Productivity of Anglicism Bases in Hyphenated German Compounds

Steven Coats[1], **Adrien Barbaresi**[2]
[1]University of Oulu, Finland, [2]Berlin-Brandenburg Academy of Sciences, Germany
steven.coats (at) oulu.fi, barbaresi (at) bbaw.de

Abstract

This study investigates hyphenated German compounds that contain English constituents, a part of the German lexicon that exhibits great diversity. We determine the frequency of a set of English noun bases in different constituent positions in three corpora of online language, and quantify the productivity of bases in these compounds using a metric based on Shannon entropy, a measure of information content. Compound entropy values for English bases reflect their diversity of use, and are unequally distributed for left-hand, internal, and right-hand constituents. The semantics of the base types with the highest frequencies and entropy values reflect contemporary cultural and technological concerns. Differences in entropy according to constituent position may be an indication of word class conversion of anglicisms in German.

Keywords: Anglicisms, German, Compounds, Corpus linguistics, Productivity, Entropy

1. Introduction

English is the most important source language for new words in German at present, accounting for the majority of lexical borrowings (Onysko, 2007). The prevalence of English in numerous public, job-related and international contexts in German society ensures a broad diffusion of language knowledge, so that a bottom-up diffusion of English-based neologisms and compounds is possible, in a sharp contrast to previous patterns of foreign lexical assimilations, for example of Latin or French words (Krome & Roll, 2017). Anglicisms exhibit a great deal of variation in written German: They can be integrated with minimal modifications (e.g. *Discounter*), assimilated in morphology and orthography and adapted to German inflectional paradigms (e.g. *geliked* or *gelikt*, 'liked on social media') (Burmasova, 2010; Coats, 2018), or serve as the basis for calques (loan translations) and syntactic constructions. Recent research has cataloged their overall diversity (Eisenberg, 2011, 2013), investigated the extent to which they overlap in meaning with existing German lexemes (Onysko & Winter-Froemel, 2011; Winter-Froemel, Onysko, & Calude 2011), or documented their assimilation to German orthography and inflectional morphology (Coats, 2018). English elements can also be joined to German constituents with hyphens to create compounds (e.g. *Urlaubs-Feeling* 'holiday feeling', *Service-Zentrum* 'service center'), a productive process which is the focus of the present research.

The processes by which composite words in German can be formed by combining indigenous and exogenous lexical material have been described in the literature (Fleischer & Barz, 2012), and the importance of relative frequencies of bases for the recognition and processing of compounds has been shown in experiments (e.g. Lüdeling & de Jong, 2002; Baayen, Wurm, & Aycock, 2007), but the productivity of English bases in hyphenated compounds has not yet been a focus of German lexicography. Most research into productivity in German has focused on the productivity of affixes, rather than word bases (Lüdeling & Evert, 2005; Lüdeling, Evert & Heid, 2000). Recent developments have highlighted the importance of corpus-based methods as well as the need to extract relevant information and adequately describe changes or findings based on the explanatory power of statistical indicators (Hein & Engelberg, 2017). It is indeed still necessary to find a suitable methodology to study the dynamics of anglicisms in German, all the more since empirical frequency-based investigations have not always been the main research focus (Burmasova, 2010). In this study, we consider the productivity of English base constituents in hyphenated German compounds. By taking into account recent web and computer-mediated communication (CMC) corpora, we hope to capture phenomena unseen in standard written German, as these corpora consist of genres which are expected to differ from commonly accepted rules. Indeed, we do not stick to the concept of rules but rather try to derive norms from empirical data (Habert & Zweigenbaum, 2002) by way of statistical indicators which are related to information theory and as such yield a particular view on language constituency and productivity.

Building upon quantitative approaches to the measure of morphological productivity developed by Baayen (1994a, 1994b, 2001) and others (Hay & Baayen, 2003; Moscoso del Prado Martín, Kostić, & Baayen, 2004), we utilize Shannon entropy (Shannon, 1948) to measure the productivity of English constituents in different internal word positions in large corpora from the web and from Twitter. In light of findings from response latency experiments, this may be evidence that English constituents increasingly take part in productive word formation processes in German. In addition, the semantic values of the most productive English bases may shed light on broader developments in the German lexicon and in German-speaking society as a whole. The study addresses the following questions: 1) Which constituent base anglicism types are most frequent in hyphenated German compounds, and 2) What can morphological diversity measures such as Shannon entropy tell us about the dynamics of anglicisms borrowed into German compounds?

