

---

# Tuning Co-occurrences of Higher Orders for Generating Ontology Extension Candidates

---

**Marek Mahn**

University of Leipzig, Ifi, NLP Department, Augustusplatz 10/11, 04109 Leipzig, Germany

MAREKMAHN@GMAIL.COM

**Chris Biemann**

University of Leipzig, Ifi, NLP Department, Augustusplatz 10/11, 04109 Leipzig, Germany

BIEM@INFORMATIK.UNI-LEIPZIG.DE

## Abstract

This work introduces the notion of higher order co-occurrences as a data source for corpus-based ontology extension. After describing various parameter settings for the process we evaluate the approach on a manually annotated dataset. An outlook on further applications is given

co-occurrences to be significant if their significance score is above a certain threshold. This measure, belonging to the log-likelihood family, has proven to be the most performant for extracting PP-verb-collocations in (Krenn & Evert 2001), especially in comparison to the Mutual Information measure (as used in Church & Hanks 1990).

## 1. Introduction

For a corpus-based extension of ontologies, statistical methods to find related words are needed. In this work we describe and evaluate co-occurrences of higher orders as a source for this task. Having a method at hand that finds semantically related words automatically from a corpus, an ontology can be extended by manually labelling the candidates with the appropriate relation.

### 1.1 Co-occurrence Statistics

A well-known approach in statistical NLP to gain related terms to a given term is the calculation of (statistically significant) co-occurrences (see e.g. Charniak (1994); Manning and Schütze (1999)).

The occurrence of two or more words within a well-defined unit of information (word window, sentence, document) is called a co-occurrence. For the selection of meaningful and significant co-occurrences, an adequate co-occurrence measure has to be defined: Our significance measure compares the actual co-occurrence numbers with a Poisson distribution: Given two words  $A$ ,  $B$ , each occurring  $a$ ,  $b$  times in sentences, and  $k$  times together, we calculate the significance  $sig(A, B)$  of their occurrence in a sentence as follows. We set  $x=ab/n$  (here,  $n$  is the number of sentences or documents) and use the following approximation (taken from Lauter and Quasthoff 1999):

$$sig(A, B) = x - k \log x + \log k!$$

The measure gives us the possibility to order the co-occurrences of a term by their significance. We consider

### 1.2 Iteration of Co-occurrences

In this section we describe how to get co-occurrences of higher order by iterating the co-occurrence calculation process (see Biemann, Bordag and Quasthoff 2004). Recall that for calculating co-occurrence significance values, we basically counted how often two words appear together and without each other in one sentence (or window) and applied our measure to this.

If we regard the terms with the  $N$  highest ranked co-occurrences of a source term as a new 'pseudosentence' (leaving out the significance values, which showed to have almost no effect in preliminary experiments) and apply the same calculation to the corpus of all those new sentences, we get co-occurrences of second order. In this way, we operate on a corpus not consisting of natural language sentences but consisting of co-occurrences, taking as windows the co-occurrences of single terms.

High significance values in co-occurrences of second order mean that these words often appeared together in co-occurrence sets of first order. This means in turn that the respective words appear together in similar contexts.

Iterating again, we can arrive at co-occurrences of  $(n+1)$ -th order by taking co-occurrence sets of  $n$ -th order as input and applying the co-occurrence calculation.

While second-order co-occurrences usually reflect context similarity, or to put it in other words, contain paradigmatic relations (cf. de Saussure 1916, Rapp 2002), it is at the first glance not intuitively clear what happens in the higher orders. In data up to the order of 10 it is observable that the top  $N$  co-occurrences of high order exhibit a certain kind of semantic homogeneity amongst their member terms, but it cannot be foreseen in which attribute this homogeneity shows up. Differences can be

obtained by varying the N for choosing pseudosentence length and by starting with different window sizes (sentence-based vs. document-based): broader windows tend to extract broader dependencies and relations. In general, most of the terms belong to the same subject area as the reference term.

Table 1 shows some examples for co-occurrences of higher orders for an English corpus. Recall that the first order is the usual notion of co-occurrence, co-occurrences of (n+1)st order are words that appear often together in co-occurrence sets of n-th order.

