Measuring syntagmatic Fixedness of Multi-Word Expressions

Axel Herold and Katerina Stathi
Berlin-Brandenburgische Akademie der Wissenschaften and Freie Universität Berlin
herold@bbaw.de and *stathi@zedat.fu-berlin.de*

1 Introduction

Syntagmatic fixedness is an important feature of multi-word expressions (MWE). However, syntagmatic fixedness is gradual and various semantic and syntactic relations hold among the parts of MWEs. This poses intriguing problems for lexicography, linguistic description and language processing.

In this paper we propose a computationally inexpensive and intuitive approach to the measurement of syntagmatic fixedness based on positional co-occurrence data that is not easily captured by simple statistical significance tests.

The types of relations between frequently and systematically co-occurring lexical items have been the subject of studies in phraseology, corpus lexicography and corpus linguistics. MWEs have been classified according to different criteria: the compositionality of their meaning, their syntactic structure (phrasal vs. sentential), their internal structure, their grammatical well-formedness, their communicative function, their metaphoricity. For classifications of MWEs *cf.* (Moon, 1998; Cowie, 1998; Burger, 1998).

Types of MWEs most frequently cited include named entities, idioms, proverbs, similes, routine formulae and sayings. To this we might add conventional metaphors which often involve the co-occurrence of lexical items.

In the following we will focus on three important concepts in this discussion, namely co-occurrence, collocation, and idiom.

Our emphasis is on co-occurrence between tokens of the underlying corpus but other notions are possible. In general, co-occurrence is seen as the existence of words in structural or positional adjacency or proximity (Evert, 2005). The concept of co-occurrence is discussed in more detail in section 3.

Whereas the term co-occurrence is basically used in a rather uniform way, the term collocation has been used in a variety of ways, which can be summarised as follows:

1. frequently repeated or statistically significant co-occurrences, whether or not there are any special semantic bonds between collocating items (Sinclair, 1991; Moon, 1998). Fellbaum (2007) distinguishes between *collocation* in this sense and *collocations* (*cf.* the next point).
2. sequences of lexical items which habitually co-occur and which are nonetheless fully transparent. Usually, there is a semantic or structural relationship between these lexical items (adjective + noun, verb + noun *etc.*). Examples of collocations in this sense are *fine weather*, *torrential rain*, *etc.* A typical example of collocation is the noun phrase *heavy drinker*, where one of its constituent elements (heavy) is highly restricted contextually, and different from its meaning in more neutral contexts (Cruse, 1986; Church and Hanks, 1991).
3. the combination between e. g. a (basic-level) noun ("base") and an unpredictable verb ("collocator"), such as German *Zähne putzen* (literally *to clean teeth*, "to brush the teeth") or *Tisch decken* (literally *to cover table*, "to lay/set the table"). This notion of collocation

focusses on speech production and the needs of language learners to express themselves appropriately (idiomatically). The relation between the elements of a collocation in this sense is asymmetric: the noun is usually the known item, whereas the verb is the unknown variable which has to be learned (e. g., you have to know that German uses *decken* for "to lay/set the table", not *legen* or any other verb) (*cf.* Hausmann (2004)).

It is the task of the linguist to interpret collocation phenomena – a task that cannot remain independent of theoretical framework and assumptions (section 4).

Idioms are defined in linguistic theory as complex expressions whose meaning cannot be derived from the meanings of their parts (Weinreich, 1969; Fraser, 1970). For example, the meaning "die" of the phrase *kick the bucket* cannot be inferred on the basis of the meanings of the parts *kick* and *the bucket*. Whereas the traditional view on idioms saw them as non-compositional phrases with stipulated meanings, more recent research has challenged the notion of non-compositionality of idioms and has stressed their analysability, their figurative character, their flexibility in discourse and the contribution of the idioms' parts to the overall meaning (*cf.* Gibbs and O'Brien (1990); Cacciari and Glucksberg (1991); Nunberg, Sag and Wasow (1994)).

2 Related work

Fazly and Stevenson (2007) follow a much broader approach than the one we present here. They do not only cover syntagmatic fixedness but also focus on measures of lexical fixedness (substitutability by semantically very similar words), institutionalisation (by statistical significance of MWE constituents appearing as MWE) and the compositionality of MWEs. The overall accuracy of their approach by combining all proposed measures reaches 58.3 percent. Their strongest measure appears to be the combined lexical and syntactic fixedness measure that reaches 50.0 percent accuracy. All their tests were run on verb-noun MWEs. Obviously, much effort was put into the modelling of known characteristics of MWEs. In this respect our approach represents a completely different strategy: Starting from raw corpus data that has been prefiltered only by statistical measures of collocational significance we tried to keep the rating as theory neutral as possible (apart from tokenisation which is inherently theory dependent). The clustering results of our approach do not rely on any explicit syntactic or semantic information about the collocation partners under research.

There have been other attempts to extract MWEs from corpora often focussing on idioms. Widdows and Dorow (2005) for example propose a technique that exploits semantic graphs. Noun constituents of coordinated phrases that do not occur in reverse order within the corpus are assigned unidirectional links in the graph, while all others are assigned bidirectional links. They conclude that asymmetric links may exist due to the idiomatic nature of the MWE that contains the linked nodes but they also found that "[m]any other phrases were extracted which exhibit a typical directionality that follows from underlying semantic principles" (Widdows and Dorow, 2005, 55).

3 Measuring fixedness

In this section we describe a computationally inexpensive and intuitive approach to the measurement of syntagmatic fixedness of a set of collocations based on positional co-occurrence data. Unlike simple statistical significance tests this measurement also takes every other collocation under consideration into account at the same time. It helps to uncover syntagmatic similarity between collocations.

The measure we propose is based on positional co-occurrence data extracted from a snapshot of the DWDS corpus of German language of the 20th century (Geyken, 2007). Positional co-occurrence is based on the number of intervening tokens (the *span*) between two co-occurring items (Evert, 2005, 18f). This is the most basic approach to co-occurrence and is easily applicable to almost any linguistic corpus. Relational in contrast to positional co-occurrence depends on further linguistic annotation – and interpretation – besides tokenisation. Often relational co-occurrence tests are run on corpora annotated with phrase structure or on chunk parsed corpora. Only few of today's publicly available corpora provide this kind of structural annotation. Restricting our approach to purely positional co-occurrence data thus makes it applicable to any available linguistic corpus.

The approach consists of four steps:

1. extraction of positional co-occurrence data from the underlying corpus (histogram data),
2. translation of histograms into character sequences,
3. pairwise comparison of the resulting sequences and
4. visualisation of syntagmatic similarity.

The procedures used in each step are independently exchangeable and others may be introduced as well. Figure 1 gives a brief overview of the procedures discussed in the following sections. Within each layer, only one procedure may be activated at a time except for the sequence comparison step where distance measures may be combined.

