Linguistic experience and productivity:
corpus evidence for fine-grained distinctions

Anke Lüdeling
Humboldt-Universität Berlin
anke.luedeling@rz.hu-berlin.de

Stefan Evert
Universität Stuttgart
evert@ims.uni-stuttgart.de

1.   Morphological productivity: qualitative and quantitative aspects

In this paper we show how qualitative and quantitative analyses of morphological productivity interact. A qualitative analysis alone does not match our intuitive notion of productivity. The quantitative analysis which must complement it, on the other hand, crucially depends on a detailed intensional understanding of the morphological process studied and its sub-processes.

Morphological productivity (roughly: the readiness with which a word formation process forms new words) has long been one of the central mysteries of morphology. There are many detailed *qualitative* descriptions of word formation processes that list the restrictions for the possible bases for an affix (like the restriction that -*able* attaches to verbs in English) or the restrictions on a given compounding process (like the fact that noun-noun compounding in German is recursive while noun-verb compounding is not). Most of these works are not concerned with productivity or equate productivity simply with type frequency (see for example the standard descriptive works for German word formation Kühnhold, Putzer & Wellmann 1978, Fleischer & Barz 1992).

Although we shall see below that qualitative studies of word formation processes are a necessary basis for all further research on productivity, they cannot account for our intuition that some word formation processes produce new words more easily or more frequently than others. How can we describe such differences morphologically? The underlying cognitive question is, of course: How does a speaker know that she can form more new words using process X than using Y? Speakers of English, for example, know that they can form nominalisations of new adjectives with the suffix -*ness* and not e.g. with –*th*. How can the different 'degrees' or measures of productivity be coded into morphological rules? To express their intuition of different degrees of productivity many authors use quantitative terms such as 'marginally productive', 'very productive', 'semi-productive' etc. (a list is given in Plag 1999). A number of suggestions on how to quantify productivity were made (see Aronoff 1976 and Booij 1976 for examples), but these cannot be operationalised and would often give unintuitive results.[1]

In the last 10 years or so (essentially since Baayen 1992), there are also corpus-based quantitative studies of morphological productivity based on a simple statistical model of text production (the model has been used in many studies in quantitative linguistics and related areas of statistics since the 1950's; a full mathematical definition and a discussion of the most important properties of this model can be found in Baayen 2001). These studies usually define productivity as the likelihood of finding a new type (produced by a given word formation process) after a certain amount of text has been sampled. In the statistical model of (Baayen 2001), this likelihood, called the productivity index *P,* can be estimated from the number of hapax legomena (types seen only once in the sample). Note that the definition of *P* is not a direct operationalisation of the intuitive concept of productivity above (the ease with which new types are formed), but has to be understood as an independent, quantitative measure of the degree of productivity. (Plag 1999) and (Bauer 2001) formulate comprehensive theories of productivity that integrate qualitative and quantitative aspects.

The approach adopted by Baayen need not be seen as a purely statistical approach, though. It can also be understood as a cognitive model which assumes that a speaker somehow 'knows' that she can use process X to produce new types precisely *because* she has seen many new types produced by X (i.e., she knows that other speakers have also produced new words by X). In addition to a number of assumptions about the storage of complex words in the mental lexicon that we cannot discuss here, this

---

[1] (Aronoff 1976), for example, divides the number of possible types that could be formed by a given word formation process X into the number of existing types. Even if one could determine all existing and possible types for process X, the resulting number (e.g. 50%) would not tell us whether we would ever expect to see a new type formed by X. (Booij 1977) attempts to formulate a productivity measure based on the number of restrictions on the morphological elements involved: the more restrictions there are, the less productive the process is.

approach is based on the conception of the corpus as a model for linguistic experience. The distribution of, say, -*lich*-adjectives in the corpus 'reflects' the linguistic experience of a speaker (we are aware of the problems related to such an assumption such as representativeness and pre-processing but will ignore these for the present purpose).

The case study in this paper deals with German adjective formations involving the suffix -*lich*. The observed data are taken from a 3 million word newspaper corpus (2 years of the *Stuttgarter Zeitung,* henceforth STZ corpus), which has been automatically part-of-speech tagged and lemmatised.[2] All adjectives ending in the string 'lich' (and corresponding inflected forms) are extracted from the corpus. Then the data is semi-automatically 'cleaned up', according to the principles given in (Evert & Lüdeling 2001).

