

Putting corpora into perspective: Modeling collective linguistic experience as a function of time

One of the main tenets of cognitive linguistics states that “language structure emerges from language use” (Tomasello, 2003:5) and that this emergence is an ongoing process, contributing as one factor to continuous language change. This blurs the distinction between diachronic and synchronic linguistics: One may still take a synchronic perspective on language—in fact, most studies in cognitive linguistics do so—but what one gets to see is a dynamic object whose structure is intrinsically unstable and becomes manifest only statistically. Language in this view is a moving target.

Over the past two decades, a growing body of analytical methods and theoretical frameworks has been developed that take the unstable, adaptive nature of language into account. Given these advances, surprisingly little research has addressed the question of how a corpus would have to be devised in order to provide an appropriately “synchronic” view on the ever-changing language structure. Assumptions such as the correlation between frequency of use and degree of entrenchment should in our opinion be reflected and investigated with respect to time. Especially for language phenomena whose observed usage frequencies change substantially over time, it is not clear how to conceive a valid notion of frequency that corresponds to their *present* degree of entrenchment.

If language use intrinsically influences the degree of entrenchment and if it is thereby also a source of language change, then, intuitively, the cognitive and linguistic significance of usage events and usage frequencies should fade with increasing “age”. A synchronic corpus should therefore emphasize more recent usage events. Starting from this basic assumption, the following questions are discussed (and not necessarily answered):

- ◆ How can the fading cognitive significance of usage events be modeled?
- ◆ What are the underlying assumptions with respect to individual and collective “memory”?
- ◆ What is an appropriate formal fading function?
- ◆ Is this function the same for all kinds of language domains (probably not) and within a given domain for all types of language phenomena?
- ◆ How can these issues be addressed experimentally, and what experimental evidence in the existing literature may be relevant in this context?
- ◆ In particular, does a decreasing frequency of use in fact result in reduced entrenchment?
- ◆ How do language change and the fading cognitive significance interact?
- ◆ What other potential factors are to be incorporated into a synchronic view on language?

Taking advantage of DEREKO, the Archive of General Reference Corpora of Contemporary Written German with 3.4 billion running words, we explore the empirical consequences of using a synchronic corpus in the above sense for a range of phenomena: from simple word frequencies to collocational patterning, and on to yet more complex corpus-driven analyses, e.g., inferring global contexts from collocation profiles. Implications for synchronic, diachronic and applied linguistics are discussed.

References

Tomasello, Michael (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.