In the next section, a review of previous research in morphological productivity is provided, followed by a brief overview of German usage norms for hyphenated compounds. In Section 3, the data and methods used to calculate

Shannon entropy from the corpora are presented and German compounds briefly reviewed. Section 4 presents the results, and Section 5 closes the paper with a summary and outlook for future research.

2. Previous work

2.1. Productivity measures

Baayen (1993, 1994a, 1994b, 2001) proposed several measures of morphological productivity, including the category-conditioned degree of productivity: the ratio of *hapax legomena* (words that occur once in a text or a corpus) for an affix to the size of its morphological category, which represents that the probability a new word encountered in a text or corpus will be a type that has not yet been encountered, given that it belongs to a particular morphological class. For example, for the German deadjectival nominal suffix *-keit* (e.g. *Sparsamkeit* 'thriftiness'), the category-conditioned degree of productivity is the ratio of the sum of all *hapax* words ending in *-keit* to the sum of all words ending in *-keit*. Because *-keit* is more productive than suffixes such as *-nis* or *-tum*, it will have a higher value when using this measure.

For word bases in compounds, frequencies can be assessed by the morphological family size (the number of distinct types in which a base appears) and the cumulative family frequency (the sum of token frequencies for all types in which a base appears). Morphological family size is negatively correlated to reaction times in lexical recognition experiments (Baayen, Wurm, & Aycock, 2007). For example, because a German base like *Schrift* 'writing' may occur in a large number of compound words (e.g. *Schreibschrift* 'handwriting', *Schriftsteller*, 'writer', *Unterschrift* 'signature', etc.) with relatively high frequencies, compounds containing the base are recognized more quickly than are compounds that contain constituents with lower frequencies or smaller family sizes, such as *Schund* 'rubbish', which is a constituent in a smaller number of words (e.g. *Schundliteratur* 'trashy writing').

Like morphological family size, cumulative family frequency has been shown to correlate negatively with reaction times in lexical recognition experiments (Baayen & Hay, 2002; Baayen, Lieber, & Schreuder, 1997; De Jong, Schreuder, & Baayen, 2000; Schreuder & Baayen, 1997). Hay (2001) found that for English compounds, frequent words are processed more quickly, and thus likely to be stored in the mental lexicon as opaque single units of meaning, whereas infrequent compounds may be stored as decomposable items. For German verbs, Lüdeling and De Jong (2002) found a negative relationship between morphological family size and response latencies in an experimental task.

Moscoso del Prado Martín, Kostić, and Baayen (2004) utilized a metric based on Shannon entropy to calculate an "information residual" for a word, or the difference between the logarithm of a word's frequency to the Shannon entropy for all inflected forms of the word. They found that in regression models of word response latencies, the word information residual provides a better fit than morphological family size or cumulative family frequency, meaning that from a statistical standpoint this additional indicator yields more fine-grained information and is thus suitable to draw conclusions on lexical use.

2.2. Hyphenation in German compounds

In standard German orthography, constituent elements in composite words are typically linked without a hyphen. Hyphens can be optionally used in composite words in order to emphasize particular constituents or to enhance the legibility of long composite words with multiple constituents (Duden, 2006, p. 39; Fleischer & Barz, 2012, p. 193). Hyphenation is recommended if a composite form contains a constituent that is an abbreviation or initialism (*Fussball-WM* 'football/soccer world cup'), and is preferred if the first constituent element of a compound is a proper noun, especially a personal name (Fleischer & Barz, 2012, p. 193; e.g. *Merkel-Regierung* 'Merkel government'). Hyphenation is also used in longer composite phrasal word forms (*Pro-Kopf-Verbrauch* 'per capita use') (Duden, 2006, p. 41; Fleischer & Barz, 2012, p. 175). Fleischer and Barz note that composite word formations in German can incorporate foreign constituents as first elements, internal elements, or final elements, "without restrictions" (2012, p. 111).