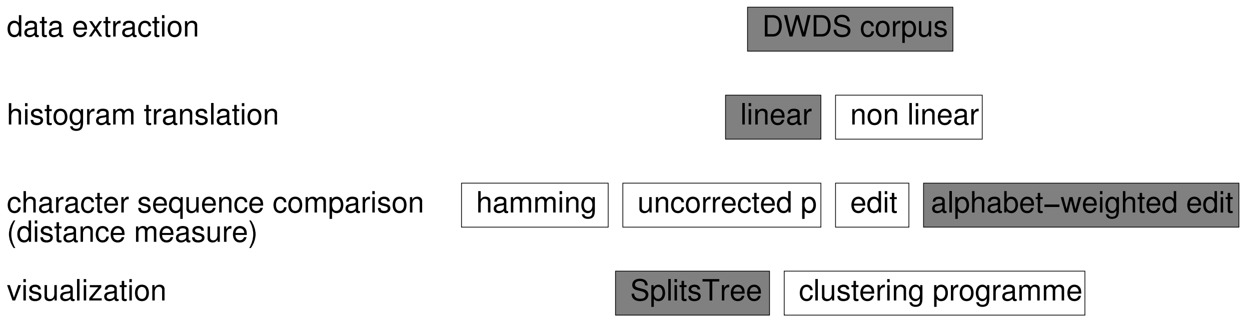


Figure 1: Different steps of the approach. One procedure is chosen for each layer. One possible configuration is given by the gray boxes.

3.1 Data extraction

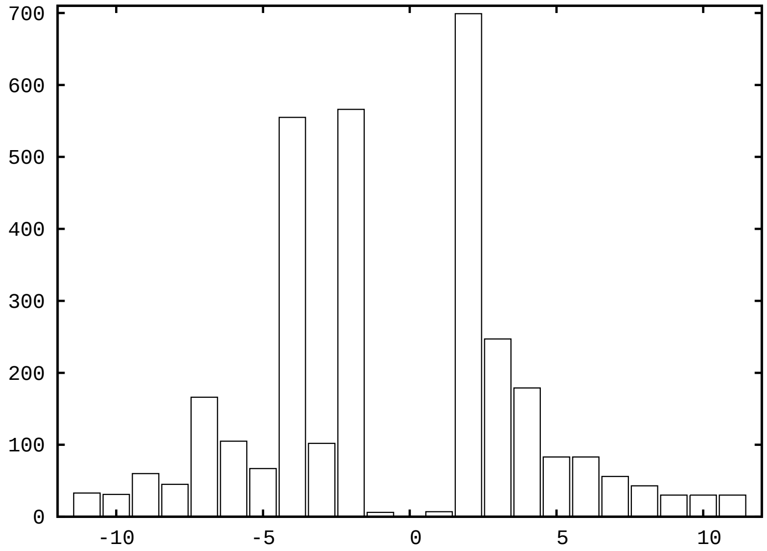
We tested our approach for noun–noun collocations so far. First, we compiled a list of target words L_{target} based on the well known Swadesh-200 list of basic nouns. Most of these nouns appear frequently within our corpus. Choosing Swadesh-200 as L_{target} remains an arbitrary choice, others were possible as well, e. g. the most frequent nouns of our corpus (corpus-driven target selection). By restricting ourselves to an *a priori* fixed set of target nouns not implicitly dependent on the underlying corpus we can easily switch to other corpora. By translating L_{target} using semantically corresponding words of its members one can also apply our approach to cross-linguistic research.

For each member l of L_{target} two sets of the most frequent collocates L^l_{MI} and L^l_{LL} were computed according to the mutual information (MI) and log likelihood (LL) association measures. A regular expression was used to restrict the collocates to tokens beginning with a capital letter as German nouns are capitalised and part-of-speech information was not available at that time. Both L^l_{MI} and L^l_{LL} were limited to at most 200 types each. We then merged the resulting lists to form a single set of collocates. Both L^l_{MI} and L^l_{LL} where largely identical except for their ranking order. The resulting List $L^l_{coll} = L^l_{MI} \cap L^l_{LL}$ was then manually cleansed of false positives (mostly adjectives that were capitalised due to their sentence initial position).

Data extraction was run automatically on the DWDS corpus using a pipeline of Python scripts. For each member w_1 of L_{target} and for each member w_2 of the associated list L^l_{coll} the corpus was queried for co-occurrences of w_1 and w_2 within a window of ±10 tokens (distance based positional co-occurrences (Evert, 2005, 68ff.)). This was accomplished by queries like (1). Each query is strictly sentence based so (1) would read “each occurrence of the token *Wasser* (water) followed by at most *n* tokens and followed by the token *Luft* (air) while suppressing those occurrences with at most $n-1$ intervening tokens”.[1]

```
(1)    "@Wasser #(n) @Luft"&& !"@Wasser #(n-1) @Luft"
```

The result of the corpus queries is the data given in figure 2. It is common to display this data in form of a histogram.



token distance	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
occurrences	33	31	60	45	166	105	67	555	102	566	6

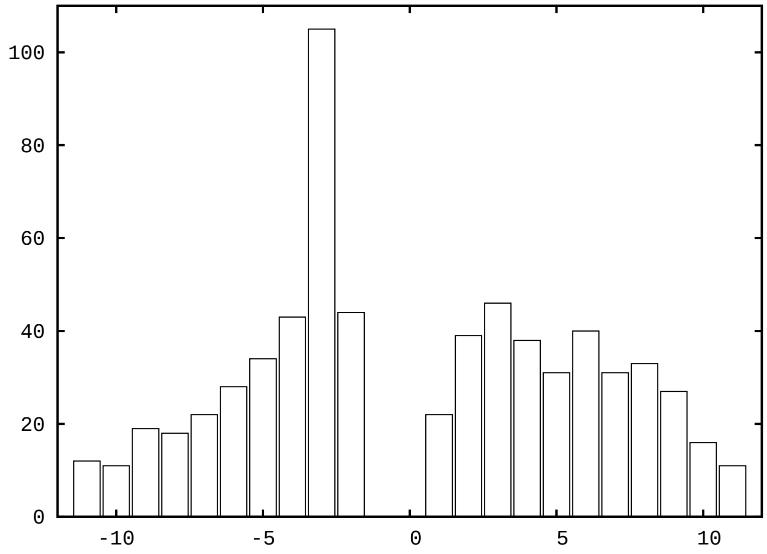
token distance	1	2	3	4	5	6	7	8	9	10	11
occurrences	7	699	247	179	83	83	56	435	30	30	30

Figure 2: Sentence based co-occurrence data of the tokens $w_1 = Wasser$ (water) and $w_2 = Luft$ (air) and the corresponding histogram plot.

Another data set and corresponding histogram is given in figure 3. The collocates here are *Wasser* (water) and *Brunnen* (well/fountain). Note the striking difference in the shape of the histograms. In figure 2 we find more evidence for fixedness which manifests itself in the strong peaks at token

[1] There is no possibility to specify arbitrary token sequences of an exact length. For a description of the query language used to access the DWDS corpus *cf.* (Sokirko, 2003).

distances -4, -2 and $+2$, whereas in figure 3 no notable distance preference can be found besides the one for token distance -3.



token distance	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
occurrences	12	11	19	18	22	28	34	43	105	44	0

token distance	1	2	3	4	5	6	7	8	9	10	11
occurrences	22	39	46	38	31	40	31	33	27	16	11

Figure 3: Sentence based co-occurrence data of the tokens $w_1 = Wasser$ (water) and $w_2 = Brunnen$ (well/ fountain) and the corresponding histogram plot.