Table 1: Examples for co-occurrences of higher orders for an English corpus. Here, a sentence was chosen as window.

order	reference word	top 10 co-occurrences
10	wine	wines, grape, sauvignon, chardonnay, noir, pinot, cabernet, spicy, bottle, grapes
1	ringing	phone, bells, phones, hook, bell, endorsement, distinctive, ears, alarm, telephone
2	ringing	rung, Centrex, rang, phone, sounded, bell, ring, FaxxMaster, sound, tolled
4	ringing	sounded, rung, rang, tolled, tolling, sound, tone, toll, ring, doorbell
10	aluminum	chromium, nickel, Aluminum, cans, Alcoa, Aluminium, Fetterolf, Metals, Parry, smelter
5	warm	temperatures, weather, winter, cold, winds, snow, Plains, degrees, rain, nation
10	pressing	Ctrl, Shift, press, keypad, keys, key, keyboard, you, cursor, menu, PgDn, keyboards, numeric, Alt, Caps, CapsLock, NUMLOCK, NumLock, Scroll

The idea of using second order co-occurrences has been applied before e.g. to query expansion by (Ruge 1992), and word sense discrimination by (Schütze 1998). However, we could not find literature on co-occurrence statistics for higher orders than the second.

## 2. Co-occurrences of Higher Order in the dpa-corpus

### 2.1 The dpa Corpus

For our experiments, we used newswire stories obtained from the "Deutsche Presseagentur" (German press agency). Our portion of the dpa corpus contains all 229500 dpa news items of the year 2000. These documents consist of 4 up to 250 words (after eliminating stopwords), the average document length is 110 words.

The scope of the co-occurrence calculations in the following experiments is one document – so the scope is chosen quite wide here. This is due to the fact, that the calculations shall be comparable with those of Probabilistic Latent Semantic Analysis (Hofmann 1999) as applied by (Paass et al. 2004).

### 2.2 Evaluation Resources

Since we can obtain a ranked list of words for a given input word (e.g. higher order co-occurrences ordered by significance), it is possible to evaluate, how many of these words are semantically related to the input word. For evaluation we use a program called WordSetAnalyzer (see Bordag et al. 2005), which is especially suited for this task and evaluates the pair of input word and ranked word list against manually developed resources like WordNet (Fellbaum 1998, Budanitski & Hirst 2001), GermaNet (Hamp & Feldweg 1997), or others. In this work we use a manual annotation that was carried out at the University of Leipzig for German language.

Table 2: Most frequently assigned relations. Part of speech is denoted by A (for adjective), N (noun), and V (verb).

Relation	Fraction
N-N-Co-hyponym	28%
N-N-Hyponym / Hyperonym	10%
N-N-Synonym	9%
A-N-Typical Property	6%
N-N-Typical Location	6%
A-A-Co-hyponym	5%
V-V-Co-hyponym	4%
N-N-Part or Substance	4%
A-A-Synonym	3%
N-N-Antonym	2%

In total, 46 different semantic labels for single words and 57 different relations (including relations between equal and different part-of-speech) were assigned to in total 90'000 different single words and 110'000 pairs of words. The set of relations used is based on Meaning-Text-Theory (Steele 1990) and was reduced to the most frequent relations found between co-occurring words. Table 2 shows the amount of the most frequently assigned relations. Co-hyponyms (words that have a common hypernym like "beer" and "wine") account for over one forth of annotated relations.

### 2.3 Experiments

In this section we describe various experiments on the dpa-corpus and evaluate them against the manually annotated data. Our main evaluation measure for the quality of our results is the portion of co-occurrences for which annotated relations exist.

On the other hand, we also have to keep an eye on the quantity of our results measured by the number of co-occurrences and the number of words for which co-occurrences exist. As only a small fraction of all possible relations is annotated, and the resource was not designed as a gold standard for the dpa corpus, the values of our quality measure will be quite low. However, as we do not use it for universal judgment about the quality of our method, but for comparison of different of its configurations this measure is sufficient.

