Cross-lingual Data Linking in e-Lexicography

Sina Ahmadi Mihael Arcan John McCrae

Insight
Centre for Data Analytics

Introduction

Cross-lingual knowledge linking is the task of **building relationships** between entities in different languages. It aims at extending knowledge across different languages and various resources, in particular **lexicons**, and improving applications in fields such as Information Retrieval and Machine Translation.

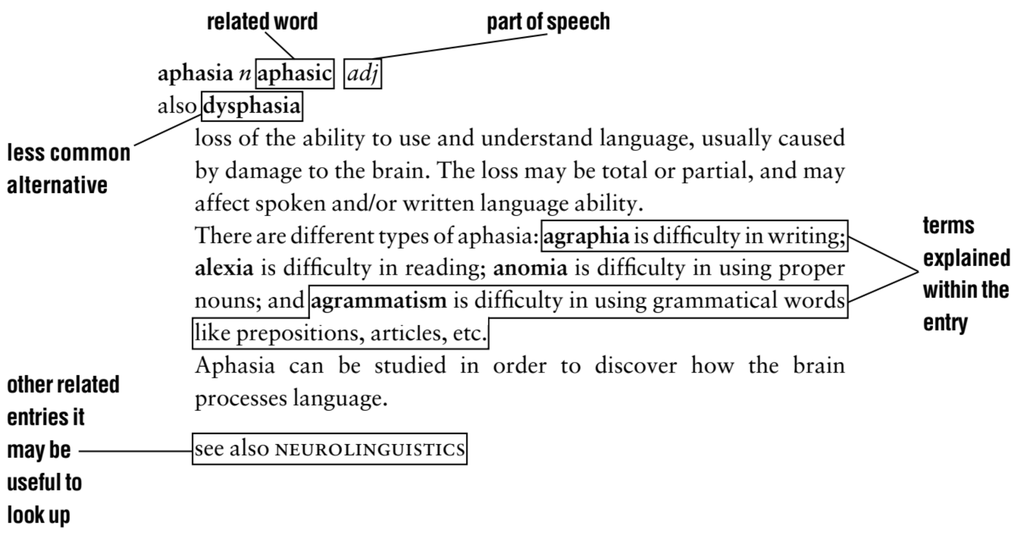

Figure: *Dictionary as a source of data*

Motivations

The lexicographic landscape in Europe is currently rather **heterogeneous**. On one hand, it is characterised by **stand-alone lexicographic resources**, which are typically encoded in **incompatible data structures** and on the other hand, there is a significant **variation in the level of expertise and resources** available to lexicographers. This forms a major obstacle to more ambitious, innovative, transnational, data-driven approaches to dictionaries, both as tools and objects of research.

We are specifically aiming at understanding how **linguistic and non-linguistic features** can help connecting dictionaries. And to what extent semantic resources and data structures can enhance accuracy in the linking task?

Objectives

This project is a part of the **ELEXIS** project (**https://elex.is/**). ELEXIS is an infrastructure to integrate, extend and harmonise national and regional efforts in the field of lexicography, both modern and historical, with the goal of creating a sustainable infrastructure. In the next four years, we are aiming at:

- Building the infrastructures for sharing language resources
- Integrating and enriching lexicographic data
- Motivating interoperability by promoting common lexicographic concepts
- Defining a minimal common data model capturing the basic concepts in lexicography
- Providing conversion tools for different data formats
- Creating a huge multilingual registry that connects dictionaries across common concepts
- Bridging the gap between lesser-resourced and advanced communities

