Scientific Report from Short Term Scientific Mission

COST STSM Reference Number: COST-STSM-IS1305-26496
Period: 2015-04-07 to 2015-04-17
Duration: 9 working days
COST Action: IS1305

STSM type: Regular (from Ireland to the Netherlands)
STSM Title: European Dictionary Portal
Guest/STSM applicant: Michal Měchura, Fiontar, Dublin City University, valselob@gmail.com
Host: Bob Boelhouwer, Instituut voor Nederlandse Lexicologie, bob.boelhouwer@inl.nl

1. Purpose of the STSM

To work with computational linguists at INL to build an initial working version of a portal of European dictionaries, as envisaged by ENeL's Working Group 1.

2. Description of the work carried out during the STSM

2.1. Summary

The STSM's purpose has been achieved and an initial version of the portal is now available [1]. Members of Working Group 1 will now be polled for feedback and, following a discussion at ENeL's next meeting in August 2015, the portal will hopefully be developed into a fully functioning website.

2.2. Intended audience

During the STSM we have established that:

- The website will be intended primarily for the general public, especially for people who need to use dictionaries often: translators, language instructors, language students and people who work with texts (journalists, editors). Only secondarily will the website be intended for lexicographers.

- The goal will be to make European dictionaries easy to find and convenient to search for the ordinary internet user, while guiding the user's attention to high-quality trustworthy dictionaries. The website will make dictionaries searchable from one place, in a single coherent user experience, thereby improving on the existing situation where internet users need to search multiple dictionaries separately. It will act as a specialized, federated search engine for dictionaries, hoping to eventually replace general search engines like Google and Bing as the tool of choice for people looking for dictionary content on the web.

2.3. Website features

The initial working version of the portal that has been produced during the STSM has two main features:

- **Dictionary catalog.** This is accessible under the heading 'Find a dictionary' and it is a

searchable, filterable catalog of online dictionaries that ENeL considers trustworthy. The goal here is to help the user find a dictionary (as opposed to find something *in* a dictionary). For demonstration purposes, the catalog has been populated with data from the 'inventory of scholarly dictionaries' spreadsheet, a result of Gerbrich de Jong's STSM earlier in 2014 [2].

- **Metasearch.** This is accessible under the heading 'Search in multiple dictionaries'. The goal here is to help the user find something in multiple dictionary websites, quickly and conveniently from a single user interface. Note that this feature essentially provides direct links to search results; it does not scrape, process or filter data from other websites in any way. For demonstration purposes this feature has been populated with a handful of dictionaries in four languages (English, Dutch, Irish, Czech).

  Every dictionary website is represented here by a small JavaScript file, essentially a 'plug-in' which tells our portal everything it needs to know about the website: its name and URL, what languages and language pairs it has and how one is supposed to compose the URL to link to search results. This architecture is inspired by *OpenSearch description documents* [3] but with more detail for representing facts about the search engines we link to.

2.4. Non-features

A long-term goal of ENeL's Working Group 1 is to build a portal with a deeper linking mechanism between dictionaries: the portal website would recombine and 'mash up' content from various dictionaries and present it to the user on a single page.

We have concluded during the STSM that this is not feasible at the moment as it would require an infrastructure of machine-readable web services, downloadable datasets and mutually agreed-on data formats that does not exist in the dictionary sector yet.

However, a precedent can be found in the terminology sector, for example in the EuroTermBank project [4], a federated search engine which pools search results from several terminological databases on the Internet. The building of such a portal has been made possible because terminology data often comes in standardized machine-readable formats such as TBX [5] as a result of a frequently occurring requirement by translators and other language professionals to exchange terminological data between various CAT (Computer-Aided Translation) software packages such as Trados and Wordfast.

3. References

[1] http://dictionaryportal.eu/

[2] http://www.elexicography.eu/wp-content/uploads/2014/12/Scientific-Report-STSM-Gerbrich_de_Jong.pdf

[3] http://en.wikipedia.org/wiki/OpenSearch

[4] http://www.eurotermbank.com/

[5] http://en.wikipedia.org/wiki/TermBase_eXchange