The main parameters of the iteration process are:

- *sentence length*: number of words in pseudosentences
- *significance threshold*: minimum significance for a term to appear in pseudosentence, value 0 means: no significance threshold
- *include source term*: binary attribute – the source term whose co-occurrences are used to generate the pseudosentence is/is not included in the pseudosentence.

For different settings of this parameters the first to tenth orders were calculated in order to examine the influence of the parameters to the semantic relatedness of the resulting word sets.

Figure 1 shows the results for *sentence length* 10, 20 and 50, *significance threshold* 0 and 20 with *source term included*. There seems to be no further improvement of the results after the 3<sup>rd</sup> order – in the cases without significance threshold even after the 2<sup>nd</sup> order. Furthermore it seems that a lower *sentence length* and a higher *threshold* lead to better results.

But before looking closer at these parameters we will first examine the influence of including the source term.

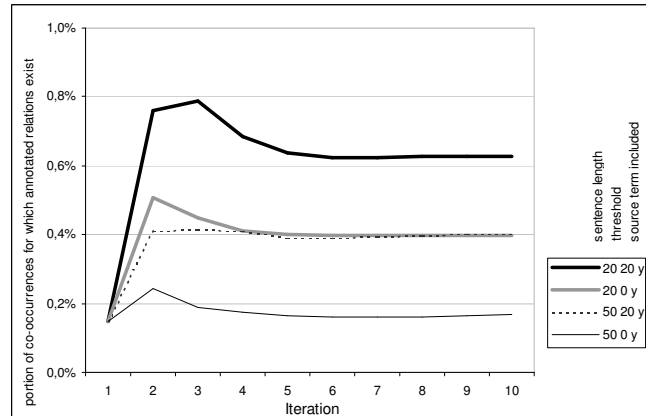


Fig. 1: Evaluation for different settings on annotation data

Figure 2 shows the best configuration from figure 1 – sentence length 20 and threshold 20 – with and without the source term included in the pseudosentence. While the result in the 2<sup>nd</sup> order is nearly the same, in order 3 to 5 not including the source term leads to much better results. However, from the 6<sup>th</sup> order on, no co-occurrences exist at all, as there are no more terms with significances exceeding the significance threshold. As the best results are found in lower orders, further experiments were made without the source term in the pseudosentences. To verify our observation, that lower sentence length leads to better results, the co-occurrences for a sentence length of 10 were calculated. This configuration shows the best results in Figure 2.

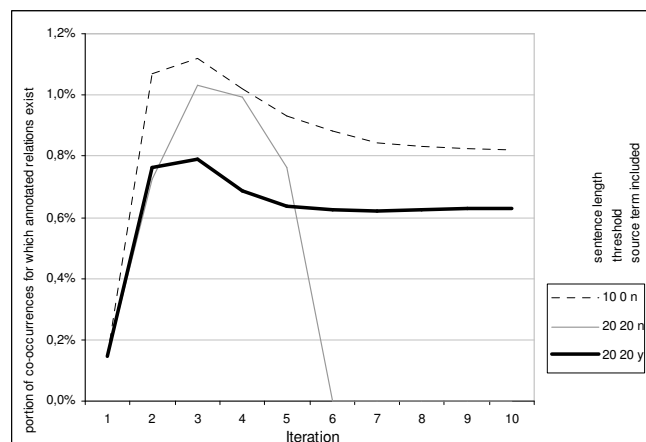


Fig. 2: Quality improvement by not including source term and smaller sentence length

Applying a high significance threshold to this configuration should lead to further improvement. Figure 3 shows the best configuration of Figure 2 in comparison to a configuration with sentence length 10 and threshold 50. Our quality measure rose to 2,2% in the 2<sup>nd</sup>, 5,2% in the 3<sup>rd</sup>, 8,8% in the 4<sup>th</sup> order. No co-occurrences at all were obtained in the higher orders with this parameter setting.

