PREVALENCE BASED QUANTITATIVE ANALYSIS OF INNOVATIONS IN LEXICAL EVOLUTION

ANDREAS BAUMANN[*1] and KLAUS HOFMANN[1]

[*]Corresponding Author: andreas.baumann@univie.ac.at
[1]Faculty of Philological and Cultural Studies, University of Vienna, Vienna, Austria

The spread of linguistic innovations has been modeled by means of logistic growth (Altmann, 1985; Labov, 1994; Ghanbarnejad et al., 2014). The epidemiological argument underlying this type of growth is that (a) innovations are transmitted from one individual to the next and that (b) growth rate depends on both the number of individuals that already know the innovation and the number of individuals that do not yet know it (Figure 1b, left; Cavalli-Sforza & Feldman, 1981). The resulting dynamics are S-shaped (as in Figure 1c; cf. Blythe & Croft, 2012). Interestingly, although this argument is inherently population dynamic and involves interacting individuals, empirical accounts typically assess S-shaped linguistic spread by means of token frequency (Denison, 2003).

In this contribution, we promote the use of *prevalence* for studying lexical change, i.e., the fraction of individuals that know and use a word. Lexical prevalence was employed in mathematical accounts of lexical spread (Nowak, 2000) and, more recently, considered in psycholinguistic research where prevalence was argued to function as relevant (control-)variable in experimental setups (Brysbaert et al., 2019). Here, we show (a) that lexical prevalence is relevant for studying the spread of lexical innovations in that prevalence trajectories typically precede frequency trajectories, and we propose (b) the use of coupled dynamics of prevalence and frequency. Our study unfolds in two steps.

In the first part, we use a diachronic corpus of German newspapers spanning a period of two decades (Ransmayr et al., 2017) and derive trajectories of token frequency and prevalence (by using the fraction of authors using a word as a proxy for prevalence; cf. Johns et al., 2020) for a set of about 700 words that have been strongly increasing in this period (e.g., *googeln*, 'to google'). By fitting logistic models to each trajectory, we determine (i) the intrinsic growth rate, (ii) the inflection point, and (iii) goodness-of-fit for each word. We show that prevalence

curves have (i) a higher intrinsic growth rate (cf. Figure 1a), (ii) earlier inflection points, and (iii) a better fit that frequency curves. We take this to show that logistic prevalence curves detect lexical innovations earlier and more reliably than this is the case for logistic frequency curves.

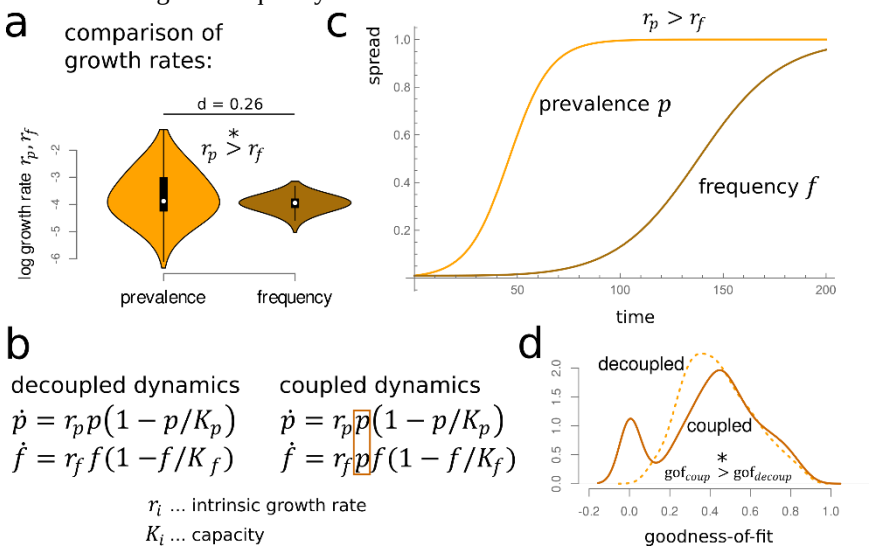

Figure 1. (a) Empirical comparison of prevalence and frequency trajectories; (b) studied dynamical systems; (c) solution for coupled dynamics; comparison of goodness-of-fit of the systems in (b).

In the second part of our study, we test if goodness-of-fit improves if we couple the dynamics of prevalence and frequency rather than keeping them separate as before. This is done by letting the growth of frequency at a certain time not only depend on token frequency but also on prevalence (Figure 1b, left). This is motivated by the assumption that the production of words depends on whether an individual knows a word and on how entrenched it is (Bybee, 2006). Given that intrinsic growth is higher for prevalence than for frequency, the frequency trajectory is shown to be preceded by the prevalence trajectory in this model (Figure 1c). Subsequently, we use the empirical trajectories analyzed before and compute goodness-of-fit for the respective solutions of both systems, coupled and decoupled. We show that although coupled dynamics are in some cases uninformative (goodness-of-fit close to zero), they generally show a slightly better fit than decoupled systems (Figure 1d; our study is complemented with a sensitivity analysis of goodness-of-fit differences). We conclude that coupled dynamics of prevalence and frequency represent a reasonable model for studying lexical evolution that, in addition, help to reconstruct diachronic prevalence trajectories from frequency data.

References

Altmann, G. (1985). On the dynamic approach to language. In *Linguistic Dynamics* (pp. 181-189). de Gruyter.

Blythe, R. A., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, 269-304.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior research methods*, *51*(2), 467-479.

Bybee, J. (2006). *Frequency of use and the organization of language*. Oxford University Press.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton University Press.

Denison, D. (2003). Log (ist) ic and simplistic S-curves. *Motives for language change*, *54*, 70.

Ghanbarnejad, F., Gerlach, M., Miotto, J. M., & Altmann, E. G. (2014). Extracting information from S-curves of language change. *Journal of The Royal Society Interface*, *11*(101), 20141044.

Johns, B. T., Dye, M., & Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, *73*(6), 841-855.

Nowak, M. A. (2000). The basic reproductive ratio of a word, the maximum size of a lexicon. *Journal of theoretical biology*, *204*(2), 179-189.

Ransmayr, J., Mörth, K., & Ďurčo, M. (2017). AMC (Austrian Media Corpus) – Korpusbasierte Forschungen zum österreichischen Deutsch. In Claudia Resch and Wolfgang U. Dressler (eds.), *Digitale Methoden der Korpusforschung in Österreich*. 27-38. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.