The paper is organised as follows: the morphological and statistical models and assumptions from which our case study departs are explained in Section 2. Section 3 reviews some problems of the preliminary quantitative analysis and shows that a more detailed analysis is required. A summary is given in Section 4.

2. Underlying qualitative and quantitative models

In this section we formulate the models and assumptions that we take as our starting point. Some of them will have to be revised later.

2.1 The morphological model

Our study is based on two assumptions that are not necessarily standard among morphologists. Firstly, we use a simple rule-based item-and arrangement model. Stem changes such as the insertion of 'linking' elements, the fronting of certain vowels (umlauting), or the elision of some elements are considered lexical specifications that are listed as 'word formation stem forms' together with the lexical entry of each morphological element (Fuhrhop 1998, Lüdeling & Fitschen 2002).

Secondly, we use the so-called compounding theory of derivation which means that derivation in German is conceived of as pure concatenation and that suffixes are morphological heads with part-of-speech categories. The consequence is that there is no categorical difference between stems and affixes, and also no difference between compounding and derivation (cf. Höhle 1982; see also Williams 1981 for a similar approach). Rules can be maximally specified – that is, each morphological element can be described on all linguistic levels. Some examples for such morphological rules are given in (1), where (1a) describes a very general adjectivisation rule that derives adjectives from verbs and any adjective suffix, (1b) describes the formation of adjectives from transitive verbs and the adjective suffix -*bar* 'able', (1c) describes the formation of adjectives from verbs and the adjective suffix -*lich*, and (1d) describes the derivation of the adjective *lesbar* 'readable' from the verb stem *les*- 'read' and the suffix -*bar*. (1b) and (1c) are specialisations of the general rule (1a), with (1d) as a further specialisation of (1b).

(1a) ADJ → V + ADJ_SUFFIX
(1b) ADJ → V[transitive] + -bar
(1c) ADJ → V + -lich
(1d) lesbar → les- + -bar

Note that some rules contain intensionally defined variables, while others do not. In this framework, all rules that contain variables are maximally productive, that is, we expect that *all* elements that fit the description can be inserted here (for example, as it stands, rule (1b) states that all transitive verbs form bar-adjectives). This gives a simple distinction between productive rules (rules that contain variables, such as (1a-c)) and unproductive rules (rules that contain no variables, such as (1d)). Every productive rule should produce an in principle unlimited number of types, otherwise it could simply be replaced by a finite number of unproductive rules like (1d). We have no way of expressing our intuition that some

---

[2] The work was inspired by a detailed qualitative and quantitative study of about 200 word formation processes in German (collected in the DeKo project at the IMS, University of Stuttgart; see Schmid et al. 2001 for a details on the qualitative description), based on a 200 million word newspaper corpus; since in this case a manual correction of all the types was necessary we had to rely on a smaller corpus.

rules produce new words more readily than others. Therefore we have to complement this approach with a quantitative model of productivity.

2.2 The productivity model

The statistical model of (Baayen 1992, 2001) describes the linguistic output of a speaker (or the collective output of a homogeneous group of speakers) as the result of a simple stochastic process, where words are chosen randomly from the speaker's vocabulary. Each word type in the vocabulary is associated with a characteristic type probability (π_i) corresponding to its average frequency of use (by the modelled speaker). Consecutive word tokens produced by the stochastic process are assumed to be independent, hence Baayen's model is completely oblivious of syntax. Although a fairly crude approximation of natural language, the model provides a satisfactory explanation for many word frequency distributions, as the examples in (Baayen 2001) show. The output of the stochastic process can be visualised as a vocabulary growth curve, plotting V(N), the number of types found among the first N tokens, against N (see Figure 1).

When this model is used to describe morphological productivity, the vocabulary is usually restricted to types formed by one particular process (represented by a rule as in (1a-d)), so that the output of the stochastic process will only contain such formations. It is important to understand that even for a productive process, the statistical model assumes a fixed vocabulary, which must not be interpreted as a kind of 'mental lexicon'. Rather, the vocabulary includes all types that can be productively formed by the speaker in addition to lexicalised types.