3.2 Histogram translation

The actual number of occurrences of w_1 and w_2 is not of interest. Only the distribution of occurrences within the window is considered. First, all distributions are normalised against the total co-occurrence of w_1 and w_2 in the given window. This allows for direct comparison of peaks within the histograms. In figure 2 the span -4 accounts for 15 percent of all co-occurrences, the spans -2 and $+2$ for 16 percent and 19 percent respectively. Span -3 in figure 3 accounts for 16 percent of the co-occurrences.

Second, a discrete function $f_n : \mathbb{R} \longrightarrow C$ where C is a fixed set of n characters is applied to the normalised histogram data to split the possible range into n sub-ranges of equal size. Each possible collocational span in the window is thus assigned a character according to its normalised co-occurrence count. We chose a linear function f_n that creates intervals of equal width for each character in C. (Using a non-linear function would permit to amplify the peaks in the histogram.)

Given the small set $C_1 = \{a,b,c,d,e,f,g,h,i,j,k,l\}$ of twelve characters the resulting character sequences s_1 and s_2 describing the histograms from figures 2 and 3 are:

(2) $s_1 =$ aaaaaaababaaacaaaaaaaaa
$s_2 =$ aaaaaaaabaaaaaaaaaaaaaaa

Note the bs and the single c in the sequences that represent the peaks in the histograms.

The character set's size determines the discriminatory strength of the algorithm. The more characters provided, the more different classes of peaks can be recognised. Put differently, the more characters provided, the more different histograms can be described. But the choice of the size of the character set also depends – at least partly – on the algorithm that is applied for string comparison. This is discussed in the following section.

3.3 Character sequence comparison

The core element of our approach is the computation of relatedness of the character sequences. We have applied four different string distance measures to our data. All of them are rather basic and well known.

Hamming distance: This measure returns the number of unequal characters when comparing pairwise all characters of two sequences having the same index. Consider the sequences given in (2) and repeated in (3) with character mismatches typeset boldface. The Hamming distance $d_{hamm}(s_1, s_2)$ in this example is four.

(3) $s_1 =$ aaaaaaa**bab**aaa**c**aaaaaaaaa
$s_2 =$ aaaaaaa**aba**aaa**a**aaaaaaaaa

Obviously, a peak in s_1 that is non-existent in s_2 at the same index or that is represented by a different character will result in a mismatch and add to the overall distance between s_1 and s_2. Using the Hamming distance as a measure for the relatedness of two sequences is thus not sensitive to sequences that differ only slightly.

Uncorrected p distance: This measure returns the proportion of character mismatches in relation to the length of the character sequences. It is thus the Hamming distance normalised by the sequence length. For our data the uncorrected *p* distance behaves similar to the Hamming distance because all the sequences are of equal length.

Minimum edit distance: The minimum edit distance measure is based on the concept of minimum edit scripts. Such scripts describe the transformation of one sequence into the other using only the basic operations *match*, *delete*, *insert* and *substitute*. *Match* operations do not increase the distance value and are usually left out of the script. All other operations increase the distance value by one. Usually there is more than one script of minimum length. For (3) one minimum length transformation is *deleting* the first mismatched character (b, this leads to matching of the following two formerly mismatched characters) then *substituting* the mismatched c and finally *inserting* an a at the end of the sequence.

In the rest of the paper we will simply use the term *edit distance* as we are only interested in the smallest possible edit distance.

When comparing histograms the edit distance measure accounts for peaks of equal height that are shifted locally. That is, histograms that differ only by the spans of their respective peaks are rated more similar than they would be by the Hamming distance measure.

For a detailed description of the mathematical basis of the measure as well as the computational implications and algorithms *cf.* Gusfield (1997).

Alphabet-weighted edit distance: This is a variant of the basic edit distance measure. Instead of simply checking for exact matches it evaluates the similarity of mismatched characters. Character similarity is modelled by a function $f_{sim}(c_1, c_2) : C \times C \longrightarrow [0,1]$ that returns the linear distance between c_1 and c_2. Roughly speaking that is the difference between the heights of two peaks being compared. For maximally dissimilar characters – one minimum and one maximum peak – the distance value increases by one and for identical characters it remains unchanged.

We used an adoption of the Levenshtein algorithm for the computation of the alphabet-weighted edit distance. For further discussion and mathematical background *cf.* Gusfield (1997).

All distance measures described are symmetrical. The result in any case is a matrix M_{dist} that for every collocational pair contains its pairwise distance.

As stated above, the size of the character set depends in part on the applied string comparison algorithm. For large character sets the Hamming and the classical edit distance similarity measures perform worse. There are more different intervals and slight deviations of peaks representing the same span within different histograms result in different characters being assigned to those spans. What follows is an increase in mismatches. The alphabet weighted edit distance algorithm is more robust against such deviations because small differences between characters account for small distance values. Here, mostly the increased discriminatory strength comes into effect. It cannot, though, distinguish between histograms that show slight differences between several peaks and histograms that strongly differ in one peak only. To account for this drawback we combined the alphabet weighted edit distance measure with the uncorrected p measure. The combined distance is therefore greater for histograms that differ in many peaks than the distance of largely identical histograms.

3.4 Visualisation

We used SplitsTree 4.8 for data visualisation (Huson and Bryant, 2006). This tool was originally intended for the computation of phylogenetic trees and networks based on genetic sequences and different models of evolutionary variation and development. Though these models are employed by linguists as well to examine relationships between languages, we simply use it to compute trees representing similarity relations between histograms. There is no notion of evolution in our approach.

So far we applied the computationally inexpensive tree building UPGMA algorithm (unweighted pair group method with arithmetic mean) to our distance data. UPGMA enables us to cluster our data hierarchically. In brief, the algorithm works on M_{dist} as follows:

1. find the pair $(w_1, w_2) \in L^l_{coll}$ with minimal pairwise distance in M_{dist},
2. connect w_1 and w_2,
3. delete w_1 and w_2 in M_{dist},
4. insert w_{new} into M_{dist} where the distance to all other collocational partners in L^l_{coll} is the mean of their respective distances to w_1 and w_2,
5. go back to step 1 if $|M_{dist}| > 1$.

Keeping the length of the twigs and branches of the resulting tree in sync with the distances from M_{dist} permits us to interpret the overall length of the shortest path between two nodes (collocators) as the relative degree of similarity of the histograms they represent.

4 Discussion

The discussion will focus on two examples – the noun collocators for *Auge* (eye) and *Feuer* (fire). For reference and orientation we also included an empty histogram (EMPTY) without any co-occurrence data for analysis. As there are only significantly co-occurring nouns in L^l_{coll} we expected EMPTY to cluster with those collocators that are syntactically most fixed and therefore exhibit only one distinct peak.

