Statistical Variations of German Support Verb Constructions in very large Corpora

Alexander Geyken

Abstract

Unlike the vocabulary diversity in large document collections the growth of multi-word-expressions in large corpora has hardly been studied. The resources of this empirical study are a balanced 100 million and an opportunistic 1 billion token corpus of German. The target data are three verbs that act as light verbs in verb-noun expressions. A comparison between both corpora is followed by the investigation of the growth of linguistically relevant verb-noun-expressions with increasing corpus size. Also the corpus data are compared with a large monolingual dictionary. The results show that (almost) all verb-noun expressions in the dictionary are statistically salient in a 1 billion word corpus; however the same is not true for a 100 million word corpus. Conversely, a considerable number of statistically salient constructions are missing in the dictionary.

1. Introduction

Measurements of vocabulary diversity are well studied and play an important role not only in linguistics but also in information retrieval (e.g. Manning et al. 2008). The common measures used are based on the ratio of different words (Types) to the total number of words (Tokens), known as the Type-Token Ratio (TTR). With growing text this number decreases. The TTR growth rate can be described by Heap's law, an empirical law, which states that the vocabulary of a corpus of size n, V(n), can be formulated as $V(n) = k*n^{\beta}$ where k and β are determined empirically (Heap 1978). For English text corpora, typically K is between 10 and 100, and Beta is between 0.4 and 0.6. Regardless of the values of the parameters for a specific text corpus, Heap's law indicates that the number of words continues to increase with corpus size, rather than to converge to a maximum vocabulary size. Moreover, the number of types can reach large numbers. For example, the British National Corpus with 100 million tokens V(n) has appr. 660.000 types. For the DWDS-Kernkorpus (s. below), a balanced 100 million token corpus of the German texts of the 20th century, the number of types is 2.1 million. These greater numbers are due to the almost unlimited productiveness of noun compounding in German.

Even though there is some evidence that this law can be extended to multi-word phenomena (Köhler 2003), there are to our knowledge no empirical studies about the growth rate of particular syntactic phenomena in very large corpora.

Here we investigate the growth of a specific class of support verb constructions, the so called verb-nominalization constructions (NVG). We follow the terminology of Eisenberg (1999) who calls constructions like *in Erscheinung treten* ('make [one's] appearance') and *eine Entscheidung treffen* ('make a decision') Nominalisierungsverbgefüge or verb-nominalization constructions. A related class of constructions called Funktionsverbgefüge or function verb constructions (FVG) is usually considered a subclass of NVG (von Polenz 1987, Eisenberg 1999). For FVG, the meaning of the entire construction has specific grammatical properties that are absent in the corresponding simplex verb, e.g., a change of aspect or passivization (Storrer 2007). For the most part, FVG have the structure verb +

preposition + predicative noun; hence, the example given above *in Erscheinung treten* is an FVG, whereas *eine Entscheidung treffen* is an NVG but not an FVG. To distinguish FVG and NVG linguists have considered the syntactic restrictions (choice of determiner, attribution, pronominalization, etc.), the degree of grammaticalization or lexicalization, the relation between the NVG/FVG and the base verb or the classification of NVG/FVG according to formal and informal communication style.

In this paper we raise two questions: first, does the number of distinct NVG grow with a growing corpus due to the potentially infinite number of German compounds, and second how do NVGs in corpora compare to large monolingual print dictionaries?

In the following two sections we shortly describe the different resources as well as the target data that we took to conduct our study. In section 4 we present our method including an approximation to investigate the number of occurrences of NVGs in the different corpora for our target data. In section 5 we discuss the results with respect to our two initial questions.

2. Resources

The resources for our corpus study are two large corpora as well as a large monolingual dictionary.

Over the course of three years, between 2000 and 2003, the project *Digitales Wörterbuch der Deutschen Sprache* (Digital Dictionary of the German Language, DWDS) at the Berlin-Brandenburg Academy of Sciences created two different corpora of the 20th century German language: the balanced DWDS-Kerncorpus (henceforth DWDS-C) and the opportunistic DWDS-Ergänzungscorpus (henceforth DWDS-E) (Geyken 2007, Klein and Geyken, 2000).

The DWDS-C corpus consists of 100 million tokens, thus matching in size the British National Corpus. It is a balanced corpus of German texts of the 20th century, i.e. it is equally distributed over time and over the following five text types: journalism (approx. 27% of the corpus), literary texts (26%), scientific literature (approx. 22%) and other non-fiction (approx. 20%), as well as transcripts of spoken language (5%). This classification has mostly practical reasons. As all texts are encoded following the TEI-guidelines (cf. http://www.tei-c.org) it would be comparatively easy to introduce more fine-graded text types or to classify the texts according to a different classification code. Besides journalistic texts (newspaper reports and articles from periodicals taken from more than 50 different newspapers and magazines), there are literary monographs, poetry and dramatic works. Non-fiction texts such as cook books, maintenance manuals, and guides to etiquette are found in the corpus as well as important scientific works (e.g. by A. Einstein, R. Koch, J. Habermas). DWDS-C can be consulted on the DWDS project's website: http://www.dwds.de .

