On the Relation between syntactic form and statistical association measures

Extended Abstract
Christiane Hümmer, Alexander Geyken, Rita Finkbeiner
E-Mail: (huemmer|geyken|finkbeiner)@bbaw.de

The automatic extraction of linguistically relevant collocations from corpora relies on a variety of statistical association measures (AM). Candidate collocations are extracted along with a score produced by the AM and recall is evaluated against the judgement of a human annotator.

Previous studies have failed to show a clear relation between AM and recall with respect to a given test set (e.g. Dunning 1993). The best AM is the one that predicts the maximum number of collocations in terms of recall and precision. Collocations by this evaluation standard are defined as having a „typical" relation between two candidate terms. This absence of a clear relation between AM and recall, however, does not depend so much on the question of how to evaluate the „typicality" of the relation between two candidate terms. Neither does it prove that there is no „best" statistical test available for the detection of collocations. We argue that, instead, these negative results reveal a problem of a different nature, namely the absence of an adequate characterization of collocations.

Krenn and Evert (2001) suggest to set up narrowly defined classes of collocations and to test the recall statistical AM against these narrower collocation classes. In the same paper they give some evidence for this approach and show that Mutual Information (MI; Hanks and Church, 1989) provides significantly better results in terms of precision for the detection of light verb constructions than mere cooccurrence frequency would have predicted.

Following this idea, we established a linguistic classification of German collocations based on fine-grained distinctions among construction classes. The term "collocation" is used in a broad sense, and constitutes the focus of our project Collocations in the Dictionary at the Berlin-Brandenburg Academy of Sciences. We include idiomatic phrases (*kick the bucket*), narrow collocations (*pay attention* or *blind alley*) as well as selectional preferences such as (*brush teeth*). The latter types of collocations differ from idiomatic expressions by their semantic transparency. Idiomatic expressions are said to comprise only semantically opaque components, whereas collocations contain at least one transparent component (e.g. Aisenstadt 1981, Mel'čuk 1998).

The main idea of our classification is to use a finer granularity than just `N-V` od `V prep N`. For example, *alle Register ziehen* (*to pull out all the stops*) would belong to a class `V Quant N`, *im Grunde meines Herzens* (lit. ‚deep in my heart') would be classified as `Prep N + Poss Ngen` or auf *seine Kosten kommen* (‚get one's money's worth') as `prep POSS N V`. Additionally, our classes are refined with syntactic properties. For example, we encode that *in Rechnung stellen* (*to invoice*, lit. ‚charge to one's account') can only be used in a singular form **in Rechnungen stellen* or that modal passive is possible: *in Rechnung ... zu stellen*. Other properties encode the syntagmatic order: some `ADJ Noun` collocations can only be used attributively but not predicatively without losing their idiomatic character: *blind alley* but **the alley is blind*. Note however, that these properties do not hold for all `ADJ Noun` collocations: *heikles Thema* (‚sensitive issue') and *das Thema ist heikel* (‚the issue is sensitive').

At present, we have identified some 65 different construction classes and 10 syntactic properties. These construction classes can be subdivided into phrasal (AdjP, AdvP, NP, and PP) verbal

constructions[1]. The latter are more interesting for us since they imply more syntactic restrictions and consist of longer sequences of PoS-strings. The length of the sequences of PoS-strings vary from 3[2] (e.g. Det-N-V, *eine Entscheidung treffen*, ‚take a decision') to 6 (e.g. Pronrefl-Prep-N-Conj-N-V, *sich um Kopf und Kragen reden*, risking life and limb by saying s.th.). The sequences are neither completely frozen nor adjacent. The degree of frozenness is determined by the number of components and their variation according to adjacency and paradigmatic substitution. Generally, we would expect the degree of frozenness to increase with the length of the sequence. Hence, the shortest sequence Det-N-V which covers support verb constructions as well as figurative expressions is quite „free". Its components are discontinuous since there can be an adjective phrases between the determiner and the noun as well as almost arbitrary material between the noun and the verb. Also variation of the verb form (*Entscheidung treffen | trafen | trifft*) and of the determiner (*die | eine Entscheidung*) is possible. On the other hands, PoS-strings with the length of 6 contain almost certainly frozen components. In the example above, *um Kopf and Kragen* is the longest frozen substring.

