The Use of Corpora in Word Formation Research

Pius ten Hacken et Renáta Panocová
The study of word formation is concerned with the rule-based extension of the lexicon. The position of corpora in this research depends on the assumptions made about the nature of the lexicon and the position of word formation in the general organization of language. The use of corpora is particularly relevant in the study of productivity, which is arguably a (if not the) central aspect of word formation research.

1. Productivity

Productivity is the property of rules according to which they can produce new items. In the domain of word formation, the concept of productivity has been interpreted in different ways, two of which are particularly interesting. The first sees productivity as an absolute notion. A rule or process is productive if it can be used to produce new expressions. This notion corresponds to Corbin’s (1987) *disponibilité* (‘availability’). The second interpretation sees productivity as a matter of degree. The productivity of a rule or process can be placed on a scale from rarely used to generally used. This corresponds to Corbin’s (1987) *rentabilité* (‘profitability’).

When introducing the terms of *disponibilité* and *rentabilité*, Corbin argued that the former is the underlying notion and the latter derived from it. As argued by ten Hacken (2012), *disponibilité* is more connected to the individual speakers’ competence, whereas *rentabilité* is more a result of social interaction in the speech community and linked to performance. As indicated by ten Hacken & Panocová (2011), these two perspectives complement each other and should be studied in combination.

The study of *rentabilité* encounters a number of problems in relation to determining the correct basis for measuring it. This can be illustrated with reference to the examples in (1).
The rule illustrated in (1a) turns verbs into nouns, whereas the one illustrated in (1b) turns nouns into verbs. How can we compare the rentabilité of these two processes? The most straightforward approach seems to be to determine how many items were formed by the two rules and comparing this to the number of possible items the rules could have applied to.

One set of problems relates to the basis for comparison. It is well known that there are more nouns than verbs in a language. How does this affect the way we measure rentabilité for the processes in (1)? Aronoff (1976) considers taking as a basis for determining productivity the set of items that may serve as input, but he notices immediately that this is problematic. In fact, it is not only difficult to measure the number of nouns and verbs in a language, it is also open to argument whether these are in fact the right classes. We could also imagine smaller, more specialized classes that are defined in part semantically, e.g. verbs expressing actions or nouns expressing tools.

Another set of problems in measuring rentabilité concerns determining the number of output items. Here there are two main methods available. One is to consult a dictionary. In electronic dictionaries that offer wild card searches, it is straightforward to retrieve all words in -ation and in -ize. We then eliminate the irrelevant cases (e.g. nation, prize) and end up with a full list. The problem with this method is that the result depends on the decisions taken by the lexicographers compiling the dictionary. For each entry in the dictionary, a lexicographer has decided to include it. We can assume that the decision is well-informed, but the criteria are not transparent and may well be very different from the ones relevant in word formation research.

The alternative method is using a corpus. This method is more "objective" in the sense that no conscious decisions are taken about the inclusion of individual cases. They emerge from the compilation of the corpus. Another difference is that the dictionary includes types, whereas the corpus includes tokens. In this sense, the corpus gives more information, because we can easily reduce the output from a corpus to a list of types, but we cannot find frequency information in a dictionary.

2. Corpus-based measuring of productivity

The most influential proposal to use corpora in measuring productivity (in the sense of rentabilité) is the one by Baayen (1992). The formula in (2) is the one he proposes as an approximation of the "growth rate of [the vocabulary] \( V \)", which he takes as a "quantitative formalization of the linguistic notion of morphological productivity" (1992: 115).

\[
P = \frac{n_1}{N}
\]

In (2), the productivity measure \( P \) is the quotient of two quantities that are calculated for a particular corpus. Baayen (1992: 111) defines \( N \) as the number of tokens for the rule under consideration, e.g. all occurrences of deverbal nouns in -ation. He uses the notation \( n \) to refer to all types \( n \) that have a frequency \( r \) in the corpus (1992: 112). Hence, (2) determines the productivity of the word formation rule illustrated in (1a) as the number
of words resulting from this rule with a single occurrence in the corpus, divided by the overall number of occurrences of outputs of this rule.