The much larger DWDS-E corpus contains around 1 billion words of running text, mainly from daily and weekly newspapers of the 1990s such as *konkret*, *Frankfurter Allgemeine Zeitung*, *Frankfurter Rundschau*, *Neue Zürcher Zeitung*, *Spiegel*, *Süddeutsche Zeitung*, *taz*, and *Die ZEIT*. More than 2 million newspaper articles have been gathered in this corpus. Due to copyright restrictions this corpus can only be used for internal purposes.

The "Wörterbuch der deutschen Gegenwartssprache" (WDG, en.: 'Dictionary of Present-day German) published between 1952 and 1977 was compiled at the Deutsche Akademie der Wissenschaften (since 1972: Akademie der Wissenschaften der DDR). It comprises six volumes with over 4,500 pages and contains more than 90,000 headwords (120,000, if compounds are counted separately). The term "deutsche Gegenwartssprache" ('German present-day language') is understood in a broad sense by the lexicographers; the dictionary is not restricted to the language spoken and written in the middle of the 20th century, but also incorporates sources form the 18th and 19th centuries as far as these are still widely read (cf. Malige-Klappenbach 1986, Wiegand 1990). The WDG is particularly known for its quality of the description of collocations, specifically of NVG. Like DWDS-C the WDG can be consulted in its electronic form on the DWDS project's website.

3. Target data

Our target data are three German verbs *leisten* ('render, perform'), *erteilen* ('give') and *hegen* ('preserve') that act as NVG as in the following examples:
1) *leisten* + noun (direct object)
examples: *Widerstand leisten* ('offer resistance'); *Arbeit leisten* ('to work')*; Schwur leisten* ('swear')
2) *erteilen* + noun (direct object)
examples: *Unterricht erteilen* ('give lessons'); *Weisung erteilen* ('give an order') ...
3) *hegen* + noun (direct object) [elevated style]
examples: *Hoffnung hegen* ('harbour hope'); *Zweifel hegen* ('have some doubt')
The linguistic properties of these constructions were studied in more detail in Hanks et al. 2006.

4. Method

In general NVG are characterized by a frequent – light – verb with a direct object. Almost every sentence contains a noun in the proximity (left or right) to the verb thus being a potential candidate for a NVG. For very large corpora these light verb frequencies reach from ten thousand to more than 1 million occurrences. Obviously a manual analysis of all the concordance lines containing the light verb is neither feasible nor efficient.

Statistical methods help to summarize the data. For this reason we extracted all statistically significant co-occurrences for the target word regardless of their syntactic property and we filter the corresponding NVG manually. Thus we reduce the search space, i.e. the number of cases to be analyzed by more than 90% with respect to concordance lines (cf. section 5). The co-occurrences are computed as bigrams in a word distance between one and five tokens for each sample in order to keep the computational complexity manageable. As we will see below, the large corpus size of the underlying corpora compensates the loss of NVGs that occur in a distance greater than 5. It is necessary to choose the symmetric distance because of the free word-order property of German that also holds for NVG constructions such as in the following example:

- ... *sie leisten ganze Arbeit*
- ... 'they do [a] good job'
- ... *weil sie ganze Arbeit leisten*
- ... 'because they good job do'

We compute statistically salient bigrams by calculating pointwise Mutual Information (MI) as described in Church and Hanks (1991):

$$\mathrm{MI} = \log_2((P(x.y)/P(x)*P(y)) \sim \log_2((f(x.y)/f(x)*f(y))$$

Here f(x) = x/N, f(y) = y/N and N denotes the corpus size expressed in tokens. The different thresholds were: f(x.y) >= 5, MI>0. The MI measure is known to be very sensitive to very low frequencies thus producing a lot of false positives in an n-best list of co-occurrences. A threshold of 5 is considered to be safe for producing a reasonable n-best list (Evert 2008).

In order to study the variation of distinct NVGs among corpora we subdivided our 1 billion token corpus, the DWDS-E, into different samples. To this end, we split the DWDS-E corpus into sentences and performed a random extraction of sentences in an iterative process until the number of tokens exceeds the sample size. The different sample sizes are of 100 million, 250 million, 500 million, 750 million tokens. In order to study the variability of the corpora we extracted 5 different samples of 100 million tokens.