Mathematically, a productive rule is usually equated with a very large or infinite vocabulary size S (just as every productive rule should give rise to infinitely many types, according to the definitions in 2.1) and a large number of low-probability types in the vocabulary. Such a distributions of type probabilities is known as an LNRE distributions (for large number of rare events).[3] Therefore, as more and more text is generated by the model, the number of types observed in the output will continue to grow indefinitely. By contrast, for an unproductive process the vocabulary growth curve will flatten out until the vocabulary is exhausted (i.e. V(N) = S) and no more new types can appear. So far, this gives us a distinction between productive and unproductive processes. In order to capture the intuition that different processes exhibit different degrees of productivity, (Baayen 1992) suggests the productivity index *P*, measuring the rate at which previously unseen words occur in the output of the model. Intuitively, *P* can be interpreted as the slope of the vocabulary growth curve. (Baayen 2001) shows that *P* can be estimated from the number of hapax legomena $V_1(N)$ using the formula $P = V_1(N)/N$. Note that *P* is dependent on N: even for a productive process, *P* is a decreasing function of N, although it never reaches 0.

In order to make inferences about the unknown vocabulary and type probabilities of a word formation process, the observed data from a corpus is assumed to be the output (or, more precisely, a plausible output) of the stochastic process described above. Figure 1 shows the observed vocabulary growth curve for lich-adjectives in the STZ corpus. A point on the x-axis corresponds to the first N occurrences of lich-adjectives in the corpus. On the y-axis, the thick line indicates the number of different types at this point, while the thin line indicates the number of hapax legomena. At the end of our corpus, after sampling close to 140,000 occurrences of lich-adjectives, both lines continue to grow. This suggests that lich-adjectivisation is indeed a productive process: if we added more text to the corpus, we would expect to see more new lich-types. However, Figure 1 also shows that the slope of the vocabulary growth curve, and hence *P* decreases, just as the statistical model predicts. Therefore, we cannot take *P* as an 'absolute' measure of the degree of productivity, but rather as a measure of the productivity rate after a certain number of lich-tokens have been sampled. For the same reason, it is problematic to use *P* to compare the productivity of different processes, at least when the numbers of tokens differ substantially. (Evert & Lüdeling 2001) visually compare vocabulary growth curves, but can only make the coarse distinction between productive and unproductive processes in this way.

---

[3] Since the probabilities of all types in the vocabulary must sum up to 1, an LNRE distribution is a necessary consequence of an infinite or extremely large vocabulary. Note that the original definition of LNRE distributions by (Khmaladze 1987) explicitly requires an infinite vocabulary.
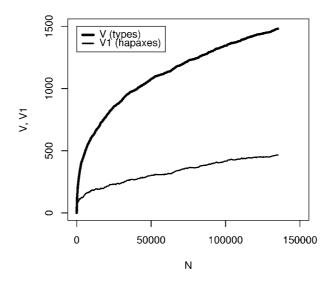
Figure 1. Vocabulary growth curve for lich-adjectives

The statistical model described above can be used to extrapolate vocabulary growth curves and thus predict the development of *P* if more text were added to the corpus, allowing for the comparison of different processes. The behaviour of the extrapolated vocabulary growth for N → ∞ can also be understood as an 'absolute' measure for the degree of productivity.

However, it is impossible to estimate all the type probabilities, which are needed for extrapolation, directly from the observed sample. This is especially true for the unseen types, which do not occur in the sample at all. Therefore, (Baayen 2001) introduces several LNRE models for the distribution of type probabilities in the vocabulary, thereby reduceing the number of parameters that have to be estimated. The remaining two or three parameters are estimated from the frequency spectrum, which is essentially a listing of all types and their frequencies in the sample. The quality of an LNRE model, its goodness of fit, is determined by how closely the predicted frequency spectrum matches the observed one. Alternatively, the vocabulary growth curve predicted by the model can be compared to the observed curve. For our data, none of Baayen's LNRE models resulted in a fully satisfactory fit, which indicates that the vocabulary underlying our data does not have the simple, homogeneous probability distribution implied by these models.

3. Consequences for the treatment of productivity

Before we come back to the problem of improving goodness of fit, we need to review our goals. We wanted to find out more about the productivity of the word formation process that generates lich-adjectives. The vocabulary growth curve in Figure 1 indicates that lich-adjectivisation is a productive process, but it does not seem to be generated by a vocabulary consistent with any single one of the LNRE models described above.

The rule that was implicitly assumed to generate the data in Figure 1 is

(2) ADJ → X + -lich

A possible reason for the fact that the lich-words do not seem to follow a single model is that they are not generated by one homogeneous process but rather by several processes – each of them contributing in a different way to the final curve. (Baayen 2001) discusses a similar case where he argues that lexicalised words that are stored and can simply be retrieved from the mental lexicon do not necessarily follow the same model as words that have to be generated ad hoc by a morphological rule. The vocabulary growth curve is thus generated by two different models. In the next section we show that in the case of lich-adjectives the assumption of a single morphological process is indeed linguistically implausible.

