On the usefulness, reliability and efficiency of corpus-based data

Edmond Cane
Academy of Sciences of Albania
edmond.cane@akad.gov.al
ecane2000@yahoo.com

The argument in this paper is based on a CL (Cognitive Linguistics) account, which highlights the usage-based nature of language structures. Empirical quantitative evidence is useful to prove and/or to complement the hypotheses rising from the cognitive linguists’ skilled intuition about language. This paper presents the range of quantitative data relevant to a CL account of the set of verbs, having their non-progressive or progressive lexical content (stative verbs for instance: *be*, *have*, *think*, *understand*, *hear*), which either blocks their use with a progressive tense, or renders it unnecessary. Cross-linguistic evidence argues against any pre-determined nature of the shaping or structuring of such feature, positing it as a language-specific one. Languages like English, Albanian, Italian, have different sets of such verbs, which implies a different map of the distribution of such specific content in different languages. There is sufficient cross-linguistic data to confirm this different mapping not only through different languages, but even within the dialects of the same language (Cane 2016). Further evidence suggests that there is evolution of particular verbs in chronological time, with non-progressive verbs developing by ’opening their lexical content” towards progressive content.

The advent of corpora has opened the way to relatively reliable quantitative examination. The data from BNC and Albanian Corpus show the significant quantitative difference in the use with the progressive tenses of the particular set of verbs as compared to the rest of the (common) verbs, thus linking the said qualitative feature to the quantitative data from usage. The data goes further to show that English and Albanian have a lot in common, but also significant differences. Furthermore, the Albanian corpus shows a small set of verbs that have a clearly significant number of uses in progressive tenses only in the Gheg dialect.

On the other hand, some of the respective verbs in English and Albanian which have a clear non-progressive or prevalent progressive content, and for either reason do not generally take progressive tenses, have their marked exceptions, with a marked form (atypical progressive tense) and marked content. The data from corpora show further differences in English and Albanian as well as within the dialects in Albanian. Although the quantitative data shows and confirms different distribution of the lexical content of the respective verbs, the linear numbers need to be further elaborated (using correspondence analysis, including its additional variants – Gries (2013); Stefanowitch & Gries (2003). This may imply the need to evaluate the standard deviation (SD) for such feature as a language-specific value and then estimate the actual individual SD of verbs as a function to their own language-specific SD value.

Another way to make the corpus-based data more efficient/reliable is to complement that with data from other sources, such from the multilingual corpora or parallel aligned corpus of translated texts. This has been provided for many European languages, but not for Albanian. However, it may be used indirectly.

Further, one interesting perspective is to evaluate the SD value of certain verbs against the metatext data, which will further hint and point to the discourse register where such change is being introduced or is actually occurring.
References


Corpora

- British National Corpus
- Linguee Dictionnaire anglais-français et recherche via un milliard de traductions: https://www.linguee.fr