3. Data and methods

A list of potential English constituents was created by combining the 3,262 most common nouns in the British National Corpus (Kilgarriff, 1997) with the 10,000 most common nouns in the 9.6b-token ENCOW16ax corpus, a corpus of English texts from the web (Schäfer, 2015; Schäfer & Bildhauer, 2012), then converting to lower case and removing types containing punctuation or shorter than 4 characters. The list thus combines attested data from a stratified reference corpus and more current utterances from a large web sample. The frequencies of these 8,313 unique types as left-hand, central, or right-hand constituents in hyphenated German words were determined in three corpora of online German: A German Twitter corpus of 534m tokens (Coats, 2018), the DECOW16bx corpus, a German web corpus of 11b tokens (Schäfer, 2015; Schäfer & Bildhauer, 2012), and a corpus of German WordPress blogs with 2.1b tokens (Barbaresi, 2016). In order to account for non-standard capitalization (common on Twitter and in other informal online genres), all words were converted to lowercase. A regular expression was used to additionally target plural and genitive forms while taking potentially unknown forms into account.

For each of the 8,313 English potential base types, we used token counts for individual compounds and the cumulative family frequency for the base (i.e. the summed frequencies of all compound types containing the base) to calculate an entropy measure. The compound Shannon entropy of a base can be calculated according to the formula

$$H(B) = -\sum_{i=1}^{n} \frac{F(x_i)}{F(B)} \cdot log_2 \frac{F(x_i)}{F(B)} \qquad (1)$$

where B represents an anglicism base, $F(x_i)$ the frequency of a particular compound type containing B, n the morphological family size for the base, and $F(B)$ the cumulative family frequency for the base. The value can range from zero to $\log_2 n$. As an example, the entropy for the English base *payment* can be calculated in a corpus of 8 tokens in which *Payment-Taste* 'payment button' occurs 5 times, *Crypto-Payment-App* 'crypto payment app' twice, and *Online-Payment* 'online payment' once. In this example, the English base 'payment' occurs as a left-hand constituent, as an internal constituent, and as a right-hand constituent. Entropy can be calculated by constituent position, or a total entropy score can be calculated that takes into account all possible configurations: In this example, the left-hand, internal, and right-hand compound entropies would be zero (because only one type occurs at each word-internal position), while the total entropy would be 1.30. In general, low entropy values indicate skewed frequency distributions, while high values indicate more uniform distributions.

4. Results and discussion

The corpora feature a large number of hyphenated compounds containing anglicisms: the Twitter corpus 619,338 unique types, the most frequent of which are *youtube-video*, *live-tracking*, and *start-up*; the DECOW16bx corpus 6,567,984 types (*online-shop*, *html-code*, and *bb-code* are the most frequent) and the WordPress corpus 808,648 distinct types (*us-dollar*, *online-shop*, and *after-sales-service* are the most frequent).

Tables 1, 2, and 3 show the base types with the highest cumulative family frequencies for the three corpora, their frequencies, left-hand, internal, right-hand and total entropy calculations for the types, and the three most frequent types containing the base elements.

Many of the most frequent bases in the three corpora denote entities related to technology or the internet (*video*, *twitter*, *facebook*, *youtube*, *blog*, *internet*, *code*, *software*). Left-hand, internal, right-hand, and total entropy values (indicated by H_L, H_I, H_R, and H_T in the tables) provide an overall indication of the diversity of the frequency distributions for compounds containing bases in that constituent position. Entropy values for the bases in different hyphenated word positions vary from low (*dollar* in right-hand position) to high (*team*, *system* in right-hand position). Lower entropy values can indicate that a base has been lexicalized in a hyphenated word (*us-dollar*), resulting in far higher frequencies of that type compared to other types in the morphological family. For the Twitter and DECOW16bx corpora, internal entropy values are the lowest, right-hand entropies intermediate, and left-hand entropies the highest. For the WordPress blogs corpus, right-hand entropy values are highest.[1] Because the base types considered in the study are primarily English nouns, this value may also provide a preliminary indication of word class conversion of borrowings (Figure 1). This possibility, however, needs further exploration, for example by comparison with English-language compounds.

