Automatic example sentence extraction for a contemporary German dictionary

Jörg Didakowski, Lothar Lemnitzer & Alexander Geyken

Keywords: *example extraction, digital dictionary, practical lexicography, natural language processing*.

Abstract

The integration of illustrative examples into monolingual dictionaries provides an intuitive means for grasping the meaning of a word. Tight space constraints of print media no longer apply with online dictionaries. Thus, the inclusion of examples is obviously a useful complement or substitute for the traditional ways of meaning exemplification. In this article, an approach is presented to automatically extract example sentences from a large German corpus collection. The extraction is done on the basis of the notions of sentence readability and complexity and word usage. The extracted examples are a good pre-selection for further integration into a digitized version of a contemporary German dictionary by lexicographers. A quantitative and qualitative evaluation of the extraction results is presented in the article. The work is related to the dictionary project *Digitales Wörterbuch der deutschen Sprache* (The Digital Dictionary of the German Language, DWDS in short) which integrates multiple dictionary and corpus resources and language statistics on the German language in a digital lexical information system which can be accessed on-line.

1. Introduction

The work presented in this paper is based on the DWDS dictionary, a digitized and enhanced version of the *Wörterbuch der deutschen Gegenwartssprache* (Dictionary of Contemporary German, WDG in short). The WDG is a large monolingual general language dictionary which comprises rich lexicographical information for approximately 90,000 headwords. Additionally, the WDG lists approximately 30,000 compound words under the articles of the heads of these compounds and without further exemplifying their meaning, cf. Herold and Geyken 2008. For the DWDS dictionary, the meaning(s) and uses of these compound words should be exemplified at least by a selection of examples from our corpora. We want to provide example sentences with the help of an example extractor tool. Based on an operational definition of ‘good example’, this tool presents to the lexicographer a limited number of sentences which are ranked highest with respect to goodness criteria.

2. Towards an operational definition of good example

A prerequisite for the extraction of examples and their subsequent assessment is to find an operational definition of goodness in the form of criteria which an example should meet. In our effort to define such criteria we refer to the work of Gisela Harras (1989) and Adam Kilgarriff (2008). The criteria must be operational in the sense that they can be matched by parameters of the extraction process and they can serve as guidelines for the intellectual assessment of the extracted examples. Harras mentions four basic criteria which a good example should meet; it should a) illustrate the prototypical features of the object or activity which the headword signifies; b) should present words with which the headword typically co-occurs; c) be authentic and d) contain words which are lexically-semantically related to the headword. Not all criteria are equally well-suited for our task: criterion a) is too vague to be operational and criterion c) is trivial in the sense that all our examples are from corpora and therefore authentic. However, criterion b) and d) will play an important role in the qualitative evaluation of the data set (see section 5.2). In addition, we define the following criteria: e) an example should be a complete, well-formed and not too complex sentence; f) the sentence should be self-contained, that is its content should be graspable without the larger context; g) the headword should not be used as a proper name and h) the set of extracted good examples should exemplify all meanings of the headword. Some of the criteria, in particular e) and f), are also mentioned in Kilgarriff et al. 2008. Kilgarriff and Rychlý 2010 provide an advanced approach to bridge the gap between readings in a dictionary and evidence in corpora, or in other words, to reconcile clean meaning delimitation in dictionaries and the bewildering variety of word usage in texts. Criterion h) is particularly important since we want to provide examples as the sole mean of exemplification. Some of these criteria are hard ones in the sense that if an example does not meet the criterion it will be dismissed (criteria e and g). Other criteria are soft ones in the sense that they only influence the quality score that is assigned to each example (criteria b, d and f). Criterion h) does not apply to single examples, but to sets of them. It will therefore play a prominent role in the evaluation (see section 5).

3. The corpora

The core of our corpus collection is the so-called DWDS-Kernkorpus. It consists of 100 million tokens and is a balanced collection of German texts of the twentieth century, that is it is roughly equally distributed over time and over five genres: journalism, literary texts, scientific literature, other nonfiction and transcripts of spoken language. Among others it contains literary monographs, poetry and dramatic works from major German writers (e.g. literary works of Franz Kafka, Günter Grass, and Martin Walser). In addition to the DWDS-Kernkorpus, five German newspapers from 1946 to 2008 are used: DIE ZEIT, DIE WELT, Bild, Süddeutsche Zeitung and Der Tagesspiegel. These corpora, which are a subset of the DWDS corpora, comprise approximately one billion tokens.

