as time and topic. Any major progress concerning both data and algorithms is regularly included in the CCDB.

5 References

Kohonen T. (1984). Self-Organization and Associative Memory. Berlin: Springer.