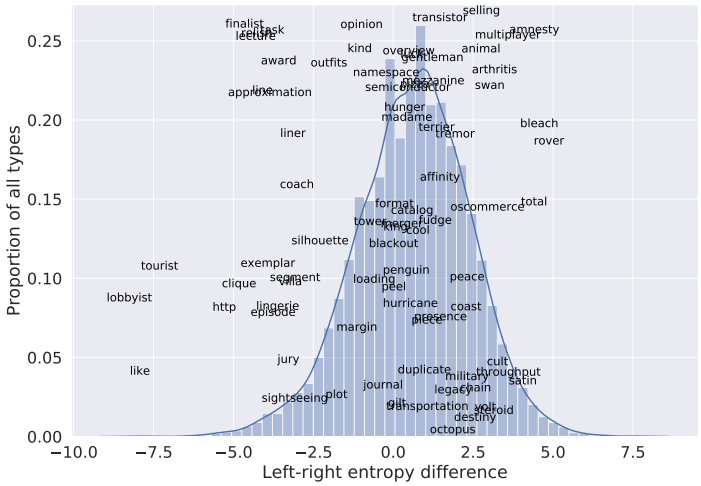

Figure 1: Left-right entropy difference distribution, DECOW16bx corpus

Figure 2 shows the total compound morphological entropy versus the morphological family size for the Twitter corpus, the DECOW16bx corpus, and the WordPress blogs corpus. In the plots, each point represents an English base, the red line represents the maximum entropy, and the magenta line represents a nonparametric locally-weighted regression.

For many of the 8,313 English-language bases analyzed in the study, entropy values are close to the maximum possible entropy curves in the three corpora, which can be seen as an indication of the relative lack of syntactic or semantic constraints on hyphenated word formation in German. Comparing the regression curves (magenta lines in the figures) to the maximum entropy curves (red lines) shows overall higher entropy values for the WordPress blogs corpus and for the Twitter corpus, compared to the DECOW16bx corpus. This is likely due to the more informal nature of written language on Twitter and in blogs, genres that exhibit relatively high rates of creative textual features such as non-standard orthography, expressive lengthening, or emoticon and emoji use (Argamon et al., 2007; Coats, 2016), and that therefore may also exhibit a greater diversity of hyphenated compound types. In contrast, many of the genres that comprise the DECOW16bx corpus, such as news articles, consist of relatively formal, conventionalized writing in which non-standard usages are uncommon.

The bases with the highest total entropy values in the Twitter corpus are the forms *team* and *chef*, followed by *media*, *twitter*, and *video*. For the web corpus, the types with the highest total entropy are *team* and *system*, followed by *forum*, *software*, and *service*. For the blogs cor-

[1] Because the anglicism constituents are nouns and in German compound nouns are almost exclusively right-headed, it is conceivable that right-hand entropies may be lower. The reason for the reversal of this pattern in the WordPress blogs corpus is unknown at this stage. Although different text processing procedures may play a role, compounding creativity may explain this behavior.

Table 1: Most frequent types, Twitter Corpus

Base	Freq.	H_L	H_I	H_R	H_T	Most freq. types
video	29412	8.72	7.98	6.91	8.67	'youtube-video', 7285, 'video-interview', 803, 'youtube-videos', 763
twitter	27459	8.99	8.96	7.58	9.25	'twitter-account', 2602, 'twitter-zufallsstory', 1510, 'twitter-app', 772
facebook	23901	8.40	8.14	5.98	8.56	'facebook-seite', 2740, 'facebook-gruppe', 751, 'facebook-fans', 606
team	18113	8.63	6.90	10.68	11.02	'orga-team', 553, 'dfb-team', 335, 'social-media-team', 305
chef	17818	7.81	4.90	9.63	9.88	'spd-chef', 573, 'fdp-chef', 452, 'ex-chef', 356
news	17553	6.98	7.83	5.91	7.02	'heise-news', 1925, 'it-news', 1812, 'fake-news', 1189
youtube	15435	4.71	6.44	5.61	4.87	'youtube-video', 7285, 'youtube-kanal', 1312, 'youtube-videos', 763
blog	15396	5.78	3.85	8.88	8.70	'blog-eintrag', 1212, 'blog-beitrag', 838, 'blog-artikel', 668
marketing	14121	8.64	8.42	4.80	8.02	'online-marketing', 2349, 'content-marketing', 1339, 'influencer-marketing', 414
media	14097	8.34	8.95	2.45	9.38	'social-media', 1051, 'social-media-team', 305, 'social-media-marketing', 286