5. Results

The number of different bigrams in both corpora exceeds by far their number of types. The DWDS-C with 100 million tokens has 2.1 million types, but contains 9.135.316 bigrams that occur more than 5 times and 860.829 bigrams that occur more than 50 times. The DWDS-E corpus with its 1 billion tokens and 8.9 million types contains 66.667.626 bigrams that occur more than 5 times and 7.029.606 occur more that 50 times (cf. Table 1).

Table 1: bigram frequencies of DWDS-C and DWDS-E

Corpus	DWDS-C	DWDS-E
f(w1,w2) >= 5	9.135.316	66.667.626
f(w1,w2) >= 10	4.340.056	32.667.626
f(w1,w2) >= 50	860.829	7.029.606

In the following we focus on type and bigram frequencies of our three target verbs *leisten*, *erteilen* and *hegen* (cf. section 3). Their type frequencies in DWDS-C (resp. DWDS-E corpus) is given in Table 2:

Table 2: verb frequencies in DWDS-C and DWDS-E

Verb	DWDS-C	DWDS-E
leisten	6990	70344
erteilen	1476	9619
hegen	461	3285

NVGs with *leisten* ('render, perform')

The WDG dictionary, well known for its quality in the description of verb-noun constructions, lists 16 NVGs for *leisten*:

Abbitte	('offer one's apologies')
Arbeit	('to work')
Bürgschaft	('act as a guarantor')
Dienst	('perform [oder render] a service')
Ersatz	('pay compensation for sth')
Folge	('obey an order')
Gesellschaft	('keep sb company')

Gehorsam ('to obey')
Gewähr ('to guarantee')
Hilfe ('provide help')
Schwur ('make a vow')
Unterschrift ('give one's signature')
Verzicht ('make sacrifices')
Vorschub ('aid or to abet sth')
Widerstand ('offer resistance')
Zahlung ('make a payment')

The frequencies of distinct co-occurrences of *leisten* that are statistically salient (MI >0) are listed in Table 3 (column 2) for each selected corpus (column 1). The frequency treshold is either 5 or 50 (column 3). The last column contains the number of NVG that we manually filtered out of the candidate list of salient co-occurrences. The numbers show that the percentage of statistically salient co-occurrences that can be considered candidates for NVG falls with the growing corpus. In DWDS-C it is around 20% whereas in the DWDS-E it is only around 6%. If we increase the threshold by a factor 10 the percentage rises again to more than 10% in DWDS-E.

Table 3: co-occurrence statistics of *leisten*

Corpus	Co-occurences	Threshold	NVG
DWDS-C	157	5	38
Sample 100m(1)	185	5	42
Sample 100m(2)	176	5	45
Sample 100m(3)	179	5	39
Sample 100m(4)	182	5	41
Sample 100m(5)	187	5	42
Sample 250m	425	5	80
Sample 500m	714	5	131
DWDS-E	3161	5	185
DWDS-E	528	50	60

There are different findings for the comparison of the WDG dictionary with the different corpora as well as among the corpora themselves:

- all the 16 WDG nouns listed with *leisten* are in DWDS-E. Only four out of these 16 nouns co-occur less 50 times with *leisten* in DWDS-E: *Zahlung* (39), *Bürgschaft* (26), *Gewähr* (15), and *Schwur* (12).
- Conversely 169 NVG constructions of DWDS-E are missing in the WDG. The majority of them are transparent compounds, but many of them are either not transparent or no simple nouns. Some of the more eye-catching examples are: *Amtshilfe* ('administrative assistance'), *Eid* ('oath'), *Garantie* ('guarantee'), *Knochenarbeit* ('back-braking work'), *Meineid* ('false oath'), *Obolus* ('small sum of money'), *Offenbarungseid* ('declaration of bankruptcy') ...
- Only 12 out of 16 nouns of WDG nouns can be found as statistically salient co-occurrences of *leisten* in DWDS-C. The missing four nouns are: *Gewähr*, *Bürgschaft*, *Unterschrift*, *Schwur*. If we took a lemma based co-occurrence statistics instead, we would have found *Unterschrift* since the participle *Unterschrift geleistet* is a statistically salient co-occurrence in DWDS-C.
- In DWDS-E but not in DWDS-C are 147 nouns (cf. Appendix A for the complete list).

Examples for simple nouns are: *Abhilfe* (‘relief/remedy’), *Ausgleich* (‘adjustment’), *Nachzahlung* (‘subsequent payment’), *Obolus* (‘contribution’) *Überstunden* (‘overtime’). Important compounds are: *Amtshilfe* (‘obligatory exchange of information’), *Ersatzdienst* (‘non-military service’), *Frondienst* (‘compulsory labour’), *Kärrnerarbeit*, *Knochenarbeit* (both: ‘back-braking work’), ... *Widerruf* (‘revocation’)

- Even among the DWDS-E and the 500m-sample there are some significant differences. The following salient and linguistically interesting NVGs are present in DWDS-E but not in the 500m-sample: *Fronarbeit*, *Frondienst* (‘both: compulsory labour’), *Knochenarbeit* (‘back-braking work’), *Kompensation* (‘compensation’), *Vorauszahlungen* (‘advance payment’), *Starthilfe* (‘jump-start’), *Widerruf* (‘revocation’), *Beziehungsarbeit* (‘relationship building’), *Abschlagszahlungen* (‘payment on account’).

