**Hapaxes in word formation: heralds or happenstance?**

*Muriel Norde, Humboldt-Universität zu Berlin*

If we consider the inventory of all utterances at a specific point in time a corpus, we may say that all actuation in word formation starts with hapaxes. The reverse, however, is not true – not all hapaxes foreshadow the emergence of a new word formation pattern, indeed, the larger the corpus, the more "exotic" the hapaxes we find (De Smet 2020). Even regardless of obvious typos, hapaxes in billion token corpora may well be one-offs that never spread across a speech community. In other words, hapaxes are "an indication of the potential to form neologisms but should not be identified with neologisms themselves". (Cappelle 2010: 343).

This paper aims to explore the question of whether it is possible to predict which hapaxes herald a new word formation pattern. As a methodological reflection on how to solve the actuation problem, we present a corpus-based study of the (relatively recent) neologism *selfie* and related words, i.e. compounds like *koalaselfie* (picture of self with koala), or blends like *felfie* (selfie taken by or with farmer). Of particular interest are a subset of blends where the first input word is not shortened and *selfie* is reduced to *fie*, as in *carfie* (picture of self in or next to car), which might indicate the rise of a productive libfix pattern (cf. Norde & Sippach 2019). Data are drawn from the English, Dutch and Swedish JSI Timestamped corpora available at Sketch Engine (Bušta & Herman 2017), covering six years (2014-2019) of news articles from RSS-enabled sites across the world, which allows for a comparative analysis of productivity in three closely related languages. In all, our data base consists of 9 data sets (three construction types x three languages). As mere hapax frequencies naturally do not suffice to answer the research questions below, we consider various measures of productivity: Type / Token Ratio (TyToR), Hapax / Token Ratio (HaToR) (also known as "potential productivity", Baayen 2009) and global productivity (Baayen & Lieber 1991). However, since token-based productivity measures can be skewed by high token frequencies of just a few types, we also factor in the Hapax / Type Ratio (HaTyR), which is a more reliable measure for the significance of hapaxes (Gyselinck 2018: 177). In addition, we consider the frequency distributions in all sets. The research questions in this case study are the following:

RQ1: How are compounds, blends and (potential) libfixes distributed in the three languages?
RQ2: How are (ranked) token frequencies distributed in the three data-sets?
RQ3: How productive are the three construction types in the three languages?
RQ4: Do productivity and frequency distributions suggest (the emergence of) a libfix pattern?

A preliminary analysis of the data suggests the following:

RQ1: All three construction types are far less common in Dutch and Swedish than they are in English. This may be related to a far lower relative frequency (per million words) of *selfie* in the three subcorpora generally. RQ 2: Some token frequencies show a Zipfian distribution (e.g. Dutch compounds or English blends), others are very unevenly distributed (for instance, the Dutch (potential) libfix *stemfie* (*stem* 'vote' + *fie*) 'picture of self in voting booth' accounts for about 65% of all tokens). RQ 3: Compounds have the highest TyToR and HaToR in all languages, which is what one would expect from a pattern (compounding) that is already fully productive. Productivity for blends and (potential) libfixes is much lower, but with an interesting twist: where blends have low scores on all three productivity measures, (potential) libfixes have low TyToR and HaToR, but a high HaTyR. This is most evident in Swedish, where almost all (potential) libfix constructions are hapaxes. The answer to RQ4, we argue, depends on how the findings to RQ 1-3 are weighted, but a productive word formation pattern seems decreasingly likely for English, Dutch and Swedish (in that order).

**Keywords:** productivity, hapax legomena, blends, libfixes, Germanic

**References**


Baayen, Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Lüdeling, Anke & Merja Kytö (Eds.) *Corpus linguistics: An international handbook*, 900-919. Berlin / New York: De Gruyter Mouton.

Baayen, Harald & Rochelle Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 29, 801-843.

Bušta, Jan & Ondřej Herman. 2017. JSI Newsfeed Corpus. *The 9th International Corpus Linguistics Conference*, University of Birmingham, 25-28 July 2017.

Cappelle, Bert. 2010. Doubler-upper nouns: A challenge for usage-based models of language? In Onysko, Alexander & Sascha Michel (Eds.) *Cognitive Perspectives on Word Formation*, 335-374. Berlin / New York: De Gruyter Mouton.

De Smet, Hendrik. 2020. What predicts productivity? Theory meets individuals. *Cognitive Linguistics* 31(2), 251-278.

Gyselinck, Emmeline. 2018. *The role of expressivity and productivity in (re)shaping the constructional network*. PhD thesis, Universiteit Gent.

Norde, Muriel & Sarah Sippach. 2019. *Nerdalicious scientainment.* A network analysis of English libfixes. *Word Structure* 12(3), 353-384.