Table 2: Most frequent types, DECOW16bx Corpus

Base	Freq.	H_L	H_I	H_R	H_T	Most freq. types
system	582231	9.91	9.68	11.97	12.22	'it-system', 16797, 'erp-system', 11115, 'content-management-system', 9506
internet	507073	8.90	10.62	6.33	9.15	'internet-seite', 36068, 'internet-adresse', 19172, 'internet-auftritt', 16666
forum	412839	8.96	10.27	10.24	10.63	'feuerwehr-forum', 17621, 'fan-forum', 10562, 'hifi-forum', 9812
code	398698	8.46	9.61	3.57	4.17	'html-code', 151710, 'bb-code', 132344, 'qr-code'
team	359205	9.31	10.36	12.64	12.85	'top-team', 5455, 'orga-team', 3963, 'support-team', 3307
shop	344979	6.51	8.66	5.22	6.07	'online-shop', 171998, 'internet-shop', 9308, 'shop-system', 5905
version	337625	7.07	7.67	9.85	9.89	'beta-version', 13907, 'pc-version', 10173, 'windows-version', 9784
software	325250	8.07	10.93	10.55	10.47	'software-entwicklung', 9605, 'software-update', 7840, 'software-lösung', 7746
service	321222	8.12	9.81	9.65	10.46	'full-service', 6824, 'it-service', 6728, 'service-center', 6591
video	254431	9.32	10.80	8.53	10.27	'youtube-video', 15269, 'hd-video', 5563, 'video-kritik', 5517

Table 3: Most frequent types, WordPress blogs corpus

Base	Freq.	H_L	H_I	H_R	H_T	Most freq. types
blog	47324	6.86	9.07	11.86	11.84	'blog-post', 743, 'satire-blog', 555, 'blog-event', 522
shop	39271	6.29	8.02	4.91	5.18	'online-shop', 13952, 'online-shops', 9930, 'web-shops', 2969
team	31366	7.16	7.97	12.00	12.10	'orga-team', 659, 'dream-team', 239, 'blog-team', 224
system	26528	6.81	7.30	11.53	11.63	'erp-system', 481, 'herz-kreislauf-system', 318, 'crm-system', 256
video	23244	7.31	9.49	9.83	10.26	'youtube-video', 1478, 'youtube-videos', 1286, 'video-interview', 478
chef	21689	5.41	6.89	9.84	9.94	'spd-chef', 875, 'fdp-chef', 627, 'ex-chef', 520
version	20518	4.50	4.51	9.74	9.76	'beta-version', 1125, 'online-version', 564, 'pc-version', 564
service	19351	6.19	8.54	6.01	6.98	'after-sales-service', 7669, 'euro-finanz-service', 339, 'shuttle-service', 300
film	19273	6.52	9.18	10.25	10.52	'film-reviews', 575, 'science-fiction-film', 433, 'bond-film', 328
dollar	17667	5.48	9.43	1.07	2.36	'us-dollar', 14271, 'us-dollars', 572, 'petro-dollar', 112

pus, the highest-entropy types are *team* and *blog*, then *system*, *film*, and *video*. Several of the most-attested types are words used in domains of interaction that have been particularly affected by the influx of English, and may represent examples of *Bedürfnislehnwörter* ('necessary borrowings', Carstensen, 1965), or lexical elements whose denotation is not well-represented by existing German lexemes. This is the case for the brand names among the most frequent and highest-entropy bases (*twitter*, *facebook*, and *youtube*), and may also be true for workplace-related elements (*team*, *chef*, *service*, and *marketing*). Types with high total compound morphological entropy values represent those English-language elements that have been borrowed into the German lexicon and are the most flexible in terms of their potential productivity. Types with low values, on the other hand, are typically used only in one or a few set formulations.