Another question concerns the variability of NVGs within corpora. We can deduce from the figures in table xyz above that there is a significant growth between statistically salient co-occurrences of DWDS-C and DWDS-E. Moreover, we found that there is a considerable variation of samples with the size of 100 million tokens. Therefore we extracted 5 different samples of 100 million tokens, denoted 100m(1), 100m(2) ... 100m(5). If we look at the intersection of all 100m(i), where i varies between 1 and 5 we can make the following observations:

- Only 9 out of the 16 NVG of the WDG are in the intersection of 100m(i), 0 <= i <= 5, and DWDS-C.
- Only 20 NVGs out of about 40 NVGs in each sample are the intersection of DWDS-C and 100m(i). This list is as follows: *Dienst*, *Hilfe*, *Widerstand*, *Arbeit*, *Verzicht*, *Vorschub*, *Folge*, *Abbitte*, *Gesellschaft*, *Dienste*, *Hilfestellung*, *Beitrag*, *Beistand*, *Entschädigung*, *Unterstützung*, *Beiträge*, *Schadenersatz*, *Offenbarungseid*, *Schadensersatz*, *Zahlungen*, *Eid*, *Zwangsarbeit*. Missing in WDG but important are: *Beitrag* (‘make a contribution to sth’), *Beistand*, *Unterstützung* (‘give so. support’), *Eid* (‘swear an oath’), *Offenbarungseid*, *Meineid*.

This means that approximately one half of the statistically salient NVGs vary among corpora of the sample size of 100 million tokens. As a consequence 100 million tokens do not seem to be sufficient as a stable basis for a profound lexicographic analysis of NVG constructions. This also demonstrates the necessity to extend corpora like the BNC oder DWDS-C for the study of multi-word-expressions.

NVGs with *erteilen* (‘give’)

The verb *erteilen* is by a factor 5 less frequent in the DWDS-C corpus than the verb *leisten*. It occurs even 7 times less in the DWDS-E corpus. The WDG dictionary lists 12 (14 with optional parenthesis) different NVGs for *erteilen*:

(Ausreise)-erlaubnis (‘give sb permission [an exit permit]’)
Abfuhr (‘give sb the brush-off’)
Aufenthalts(genehmigung) (‘give sb permission [residence permit] ’)
Auftrag (‘put in an order’)
Auskunft (‘give sb some information’)
Lehre (‘teach sb a lesson’)

Lektion (‘teach sb a lesson’)
Rüge (‘reprimand ’)
Tadel (‘give sb. a blame’)
Unterricht (‘give lessons’)
Verweis (‘rebuke ’)
Weisung (‘instruct ’)

The frequencies of distinct co-occurrences of *erteilen* that are statistically salient (MI >0) are listed in Table 4:

Corpus	Co-occurences	Threshold	NVG
DWDS-C	58	5	29
Sample 100m(1)	36	5	24
Sample 100m(2)	40	5	25
Sample 100m(3)	41	5	28
Sample 100m(4)	36	5	23
Sample 100m(5)	40	5	25
Sample 250m	93	5	47
Sample 500m	201	5	71
DWDS-E	460	5	107
DWDS-E	49	50	32

Table 4: co-occurrence statistics of *erteilen*

The main observations concerning the verb erteilen follow largely those of the verb leisten above. More specifically, the comparison among the different corpora and the WDG results in the following remarks:

- Even though DWDS-E contains 107 NVG with respect to 12, listed in the WDG, there are two nouns in the WDG that are not in DWDS-E: *Tadel* (‘to reprimand’), *Ausreiseerlaubnis* (‘exit permit’). *Tadel* is a historical use, today it would be *Verweis erteilen*, whereas *Ausreise erteilen* had been typical only in an East German context.
- 95 NVG are in DWDS-E but not in the WDG, the more eye-catching of them being *Absage* (‘cancellation’), *Lektion* (‘lesson’), *Ratschlag* (‘advice’)
- In WDG, but not in DWDS-C are four nouns: *Tadel*, *Verweis*, *Rüge*, *Ausreiseerlaubnis*
- 7 out of 12 WDG entries are not in the intersection between DWDS-Kern and 100m
- DWDS-E and DWDS-C: there are 79 statistically salient NVG in DWDS-E that are not statistically salient in DWDS-C (Appendix B).