4.1 Auge (eye)

In figure 4 (*Auge*) one can easily spot a narrow cluster in the lower right corner consisting of the collocators *Lesers* (reader's), *Gesetzes* (law's), *Kindes* (child's), and others. Note that all collocators in this cluster except for *Zahn* (tooth) and *Eros* are singular genitive forms. Another two notably narrow clusters are situated in the lower left corner of figure 4. They contain *Tod* (death) and other collocators and *Fest* (feast) and others.

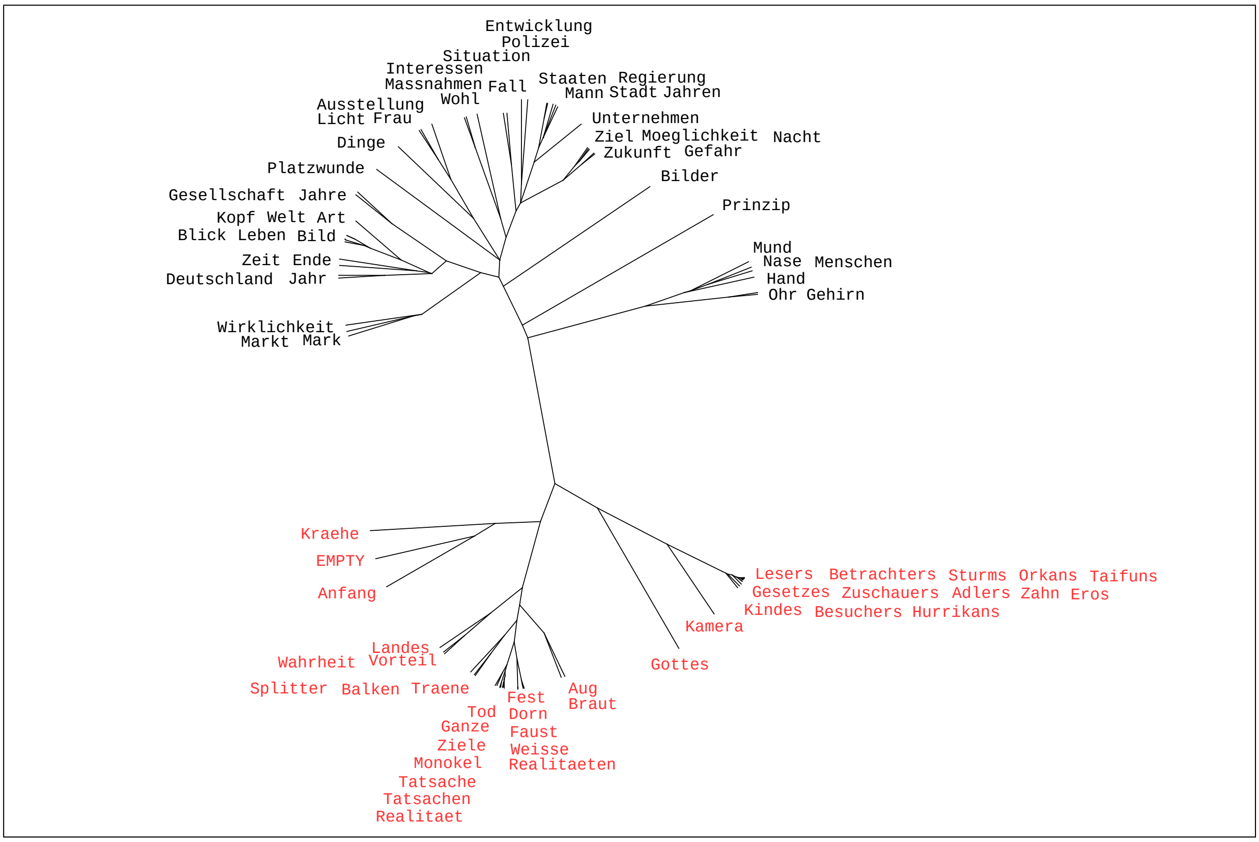


Figure 4: Noun collocators for *Auge* within the tree derived from the distance matrix by UPGMA hierarchical clustering. The nodes discussed in detail are marked red.

All collocators of *Auge* marked red in figure 4 are located in a distinct subtree. For each two terminal nodes within that subtree their pairwise distance is shorter than the distance to any terminal node outside the subtree. As the EMPTY element lies also within that subtree it becomes clear that the subtree contains the elements that show a high degree of syntagmatic fixedness. The other collocators do not form any cluster as the distances between adjacent nodes is greater than the distance between their parent nodes. Interestingly still, `EMPTY` is not located within any of the three narrow clusters. Instead, it behaves like an outlier together with its nearest neighbors *Gottes* (God's), *Krähe* (crow) and *Braut* (bride) without their forming a cluster of their own.

The marked subtree is divided into two further subtrees. These differ with respect to the order of the tokens: In the subtree on the right *Auge* precedes the nouns, whereas in the left subtree *Auge* occurs after the nouns. What is the relation between *Auge* and these nouns? Among the nouns we find some that occur in different MWEs with *Auge*:

Idiomatic expressions: *Dorn* (thorn) and *Faust* (fist) are part of idiomatic expressions, namely *etw. ist (jmdm.) ein Dorn im Auge* (literally, *to be (to sb.) a thorn in the eye*, "to be a thorn in sb.'s flesh/side") and *etw. passt wie die Faust aufs Auge* (literally, *something fits like a fist on the eye*, "sth. does not fit at all"). The first expression is a figurative expression with predicative structure, the second expression consists of a verb and a simile.

Sayings: *Zahn* (tooth) occurs in the expression *Auge um Auge, Zahn um Zahn* (literally, *eye for eye, tooth for tooth*, "an eye for an eye, a tooth for a tooth").

Another saying with Biblical origin in which the constituent parts are represented in the tree – namely *Splitter* (splinter) and *Balken* (beam) – is *den Splitter im fremden Auge, aber den Balken im eigenen nicht sehen* ("to see the speck in the others' eye, but don't see the beam in one's own"). This saying is contextually rather variable, as is illustrated in (4):

(4) (a) ... sehe den Splitter im Auge seines Bruders, den Balken im eigenen nicht.
("sees the speck in his brother's eye, but not the beam in his own.")
(b) Die Grünen forderte Corts auf, "die Splitter im eigenen Auge zu sehen".
("Corts urged the Green Party to 'see the speck in their own eye'.")
(c) Sie widmen alle Aufmerksamkeit dem Splitter im eigenen Auge (Fall Kohl), übersehen oder verzeihen den Balken im Auge des anderen (Fall Berlusconi).
("They pay all attention to the speck in their own eye (the Kohl affair), but they overlook or forgive the beam in the others' eye (the Berlusconi affair).")

Nevertheless, the syntagmatic fixedness between the nouns *Splitter* (splinter) and *Auge* and *Balken* (beam) and *Auge* is not influenced by its syntactic and lexical flexibility. Finally, *Splitter im Auge* (literally, *splinter in the eye*) is also frequently encountered with literal meaning.

