Digital Geolinguistics: On the use of Linked Open Data for Data-Level Interoperability Between Geolinguistic Resources

Giorgio Maria Di Nunzio

Dept. of Information Engineering – University of Padua dinunzio@dei.unipd.it

Abstract. The Open Language Archives Community which recently celebrated its first 10 years of activity, is a worldwide network dedicated to collecting information on language resources and developing standard protocols for interoperability. In this context, Linked Open Data paradigm is very promising, because it eases interoperability between different systems by allowing the definition of data-driven models and applications.

In this talk, we give an overview of present geolinguistics projects and an approach which moves the focus from the systems handling the linguistic data to the data themselves. As a concrete example, we present a geolinguistic application build upon a real linguistic dataset which provides linguists with a system for investigating variations among closely related languages.

1 Introduction

The research field of linguistics studies all aspects of human language, including morphology (the formation and composition of words), syntax (the formation and composition of phrases and sentences from these words) and phonology (sound systems) [1]. Research in the variations in languages allows linguists to understand the fundamental principles that underlie language differences, language innovation and language variation in time and space.

Geolinguistics is an interdisciplinary field that incorporates language maps depicting spatial patterns of language location or the results of processes that lead to language change [2]. In this context, the linguistic atlas has proved to be a vital tool and product of geolinguistics since the earliest stages of the field, and it has provided a stage for the incorporation of modern GIS.

In the last two decades, several large-scale databases of linguistic material of various types have been developed worldwide. The Open Language Archives Community, which recently celebrated its first 10 years of activity, is a worldwide network dedicated to collecting information on language resources (field

¹ http://www.language-archives.org

notes, grammars, audio/video recording, descriptive papers, and so on) and developing standard protocols for interoperability. GOLD² was the first ontology to be designed specifically for linguistic description on the Semantic Web [3]. It proposes a solution to the lack of interoperability between linguistic projects and projects designed specifically for NLP applications. It can act as a kind of lingua franca for the linguistic data community, provided that data providers are willing to map their data to GOLD or to some similar resource. In [4], the authors present a framework for producing multi-layer annotated corpora: a pivot format serving as "interlingua" between annotation tools, an ontology-based approach for mapping between tag sets, and an information system that integrates the various annotations and allows for querying the data either by posing simple queries or by using the ontology.

Language resources that have been made publicly available can vary in the richness of the information they contain: on the one hand, a corpus typically contains at least a sequence of words, sounds or tags; on the other hand, a corpus may contain a large amount of information about the syntactic structure, morphology, prosody and semantic content of every sentence, plus annotations of discourse relations or dialogue acts [5]. However, the quality of such corpora may have been reduced by the intense, and often poorly controlled, usage of automatic learning algorithms [6].

The heterogeneity of linguistic projects has been recognized as a key problem limiting the reusability of linguistic tools and data collections [7]. The rate of re-use for linguistic database technology together with related processing tools and environments is still too low. For example, the Edisyn search engine – the aim of which was to make different dialectal databases comparable – "in practice has proven to be unfeaseable".³ In order to find common ground where linguistic material can be shared and re-used, the methodological and technological boundaries existing in each research linguistic project needs to be overcome.

2 Linked Open Data for Geolinguistic Resources

The research direction we want to discuss in this talk is to move the focus from the systems handling the linguistic data to the data themselves. For this purpose the LOD paradigm [8] is very promising, because it eases interoperability between different systems by allowing the definition of data-driven models and applications. LOD is based on the definition of real-world objects, identified by means of a dereferenceable URI⁴. Objects are related to one another by means of typed links. Interoperability is achieved by a unifying data model (i.e. RDF⁵), a standardized data access mechanism (i.e. HTTP), hyperlink-based data discovery (i.e. URI), and self-descriptive data (based on shared open vocabularies from different namespace) [8].

² http://linguistics-ontology.org/

³ http://www.dialectsyntax.org/

⁴ http://tools.ietf.org/html/rfc3986

⁵ http://www.w3.org/RDF/

In this context, a relevant initiative is ISOcat,⁶ a linguistic concept database developed by ISO Technical Committee 37, Terminology and other language and content resources, to provide reference semantics for annotation schemata. The goal of the project is to create a universally available resource for language-related metadata that can be used in a variety of applications and environments. It also provides uniform naming and semantic principles to facilitate the inter-operability of language resources across applications and approaches [9].

In this talk, we discuss the steps of a possible approach for exposing geolingistic data into LOD [10–12] by presenting:

- the ASIt⁷ linguistic project which is based on micro-variations of Italo-Romance dialects;
- a geolinguistic Web application that provides functionalities for accessing, browsing, searching the linked open data by means of linguistic features, and visualizing the data on dynamically generated maps.

References

- 1. Akmajian, A., Demers, R.A., Farmer, A.K., Harnish, R.M.: Linguistics, Sixth Edition An Introduction to Language and Communication. The MIT Press (2010)
- 2. Hoch, S., Hayes, J.J.: Geolinguistics: The Incorporation of Geographic Information Systems and Science. The Geographical Bulletin **51**(1) (2010) 23–36
- 3. Farrar, S., Langendoen, T.: A Linguistic Ontology for the Semantic Web. Glot International **7**(3) (March 2003) 97–100
- 4. Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., Stede, M.: A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. TAL 49(2) (2008) 217–246
- Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
- 6. Spärck Jones, K.: Computational linguistics: What about the linguistics? Computational Linguistics $\bf 33(3)$ (2007) 437–441
- Chiarcos, C.: Interoperability of corpora and annotations. In Chiarcos, C., Nordhoff, S., Hellmann, S., eds.: Linked Data in Linguistics. Springer Berlin Heidelberg (2012) 161–179
- 8. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers (2011)
- Kemps-Snijders, M., Windhouwer, M.A., Wittenburg, P., Wright, S.: ISOcat: Remodelling Metadata for Language Resources. IJMSO 4(4) (2009) 261–276
- Di Buccio, E., Di Nunzio, G.M., Silvello, G.: A system for exposing linguistic linked open data. In Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F., eds.: TPDL. Volume 7489 of Lecture Notes in Computer Science., Springer (2012) 173–178
- 11. Di Buccio, E., Di Nunzio, G.M., Silvello, G.: An open source system architecture for digital geolinguistic linked open data. In Aalberg, T., Papatheodorou, C., Dobreva, M., Tsakonas, G., Farrugia, C.J., eds.: TPDL. Volume 8092 of Lecture Notes in Computer Science., Springer (2013) 438–441
- 12. Di Buccio, E., Di Nunzio, G.M., Silvello, G.: A curated and evolving linguistic linked dataset. Semantic Web 4(3) (2013) 265–270

⁶ http://www.isocat.org/

⁷ http://svrims2.dei.unipd.it:8080/asit-enterprise/