NVGs with *hegen* (preserve)

Less common - the frequency of *hegen* is 461 in DWDS-C and 3285 in DWDS-E - but not less interesting is the use of *hegen* as a light verb such as in the following constructions: *Hoffnung hegen* (‘harbour hope’) or *Zweifel hegen* (’have some doubt’). It is remarkable that the WDG lists more nouns with *hegen* (29 nouns) than with *leisten* (16) even though the latter is almost 20 times more frequent in DWDS-E. The WDG list with *hegen* is as follows:

> *Abneigung* (‘to have an aversion to sb/sth’), *Abscheu* (‘abhorrence’), *Absicht* (‘intention’), *Achtung* (‘esteem’), *Argwohn* (‘to be suspicious’), *Bedenken* (‘misgivings’), *Befürchtungen* (‘to fear’), *Besorgnis* (‘concerns’), *Bewunderung* (‘admiration’), *Ekel* (‘disgust’), *Erwartung* (‘expectation’), *Freundschaft* (‘friendship’), *Furcht* (‘fear’), *Gedanken* (‘thoughts’), *Gefühle* (‘feelings’), *Gesinnung* (‘attitude’), *Groll* (‘bear a resentment’), *Haß* (‘hatred’), *Hoffnung* (‘hope’), *Liebe* (‘love’), *Meinung* (‘opinion’), *Mißtrauen* (‘mistrust’), *Plan* (‘plan’), *Verdacht* (‘entertain a suspicion’), *Vermutungen* (‘assumption’), *Zorn* (‘anger’), *Zuneigung* (‘affection’), *Zweifel* (‘have doubts’)

The frequencies of distinct co-occurrences of *hegen* that are statistically salient (MI >0) are listed in Table 5:

Corpus	Co-occurences	Threshold	NVG
DWDS-C	13	5	8
Sample 100m(1)	7	5	4
Sample 100m(2)	4	5	3
Sample 100m(3)	6	5	3
Sample 100m(4)	9	5	5
Sample 100m(5)	6	5	5
Sample 250m	23	5	11
Sample 500m	69	5	27
DWDS-E	155	5	47
DWDS-E	17	50	11

Table 5: co-occurrence statistics of *hegen*

The comparison among the corpora as well as between the corpora and WDG leads to the following observations:

- WDG and DWDS-E: 20 out of 29 WDG-entries are in DWDS-E. The following are not: *Abscheu*, *Achtung*, *Bewunderung*, *Ekel*, *Freundschaft*, *Gesinnung*, *Vermutungen*, *Zorn*, *Zuneigung*. We could further reduce this list if we took lemmatized co-occurrences since 7 of these 9 co-occurrences would be statistically salient in DWDS-E: *Abscheu* (with a frequency of 11), *Achtung* (10), *Bewunderung* (59), *Freundschaft* (21), *Gesinnung* (17), *Vermutungen* (87), *Zorn* (2), *Zuneigung* (15). In other words, the only two co-occurrences that are statistically non-salient remain *Ekel hegen* and *Zorn hegen*.
- Conversely, in DWDS-E, but not in WDG are 12 entries: *Ambitionen* (‘ambitions’), *Feindschaft* (‘enmity’), *Glauben* (‘belief’) *Illusionen* (‘illusion’), *Interesse* (‘interest’), *Sorge* (concern’), *Sympathie (*‘sympathy’*)*, *Träume* (‘dreams’), *Vorbehalte* (‘to have reservations’*)*, *Vorliebe* (‘preference’), *Vorstellung* (‘imagination’), *Vorurteil* (‘prejudice’)
- WDG and DWDS-C: only 6 out of 29 WDG verb-noun-pairs are statistically salient co-occurrences in DWDS-C. Again the reason for this is that most of the NVG vary with verb inflection, e.g. if we took the co-occurrences with *hegen* as a lemma, we would find 12 WDG nouns instead.
- No NVG is in the intersection of 100m(i), 0 <= i <= 5, and DWDS-C.

- DWDS-E and DWDS-C: there are 39 statistically salient NVG in DWDS-E that are not statistically salient in DWDS-C (cf. Appendix C).

6. Discussion and Conclusion

In this section we discuss the two questions initially raised, namely to what extent growing corpora provide us with more information about support verb constructions and related to that how large these corpora need to be in order to contain at least all constructions listed in a large monolingual dictionary. On the one hand it is certainly true that "*the more data we have, the more we learn*" (Atkins and Rundell 2008,p. 61), on the other hand we know that the more data we have, the more time we spend to analyze the data. For example support verbs such as *geben* ('to give') or *kommen* ('to come') have both more than a million occurrences in the 1 billion word corpus (DWDS-E) thus making the manual analysis of concordance lines an intractable task. Statistical methods help to combine the data richness of very large corpora with a minimal analysis effort for the lexicographer. Indeed the simple statistics we have adopted here reduce the search space by more than 95% for the target verbs of our case study. For example the verb *leisten* with 70,344 occurrences (that would amount to the same number of concordance lines) has only 3,161 statistically salient co-occurrences (with a threshold of 5). Since German nouns are always capitalized the number of candidates can be further reduced to 1,788. The question is now how much we lose by this reduction.