The intuitive motivation for using hapaxes, i.e. \( n_1 \), in determining productivity is that the strength of a rule is demonstrated not so much by many lexicalized items, but by the formation of new words. However, the way (2) is formulated makes it highly sensitive to the existence of frequent items. It is counterintuitive to make the frequency of organization a factor in determining the productivity of the rule illustrated in (1a). Yet, each additional occurrence of organization increases \( N \) and therefore reduces \( P \).

It is interesting to note that Baayen (1992: 117-19) only discusses a completely different type of problem with (2). He observes that for all word formation processes, the limit for \( N \) is 0. Increasing corpus size will at first make the distinctions between more and less productive rules bigger, but as \( N \) is the denominator in (2), increasing it to infinity ultimately makes \( P \) zero.

The question why this problem is so serious casts some light on Baayen’s underlying conception of language and of the position of a corpus in relation to language. A corpus is a collection of texts and spoken language collected from various sources. Often it is used as a sample of the language it is taken from. Baayen’s (1992: 114) statement in (3) indicates the importance of considering a sample of unlimited size.

(3) “The population number of types \( S \) […] is estimated by considering the limit of \( V^{(N)} \) for \( N \rightarrow \infty \).”

Baayen (1992: 111) introduces \( V \) as the number of types. If we consider a particular word formation process, e.g. the one in (1a), \( V \) depends on \( N \). \( V^{(N)} \) expresses this dependency as a function, the number of types \( V \) in a set of tokens \( N \). Clearly, \( N \) and \( V^{(N)} \) are properties of a corpus. However, \( S \), as referred to in (3), is a property of the language the corpus is taken from. This means that Baayen assumes that English (and other named languages) are empirical entities. As we will show next, this assumption is problematic.

### 3. Competence, corpora and productivity

Chomsky’s (1965) distinction between competence and performance has often been construed as an argument against the use of corpora as a basis for linguistic research. However, Chomsky (1965: 18) immediately acknowledges the importance of performance in the study in (4).

(4) “Clearly, the actual data of linguistic performance will provide much evidence for determining the correctness of hypotheses about underlying linguistic structure, along with introspective reports.”

The focus of research in (4) is quite distinct from that implied in (3). In (4), the “underlying linguistic structure” is the ultimate target. Moreover, “the actual data of linguistic performance” are not equal to a corpus. Thus, Chomsky (1980: 198-99) observes the problem in (5).

(5) “A corpus may contain examples of deviant or ungrammatical sentences, and any rational linguist will recognize the problem and try to assign to observed examples their proper status.”

For Chomsky, a corpus can only be a source of inspiration, not an authority. The linguist’s competence and expertise are needed to interpret the data found in the corpus. The competence is the source of judgements and the expertise is the basis for interpreting the
corpus data. This way of embedding corpus use in linguistic research is quite far removed from Baayen's, who considers the corpus as a direct reflection of the language.

18 As shown by ten Hacken (2007), the original reason for introducing the distinction between competence and performance was rather to create a basis for a mentalist theory of language than to exclude corpora as a source of data. A side effect is that the status of named languages becomes a problem, as Chomsky (1980: 217) recognized. English is neither a competence nor a performance. Contrary to competence, English is not individual. Contrary to performance, it has a long and complex history in the course of which it spread to several continents. Its empirical status is problematic, as shown, for instance, by the impossibility to arrive at a clear concept of (a) language on a linguistic basis. It is political rather than purely linguistic reasons that determine that West Flemish is a dialect of Dutch, whereas Slovak and Czech are different languages (on the latter, cf. Jarošová & Benko, 2012). In both cases, it is clear that speakers of one and the other can understand each other, though with quite some difficulty, and there is no linguistic measure to quantify this difficulty. It is on the basis of such considerations that Uriagereka (1998: 27) claims that “English doesn’t really exist.” There is no empirical entity corresponding to the name English.