These evaluation metrics, however, do not take into account the actual number of co-occurrences obtained. In this case, we traded the number of words for which higher order co-occurrences exist at all for quality, obtaining few but good results. While for some application this strong reduction in numbers will hurt coverage dramatically, we propose how to still make use of this effect in section 3.

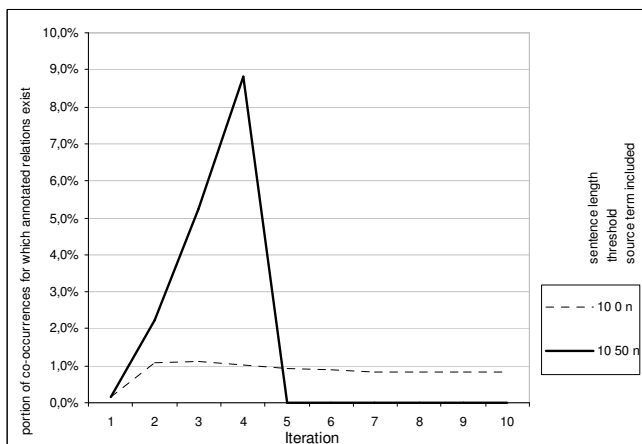


Fig. 3: Further improvement by applying a high significance threshold

To put it into a nutshell: a low pseudosentence length is in virtually all cases advisable. In other words: only the highest ranked co-occurrences of previous orders should be used to calculate the higher order's statistics. A high significance threshold may lead to better results, but it may reduce the number of existing co-occurrences too much. So the use of a significance threshold should strongly depend on what we want to use the co-occurrences for. If we need co-occurrences for an extensive set of the terms of our source documents, using a high significance threshold is not advisable – if we only need the best co-occurrences using a high significance threshold will be the right choice.

Furthermore, not including the source terms in the pseudosentences proved to be the better choice. This can be motivated as follows for the first and second order: While we mix first and second order relationships when including the source term, we consider only pure second-order relationships when omitting it, which rules out syntagmatic effects.

In most cases it is advisable to use the co-occurrences of 2<sup>nd</sup> or 3<sup>rd</sup> order for further processing, as in these orders the balance between quality and quantity of the co-occurrences is at its best.

In (Biemann, Bordag and Quasthoff 2004) a manual evaluation on co-occurrences of 2<sup>nd</sup> and 3<sup>rd</sup> order was carried out. Here, no manual evaluation resource was employed but the co-occurrence sets of single words and the disjunction of co-occurrences of two synonyms were examined. The precision values for combined co-hyponymy, hyponymy and synonymy relations to the

reference word(s) scored at 40-45% for single words (nouns) and at 70-75% for the synonyms (nouns).

### 3. Further Experiments and Outlook

Currently some experiments of combining Co-occurrences and Probabilistic Latent Semantic Analysis (PLSA) are in progress. These include two different ways of concatenating these techniques.

The first approach is the use of PLSA as a Word Sense Disambiguator and the subsequent calculation of co-occurrences on the disambiguated words. This is done by marking all occurrences of a term by the number of the most probable PLSA-class for this occurrence. This class is calculated from class-probability-vectors for the term and for the document the term occurs in, by selecting those classes which have the highest probability in the document's vector from the set of classes which have a probability in the term's vector exceeding a certain threshold.

The second approach deals with reducing the performance requirements of PLSA. In section 2.3 we showed that there are configurations of the iteration process, which lead to a strong reduction of the quantity of words, for which higher order co-occurrences exist. Could this be used for calculating PLSA equally good, but cheaper? The words, for which higher order co-occurrences exist, are most likely words with strong semantic ties in the document – therefore using only these words for calculating PLSA should lead to nearly equally good class assignment at much lower calculation costs. If this assumption should prove true, this approach will be compared to other means of word reduction based on tf/idf.

Another issue is the determination of the actual relations in order to build a taxonomy. While it is generally possible to semi-automatically extend prototypical taxonomies with a large number of words in the hierarchy nodes using co-occurrences as shown in (Biemann, Shin and Choi 2004) by attaching the newly inserted words to the most similar node, a fine-grained hierarchy as constructed in WordNet (Fellbaum 1998) is beyond the scope of statistical methods that employ the bag-of-words model on document- or sentence basis. These methods can extract high quality candidates, but their hierarchical ordering must be left to pattern-based methods that take positional and context information into account, as e.g. elaborated in (Hearst 1992 or Caraballo 1999).