4. The extraction of good examples

4.1. *The method*

In order to automatically extract good examples for headwords from a corpus collection, the software should ideally act like a lexicographer. In consequence, the computer would be confronted with the problem of completely understanding the examples with regard to their content. However, the current state of computational linguistics still falls behind solving such a complex problem. Consequently, the task has to be simplified. This can be done with the help of operational criteria which are focused on sentence readability, complexity and word usage (cf. section 2). To make the notion of readability and complexity operational we use the following computational linguistic tools: a broad-coverage German morphology (TAGH, see Hanneforth and Geyken 2006), a part-of-speech tagger (moot, see Jurish 2003) and a dependency parser (SynCoP, see Didakowski 2008).

The dependency parser builds on the analyses of the other tools. It makes use of a hand-written grammar consisting of weighted pattern-based rules. The weights are used to model preferences of syntactic structures. Left and right embeddings of sentences are replaced by iteration (see Karlsson 2010) and center embeddings of sentences are restricted to a degree of one. If too much computer resource would be needed to parse a sentence exhaustively the analysis process is terminated. The rationale behind the usage of a parser is to provide a model for measuring sentence complexity and readability. The nesting of sentences, for example, is a good indicator for reading difficulties (see Gibson and Pearlmutter 1998). If the parser is not able to analyze a sentence, the sentence can either be seen as ungrammatical or as an unusual construction not covered by the grammar or as a sentence which is too complex.

The following crisp criteria for the formal assessment of sentence quality have been defined: a) sentence length: a good example sentence should be between ten and twenty-fife words long; b) completeness: a capitalized word in sentence initial position and a punctuation mark in sentence final position are strong indicators that the sentence is complete; c) known word: all words of a sentence have to be analyzable by the morphological component; d) no free pronouns: a sentence must not have substituting, reflexive or irreflexive pronouns; e) complexity: a sentence has to be parsable by the dependency parser.

Given a concordance for the headword the application of these crisp criteria leads to a reduction of the initial set of concordance lines. This reduced set is further narrowed down by additional global criteria: the resulting set should be balanced as much as possible over the decades or other time slices into which the corpus is partitioned. Furthermore, some specific documents of the DWDS-Kernkorpus are weighted higher than others and documents of the DWDS-Kernkorpus are in general weighted higher than documents of the newspaper corpora. This has to do with the quality of the texts which is expected to be higher in literal works than in other prose texts. Texts originating in the well sampled DWDS-Kernkorpus are considered to be of higher quality than texts from the opportunistically sampled newspaper corpora. A selection process with these parameters has been implemented which dismisses and ranks example candidates following the above-mentioned criteria. Finally, the remaining examples are ordered in respect to their goodness with the help of some soft criteria which are listed in the order of their importance: a) including words should be among the 17000 most frequent words of our balanced corpus; b) including words should be no longer than 15 characters; c) finally, the keyword should be within the matrix clause.

4.2. *Implementation*

We extracted approximately 200,000 lemma forms from the corpora. For these targets, we extracted good examples following the approach mentioned above. At this, we divided the twentieth century into five time slices of twenty years each with the first decade of the twenty-first century as an additional time slice. The application of the crisp and soft criteria is calculated in advance so that for the given set of time slices the best twenty examples, if they exist, for each target word are calculated. In order to make the examples accessible we implemented an internal web service which provides the extracted good examples in ranked order on request. The number of examples can be specified in a range from one to twenty. For a given lemma form the service returns the n best examples with additional information like date and origin. The important lexicographic criteria (see section 2) can be captured only partially and imperfectly by the methods that computational linguistics provides. However, we will show in section 5 that our method of automatic extraction of good examples does a good job in presenting input to the lexicographers.

5. Related work

Kilgarriff et al. (2008) present a similar strategy for the automatic extraction of good examples. Example extraction is done, however, with a different target audience in mind, that is language learners. Scores are assigned to sentences of a concordance on the basis of ranked ‘features’. Similar to our approach, their features also focuses on the notions of readability and complexity. The main difference of their approach to ours is that all sentences of a concordance are ordered by means of scores. That is, all features are soft ones in the sense that none of them must necessarily be met. In our approach the main criteria are crisp because we are not interested in an exhaustive ordering and we want to exclude all sentences which do not meet these hard criteria at the outset. Furthermore we make use of deeper linguistic information, for example parsing information, and we take metadata of our documents such as the publication date and the origin into account. Melo and Weikum (2009) present an approach to extract example sentences for specific readings of a word. Similarly to ours their aim is to provide example sentences as one means of meaning exemplification in a digital dictionary. The approach relies on word sense heuristics and word sense databases. In order to recognize the different word senses with great accuracy they make use of aligned parallel corpora. In contrast to our approach their focus is on word sense disambiguation and not on quality of extracted example sentences, that is they do not take sentence readability and complexity and word usage into account. There are current approaches in other research areas concerning sentence complexity and readability which are worth to be mentioned: In Tanguy and Tulechki (2009) different linguistic features are collected which deal with sentence complexity. They start with a huge set of automatically measurable features drawn from different research areas. This set is reduced to a smaller set by elimination of redundancy. Their goal is to identify the latent structure of sentence complexity. Heilman et al. (2008) deal with the prediction of reading difficulty of texts by means of a grade scale. They use lexical features and grammatical features derived from parsing sub-trees. These features are evaluated on different statistical models and for different scales of measurement. Their results show amongst others that grammatical features separately can be good predictors of readability. Tanaka-Ishii et al. (2010) present an approach where readability assessment is a relative comparison and not an absolute grading. In their approach texts are ordered by means of their readability. In order to implement a comparator they train a binary classifier on the basis of two sets of texts, one difficult and the other easy. They show that the approach is promising in the area of language learning.