19 If English is not an empirical entity, a corpus cannot be representative of that entity. This has important implications for the use of corpora in the study of word formation. To the extent that Corbin’s rentabilité is to be measured for English rather than for a speaker’s competence, it does not have a basis in empirical reality. Corbin’s disponibilité is realized in an individual speaker’s competence, but it is not easily measured with the help of a corpus. Therefore, the role of corpora is limited to that of a source of inspiration. Any quantitative results obtained will have to be checked and interpreted on the basis of linguistic intuitions and expertise.

4. An onomasiological approach to word formation

20 The opposition between onomasiological and semasiological approaches is framed in the model of the Saussurean sign. A sign consists of a signifiant and a signifié. In an onomasiological approach, the signifié is the starting point and the corresponding name the target. In a semasiological approach the starting point and target are reversed. Generative approaches such as Chomsky’s, which assume that syntactic rules are central, should arguably not be classified in these terms.

21 In onomasiological approaches, the central question is how concepts are named. Here we will take as a basis the model presented by Štekauer (2005). A crucial distinction in this model is the one between linguistic and extra-linguistic components of the naming process. In terms of the Saussurean model of the sign, the extra-linguistic part of the process results in a signifié and the linguistic part assigns a signifiant to it.

22 In Štekauer’s model, the starting point of the naming process is the recognition of something that needs a name. This recognition emerges from the interaction of a speech community with extra-linguistic reality and results in a conceptual representation. The speech community consists of the speakers of a particular language. It is important to note that the object of the conceptual representation is underdetermined in many respects. It need not be concrete, it need not be genuinely new, the only real requirement is that the speech community (or a large enough part of it) agrees that it deserves a new
name. When cool emerged in the sense of ‘fashionably attractive’, speakers of English (or at least a relevant section of them) were presumably not satisfied with the existing words for similar concepts and came up with this new name for what they judged a new concept.

In this context, the references to a speech community and a language need some further discussion. A language such as English is considered to be based in the speech community and realized in the individual speakers. As shown by ten Hacken & Panocová (2011), this view is entirely consistent with the Saussurean tradition. The role of the speech community involves the identification of the need for a new name, the conceptualization leading to the signifié to be named, and the choice of a signifiant to name the concept. There is no democratic vote, but rather an at most partially conscious process based on the interaction between speakers. Therefore, there is no need to define a particular speech community and determine its boundaries. It emerges as a side effect of actual and potential communication among individual speakers. Likewise, there is no need to refer to a named language such as English as an entity. It can remain a concept with fuzzy boundaries.

The internal structure of the linguistic component of Štekauer’s model involves a lexical component that communicates with a syntactic component and a word formation component. Although Štekauer (2005) concentrates on the working and the internal structure of the word formation component, naming needs can also be fulfilled by the extension of the sense of existing words or by borrowing. In English, the former is illustrated by cool as mentioned above and the latter is currently mainly frequent in culinary and gastronomic terms (e.g. kebab).

Word formation in Štekauer’s model is a rule-governed system that provides the speech community with a mechanism for generating new words on the basis of existing words. The word formation process is divided into four levels of progressive specification spanning the gap between a semantic representation and a phonological representation. The semantic representation constitutes a linguistic representation of the extra-linguistic concept. The phonological representation serves as a basis for pronouncing (or writing) the name.

In between the semantic and the phonological levels, there are the onomasiological and the onomatological levels. At the onomasiological level, an onomasiological type (OT) is selected. Štekauer (1998) distinguishes five OTs that differ in the selection they make of properties in the semantic representation of the concept that are expressed. Thus, despite their similarity in meaning, piano player and pianist belong to different OTs, because the former but not the latter expresses the activity explicitly. The onomatological level assigns morphemes to the elements of the onomasiological representation.