3.1   Category of base

The first and most obvious distinction we can make is the distinction between different parts of speech of the base: *-lich* attaches to adjectives, nouns and verbs. In each of these sub-patterns *–lich* has different restrictions and different semantic effects (compare Kühnhold, Putzer & Wellmann 1978, Fleischer & Barz 1992) :

   (a) adjective + *-lich* means 'a little ADJ', cf. *grün* 'green' – *grünlich* 'a little green, greenish' or *süß* 'sweet' – *süßlich* 'sweetish'. The adjectives in this pattern are all native and mostly monosyllabic or bisyllabic.

   (b) The semantics of verb + *-lich* is irregular and difficult to describe. We find cases where *-lich* denotes a property of the subject as *vergeßlich* 'forgetful' which is derived from transitive *vergessen* 'to forget': *X vergißt Y* 'X forgets Y' – *X ist vergeßlich* 'X is forgetful'. In other cases, the *lich*-adjective denotes a property of an object of the transitive verb, as in *Y ist beachtlich* 'Y is remarkable' from *X beachtet Y* 'X notices Y'.

   (c) *-lich* plus a noun again falls into different semantic classes. If *–lich* attaches to nouns that denote a personal profession or relationship, the resulting adjective means 'as an N, from an N', cf. *Arzt* 'physician' – *ärztlich* 'from a physician, as a physician', *Großmutter* 'grandmother' – *großmütterlich* 'from a grandmother, grandmotherly'. With nouns denoting a time span *–lich* means 'every N': *stündlich* 'hourly' from *Stunde* 'hour'. If *–lich* attaches to other nouns it is a quite generic adjectivisation operator meaning roughly 'like N, from N' as in *Geschichte* 'history' – *geschichtlich* 'historical'.

   (d) In addition there are phrasal bases[4] (these are only mentioned in passing in the descriptive literature, but see below): *nachweihnachtlich* 'after christmas' is clearly derived from the PP *nach Weihnachten* and not a compound of *nach* 'after' and *weihnachtlich* 'Christmas-like'; *zweiwöchentlich* 'every fortnight' is derived from *zwei Wochen* 'two weeks' and not a compound of *zwei* and *wöchentlich* . For the sake of this study we collect all those cases under the heading phrasal which are not clearly derived from words (in the corresponding rule (3d) we use the category label XP although the categorical status of the phrase is debatable).

Even if we only differentiate for category of base, we have four processes instead of just one.

   (3a)   ADJ → ADJ[native, monosyllabic or bisyllabic] + -lich
   (3b)   ADJ → V + -lich
   (3c)   ADJ → N +-lich
   (3d)   ADJ → XP + -lich

As noted above, quantitative studies of *-lich*-derivation normally do not differentiate between these sub-patterns (since they only want to compute the probability of *any* new *-lich*-type). Our hypothesis is that speakers, on the other hand, do differentiate between the sub-patterns, which implies that we should calculate productivity measures separately for each sub-pattern. Figures 2 and 3 show that their productivities differ considerably.

Figure 2 shows that N + *-lich* is the most frequent pattern and clearly productive, while V + *-lich* is not: obviously nearly all possible types have already been sampled – this corresponds to the irregular semantics of the V + *-lich* pattern. A + *-lich* and XP + *-lich* produce only a small number of tokens. In order to be able to see their curves better we produced a zoomed plot in Figure 3.

---

[4] These formations are interesting because many morphological analyses assume that phrases cannot enter word formation. We cannot elaborate on this point here. Cases like *zweiwöchentlich* are sometimes called 'Zusammenbildungen' (see Leser 1990).
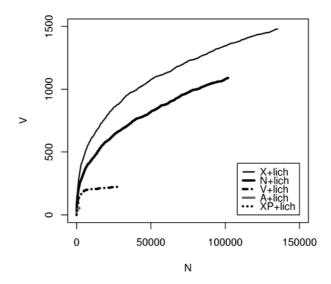
Figure 2: Vocabulary growth curves for the different lich-patterns