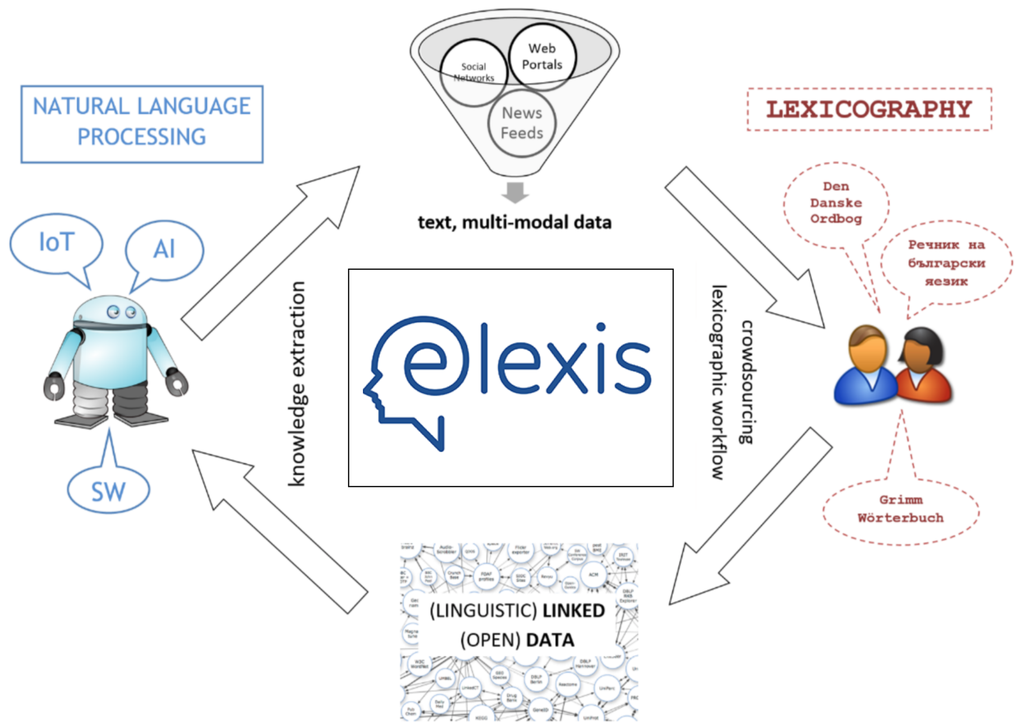

Figure: The work packages of ELEXIS project

NAISC — Nearly Automatic Integration of SChemas

We are developing **a semi-automatic system** called NAISC (means "links" in Irish) that will make the linking problem viable for large resources by using state-of-the-art semantic and natural language processing techniques and creating new methods in link discovery and data linking.

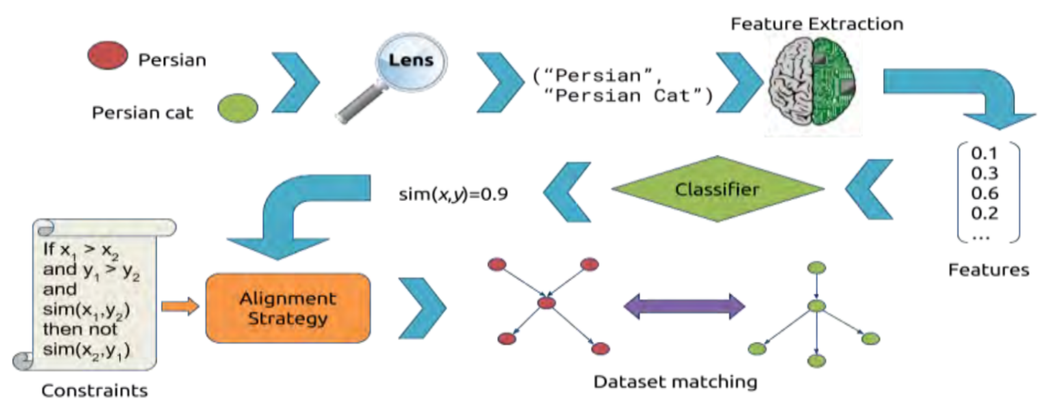

Figure: The architecture of NAISC

Conclusion

This project will allow companies **to manage large taxonomies and knowledge graphs** that describe their business. It will also be extended to work across languages allowing multinational companies to more easily collaborate and easing localization of products to **make better business decisions** and fully exploit it to **maximise business value**.

Further Reading

- McCrae, John P., and Paul Buitelaar. "Linking Datasets Using Semantic Textual Similarity." *Cybernetics and Information Technologies* 18.1 (2018): 109-123.
- Pedersen, Bolette Sandford, et al. "ELEXIS-a European infrastructure fostering cooperation and infor-mation exchange among lexicographical research communities." *Proceedings of Global Wordnet Conference 2018*. 2018.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD Dublin Science Foundation Ireland sfi For what's next