6. Evaluation

6.1. *Quantitative evaluation*

We have checked and classified 19,000 examples for 5,076 headwords qualified as good examples by our classifier. The following classes and labels were used: ‘1’ for examples which are grammatically correct and at least acceptable for their function to exemplify one meaning of the headword; ‘2’ for examples which are acceptable but would need some minor corrections; ‘3’ for examples which are not acceptable because they are malformed, the headword is used as a proper name or the content of the sentence is offending. 18,113 examples (95.3%) have been rated with class 1; 342 (1.8%) with class 2 and 543 (2.9%) with class 3. It is worth noting that, since we have got 3.7 examples at average per headword, there are only 34 headwords for which we could not get any acceptable (i.e. class 1) example.

Another feature which is worth noting is the distribution of the examples over time. We divided the twentieth century into five time slices. The distribution of examples over these slices is as follows. The program extracted 750 examples (4.0%) for slice one (i.e. for the time between 1900 and 1919), 750 (6.75%) for slice two, 1341 (7.05%) for slice three, 4342 (22.75%) for slice four, 6704 (35.28%) for slice five and 4632 (24.37%) from the twenty-first century data. From these figures we can see that there is a bias towards more recent examples. One of the reasons might be that the majority of sources, that is all the newspaper corpora, contain data from the last thirty years. Nevertheless, the first half of the century has a still a significant share which is well above its share in the base data (approx. 6%). We will now proceed with a qualitative investigation of the data.

6.2. *Qualitative evaluation*

We focus here on Harras' criterion d) and our additional criterion h) (see section 2).
Ad d): Many examples illustrate semantically related words as cohyponyms, typically in the form of conjunctions:

(1) *Der Ökotourismus schafft Arbeit für Wildhüter, Fahrer, Kellner, Administratoren und* ***Fährtensucher***. (the headword is part of a list of jobs which you find in the tourism sector)

Ad h): The most important feature of good examples is, however, that they mirror the semantic structure of the headwords.

6.2.1. *Regular polysemy*. One typical example of regular polysemy is that between an activity and the organization which is carrying it out. This kind of polysemy is inherited by a compound from its head. Let us look at the word *Mission* which signifies either an activity or an organization (both: 'mission'). This regular polysemy can be found in the following examples for *Militärmission*:

(2) *Die britische* ***Militärmission*** *besteht aus je einem Vertreter des Heeres, der Marine und der Luftwaffe. (='organization')*

(3) *Die Mehrheit der Deutschen lehnt die* ***Militärmission*** *in Afghanistan laut Umfragen ab. (='activity')*

6.2.2. *Metonymy*. The word *Waffenrock* ('tunic') is a good example for the metonymous transfer from a piece of cloth to the person wearing it (a soldier):

(4) *Neben dem blassen kleinen Witmann liegt Christopher auf den Knien , zerrt und reißt und schneidet an dessen* ***Waffenrock*** *herum. (='cloth')*

(5) *Bundeswehroffiziere hielten die Studiotribüne mit großer Mehrheit besetzt; der Anzahl sichtbarer* ***Waffenröcke*** *mußte mindestens noch einmal die gleiche Menge militärischer Staatsbürger in Zivil hinzugezählt werden. (='soldier')*

6.2.3 *Non-literal senses*. The noun *Zugpferd*, in the literal sense, denotes an animal (a kind of strong horse) and, in the non-literal sense, a person who is acting like a strong horse.