In our case study with support verb constructions of the verbs *leisten*, *erteilen* and *hegen* (henceforth called NVG, cf. section 1) we compare the candidate lists of our two corpora with the verb entries in a large monolingual German dictionary (WDG). The comparison shows that (almost) all the construction listed in WDG are statistically salient in our 1 billion word corpus (DWDS-E); however the same is not true for DWDS-C, a 100 million word corpus. For the verbs *leisten* (respectively *erteilen*) DWDS-C contains only 12 out of 16 (respectively 8 out of 12) statistically salient NVG. For *hegen* the situation is even worse: only 6 out of 29 NVG were present in DWDS-C. Moreover we examined the variation of among several 100 million word samples of DWDS-E and DWDS-C: only one half of the statistically salient NVG can be found in the intersection of these corpora. Do these results indicate that the token size of 100 million is not large enough to serve as a sufficient basis for a thorough lexicographic work with NVG? The answer is 'yes' if the analysis is done on the basis of statistically salient occurrences alone. It is 'no' if we look at individual patterns in the DWDS-C corpus. For example all missing NVG for the verbs *leisten* and *erteilen* are in DWDS-C but with frequency of only 1 or 2. For *hegen* the missing NVG patterns can be found in DWDS-C with different inflected forms. Thus Hank's claim that 'a corpus of 100 million words, a simple right – or left sorted corpus clearly shows most of the normal patterns of usage for all words except the very rare' (Hanks 2002:157) can be confirmed for the target patterns of our study.

Another finding of our case study is that the number of statistically NVG candidates for our three candidate verb *leisten*, *erteilen* and *hegen* grows to the size of 1 billion words. Even though significant new constructions can be detected between the size of 100 million and 500 million words there is much less we can "learn" in the interval between 500 million and 1 billion words. Our analysis shows that only a few compound-nouns emerge as statistically salient co-occurrences in this interval.

Finally, a closer look at the statistically salient co-occurrences in DWDS-C and DWDS-E that are not listed in the WDG shows that both corpora provide an important amount lexicographically relevant material for NVG that are not in WDG (cf. Table 6 and Appendix A-C)

Table 6: corpus NVG that are not present in WDG

verb	DWDS-C (relevant)	DWDS-E (relevant)	DWDS-E (total)
leisten	6	20	153
erteilen	8	15	77
hegen	7	14	14

Future work should focus on an analysis of a high frequent support verbs such as *geben* ('to give') or *kommen* ('to come') with 1.4 million resp. 1.1 million occurrences in DWDS-E. We expect by the analysis of a verb of this frequency class a better insight in the relationship between the frequency threshold of bigrams and the number of linguistically interesting co-occurrences. We have already stated above that a frequency of less than five will not only produce much noise in the n-best of NVG candidates list but also, what is even worse, statistics the n-best list will simply be too large in order to analyze it manually. For example, for higher frequent light verbs such as *geben* ('to give') or *kommen* ('to come') with 1.4 million resp. 1.1 million occurrences in DWDS-E we can expect a NVG candidate list of 50,000 to 70,000 . On the other hand, a frequency threshold of 50 seems too large since a significant amount of the interesting verb-noun-expressions has a smaller bigram frequency in the 1 billion DWDS-E corpus.

References:

Atkins, S. and Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Church, K. and Hanks, P. 1991. 'Word Association Norms, Mutual Information and Lexicography' *Computational Linguistics* 16.1: 22–29.

Evert, S. 2008. 'Corpora and collocations' in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*. Berlin : Mouton de Gruyter.

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, 23–40.

Eisenberg, P. 1999. *Grundriß der deutschen Grammatik. Band 2: Der Satz*. Stuttgart/Weimar: Metzler.

Hanks, P. 2002. 'Mapping Meaning onto Use' in M.-H. Corréard (ed.), *Lexicography and Natural Language Processing: a Festschrift in honour of B. T. S. Atkins*. Euralex.

Hanks, P., Urbschat, A. and Gehweiler, E. 2006. 'German Light Verbs in Corpora and Dictionaries' *International Journal of Lexicography* 19.4: 439–457.

Heaps, H. S. 1978. *Information Retrieval - Computational and Theoretical Aspects*. London : Academic Press.

Klein, W. and Geyken, A. 2000. 'Projekt ‚Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts'. *Jahrbuch der BBAW 1999*. Berlin: Akademie-Verlag, 277–289.