Proverb: The noun *Krähe* (crow) is part of the proverb *eine Krähe hackt der anderen kein Auge aus* (literally, *a crow does not peck out the eye of another*, "there's honour among thieves").

Metaphor: This involves the fixed expression *(das) Auge des Gesetzes* ("(the) eye of the law"). Additionally, *Fest* (feast) is related to *Auge* in the expression *ein Fest fürs/für das Auge* ("feast for the eye"). Finally, *(das) Weiße* occurs in the expression *das Weiße im Auge sehen* (litarally, *to see the white of the eye*) as in (5):

(5) Die Grenze umklammerte die Dörfer von drei Seiten, an manchen Stellen so eng, dass buchstäblich das Weiße im Auge der DDR-Grenzwächter zu sehen war.
("The boarder surrounded the villages from three sides, at some places so tightly, that the white of the GDR frontier guard's eye was literally visible.")

In addition to these expressions, further clusters of co-occurring items are visible in the subtrees. One cluster comprises the nouns that refer to persons, namely *Auge des Lesers/Betrachters/Zuschauers/Besuchers* ("the reader's/observer's/spectator's/visitor's eye"), which is an instance of metonymy. A similar example is the expression *Auge Gottes* ("God's eye").

A further cluster involves *Auge des Taifuns/Orkans/Sturms/Hurrikans* ("eye of a typhoon/hurricane/gale/hurricane"). These examples of syntagmatic fixedness illustrate one important aspect of polysemy: One of the senses of *Auge* – notably a specialised sense – is associated with a particular syntagmatic pattern. Syntagmatic fixedness in this case signals a metaphorical sense of the head noun (*cf. Auge des Gesetzes*). Polysemy signaled by an adnominal genitive is also the motivation for the noun *Kamera* ("camera") which occurs in *Auge der Kamera* ("camera's eye"). But contrary to the previous examples, the co-occurrences of *Kamera* and *Auge* in the corpus are not restricted to this particular expression. The syntagmatic relatedness between *Kamera* and *Auge* is variable, as their relation in real world is one of contiguity as in (6):

(6) (a) … wenn die Kamera vor das Auge gehalten wird …
("… when the camera is being held before the eye … ")
(b) … mit scharfem Auge in die Kamera blickt.
("… look into the camera with a sharp eye.")
(c) Stattdessen gleitet die Kamera vom Auge herunter zur Hand …
("Instead, the camera slips from the eye down to the hand … ")
(d) "Ich hab' keine Ahnung", antwortete er, ohne das Auge von der Kamera zu nehmen.
("'I have no idea,' he answered without taking his eye from the camera.")

The rest of the nouns that appear on the subtree are not parts of MWEs with *Auge*. The question that arises is what kind of relations do these nouns have to *Auge* and whether they can be classified in larger groups.

One cluster of nouns that is visible in the left tree consists of the abstract nouns *Realität/Realitäten/Wahrheit/Tatsache/Tatsachen* (reality/realities/truth/fact/facts). These nouns in their respective forms are typical objects of the expression *etw. (dative) ins Auge sehen/blicken* (literally, *to see/look sth. in the eye*, "to confront/face"). The open slot typically contains nouns that express unpleasant things that one does not want to face. This use is illustrated in (7):

(7) (a) Er hoffe, daß diese Politiker den Realitäten ins Auge sähen und sich nach vorn bewegten.
("He hoped that these politicians would look the realities in the eye (face reality) and move foward.")
(b) Unserer Ansicht nach besteht die richtige Politik darin, der Realität ins Auge zu sehen …
("In our view, the right policy is to see reality in the eye (to face reality) … ")
(c) Letztendlich mußte die kapitalistische Welt der Wahrheit ins Auge sehen.
("Ultimately, the capitalist world had to see the truth in the eye (had to face the truth).")
(d) Wir sind es gewohnt, … der Wahrheit ehrlich ins Auge zu blicken, mag sie auch unangenehm sein.
("We are used to look the truth honestly in the eye (to face the truth), even if it is unpleasant.")
(e) Wir müssen den Tatsachen ins Auge schauen.
("We have to look the facts in the eye (to face up the facts).")
(f) Man müsse den Mißerfolg des Völkerbundes zugeben und den Tatsachen ins Auge blicken.
("One should admit the failure of the League of Nations and look the facts in the eye (face up the facts).")

Furthermore, the noun *Tod* (death) can also be grouped in this class, as is shown in (8):

(8) Der Staatschef, … der, inzwischen 92 Jahre alt, an seinem Geburtsort Monastir dem Tod ins Auge sieht.
("The head of state … who, meanwhile 92 years of age, looks the death in the eye (faces death) at his birthplace Monastir.")

In these examples the relation is between the nouns *Realität/Realitäten/Wahrheit/Tatsache/Tatsachen/Tod* (reality/realities/truth/fact/facts/death) and the whole expression *etw. (dative) ins Auge sehen/blicken* (to see/look sth. in the eye). This relation is one of strong selectional preferences between the verb (which is itself a complex VP) and its object.

The same type of syntagmatic fixedness can be found between other metonymic expressions such as *etw. im Auge haben* ("to mean/to be interested in"), *etw. im Auge behalten* ("to keep an eye on sth."), *etw. aus dem Auge verlieren* ("to lose sight of sth.") *etc.* and the nouns that can occur in the object slot. This motivates the occurrence of the nouns *Ziele* (goals) and *Vorteil* (advantage) on the subtree as in (9):

(9) (a) Die NATO ... müsse ... der Allianz überall ... helfen, wo es notwendig sei, und niemals die gemeinsamen Ziele aus dem Auge verlieren.
("The NATO should help the alliance wherever necessary and never lose sight of the common goals.")
(b) Man lasse sich also durch unerbetene Ratschläge Unberufener nicht beeinflussen, zumal diese nur den eigenen Vorteil im Auge haben.
("One should not be influenced by uncalled-for advice of non-experts, since they are interested only in their own advantage.")

The expression *etw. im Auge behalten* ("to keep in view") also has strong selectional preferences for the nominalised adjective *(das) Ganze* ((the) whole):

(10) Wir wollen ... ein zukunftsfähiges Verkehrskonzept, das nicht mit Stückwerk zu erreichen ist, sondern das Ganze im Auge behält ...
("We want a sustainable concept of transportation, which cannot be reached with piecemeal, but by keeping the whole in the eye (in view)".)

The occurrence of the noun *Landes* (country), which seems intuitively not justified with regard to *Auge*, can also be explained with regard to these expressions, as is shown in the corpus data. But the relation between an expression such as *etw. im Auge haben* (to mean/to be interested in) and *Landes* (in the genitive case) is not one of selectional preferences of the object. Rather, *Landes* occurs as part of the object noun phrase *Interessen des Landes* ("country's interests"), as is shown in(11):

(11) ... die nicht die wahren Interessen des Landes im Auge hätten.
("... who do not have the real interests of the country in they eye (do not keep ... in view)")

In this case, the head *Interessen* (interests) is semantically related to other frequent nouns like *Ziele* (goals) and *Vorteil* (advantage), as illustrated in (9) above. Similarly, the occurrence of *Anfang* (start) seems implausible. This noun is used as an adverbial, typically associated with the expressions mentioned above containing *Auge* as in (12):

(12) ... Wir haben dabei aber von Anfang an im Auge gehabt, dass eines Tages eine Kategorie von politisch tätigen Menschen zu uns flüchten könnten, ...
("... We thereby had in the eye from the start that one day a group of politically active people would come as refugees ...")

