The advent of post-editing lexicography

Miloš Jakubíček

Miloš Jakubíček is an NLP researcher, software engineer and CEO of Lexical Computing (LC), a research company developing the Sketch Engine (SkE) corpus platform. His research interests are devoted mainly to effective processing of very large text corpora for lexicographic and linguistic tasks and syntactic parsing of morphologically rich languages. Since 2008 he has been involved in the development of SkE, in 2011 he became director of the Czech branch of LC leading the local development team of SkE, eventually becoming CEO of LC in 2014 after its founder Adam Kilgarriff. He is a fellow of the NLP Centre at Masaryk University where his interests lie mainly in morphosyntactic analysis of Czech and its practical applications.
milos.jakubicek@sketchengine.co.uk

The lexicographic landscape has been subject to two major disruptions over the past twenty years.

The first is related to the uptake of information technology and availability of text corpora. Lexicographers were on the forefront of the shift to empiricism in linguistics[1] and it was for good: a field that never seriously acknowledged any theoretic framework was starting to benefit more than any other linguistic discipline – practical needs for describing language as used were very high.

The second change, related to the first one, was without doubt the breakdown of traditional publishing business, manifested in the end of paper dictionaries (as well as by the fall of many renowned dictionary publishers). From the perspective of users, dictionaries are tools to be *used while doing something else*, to paraphrase Hilary Nesi.[2] The environment has drastically changed and so do need to change the tools.

The impacts of both of these changes are yet to be discovered: for the latter one, the status quo can be well described by quoting another heavyweight in the field, the long-time editor-in-chief of Macmillan dictionaries, Michael Rundell, whom I often heard saying: "After working in this field for 30 years, I thought I had a pretty good idea about how to create and publish a dictionary. But things have changed so dramatically in the last five years, that I have only a limited idea of what the future of lexicography will be."

The impact is a bit easier to be foreseen as regards technological innovations. Contemporary lexicography makes heavy use of corpora and increasingly also of many natural language processing tools that automate the analysis of morphology as well as syntax and semantics. Many tools for (semi-)automation of specific lexicographic tasks have been developed as well. In a review carried out in 2011, Rundell and Kilgarriff argue[3] that while word sense identification and definition writing remain to be tackled, many other tasks alongside the lexicographic workflow have been already solved with an accuracy that delivers time- (and therefore money-) saving solutions. This is deemed to be the case for devising dictionary headword lists, finding collocations and other multiword units, or extracting dictionary examples from corpora.

What is the next step? At the moment lexicographers query corpora (by means of many tools) for finding linguistic evidence in order to draft a dictionary entry which they continue working on and which is subject to a number of reviews in the lexicographic workflow.

The next step is to spare the lexicographers from such initial corpus query and entry drafting. Instead of starting with an "empty" dictionary, they will be able to begin with a dictionary database pre-populated with entries according to a big underlying reference corpus. These entries will contain suggested word sense clustering, with definitions (or explanations in alternative forms such as image media), labels and examples extracted from corpora. These entries will then be edited in an environment that includes direct links to underlying corpus evidence so as to allow manual inspection of the source texts, as well as mechanisms for easy and simple corrections of the entry (e.g. lumping and splitting of word senses, replacing dictionary examples, amending definitions and labels). Having all the evidence at hand, the next step is to leave the "easy" bits to the computer and have human editors spend their time on the more intellectually demanding parts of the job. This opens the way to *Post-Editing Lexicography*, in an analogy to the translation process. Translators used to use many independent tools (dictionaries, in the first place!) up to the moment when machine translation

1 As can be seen from early corpus development projects like COBUILD or BNC, which were both driven and devised (also) for lexicographers, who were themselves employed in empirical linguistic research (cf. Church, Ward and Hanks, 1990. Word association norms, mutual information, and lexicography. *Computational Lnguistics* 16.1, 22-29).

2 See e.g. *The Oxford Handbook of Lexicography*, 2016. Durkin, P. (ed.). Oxford: OUP, 584.

3 Rundell and Kilgarriff, 2011. Automating the creation of dictionaries: Where will it all end? In Meunier et al. (eds.), *A Taste for Corpora. In honour of Sylviane Granger*. Amsterdam: Benjamins, 257-281.

and translation memories became mature enough to be exploited for professional translation, and henceforth translators became post-translation editors.