The rules governing the representations at each level are implemented in individual members of the speech community. As these rules are largely shared among speakers, one speaker can understand the name another speaker comes up with. Acceptance of such a name by the speech community is an emergent process.
5. An onomasiological approach to productivity

Within the framework of the onomasiological approach presented in the previous section, Štekauer (1998) presents an analysis of productivity that includes some interesting, original aspects. He presents (6) as his formula for productivity (1998: 117).

\[ P = \frac{\sum \text{WFT} \times \text{NU}}{\sum \text{analysed WFTC} \times \text{NU}} \times 100 \]

In (6), \( P \) stands for ‘Productivity’ and the multiplication by 100 at the end is a cosmetic step resulting in percentages. \( \text{NU} \) stands for ‘Naming Unit’, i.e. a word or multi-word expression used for naming a concept. The numerator and the denominator each take the sum (i.e. the total number) of naming units in a particular domain. The \( \text{WFT} \) in the numerator is the Word Formation Type. A \( \text{WFT} \) is more specific than an \( \text{OT} \), in that it gives the conceptual categories expressed by the individual components. In the case of \( \text{piano player} \), the \( \text{OT} \) only specifies that there are three components, a base (realized as \(-er\)) and a mark consisting of a determined part (\( \text{play} \)) and a determining part (\( \text{piano} \)). The \( \text{WFT} \) specifies the conceptual categories that are expressed by each of the components. In this case, it is \{Object \ Action – Agent\}, i.e. \(-er\) expresses an Agent, \( \text{play} \) an Action, and \( \text{piano} \) the Object this Action is directed to. The \( \text{OT} \) groups together a number of \( \text{WFTs} \). Štekauer (1998) calculates productivity both for \( \text{OTs} \) and for \( \text{WFTs} \). In principle, it would also be possible to take a particular affix, such as \(-ation\) in (1a) as a \( \text{WFT} \), but this is not the common way of applying (6). Štekauer (2005: 221-3) distinguishes four levels at which productivity can be calculated, including \( \text{OT}, \text{WFT}, \) and word formation rules.

The \( \text{WFTC} \) in the denominator in (6) is the Word Formation Type Cluster. A \( \text{WFTC} \) is a general class of new concepts, determined in broadly semantic terms. Štekauer (1998) presents two case studies with instruments and agents as the \( \text{WFTCs} \). The productivity of a \( \text{WFT} \) is always calculated with respect to the \( \text{WFTC} \). When we consider agents, \( \text{piano player} \) and \( \text{pianist} \) will count as representatives of different \( \text{OTs} \) or \( \text{WFTs} \) rather than as a synthetic compound and a derivation. The total productivity of a \( \text{WFTC} \) is always 100%, because when the speech community needs a name for, for instance, an agent, it will always come up with one. Štekauer’s (1998) case studies only focus on word formation, so that the sum of the productivity of the five \( \text{OTs} \) is always 100%. If borrowing and sense extension are also taken into account, the 100% is only achieved for the three processes together.

Before comparing the measure of productivity in (6) to Baayen’s measure in (2), we have to contrast it with the notion of frequency, for which Štekauer (1998: 118) gives the formula in (7).

\[ F = \frac{\sum \text{NU in the text belonging to a WFT}}{\sum \text{NU in the text belonging to a WFTC}} \]

In (7), the same concepts are used as in (6), except that productivity in (6) is a property of the \( \text{langue} \), whereas frequency in (7) is a property of \( \text{parole} \). What is referred to as “the text” in (7) is a corpus. This implies that (6) is applied not to a corpus, but to the language system. In his case studies illustrating the application of (6), Štekauer (1998) uses a dictionary rather than a corpus as a basis for retrieving the basic data. He explains that
the frequency in (7) gives information about the type-token ratio (1998: 118). Therefore, NUs in (6) are types whereas NUs in (7) are tokens.

Both for productivity and for frequency, we can ask to what extent the measures obtained are representative of the actual language. As productivity is based on langue, dictionaries are a good source of information. They can be thought of as fairly complete and at least representative of the vocabulary of the langue of many speakers. Although ten Hacken (2009) argues that dictionaries cannot be a description of the language, the selection of entries is based on a well-considered methodology.

