

CCDB: A Corpus-Linguistic Research and Development Workbench

Holger Keibel¹ and Cyril Belica¹

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.

¹ Institute for the German Language (IDS)
Mannheim, Germany

e-mail: keibel@ids-mannheim.de, belica@ids-mannheim.de

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.

© Cyril Belica: Modelling Semantic Proximity - Hierarchical Clustering, version 0.22

↑ Identität ↓

click on arrows to expand/collapse

- 1.000 Herkunft Eigenständigkeit Zusammengehörigkeit Selbstachtung Selbstverständnis Verwurzelung
 - 0.900 Eigenständigkeit Zusammengehörigkeit Selbstachtung Selbstverständnis Verwurzelung Zugehörigkeit
 - 0.776 Zusammengehörigkeit Selbstverständnis Verwurzelung Zugehörigkeit Selbstbestimmung
 - 0.759 Zusammengehörigkeit Emanzipation Selbstverwirklichung
 - **Zusammengehörigkeit** Gemeinsamkeit Selbstvergewisserung Nationalgefühl Lebenssinn Intimität Schicksalsgemeinschaft Fremdheit Deutschtum Verständigung Selbstbewusstsein Selbstbewußtsein Zerrissenheit Rivalität Geborgenheit Heimatgefühl Reinheit Solidarität Befriedigung Unmittelbarkeit Homogenisierung Heimatlosigkeit Volksgemeinschaft
 - 0.736 Emanzipation Selbstverwirklichung
 - **Selbstverwirklichung**
 - **Emanzipation**
 - 0.759 Selbstverständnis Verwurzelung Zugehörigkeit Selbstbestimmung Vielfalt Pluralität
 - 0.754 Selbstverständnis Verwurzelung Vielfalt Pluralität Verschiedenheit Bezugspunkt
 - **Selbstverständnis** Bewußtsein Bewusstsein Profil Erbe Geschichtsbewußtsein Geschichtsbewusstsein Selbstbild Nationalbewusstsein Nationalbewußtsein Lebenswelt Gedächtnis Schicksal Nationalstaat Elite Psyche Programmatik Umfeld Seele Geschichtsbild Keimzelle Leben Minderwertigkeitskomplex Gedankenwelt Gemeinwesen
 - 0.736 Verwurzelung Vielfalt Pluralität Verschiedenheit Bezugspunkt Ausdrucksform
 - **Ausdrucksform**
 - **Verwurzelung** Vielfalt Pluralität Verschiedenheit Bezugspunkt Wertvorstellung
 - 0.754 Zugehörigkeit Selbstbestimmung Religion
 - **Selbstbestimmung** Rechte Unterdrückung Überfremdung Sinnstiftung Separatismus Staatsvolk Betätigung Verfolgung Repräsentanz Minderheitenrecht Selbstbestimmungsrecht Recht Individuum Versklavung
 - 0.733 Zugehörigkeit Religion
 - **Religion**
 - **Zugehörigkeit**
 - 0.776 Eigenständigkeit Selbstachtung Integrität Zusammenhalt Entscheidungsfreiheit Freiheit
 - **Eigenständigkeit** Integrität Zusammenhalt Entscheidungsfreiheit Freiheit Stabilität
 - **Selbstachtung**
 - 0.900 Herkunft Orientierung Einmaligkeit Wurzel Nationalität Tathergang
 - **Einmaligkeit** Schuld Urheberchaft Heldenat Vaterschaft
 - 0.821 Herkunft Orientierung Wurzel Nationalität Tathergang Todesursache
 - **Wurzel** Geheimnis Ursache
 - 0.812 Herkunft Orientierung Nationalität Tathergang Todesursache Unglücksursache
 - **Orientierung** Identifizierung
 - 0.794 Herkunft Nationalität Tathergang Todesursache Unglücksursache Brandursache
 - **Tathergang** Todesursache Unglücksursache Brandursache Fremdverschulden
 - **Herkunft** Nationalität

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-Synonyms²

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.

© Cyril Belica: Modelling Semantic Proximity - Contrasting Near-Synonyms (version: 0.17)

Einsamkeit/Zweisamkeit

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

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

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.

References

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