It should be noted that many of the types in the base wordlist may not be anglicisms, but rather Greek- or Romance-language-derived words common to most European languages (*system*, *service*, *version*, *video*, etc.), which may have undergone borrowing from the source language directly into German, or may also have been borrowed via English mediation. In addition, some types represent borrowings that have long been established in the German lexicon (e.g. *film*, *chef*), and thus may no longer be perceived as anglicisms or borrowings.

5. Summary and future outlook

Compounding via hyphenization is a productive word formation process in German, and we found many hyphenated compound types including English elements across three different CMC and web corpora: a Twitter corpus, a corpus of diverse web texts, and a blog corpus. We measured the tendency of 8,313 English nouns to appear as elements in hyphenated German compounds and documented a tremendous diversity of types. Many of the most frequent types overall (e.g. *online-shop*, *youtube-video*) are hyphenated compounds that have been borrowed into German in

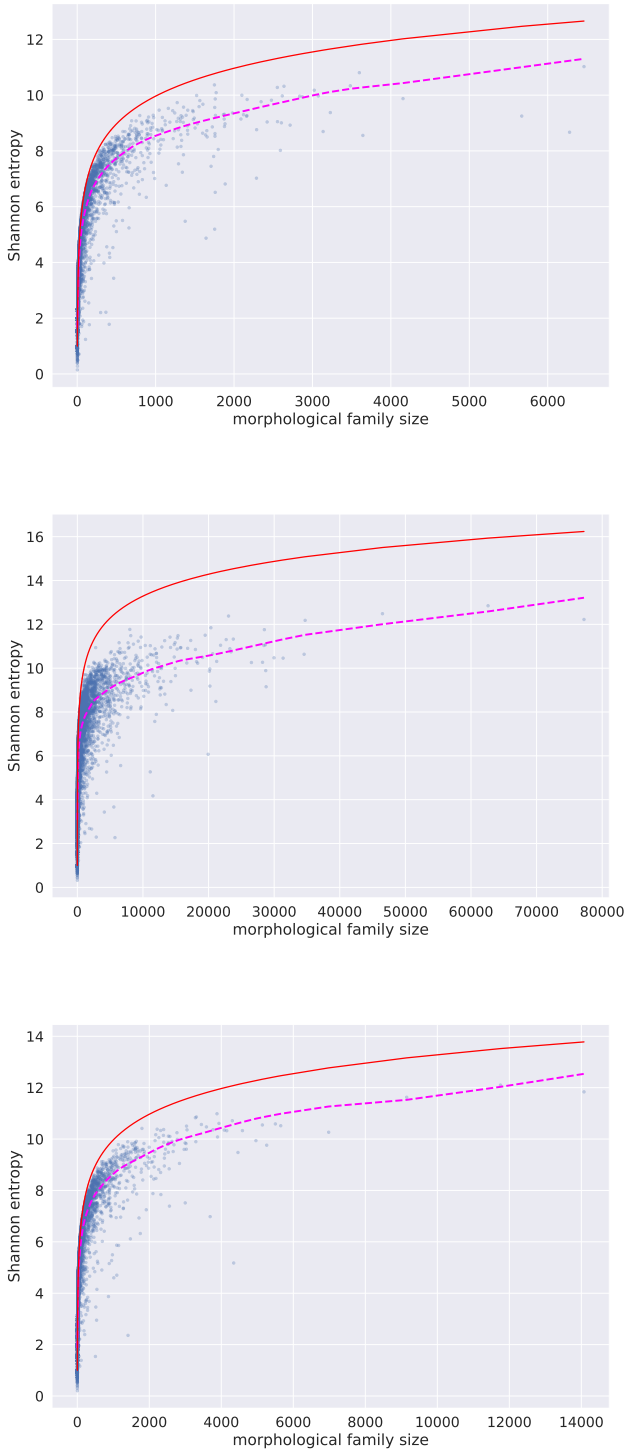

Figure 2: Shannon entropy vs. morphological family size, Twitter corpus, DECOW16bx corpus, and WordPress blogs corpus