Köhler, Reinhard: 'Zur Wachstumsdynamik (Type-Token-Ratio) syntaktischer Funktionen in

Texten' in: S. Kempgen, U. Schweier, T. Berger (eds.): *Rusistika - Slavistica - Lingvistika*. Festschrift für Werner Lehfeldt zum 60. Geburtstag. München : Sagner, 498–504.
Malige-Klappenbach, H. 1986. *Das Wörterbuch der deutschen Gegenwartssprache: Bericht, Dokumentation und Diskussion*. Tübingen : Niemeyer.
Manning, C., Raghavan, P. and Schütze H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
Polenz, Peter von (1987): 'Funktionsverben, Funktionsverbgefüge und Verwandtes. Vorschläge zur satzsemantischen Lexikographie.' *Zeitschrift für germanistische Linguistik* 15: 169–189.
Storrer, Angelika (2007): 'Corpus-based Investigations on German Support Verb Constructions' in C. Fellbaum (ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum Press, 164–187.
Wiegand, H. E. 1990. 'Die deutsche Lexikographie der Gegenwart' in F.-J. Hausmann; O. Reichmann, H. E. Wiegand and L. Zgusta (eds.), *Wörterbücher. Ein internationales Handbuch zur Lexikographie*. Berlin/New York: de Gruyter, 2100–2246.

Dictionaries:

[WDG] Klappenbach, R.and Steinitz, W. (1964-1977). Wörterbuch der deutschen Gegenwartssprache (WDG). 6 vol., Berlin : Akademie-Verlag.

Appendix A: list of nouns with *leisten* in DWDS-E

The following list of 185 nouns contains all statistically salient co-occurrences of *leisten* in DWDS-E that are NVG candidates. Only 39 of these nouns are salient in DWDS-C; they are marked with a '+'. The words followed by a '*' are present in the WDG. Nouns that are present in both the WDG and DWDS-C are marked with '*+'.

> *Abbitte*+, Abgaben+, Abhilfe, Abschlagszahlungen, Amtseid, Amtshilfe, Anschubfinanzierung, Anzahlung+, Arbeit*+, Arbeiten+, Arbeitsdienst, Arbeitsstunden, Argumentationshilfe, Aufarbeitung, Aufbauarbeit, Aufbauhilfe, Aufklärung, Aufklärungsarbeit, Ausbildung, Ausgaben+, Ausgleich, Ausgleichszahlung, Ausgleichszahlungen, Außerordentliches, Basisarbeit, Beachtliches, Beihilfe+, Beistand+, Beitrag+, Beiträge+, Beiträgen, Beratung, Besonderes, Betreuungsarbeit, Beziehungsarbeit, Bildungsarbeit, Buße, Bürgschaft*, Detailarbeit, Dienst*+, Dienste+, Eid+, Eigenbeitrag, Einsätze, Entscheidungshilfe, Entschädigung+, Entwicklungsarbeit, Entwicklungshilfe, Ergebnisbeitrag, Erinnerungsarbeit, Ersatz*+, Ersatzdienst, Erziehungsarbeit, Finanzhilfe, Finanzierungsbeitrag, Folge*+, Forschung, Forschungsarbeit, Fronarbeit, Frondienst, Frondienste, Führungsarbeit, Garantie+, Garantien, Geburtshilfe, Gefolgschaft+, Gegenwehr, Gehorsam*+, Gesellschaft*+, Gewähr*, Grundwehrdienst, Handarbeit, Hauptarbeit, Hausarbeit, Heeresfolge, Hervorragendes+, Hilfe*+, Hilfen, Hilfestellung+, Hilfestellungen, Hilfsdienste, Integrationsarbeit, Investitionen, Jugendarbeit, Knochenarbeit, Kompensation, Kostenvorschuß, Kriegsdienst, Kulturarbeit, Kärrnerarbeit, Lebenshilfe, Lobbyarbeit, Mehrarbeit, Meineid+, Militärdienst, Militärhilfe, Mitarbeit, Nachbarschaftshilfe, Nachzahlung, Nothilfe, Nützliches, Obolus, Offenbarungseid+, Oppositionsarbeit, Orientierungshilfe, Pionierarbeit, Pressearbeit, Projektarbeit, Präventionsarbeit, Präzisionsarbeit, Qualitätsarbeit+, Rechtsbeistand, Rechtshilfe, Regierungsarbeit, Rückzahlungen, Sacharbeit, Sanierungsbeitrag, Schadenersatz+, Schadensersatz+,*