An important lesson from the translation business concerning potential danger for the future of lexicography is that the transition to post-editing translation was by far not easy, partially because it may have actually begun too soon (pushed by translation and localization agencies pressing to cut costs), having machine translation do yet another "over-promised and under-delivered" U-turn.

Eventually, this adoption has further progressed as the technology became more mature, and, mainly, as the translation environment for professionals has become more suited for the task of post-editing, which is very different from translating from scratch.

However, the episode had a very undesired consequence that we want to avoid in lexicography. Translators were abandoning machine translation, for both technological reasons as well as for fear of becoming jobless. These fears remain, even though the former is being improved and the latter just did not prove to be the case: the translation industry is growing and some reports describe it as one of the fastest growing businesses today.[4] Moreover, as the whole process gets streamlined, there are good chances that lower per-word income of translators will eventually turn into a higher per-hour rate for them.

The lessons for lexicography are straightforward: the transition to post-editing must be backed by solid technology (which we believe we have), revisited workflows (which we need to work on), and with advocacy explaining that it is not meant to steal lexicographic jobs. The shift to *post-editing lexicography* might well be a fertilizer for the falling industry, showing faster (and hence, more affordable) and more effective workflows.

A specific account comes with addressing less-resourced languages – or those that are basically not resourced at all at the moment. There are plenty of them, often geographically located in areas with growing numbers of speakers. Many of these speakers live in poverty, which nevertheless does include the possession of a smartphone. Language resources will be one of the first data needed when these societies will approach information levels of the developed world. There will be a strong business need for them but no time for twenty-years-running lexicographic projects, another reason why human efforts need to be supplemented by automation as far as possible.

At the eLex 2017 conference we are going to present a proof of concept in the form of interconnecting Sketch Engine, a leading corpus query system, with Lexonomy, a new lightweight dictionary writing system.[5] We will show how a dictionary draft can be directly obtained from a reference corpus (as a One-Click Dictionary) in Sketch Engine and how it can be efficiently post-edited in Lexonomy.

The future of lexicography presents big challenges. It would be naïve not to realize that many of them pose real issues, problems and obstacles for all players in the field. However, the more so we need to look for those of them that present real opportunities – and, I believe, *post-editing lexicography* is one of them.

4 See, e.g., https://gala-global.org/industry/industry-facts-and-data.

5 Jakubíček M., Kovář V., Měchura M. and Rychlý P. One-Click Dictionary. In *Electronic lexicography in the 21st century, Proceedings of eLex 2017* (forthcoming).

Sketch Engine (SkE) started in 2004 as an academic and lexicographic product for corpus query and management and has since attracted a wide audience including translators, writers, marketers, brand naming and SEO professionals. To meet the challenge of guiding this variety of users to the functionality they need in a streamlined way, an all-new user interface is under development to not only bring in the latest Web technology but also change the way users interact with SkE. With a soft launch due in autumn 2017, users will enjoy a new friendly design which adapts to small touch screens of tablets and mobile phones. Input forms and selection screens will offer basic and advanced layouts, the former targeted at casual users without a profound knowledge of corpora or NLP and the latter serving academic and professional users. In this process, various controls have shifted to more intuitive spots, enabling the user to, for example, adjust the view on the result screen rather than make this decision beforehand as it is now, while preserving all the options and features that are currently available.
Hand in hand with developing the new look and feel, SkE will become more useful to anyone in need of glossaries or dictionaries. In addition to serving as an indispensable tool for gathering data, the new link with **Lexonomy** system will enable data conversion into a lexicographic product as part of an online dictionary writing tool, which doubles as a hosting service that can produce a dictionary and have it published online instantly. Lexonomy also features an easy-to-use XML editor suitable for users with no prior knowledge to create lexicographic products complying with current standards. Embedding Lexonomy in SkE will become vital for starting brand new lexicographic projects. Users will access a corpus to identify the most frequent words and have the list pushed to Lexonomy along with part of speech tags, usage flags, example sentences, collocations, synonyms, definitions or translation, thus generating a dictionary draft for post-editing. Likewise, a subject-specific glossary can be developed analogically from a terminology list extracted from a domain corpus. This push & pull model will dramatically change the way dictionaries are built, besides its beneficial time and money saving implications.

Ondřej Matuška
Lexical Computing