order to denote new activities, technologies, or behaviors ("necessary borrowings"). The entropy analysis shows differences in entropy according to base constituent position within hyphenated compounds, and suggests that word formation via compound hyphenization of English bases is more productive on Twitter and on blogs, compared to other online genres. This can be interpreted as a consequence of the more spontaneous nature of text as used on blogs and on Twitter, which, as a paradigmatic form of CMC, has been suggested to have a status between speech and writing in terms of communicative typology and feature frequencies (Barbaresi & Würzner, 2014; Coats, 2016; Tagliamonte & Denis, 2008).

Future work with the data can be organized along three lines. First, a more thorough analysis of entropy according to constituent position within compounds and comparison with indigenous German lexical elements and English compound words may shed light on the dynamics of how borrowings are integrated into German morphological paradigms and whether conversion is taking place for some types. Second, many hyphenated compounds can also be written without hyphenation – a frequency analysis of the semantics of hyphenated and non-hyphenated compounds may be revealing. Finally, a frequency analysis of anglicism-containing hyphenated compounds of different structural types, according to the classification proposed by Fleischer and Barz (2012), may give further insight into the productivity of this large, chaotic, and fascinating word class.

6. References

Argamon, S. et al. (2007). "Mining the blogosphere: Age, gender, and the varieties of self-expression". In: *First Monday* 12.9. http://firstmonday.org/ojs/index.php/fm/article/view/2003.

Baayen, R. H. (1993). "On frequency, transparency, and productivity". In: *Yearbook of Morphology 1991*. Ed. by G. E. Booij and J. V. Marle. Dordrecht: Kluwer, pp. 109–149.

– (1994a). "Derivational productivity and text typology". In: *Journal of Quantitative Linguistics* 1, pp. 16–34.

– (1994b). "Productivity in language production". In: *Language and Cognitive Processes* 9, pp. 447–469.

– (2001). *Word frequency distributions*. Dordrecht: Kluwer.

– (2003). "Probabilistic approaches to morphology". In: *Probability Theory in Linguistics*. Ed. by R. Bod, J. B. Hay, and S. Jannedy. Cambridge, MA: MIT Press, pp. 229–287.

Baayen, R. H. and J. Hay (2002). *Affix productivity and base productivity*. Paper presented at ESSE 6, Strasbourg. https://pdfs.semanticscholar.org/db0d/e479b9686acd21\-e8ebc5147059c46ae0ed30.pdf.

Baayen, R. H., R. Lieber, and R. Schreuder (1997). "The morphological complexity of simplex nouns". In: *Linguistics* 35, pp. 861–877.

Baayen, R. H., L. H. Wurm, and J. Aycock (2007). "Lexical dynamics for low-frequency complex words: A regression task across tasks and modalities". In: *The Mental Lexicon* 2.3, pp. 419–463.

Barbaresi, A. (2016). "Efficient construction of metadata-enhanced web corpora". In: *Proceedings of the 10th Web as Corpus Workshop*, pp. 7–16.

Barbaresi, A. and K.-M. Würzner (2014). "For a fistful of blogs: Discovery and comparative benchmarking of republishable German content". In: *Proceedings of KONVENS 2014, NLP4CMC workshop*, pp. 2–10.

Burmasowa, S. (2010). *Empirische Untersuchung der Anglizismen im Deutschen am Material der Zeitung 'Die Welt'*. Bamberg: University of Bamberg Press.

Carstensen, B. (1965). *Englische Einflüsse auf die Deutsche Sprache nach 1945*. Heidelberg: Carl Winter Verlag.

Coats, S. (2016). "Grammatical feature frequencies of English on Twitter in Finland". In: *English in computer-mediated communication: Variation, representation, and change*. Ed. by L. Squires. Berlin: de Gruyter Mouton, pp. 179–210.

– (2018). "Variation of new German verbal Anglicisms in a social media corpus". In: *Proceedings of the 6th Conference on CMC and Social Media Corpora for the Humanities*, pp. 27–32.