Schichtarbeit, Schrittmacherdienste, Schuldendienst, Schwerarbeit, Schwerstarbeit, Schwur*, Schützenhilfe, Selbsthilfe, Sinnvolles, Sklavenarbeit, Soforthilfe, Solidarbeitrag, Solidarität, Sozialabgaben, Sozialarbeit, Sozialdienst, Sozialhilfe, Sparbeitrag, Spitzeldienste, Starthilfe, Sterbehilfe, Sühne, Teamarbeit, Teilzeitarbeit, Trauerarbeit, Treueid+, Ungehorsam, Unmögliches, Unterhalt, Unterschrift*, Unterschriften, Unterstützung+, Versorgung, Verwaltungsarbeit, Verzicht*+, Vorarbeit, Vorarbeiten, Vorauszahlungen, Vorschrift, Vorschub*+, Vorsorge, Völkerverständigung, Wachstumsbeitrag, Waffendienst, Waffenhilfe, Wahlhilfe, Wahlkampfhilfe, Wehrdienst, Widerruf, Widerstand*+, Wiederaufbau+, Wiedergutmachung, Zahlung*+, Zahlungen+, Zivildienst, Zuarbeit, Zubringerdienste, Zusammenarbeit+, Zuschuß+, Zuschüsse, Zuzahlungen, Zwangsarbeit+, Öffentlichkeitsarbeit, Überlebenshilfe, Überstunden, Überzeugungsarbeit

Appendix B: list of nouns with *erteilen* in DWDS-E

The following list of 107 nouns contains all statistically salient co-occurrences of *erteilen* in DWDS-E that correspond to NVG. 79 of these nouns are not salient in DWDS-C; the rest is marked with a '+'. The words followed by a '*' are present in the WDG. Nouns that are present in both the WDG and DWDS-C are marked with '*+'.

Abfuhr*+, Abmahnung, Absage+, Absagen, Absolution, Anordnungen, Antwort+, Anweisung, Anweisungen+, Arbeitserlaubnis, Aufenthaltsbefugnis, Aufenthaltsbefugnisse, Aufenthaltsbewilligung, Aufenthaltserlaubnis+, Aufenthaltsgenehmigungen, Aufenthaltsgenehmigung*, Auflagen+, Auftrag*+, Aufträge+, Auskunft*+, Auskünfte+, Ausnahmegenehmigung, Ausnahmegenehmigungen, Aussagegenehmigung, Baugenehmigung, Baugenehmigungen, Befehl+, Befehle+, Befugnis, Belehrungen, Bescheid, Betriebsgenehmigung, Bewilligung, Bleiberecht, Dauerbetriebsgenehmigung, Denkzettel, Deutschunterricht, Duldung, Einfuhrlizenzen, Einreisevisum, Einvernehmen, Einwilligung, Empfehlungen, Entlastung, Erlaubnis*+, Ermächtigung, Fahrerlaubnis, Fahrverbot, Freigabe, Genehmigung*+, Genehmigungen, Hausverbot, Informationen+, Instruktionen+, Konzession, Landeerlaubnis, Lehre*+, Lehren+, Lektion*+, Lektionen, Lizenz, Lizenzen, Mandat, Musikunterricht, Möglichkeit, Nachhilfe, Nachhilfeunterricht, Noten, Order, Patente, Placet, Platzverweis, Platzverweise, Privatunterricht+, Projekt, Quittung, Ratschlag, Ratschläge+, Recht, Redeverbot, Regierungsauftrag, Religionsunterricht+, Richtlinien, Rüge*, Segen, Sondergenehmigung, Teilerrichtungsgenehmigung, Unterlagen, Unterricht*+, Unterrichtsstunden, Verhandlungsmandat, Verlangen+, Verweis*, Visa, Visum, Vollmacht, Vollmachten+, Vorschriften, Warnung, Weihen, Weisung*+, Weisungen+, Wort+, Zensuren, Zulassung, Zuschlag, Zustimmung+

Appendix C: list of nouns with *hegen* in DWDS-E

Appendix C contains the complete list of 47 statistically salient co-occurrences that were manually classified as NVG candidates. Only 9 of them are salient in DWDS-C; they are marked with a '+'. The words followed by a '*' are present in the WDG. Nouns that are present in both the WDG and DWDS-C are marked with '*+'.

Abneigung*, Abscheu*, Absicht*, Absichten*+, Achtung*, Ambitionen, Argwohn*, *Bedenken*+, Befürchtung*, Befürchtungen*+, Bewunderung*, Erwartung*, Erwartungen, Feindschaft, Freundschaft*, Gedanken*+, Gefühle*, Gesinnung*, Glauben, Groll*, Haß*, Hoffnung*+, Hoffnungen*, Illusion, Illusionen, Interesse, Liebe*, Mißtrauen*, Plan*, Pläne*, Sorge, Sympathie, Sympathien, Traum, Verdacht*, Vermutung*, Vermutungen, Vertrauen+, Vorbehalte, Vorliebe, Vorstellung, Vorurteile, Wunsch*+, Wünsche, Zuneigung*, Zweifel*+, Überzeugung*