### 4. Conclusion

We introduced the notion of word co-occurrences of higher order as an iteration of co-occurrence calculation. Experiments showed that co-occurrence significance is an adequate measure to sort co-occurring words in terms of quality. Further, the 2<sup>nd</sup> and 3<sup>rd</sup> order co-occurrences

showed to be most adequate for finding related words in a corpus-based way. While the approach is over-generating in a way that a lot of unrelated words are produced, it still aids ontology creation by offering candidates and appropriate regions to put them into the hierarchy.

### References

- Biemann, C.; Bordag, S.; Quasthoff, U. (2004): Automatic Acquisition of Paradigmatic Relations using Iterated Co-occurrences, *Proceedings of LREC2004*, Lisboa, Portugal
- Biemann, C., S.-I. Shin, und K.-S. Choi (2004): Semiautomatic extension of corenet using a bootstrapping mechanism on corpus-based co-occurrences. In: *Proceedings of the 20th International Conference on Computational Linguistics, COLING04*, Genf, Switzerland, 2004.
- Bordag, S., Witschel, F.H., Wittig, T. (2005): Evaluation of Lexical Acquisition Algorithms. In W. Hess und W. Lenders (Hrsg.) "Sprache, Sprechen und Computer/Computer Studies in Language and Speech" and *Proceedings of GLDV-Frühjahrstagung 2005*, Bonn, Peter-Lang-Verlag, Frankfurt am Main.
- Budanitski, A. and Hirst, G. (2001): Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, *Workshop in WordNet and Other Lexical Resources, in NAACL-2001*, Pittsburgh, USA
- Carballo, S.A. (1999): Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. In *37th Annual Meeting of the Association for Computational Linguistics*, pp. 120-126.
- Charniak, E. (Ed.) (1994): *Statistical Language Learning*, MIT Press, Cambridge, Massachusetts, USA
- Church, K.W. and Hanks, P. (1990): Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22--29, 1990.
- Fellbaum, C. (Ed.) (1998): *WordNet – An Electronic Lexical Database*, MIT Press, May 1998
- Hamp, B., H. Feldweg (1997): GermaNet - a Lexical-Semantic Net for German. In: *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, 1997.
- Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of COLING-92*, Nantes, Vol. 2 (1992) 539-545.
- Hofmann, T. (1999): Probabilistic Latent Semantic Analysis. In: *Uncertainty in Artificial Intelligence, UAI*, 1999
- Krenn, B. and Evert, S. (2001): Can we do better than frequency? A case study on extracting PP-verb collocations. *Proceedings of the ACL-2001 Workshop on Collocations*, Toulouse, France
- Läuter, M. and Quasthoff, U (1999): Kollokationen und semantisches Clustering. In *11. Jahrestagung der GLDV*, Enigma Corporation, Prague
- Manning, C., Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, Massachusetts, USA
- Paaß, G., Kindermann, J. and Leopold, E. (2004): Learning Prototype Ontologies by Hierarchical Latent Semantic Analysis. *Proc. ECML/PKDD 2004 Workshop on Knowledge Discovery and Ontologies*. Pisa: University Press, 2004, p.49-60.
- Rapp, R. (2002): The Computation of Word associations: Comparing syntagmatic and Paradigmatic Approaches, *Proceedings of COLING-02*, Taipei, Taiwan
- Ruge, G (1992): Experiments on linguistically-based term associations. *Information Processing and Management*, 28(3):317-322
- Saussure, F de. (1916) : *Cours de Linguistique Générale, Paris, Payot*
- Schütze, H. 1998. Automatic Word Sense Discrimination. *Computational Linguistics* 24:1-40, pp. 97-124
- Steele (1990): *Meaning Text Theory: Linguistics, Lexicography and Implications* by Steele, J (ed.). University of Ottawa Press, 1990