Here we see that the A and XP curves seem to be in the 'early' part of the development and are therefore difficult to interpret (note that productive and unproductive processes alike have a phase of fast growth at the beginning of their vocabulary growth curve when almost every new token is also a new type). Therefore, the curves were extrapolated as thin lines using the GIGP model described below; the extrapolation suggests that XP but not A may be productive.
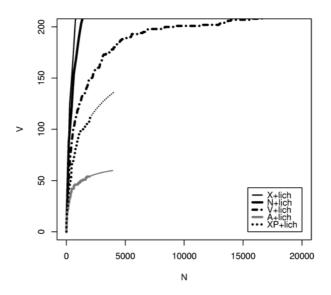

Figure 3: Zoomed-in vocabulary growth curves for the different lich-patterns. The thin lines show the model extrapolation (GIGP) for XP + *-lich* and A + *-lich*

We applied the generalised inverse Gauß-Poisson (GIGP) model with $\gamma = -0.5$ (Baayen 2001, 89-93).[5] Interestingly, unlike most other LNRE models, the GIGP model assumes a finite vocabulary, whose size S depends on the model parameters. This offers a new possibility for measuring and comparing degrees of productivity: by looking at the predicted vocabulary size and, especially, the predicted number of unseen types $V_0(N) = S - V(N)$. Table 1 shows the observed number of types V and tokens

---

[5] Although Baayen reports better results with the Yule-Simon model (Baayen 2001, 124f), the GIGP estimation is more robust and often achieves a better fit for our data.

N for all lich-adjectives and separately for the four sub-patterns (3a-d), together with the vocabulary size S predicted by the GIGP model.

After what has been said in Section 2, the predicted vocabulary size (of roughly 2,500 types) may seem far too small for a productive pattern. This is due to the fact that the model parameter were estimated from a corpus containing only newspaper articles from a single newspaper and from a relatively short time span. The model can thus only predict what happens when we add more text of the same kind; the predicted vocabulary includes only words that might have been used by the same journalists writing about the same topics.

The rightmost column of Table 1 shows the goodness-of-fit estimation calculated for the GIGP model, using the multivariate χ^2 test described in (Baayen 2001, 118f). The overall fit printed in the first row is bad, as noted above. The model gives a reasonable fit for A and XP, where a p-value of 23% indicates that the observed data is entirely consistent with the GIGP model. The estimated vocabulary size for A is only marginally larger than the observed vocabulary size, which suggests that almost all possible types have already been sampled. On the other hand the GIGP model predicts quite a number of new XP + -*lich* adjectives. The fit for N and V, however, is still poor. While V cannot be expected to fit well because it is not a productive pattern and hence does not follow an LNRE distribution, the bad fit for N is unsatisfactory. We will suggest an explanation below.

base	N	V	S (GIGP)	Goodness of fit
X	135252	1481	2532.75	X2(14) = 3106.66, p = 0
N	101953	1090	2394.59	X2(14) = 1644.47, p = 0
V	29287	225	239.96	X2(14) = 785.48, p = 0
A	1959	54	64.86	X2(14) = 32.66, p = 0.00325
XP	2053	112	188.60	X2(14) = 17.38, p = 0.23673

Table 1: Overview over the different lich-categories.

3.2 Subpatterns

The differences in productivity shown for the different base categories do not challenge our morphological model as such. However, if we look more closely at the -*lich* derivations from nouns, which are not described well by our statistical model, we find that there seem to be subpatterns involving complex nouns that cannot be distinguished categorically but that nevertheless show more types than would be otherwise expected. One example: we find 71 types of lich-adjectives that are derived from complex nouns which have the head noun *Geschichte* 'history': *naturgeschichtlich* from *Naturgeschichte* 'natural history', *literaturgeschichtlich* from *Literaturgeschichte* 'literary history', *architekturgeschichtlich* from *Architekturgeschichte* 'history of architecture', or *religionsgeschichtlich* from *Religionsgeschichte* 'history of religion' are just some examples. Similar 'clusters' can be found with a number of other head nouns such as *Recht* 'law', *Wissenschaft* 'science' or *Wirtschaft* 'economy'. All of these can be generated by rule (3c) since the variable N is not morphologically restricted there and NN compounding is productive. We thus seem to have a number of sub-rules of rule (3c), as seen in (4). These rules are strictly speaking superfluous since they cannot generate anything that would not also be generated by (3c).

(4a)   ADJ → (X+Geschichte)+-lich
(4b)   ADJ → (X + Recht) + -lich
(4c)   ADJ → (X + Wissenschaft) + -lich
(4d)   ...