Our hypothesis for which we got preliminary evidence is that it is easier to find appropriate AM's for fine-grained collocation classes than it is to identify one appropriate AM for all collocation types. As an example, we take the class `(Prep) POSS N * V` which comprises collocations like *auf seine Kosten kommen* (‚get one's money's worth') or *unter seine Fittiche nehmen* (to take s.o. under one's wing).

[1] We do not consider sentential constructions and formulae

[2] In some rare cases even 2 is possible: N-V, *Bauklötze staunen*, to be flabbergasted

In a preliminary case study, we used four different statistical AM (MI, log likelihood, Dice, entropy) and tested them against a balanced corpus of newspapers, fiction, and non-fiction texts of about 12 million words (http://www.dwds.de). Our first goal is to test for differences with respect to syntactic form. Previous work (Kreand Evert, 2001) have shown that MI yield higher scores for support verbs with scores between 4.0 and 7.5. On the other hand, candidate (pairs) where one word has a very high marginal frequency get better scores for log likelihood and entropy.

In order to test the hypothesis we use a very simple method: The collocation class is decomposed in its subsequent pairs, i.e. `prep POSS`[3], `POSS N` and `N V`. According to the observations above we apply entropy (log-likelihood) to the second bigram and MI (Dice) to the third bigram since we suppose `POSS` to be highly frequent (*der* is 2^2 more frequent than `POSS`) with respect to verbs which are less frequent (i.e. the article *der* is 2^6 more frequent than *kommen*). We observe that `POSS N` and `N V` pairs are rated higher than its rating by mere co-occurrence frequency[4]. If we take the product of its scores as the score of the combination of two AM scores then collocation candidates should get a better ranking than they would according to co-occurrence frequency. Indeed, when applying this simple method we were able to find the collocation *(auf) seine Rechnung kommen* (‚get one's money's worth') which was ranked better than by mere frequency.

In the next two months we will test and evaluate our hypothesis against a larger corpus of about 100 m running textwords, try different collocation classes, and expand the AM to non adjacent examples of construction classes.

[3] We do not compute the AM of function words.

[4] We presuppose our corpora to be POS tagged.

References:

Aisenstadt,Ester (1981): Restricted Collocations in English Lexicology and Lexicography. In: Review of Applied Linguistics 53, S. 53-61.

Church, Ken, and Hanks, Patrick (1989). Word association Norms, Mutual Information, and Lexicography. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, p. 76-83, Vancouver, Canada.

Dunning,Ted (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1), 61 – 74.

Krenn, Brigitte and Evert, Stefan (2001) Can we do better than frequency? A case study on extracting PP-verb collocations in Proceedings of the ACL Workshop on Collocations, Toulouse, France.

Schemann, Hans (1993): Deutsche Idiomatik. Die deutschen Redewendungen im Kontext. Pons:Stuttgart.

Appendix: List of verbal construction classes

1. Det Adj N V
2. Det N Adj Conj Det N V
3. Det N V
4. Det N Prep V
5. Det N Pron V
6. Det N V
7. N V
8. Adj Conj Adj V
9. Adj Conj Det N V
10. Adj N V
11. Adv Adj V
12. Adv Conj Adv V
13. Det Adj N Adv Adj V
14. Det Adj N Prep V
15. Det Adj N V
16. Det N Prep N V
17. Det N Prep V
18. Det N Ptc Prep Det N V
19. Det N V
20. Conj Adj V
21. Conj Det Adj N V
22. Conj Pron Prep N V
23. N Conj N V
24. N Prep V
25. Adj V
26. Det N Prep Det N V
27. Det N V
28. Prep N Conj N V
29. Prep Det N V
30. Prep N V
31. Prep Det Adj N V
32. Prep Det N V
33. Prep N Adj V
34. Prep N Det N V
35. Prep N Prep V
36. Prep N V
37. Prep Det N Prep N V
38. Prep Prep N V
39. Prep Pron Prep V
40. Pron Det N V
41. Pron Prep Det N V
42. Pron V
43. Pron Prep Det N V
44. Pron Prep N Conj N V
45. Pron Prep N V
46. Pron Pron N Conj N Adj N V V
47. Pron Adj Prep Adj V
48. Pron Conj Prep N Prep N V
49. Pron Conj Prep N V
50. Pron N V
51. Pron Prep Det N V
52. Pron Prep N V
53. Pron Prep Prep V