For frequency as defined in (7), the problem of selecting representative texts is much larger. The selection of texts directly influences the value of $F$ in (7). Štekauer (1998: 119) introduces the notion of ideal frequency, where the corpus size is increased to $F_0$. Obviously, this is not a realistic aim. We can at most aim for a very large corpus. However, just increasing the corpus to what we can obtain is not the best possible approximation of this aim. In the compilation of large national corpora, much attention has been paid to representativity. Thus, Górski & Łaziński (2012) describe how in the Polish National Corpus the distribution of text types was based on a large survey of reading behaviour by the Polish National Library. Even so, many questions remain and the issue of the proportion of a corpus to be assigned to spoken language cannot be addressed in this way. Arguably, the problem is due to the nature of parole, which is not an entity that can be collected comprehensively.

Rather than the bulk use of large corpora, the use of corpora in the onomasiological approach has much more attention for details of the individual texts. This includes qualitative considerations of a type very similar to the ones used in lexicography, such as the context of use, the register, the writer’s or speaker’s background, and the time in history. Of course these considerations are combined with quantitative data, as is also done in lexicography. In our example of piano player and pianist, a question is when which name is used for a person playing the piano. This question can then be embedded in the larger question of whether there are any regularities as to which conditions determine the preference for a particular OT (or morphological type, etc.). Crude use of large numbers will never answer such questions.

6. Comparison and conclusion

We started in section 1 by disambiguating productivity, distinguishing between what Corbin (1987) calls disponibilité and rentabilité. The Chomskyan approach outlined in section 3 favours the study of disponibilité, a position also taken by Corbin (1987: 177). However, the strict implementation of this approach makes the study of rentabilité all but impossible. This is unfortunate, because it constitutes an important aspect of the speech community’s influence on the language. As argued by ten Hacken & Panocová (2011), the perspectives of language as an individual speaker’s competence and as a property of a speech community complement each other. Therefore, it is worth looking for a framework that makes the study of rentabilité possible.

In sections 2 and 5, we presented two formal definitions of productivity, (2) and (6), formulated from different theoretical backgrounds. We repeat them here to facilitate comparison.

(2)
Baayen’s (1992) definition in (2) relates the number of types with one occurrence in a corpus to the number of tokens representing the word formation rule in that corpus. It is therefore intrinsically corpus-dependent. Štekauer’s (1998) definition in (6) does not refer to tokens at all. The naming units $NU$ in (6) are types and it is up to the linguist how to find them. In his case studies, Štekauer (1998) uses dictionaries as the source of a list of naming units, but he presents this as an admissible shortcut (1998: 93), not as the necessary method. Dictionaries are generally based on corpus analysis. There is no theoretical reason why one could not use a corpus directly as a source.

An important difference between (2) and (6) is that (2) uses token frequency whereas (6) does not. Štekauer (1998) assigns token frequency to a different measure, (7), which is a property of language use rather than of the system. Following Saussure, he considers the use as dependent on the system. Baayen (1992), by contrast, takes language use as primary. The system emerges from language use. This should not be taken in the Chomskyan sense, in which the language faculty produces a system on the basis of linguistic input (e.g. Chomsky 1999), but as an unguided, purely data-oriented emergence of the system. Computational modelling is our best hope to reconstruct this procedure. The system is inherent in the data, independently of any mental aspect of language. This contrasts strongly with the views both of Chomsky and of Štekauer, who both take language to be first of all in the speakers’ mind.

Another interesting difference is the unit for which productivity is measured. In (6), this is the word formation type $WFT$ in comparison to the corresponding $WFT$ cluster. This means that the productivity of a $WFT$ is measured as its strength in the competition with other $WFT$s in the domain (identified by the cluster) in which it operates. The productivity is the percentage of this domain that is covered by the $WFT$ (or $OT$) in question. In the case of $piano$ player and pianist, the central question is which proportion of the domain of agents is covered by the $WFT$s, $OT$s or other levels exemplified by these words. As the application of (6) is to a set of words (types) considered as items of the language, it does not take into account how long these words have been in existence. As long as the words are still in use, they are fully considered.