(6) *Bei Ankauf von* ***Zugpferden*** *ist besonders auf gleiche Größe, Stärke, Kraft und Temperament zu sehen. (='animal')*

(7) *Ohne das* ***Zugpferd*** *Platzeck hätte die Brandenburger Linkspartei die SPD womöglich längst eingeholt. (='person')*

7. Summary and future work

An approach has been presented to fully automatically extract example sentences out of a German corpus collection for lexicographic purposes. In order to identify sentences which are good examples for a headword, some operational machine tractable criteria referring to sentence readability, complexity and word usage are used. Additionally, it is tried to balance the set of examples for a headword as good as possible over the decades. The good examples are made accessible via an internal web service returning sentences for a query word.
As a result of qualitative and quantitative evaluation the extracted examples of real language use mirror the semantic structure of the headwords which they represent as well as semantic processes. They can therefore reliably exemplify the headword in such cases where a full-fledged semantic description cannot be afforded.

We will provide a web service by which a larger public can retrieve good examples for individual words. By this we hope to reach also other communities of users such as language teachers and learners.

References

A. Dictionaries

Klappenbach, R. and W. Steinitz 1964-1977. *Wörterbuch der deutschen Gegenwartssprache* 1-6. Berlin: Akademie-Verlag.

B. Other literature

Didakowski, J. 2008. ‘Local Syntactic Tagging of Large Corpora Using Weighted Finite State Transducers.’ In A. Storrer et al. (eds.), *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing*. KONVENS 2008. Berlin: Mouton de Gruyter, 65–78.

Geyken, A. 2007. ‘The DWDS Corpus: a Reference Corpus for the German Language of the 20th Century.’ In C. Fellbaum (ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum Press, 23–41.

Geyken, A. and T. Hanneforth 2006. ‘TAGH: a Complete Morphology for German Based on Weighted Finite State Automata.’ In A. Yli-Jyrä et al. (eds.), *Finite-state methods and natural language processing, 5th international workshop, FSMNLP 2005, Helsinki, Finland, Revised Papers*. Berlin/Heidelberg: Springer, 55–66.

Gibson, E. and N. J. Pearlmutter 1998. ‘Constraints on Sentence Comprehension.’ *Trends in Cognitive Sciences* 2.7: 262–268.

Harras, G. 1989. ‘Theorie des lexikographischen Beispiels.’ In F. J. Hausmann et al. (eds.), *Wörterbücher Dictionaries Dictionnaires: Ein internationales Handbuch zur Lexikographie*. Berlin/New York: de Gruyter, 1003–1114.

Heilman, M. et al. 2008. ‘An Analysis of Statistical Models and Features for Reading Difficulty Prediction.’ In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*. Columbus, Ohio: Association for Computational Linguistics, 71–79.

Herold, A. and A. Geyken 2008. ‘Adaptive Word Sense Views for the Dictionary Database eWDG: the Case of Definition Assignment.’ In A. Storrer et al. (eds.), *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing*. KONVENS 2008. Berlin: Mouton de Gruyter, 209–221.

Jurish, B. 2003. *A Hybrid Approach to Part-of-Speech Tagging*. http://www.ling.uni-potsdam.de/~moocow/pubs/dwdst-report.pdf.

Karlsson, F. 2010. ‘Recursion and Iteration.’ In H. Hulst (ed.), *Recursion and Human Language*. Berlin/New York: De Gruyter Mouton, 43–47.

Kilgarriff, A. et al. 2008. ‘GDEX: Automatically Finding Good Dictionary Examples in a Corpus.’ In E. Bernal and J. DeCesaris (eds.), *Proceedings of the XIII EURALEX International Congress: Barcelona, 15-19 July 2008***.** Barcelona: l'Institut Universitari de Lingüística Aplicada (IULA) dela Universitat Pompeu Fabra, 425–432.

Kilgarriff, A. and P. Rychlý 2010. ‘Semi-automatic Dictionary Drafting.’ In G.-M. de Schryver (ed.), *A Way with Words: A Festschrift for Patrick Hanks*. Kampala: Menha Publishers, 299–312.

Klein, W. and A. Geyken 2010. ‘Das digitale Wörterbuch der Deutschen Sprache (DWDS).’ *Lexicographica: International Annual for Lexicography* 26: 79–96.

Melo, G. and G. Weikum 2009. ‘Extracting Sense-Disambiguated Example Sentences From Parallel Corpora.’ In *Proceedings of the 1st Workshop on Definition Extraction*. Borovets: Association for Computational Linguistics, 40–46.

Tanaka-Ishii, K. et al. 2010. ‘Sorting Texts by Readability.’ *Computational Linguistics* 36.2: 203–227.

Tanguy, L. and N. Tulechki 2009. ‘Sentence Complexity in French: a corpus-based approach.’ In M. A. Kłopotek et al. (eds.), *Recent Advances in Intelligent Information Systems*. Warsaw: Academic Publishing House EXIT, 131–144.