Duden (2006). *Die Deutsche Rechtschreibung (24th ed.)* Mannheim: Dudenverlag.

Eisenberg, P. (2011). *Das Fremdwort im Deutschen*. Berlin and New York: de Gruyter Mouton.

– (2013). "Anglizismen im Deutschen". In: *Reichtum und Armut der deutschen Sprache : Erster Bericht zur Lage der deutschen Sprache*. Ed. by Deutsche Akademie für Sprache und Dichtung, Union der deutschen Akademien der Wissenschaften. Berlin: de Gruyter, pp. 57–119.

Evert, S. and A. Lüdeling (2013). "Measuring morphological productivity: Is automatic preprocessing sufficient?" In: *Proceedings of the Corpus Linguistics 2001 Conference*.

Fleischer, W. and I. Barz (2012). *Wortbildung der deutschen Gegenwartssprache (4. ed.)* Berlin: de Gruyter.

Habert, B. and P. Zweigenbaum (2002). "Régler les règles". In: *TAL* 43.3, pp. 83–105.

Hay, J. B. (2001). "Lexical frequency in morphology: Is everything relative?" In: *Linguistics* 39, pp. 1041–1070.

Hay, J. B. and R. H. Baayen (2002). "Phonotactics, parsing and productivity". In: *Rivista di Linguistica* 15.1, pp. 99–130.

Hein, K. and S. Engelberg (2017). "Morphological variation: the case of productivity in German compound formation". In: *Mediterranean Morphology Meetings* 11, pp. 36–50.

Jong, N. H. de et al. (2002). "The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects". In: *Brain and Language* 81, pp. 555–567.

Krome, S. and B. Roll (2017). "Anglizismen und andere fremdsprachige Neologismen als Indizien für Sprach- und Schreibwandel: Empirische Analysen zum Schreibusus auf der Basis von Textkorpora professioneller und informeller Schreiber". In: *Studia Germanistica* 19, pp. 53–91.

Lüdeling, A. and S. Evert (2005). "The emergence of productive non-medical –itis: Corpus evidence and qualitative analysis". In: *Linguistic evidence: Empirical, theoretical, and computational perspectives*. Ed. by S. Kepser and M. Reis. Berlin: Mouton de Gruyter, pp. 91–95.

Lüdeling, A., S. Evert, and U. Heid (2000). "On measuring morphological productivity". In: *Proceedings of KONVENS 2000*, pp. 57–61.

Lüdeling, A. and N. H. de Jong (2002). "German particle verbs and word-formation". In: *Verb-particle Explorations*. Ed. by N. Dehé et al. Berlin: Mouton de Gruyter, pp. 315–334.

Moscoso del Prado Martín, F., A. Kostić, and R. H. Baayen. (2004). "Putting the bits together: an information theoretical perspective on morphological processing". In: *Cognition* 94, pp. 1–18.

Onysko, A. (2007). *Anglicisms in German: Borrowing, Lexical Productivity, and Written Codeswitching*. Berlin: de Gruyter.

Onysko, A. and E. Winter-Froemel (2011). "Necessary loans – luxury loans? Exploring the pragmatic dimension of borrowing". In: *Journal of Pragmatics* 43.6, pp. 1550–1567.

Schäfer, R. (2015). "Processing and querying large web corpora with the COW14 architecture". In: *Proceedings of Challenges in the Management of Large Corpora (CMLC-3)*, pp. 28–34.

Schäfer, R. and F. Bildhauer (2012). "Building large corpora from the web using a new efficient tool chain". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 486–493.

Shannon, C. E. (1948). " A mathematical theory of communication". In: *Bell System Technical Journal* 27, pp. 379–423; 623–656.

Tagliamonte, S. and D. Denis (2008). "Linguistic ruin? Lol! Instant messaging and teen language". In: *American Speech* 83.1, pp. 3–34.

Winter-Froemel, E., A. Onysko, and A. Calude (2014). "Why some non-catachrestic borrowings are more successful than others: a case study of English loans in German". In: *Language Contact Around the Globe*. Ed. by A. Koll-Stobbe and S. Knospe. Frankfurt am Main: Peter Lang, pp. 119–142.