Table 2 shows a list of ten such sub-patterns that we found in the N + -*lich* pattern. Several others such as *kirchlich* 'of the church' and *gerichtlich* 'of or by a court' had to be excluded because the small number of types found in our corpus did not provide enough training material for the GIGP model. Together, these patterns account for 434 of the 1090 N + -*lich* types in the corpus. For three patterns

with V > 50, the GIGP model gives an excellent fit, predicting a relatively high degree of productivity: only half of their vocabulary or even less has already been sampled..

The results for the remaining N + -*lich* types, which do not fall into any of these sub-patterns, is shown in the last row. The fit of the GIGP model is extremely poor for this group, even worse than the overall fit of the N + -*lich*, indicating that this group is still an inhomogeneous mixture.

base	N	V	S	Goodness of fit
X+Arzt	1123	29	55.71	X2(14) = 100.91, p = 0
X+Bau	1020	11	23.82	X2(14) = 78.49, p = 0
X+Geschichte	703	71	258.78	X2(14) = 17.43, p = 0.23395
X+Polizei	450	13	31.23	X2(14) = 61.73, p = 0
X+Recht	3314	140	253.12	X2(14) = 18.12, p = 0.20121
X+Sprache	396	18	73.30	X2(14) = 74.13, p = 0
X+Tag	1620	20	32.33	X2(14) = 84.07, p = 0
X+Wirtschaft	7731	50	140.90	X2(14) = 19.56, p = 0.14476
X+Wissenschaft	2311	53	109.61	X2(14) = 50.83, p = 0
X+Zeit	656	29	33.92	X2(14) = 108.55, p = 0
N w/o the above	82629	656	1486.34	X2(14) = 2941.39, p = 0

Table 2: Subpatterns of the N + -*lich* pattern

4.   Conclusion

In this paper we have given a qualitative and quantitative analysis of the German adjective suffix –*lich*, in order to illustrate how both aspects of morphological productivity interact.

The mathematical properties of a quantitative study of the productivity of a word formation process can serve as a clue to the qualitative analysis: where statistical models fail to fit the observed data well it may be necessary to refine the qualitative analysis. The (purely qualitative) refinements in Section 3 led to a much better fit, at least for some of the sub-patterns.

References

Aronoff M 1976 *Word Formation in Generative Grammar*. Cambridge, MIT Press.

Baayen R H 1992 Quantitative aspects of morphological productivity. *Yearbook of Morphology 1991*: 109 – 150.

Baayen R H 2001 *Word Frequency Distributions*. Dordrecht, Kluwer.

Bauer L 2001 *Morphological Productivity*. Cambridge, Cambridge University Press.

Booij G 1977 *Dutch Morphology. A Study of Word Formation in Generative Grammar*. Lisse, de Ridder.

Evert S, Lüdeling A 2001 Measuring morphological productivity: Is automatic preprocessing sufficient. In *Proceedings of the Corpus Linguistics 2001 conference*, Lancaster, pp 167 – 175.

Fleischer W, Barz I 1992 *Wortbildung der deutschen Gegenwartssprache*. Tübingen, Niemeyer.

Fuhrhop N 1998 *Grenzfälle morphologischer Einheiten*. Tübingen, Stauffenburg.

Höhle, T 1982 Über Komposition und Derivation. Zur Konstituentenstruktur von Wortbildungsprodukten im Deutschen. *Zeitschrift für Sprachwissenschaft* 1: 76 – 112.

Khmaladze E V 1987 *The statistical analysis of large number of rare events*. Technical Report MS-R8804, Amsterdam, Center for Mathematics and Computer Science.

Kühnhold I, Putzer O, Wellmann H (1978) *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache. Das Adjektiv.* Düsseldorf, Pädagogischer Verlag Schwann.

Lüdeling, A, Fitschen A 2002 An integrated lexicon for the analysis of complex words. In *Proceedings of EURALEX 2002*, Copenhagen, pp 145 –152

Leser M 1990 *Das Problem der Zusammenbildungen. Eine lexikalistische Studie*. Trier, Wissenschaftlicher Verlag.

Plag, I 1999 *Morphological Productivity. Structural Constraints in English Derivation*. Berlin, Mouton de Gruyter.

Schmid T, Lüdeling A, Säuberlich B, Heid U, Möbius B 2001 DeKo: Ein System zur Analyse komplexer Wörter. In *GLDV – Jahrestagung,* Giessen, pp. 49-57.

Williams, E 1981 On the notions 'lexically related' and 'head of a word'. *Linguistic Inquiry* 12: 245 – 274.