In (2), most of these assumptions are reversed. There is no direct measure of competition involved. The value of $P$ in (2) is calculated for a word formation process, typically an affix, in isolation. Instead of comparing the productivity of the affixes in piano player and pianist, we calculate the value of $P$ for each of them separately. Neither the size of the corpus, nor the frequency of any competing processes play a role in its calculation. By contrast, (2) is highly sensitive to the diachronic aspect of new words becoming established. In a sense, as soon as any sign of establishment is discovered, the word is no longer seen as an indication of productivity. When a second occurrence in the corpus is found, the word disappears from $n_i$ and only counts towards $N$, so that it diminishes productivity. Neither piano player nor pianist are likely to be a hapax in a large corpus, so they only play a negative role in the productivity of -er and -ist.
In sum, we have argued that Chomsky’s approach to language does not leave sufficient room for the study of productivity in the sense of *rentabilité*. The reasons have nothing to do with an aversion to corpora, but with the abstraction from the social aspects of language as a property of the speech community. Baayen presents a measure of productivity that is dependent on corpus use, but also on a notion of language in which the implementation of the language system in the individual speaker is entirely circumstantial. In comparison to these, Štekauer’s approach to measuring productivity has the benefit of taking into account both the language system of the individual speaker and the language of the speech community. Although using dictionaries as the source of word lists, as Štekauer does, can be defended as a shortcut, the identification of naming units in a corpus rather than in a dictionary results in a more principled method. It makes the collection of naming units independent of decisions by lexicographers, who make their decisions on inclusion on the basis of quite different considerations than the prospective use in the calculation of productivity.

**BIBLIOGRAPHIE**


Dans le domaine de la formation des mots, la question où l'utilisation d'un corpus est la plus évidente est celle de la productivité. Nous comparons trois approches de l'étude de la productivité, avec une attention particulière à ce que Corbin (1987) appelle la « rentabilité ». Baayen (1992) calcule la productivité en ce sens sur la base du nombre de hapaxes dans un corpus. En linguistique chomskyenne, l'accent est plutôt mis sur la « disponibilité », parce qu'elle relève de la compétence. Štekauer (1998) établit une nette distinction entre la productivité et la fréquence et définit celle-là de telle façon qu'on peut se passer de l'utilisation d'un corpus. En comparant les approches, nous soutenons que la rentabilité mérite d'être étudiée, mais que l'approche de Baayen repose sur une notion du langage où l'utilisation de la langue est plus essentielle que son système. L'approche de Štekauer offre une base prometteuse pour concilier la perspective individuelle et celle de la communauté linguistique. Une utilisation systématique de corpus permettrait de renforcer les conclusions obtenues sur la productivité.

The most straightforward area in word formation where corpora can be used is the study of productivity. We compare three approaches to the study of productivity, concentrating on what Corbin (1987) calls rentabilité (‘profitability’). Baayen (1992) calculates productivity in this sense on the basis of hapaxes in a corpus. In Chomskyan linguistics, the emphasis is rather on disponibilité (‘availability’), because it is a property of the competence. Štekauer (1998) sharply distinguishes productivity and frequency and defines the former so that corpus use is not essential. In comparing the approaches, we argue that rentabilité is worth studying, but that Baayen’s approach depends on a notion of language where language use is taken to be prior to the language system. Štekauer’s approach offers a promising basis for reconciling individual and speech community perspectives on language. A systematic use of corpora would strengthen the results obtained on productivity.

Mots-clés : formation des mots, productivité, rentabilité, disponibilité, corpus
Keywords : word formation, productivity, profitability, availability

INDEX

AUTEURS

 PIUS TEN HACKEN
Swansea University, Royaume-Uni, p.ten-hacken@swansea.ac.uk

RENÁTA PANOCOVÁ
P.J. Šafárik University, Košice, Slovaquie, renata.panocova@upjs.sk