This is further evidence that although adverbials may be optional elements in a sentence from a syntactic point of view, certain verbs or verb phrases may show preferences for occurrence with particular adverbials.

Monokel (monocle) and *Träne* (tear) are associated with *Auge* in terms of contiguity. Syntactically, this relation is realised in different ways, as is shown in (13) and (14):

(13) (a) Mit Monokel im Auge und rotem Schal um den Hals ...
(“With a monocle in the eye and a red scarve around the neck ...”)
(b) ... der sich ein Monokel vors Auge klemmt ...
(“... who puts a monocle before the eye ...”)

(14) (a) ... verabschiedet sich mit einer Träne im Auge ...
(“(he) says good-bye with a tear in the eye ...”)
(b) Mit einer Träne in jedem Auge ...
(“With a tear in each eye...”)
(c) Verstohlen wischt er eine Träne aus dem Auge.
(“He covertly wiped a tear from the eye.”)

The form *Aug* is a shortened form of *Auge* which is attested in a variant of the above mentioned saying, i. e. *Aug' um Auge, Zahn um Zahn* (“an eye for an eye, a tooth for a tooth”) and of the expression *Auge in Auge* (“eye to eye”).

There are still some nouns which cannot be motivated intuitively. A check of the corpus data provides the answer that these nouns occur in titles that are very frequent in the corpus (these include the film title “Das Auge des Adlers” (“The eye of the eagle”), the titles of two exhibitions “Mit dem Auge des Kindes” (“With the eye of the child”) and “Elan Vital oder Das Auge des Eros” (“Eros' eye”), and the name of the pop band “Die Braut haut ins Auge” (“The bride beats in the eye”). Thus, it turns out that the reason for syntagmatic fixedness is that the nouns occur in these named entities.

The substree at the bottom which contains these nouns is topologically clearly distinguished from the rest of the tree. The upper subtree consists of nouns that are not related to *Auge* in terms of syntagmatic fixedness and do not occur with *Auge* in MWEs. Nevertheless, some elements are worth pointing out. Firstly, some nouns are related to *Auge* semantically, for example they belong to the same semantic field (e. g. the nouns on the right part of the upper tree that refer to other body parts). Secondly, other types of semantic relations may be distinguished (e. g. the noun *Platzwunde* (abrasion), which typically refers to an injury on the head or above the eye). Moreover, several nouns fit semantically into the class of arguments of the expressions mentioned above *etw. im Auge haben*, *etw. im Auge behalten*, *etw. aus dem Auge verlieren*, *etc.*, such as *Wirklichkeit* (truth/ reality), *Möglichkeit* (possibility), *Wohl* (welfare) *etc.* Note that the noun *Interessen* – which was commented on with regard to *Landes* in (11) above – appears on this part of the tree. Finally, a noun such as *Regierung* is semantically related to these expressions as a typical subject (*cf.* the case of *Sicherheitskräfte* (security forces) with regard to the noun *Feuer* below).

4.2 Feuer (fire)

The second noun that we discuss is *Feuer* (fire); the clustering results are given in figure 5. The subtree on the right contains nouns that in linear ordering precede *Feuer*; the subtree on the left shows nouns that follow *Feuer*.

A large number of nouns can be associated with classes of MWEs:

VP-idioms: The following VP-idioms containing the noun *Feuer* are represented on the subtree:

1. Kastanien (chestnuts) as in *die Kastanien aus dem Feuer holen* (“to pull chestnuts out of the fire”)

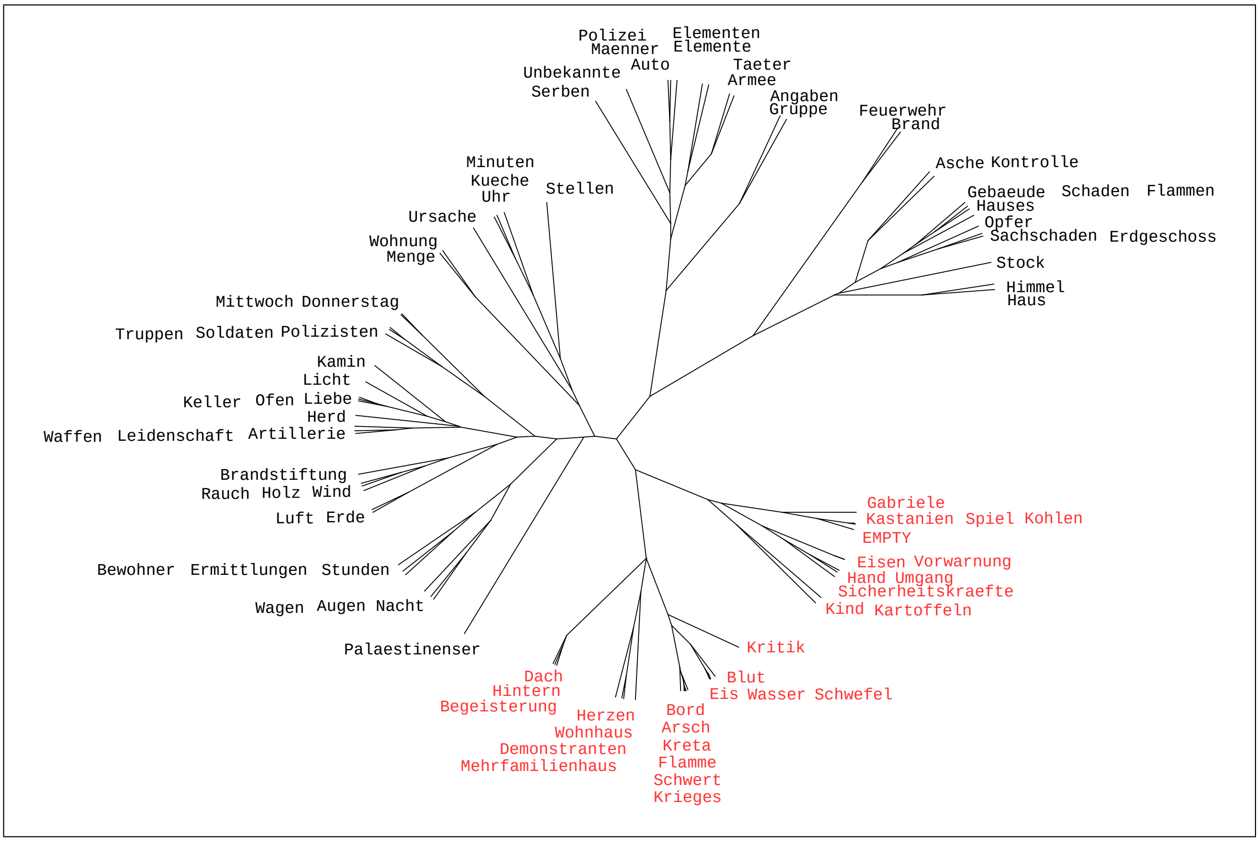


Figure 5: Noun collocators for *Feuer* (fire) within the tree derived from the distance matrix by UPGMA hierarchical clustering. The nodes discussed in detail are marked red.

2. Kohlen (coals) as in *die Kohlen aus dem Feuer holen* (literally, *to pull the coals out of the fire*, “to pull chestnuts out of the fire”)
3. Kartoffeln (potatoes) as in *die Kartoffeln aus dem Feuer holen* (literally, *to pull the potatoes out of the fire*, “to pull chestnuts out of the fire”)
4. Hand (hand) as in *(für jmdn./etw.) seine Hand ins Feuer legen* (literally, *to lay one’s hand into the fire for sb./sth.*, “to vouch for sb./to stand behind sb.”)
5. Eisen (iron) as in *(zwei/viele) Eisen im Feuer (haben)* (“to have two/many irons in the fire”)
6. Hintern (bottom) as in *jmdm. Feuer unter dem/unterm Hintern (machen)* (literally, *to make sb. fire under the bottom*, “to light a fire under sb.”)
7. Arsch (ass) as in *jmdm. Feuer unter dem/unterm Arsch (machen)* (literally, *to make sb. fire behind the the ass*, “to light a fire under sb.”)
8. Dach (roof) as in *Feuer unter dem/unterm Dach (sein)/(machen)* (literally, *to have/make fire under the roof*, “have/cause a fundamental problem”)

This list of VP-idioms contains some synonymous idioms, such as 1–3 and 6/7. Two items are at issue here. First, *Dach* (roof) is not used exclusively in the idiom but is also used literally in the data, though less frequently. Secondly, *Feuer unterm Arsch* (“fire under the ass”) is also a film title and thus represents a named entity which is very fixed.

NP-idioms: Two nouns are associated with *Feuer* in NP-idioms:

1. Spiel (game) as in *Spiel mit dem Feuer* (literally, *game with the fire*, “playing with fire”)
2. Flamme (flame) as in *Feuer und Flamme (sein)* (literally, *(to be) fire and flame*, “to be all for it”)

Binomial expressions: The nouns in binomial expressions usually occur in fixed order.

1. Schwert (sword) as in *(mit) Feuer und Schwert* (“(with) fire and sword”)
2. Schwefel (brimstone) as in *Feuer und Schwefel* (“fire and brimstone”); also used as *Schwefel und Feuer*; the expression has Biblical origins
3. Wasser (water) as in *Feuer und Wasser* (“fire and water”)
4. Eis (ice) as in *Feuer und Eis* (“fire and ice”)

Feuer und Wasser and *Feuer und Eis* are binomial expressions that are used in order to refer to two opposite things.

Metaphor: Metaphorical usage is signalled syntagmatically by the use of the genitive.

1. Kritik (criticism) as in *Feuer der Kritik* (“fire of criticism”)
2. Begeisterung (enthusiasm) as in *Feuer der Begeisterung* (“fire of enthusiasm”)
3. Krieges (war’s) as in *(im) Feuer des Krieges* (“in the fire of the war”)

Saying: *Kind* (child) occurs in the saying *(ein) gebranntes Kind scheut das Feuer* (“a burnt child dreads the fire”), which occurs also in other, modified forms.

The remaining nouns cannot be associated with MWEs and have to be explained differently. As is suggested by the data, the nouns *Sicherheitskräfte* (security forces), *Demonstranten* (demonstrators), and *Vorwarnung* (prior warning) are all strongly related to the support verb construction (SVC) *das Feuer eröffnen* (to open fire): *Sicherheitskräfte* (security forces) is a typical subject (*cf.* (15)), *Demonstranten* (demonstrators) a frequently attested noun in a prepositional phrase (*cf.* (16)), and *ohne Vorwarnung* (without prior warning) a typical adverbial (*cf.* (17)) used with this SVC:

(15) Sicherheitskräfte eröffneten das Feuer, töteten mehrere Menschen und verhafteten eine große Zahl von Anhängern der Opposition.
(“Security forces opened fire, killed many people and arrested many supporters of the opposition.”)

(16) In Lahore eröffnete die Polizei das Feuer auf die Demonstranten, wobei mindestens vier Menschen ums Leben kamen und 35 verletzt wurden.
(“In Lahore the police opened fire against the demonstrators, whereby at least four people lost their lives and 35 were injured.”)

(17) Polizisten eröffneten ohne Vorwarnung das Feuer.
(“The policemen opened fire without warning.”)

Thus, this SVC seems to be rather fixed with relation to the semantic class of its arguments. On the other hand the strong preference for these nouns and the adverbial might also reflect the way this SVC is used in newspaper texts.

A cluster of a different type is formed by the nouns *Wohnhaus* (apartment house), *Mehrfamilienhaus* (multifamily building) and *Bord* (bord). These nouns refer to the places where a fire most often breaks out. Reference to these events are typical of journalese texts:

(18) Im schleswig-holsteinischen Rendsburg starb ein Rentnerehepaar bei einem Feuer in einem Mehrfamilienhaus.
("In Rendsburg in Schleswig-Holstein a couple of pensioners died during a fire in a multifamily building.")

(19) Bei einem Feuer in einem Wohnhaus in Niederrad ist am Dienstag ein Schaden von rund 60 000 Mark entstanden.
("During a fire in an apartment hous in Niederrad on Tuesday a damage of about 60.000 marks resulted.")

(20) Der Frachter sank ..., nachdem aus ungeklärter Ursache Feuer an Bord ausgebrochen war.
("The cargo ship sank ... because a fire broke out on bord due to unknown reasons.")

The noun *Umgang* (handling) is frequently encountered in *Umgang mit Feuer* (handling of fire), a collocation that is frequently used in the police register and that occurs frequently in newspaper texts. Typical instances are given in (21):

(21) (a) ... die Behörde rief zur Vorsicht im Umgang mit Feuer auf
("... the authority appealed to caution with handling of fire")
(b) ... richtiger Umgang mit dem Feuer ...
("... right handling of fire ...")
(c) ... fahrlässiger Umgang mit Feuer ..., wie die Polizei mitteilte.
("... careless handling of fire ... as the police announced.")
(d) Das Unglück sei durch "unsachgemäßen Umgang mit offenem Feuer" ausgelöst worden.
("The accident was caused by 'improper handling of open fire'.")

Four collocates remain unexplained, namely *Gabriele*, *Kreta* (Crete), *Blut* (blood) and *Herzen* (heart (dative)). The first two are associated with the title of the book "Das Feuer von Kreta" ("The fire of Crete") written by Gabriele Beyerlein. The same is true of *Blut* (blood): Although the expression *in Feuer und Blut* ("in fire and blood") is used, the fixedness that is displayed here is due to the book title "Feuer und Blut" ("Fire and blood") which appeared very frequently in the newspapers in 1999. For *Herzen* (heart (dative)), one can find the co-occurrence of metaphorical *Feuer* (fire) and *Herzen* (heart (dative)) in various forms – *Feuer im Herzen* ("fire in the heart"), *das Feuer loderte im Herzen* ("the fire burnt inside the heart"), *etc.* –, but the fixedness that is shown in the graph again is due to a book title "Feuer in die Herzen" ("Fire into the hearts") by Jutta Ditfurth.

The remaining subtrees contain nouns that stand in a loser syntagmatic relation to *Feuer*. Some points are worth considering:

- The use of the form *Flamme* was mentioned above as part of a NP-idiom. The plural form *Flammen* appears on the upper subtree. Thus, the syntagmatic behaviour of this nouns differs in the singular and plural. This is evidence for the fact that words may not show the same behaviour in their different forms. It also confirms that syntagmatic fixedness is associated not only with particular words, but also with words in particular forms.
- Some nouns can be grouped together and correspond to classes established above according to semantic and syntactic criteria. For example, nouns that signal a metaphorical sense of *Feuer* include *Leidenschaft* (passion) and *Liebe* (love) (*cf. Kritik* and *Begeisterung*).
- Different nouns point to different domains of usage for fire:
  - Nouns such as *Armee* (army), *Artillerie* (artillery), *Waffen* (weapons) point to the domain of war.

- Nouns like *Brandstiftung* (arson), *Ermittlungen* (inquiries), *Unbekannte* (unknown), *Täter* (delinquent) *etc.* point to the scenario of arson. Apparently, this is also the explanation for the names of the days (*Mittwoch* (Wednesday) and *Donnerstag* (Thursday)) which correspond to the reports of events in newspaper texts.
- Nouns such as *Wind* (wind), *Erde* (earth), *Luft* (air) are related semantically to the sense of fire as an element of nature.

5 Conclusion

The case of *Auge* reveals that our approach can select syntagmatically fixed combinations on the given data. It serves well to visualise syntagmatic similarity between different collocations. An additionally insterted `EMPTY` element allows for locating those collocators that are positionally most fixed in regard to their target noun. However, consider the graph in figure 5 which is more difficult to interpret. Topological analysis alone does not allow for locating fixed as opposed to more freely related collocations. Here, the introduction of the `EMPTY` element is crucial for orientation within the tree.

Even without graphical visualisation the procedure described in this paper provides a means of measuring syntagmatic fixedness of collocating items. As opposed to the more sophisticated methods given in Fazly and Stevenson (2007) it is an almost knowledge-free approach and does not require any linguistic annotation of the underlying corpus besides tokenisation.

The interpretation of the semantic and syntactic relations that hold between the collocating items leads to the following conclusion: Syntagmatic fixedness is gradual, and is gradable, corresponds to a continuum of semantic and syntactic relations. At the one end of the continuum named entities exhibit the strongest degree of syntagmatic fixedness (*cf.* their occurrence in the neighbourhood of `EMPTY`). At the other end of the end we can locate selectional preferences (i. e. the semantic classes of nouns that occupy argument slots). We might also go further to nouns that belong to the same lexical field(s) as the target noun. MWEs such as idioms, proverbs, conventional metaphors *etc.* can be located between these extremes.

6 Acknowledgements

This work was carried out as part of the project “Kollokationen im Wörterbuch” at the Berlin Brandenburg Academy of Sciences and Humanities which was funded by the Wolfgang Paul prize awarded to Christiane Fellbaum by the Alexander von Humboldt Foundation and the “Zukunftsinvestitionsprogramm” of the German federal government.

We thank Christiane Fellbaum, Anke Lüdeling, Stefan Evert, Seanna Doolittle and the members of the project for discussion, helpful comments and suggestions.

References

Burger, Harald (1998): Phraseologie. Eine Einführung am Beispiel des Deutschen. Berlin: Erich Schmidt.

Cacciari, Christina and Sam Glucksberg (1991): Understanding idiomatic expressions: The contribution of word meanings. In Greg B. Simpson (ed.): Understanding word and sentence. North-Holland: Elsevier, Advances in psychology, 217–240.

Church, Ken and Patrick Hanks (1991): Word association norms, mutual information and lexicography. *Computational linguistics.* 16:1, 22–29.

Cowie, A. P. (1998): Phraseology: Theory, analysis and applications. Oxford: Oxford University Press, Oxford studies in lexicography and lexicology.

Cruse, D. A. (1986): Lexical semantics. Cambridge: Cambridge University Press.

Evert, Stefan (2005): The statistics of word cooccurrences. Word pairs and collocations. Ph. D thesis, Universität Stuttgart, Philosophisch-Historischen Fakultät, Institut für maschinelle Sprachverarbeitung, Stuttgart.

Fazly, Afsaneh and Suzanne Stevenson (2007): Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In Proceedings of the ACL workshop on a broader perspective on multiword expressions. Praha (CZ), 9–16.

Fellbaum, Christiane (ed.) (2007): Idioms and collocations. Corpus-based linguistic and lexicographic studies. London: Continuum, Research in corpus and discourse.

Fraser, Bruce (1970): Idioms within a transformational grammar. *Foundations of Language*, 6, 22–42.

Geyken, Alexander (2007): The DWDS corpus: a reference corpus for the German language of the twentieth century. In Fellbaum (2007), 23–40.

Gibbs, Raymond W. and Jennifer E. O'Brien (1990): Idioms and mental imagery. The motivation for idiomatic meaning. *Cognition*, 36, 35–68.

Gusfield, Dan (1997): Algorithms on strings, trees, and sequences. Computer science and computational biology. Cambridge: Cambridge University Press.

Hausmann, F. J. (2004): Was sind eigentlich Kollokationen? In Kathrin Steyer (ed.): Wortverbindungen – mehr oder weniger fest. IDS Jahrbuch 2003. Berlin/New York: de Gruyter, 309–334.

Huson, D. H. and D. Bryant (2006): Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, 23(2), 254–267.

Moon, Rosamund (1998): Fixed expressions and idioms in English. A corpus-based approach. Oxford: Clarendon.

Nunberg, Geoffrey, Ivan A. Sag and Thomas Wasow (1994): Idioms. *Language*, 70:3, 491–538.

Sinclair, John (1991): Corpus, concordance, collocation. Oxford: Oxford University Press.

Sokirko, Alexey (2003): DDC – a search engine for linguistically annotated corpora. In Proceedings of Dialog 2003. Protvino (RU).

Weinreich, Uriel (1969): Problems in the analysis of idioms. In Jaan Puhvel (ed.): Substance and structure of language. Lectures delivered before the Linguist Institute of the Linguistic Society of America, Univeristy of California, Los Angeles, June 17–August 12, 1966. Berkeley: University of California Press, 23–81.

Widdows, Dominic and Beate Dorow (2005): Automatic Extraction of Idioms using Graph Analysis and Asymmetric Lexicosyntactic Patterns. In Proceedings of the ACL-SIGLEX workshop on deep lexical acquisition. Ann Arbor (US), 48–56.