Towards the Construction of a Lexical Data and Technology Ecosystem: The Experience of ILC-CNR

Monica Monachini, Anas Fahad Khan
Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche
Via Moruzzi 1 - Pisa,
{monica.monachini, fahad.khan}@ilc.cnr.it

Abstract
This paper describes the activities and projects being carried on at the "A. Zampolli" Institute for Computational Linguistics (ILC) at the crossroads between computational lexicography and e-lexicography and that are intended to assist in the creation of a queryable and interconnected ecosystem of standardised lexicographic datasets and technologies.

Keywords: e-lexicography, computational lexicography, lexical resources, standards, LOD

1. Introduction

Lexicography is traditionally recognised as the branch of applied linguistics that is concerned with the design and construction of practical resources for describing the lexicon of a language. In the last few decades or so the marriage of lexicography and digital technology has resulted in the creation of two new disciplines: e-lexicography, i.e., the compilation of digitally-born dictionaries for human users, and computational lexicography, a sub-branch of computational linguistics that deals with the use of lexicons in Natural Language Processing as well as with the use of computational techniques in building and enriching lexicons (for NLP purposes).

The use of language technology has had an important impact on the task of compiling dictionaries for human use. Not only do modern day technologies allow for the easier digitization of lexical resources, but current trends in language resources and data science make it possible to imagine the fulfilment, in the very near future, of one of the most important promises of e-lexicography - namely that of a large-scale interconnected ecosystem of of open, queryable and standardised lexicographic datasets and technologies. In fact it seems as if e-lexicography's moment may finally have arrived.

In the rest of this article we will describe some of the activities, past and present, in which ILC has been involved and/or still is involved and which we believe make a strong contribution towards this ultimate aim.

2. Lexical Resources, Standards and Infrastructures

ILC-CNR can boast of a long-standing involvement in computational lexicography dating back to the pioneering work of Antonio Zampolli and others[1]. These early activities eventually resulted in the creation of influential lexical resources such as PAROLE SIMPLE CLIPS (PSC[2]) and ItalWordnet (IWN[3]).

Aside from the creation of language resources, however, another important and salient aspect of the work carried out at ILC is the participation of its members, and in particular, those of the Language Resources and Infrastructures group (LaRI[4]) within the institute, in important standardisation projects and initiatives, such as LIRICS[5] and LMF[6] (Francopoulo 2013).

LMF for instance is an influential standard within the field of computational linguistics and language technologies; it is also important for lexicographic resources intended for human users. The LMF core model is currently being revised as a multipart standard. One of the other parts of the standard aims at a level of higher interoperability with TEI through the production of a TEI-XML serialisation of LMF[7]. In addition, a new module for etymology is being added to the new version of the LMF core[8].

Two important infrastructural projects in which the LaRI group is involved are PARTHENOS[9], and ELEXIS[10]. The former project includes the presence of various European infrastructures, such as DARIAH and CLARIN and has the goal of consolidating shared practices and data models among various domains within the humanities. A number of different standardisation initiatives are currently taking place within Parthenos with the aim of improving the interoperability of lexical resources including digitized dictionaries. The latter project -- ELEXIS -- begins in February 2018 and is an ambitious project within the domains of NLP and e-lexicography with the aim of creating a European wide lexicographic infrastructure. Several different standardisation efforts are likely to converge within the ambit of this project, in particular those carried out under the banner of the International Organization for Standardization (such as LMF) along with those newly emerging standards for Linked Open

[1] For an overview see (Calzolari, Monachini, and Soria 2013).
[2] http://hdl.handle.net/20.500.11752/ILC-88.
[3] http://hdl.handle.net/20.500.11752/ILC-65
[4] http://lari.ilc.cnr.it/
[5] http://lirics.loria.fr/
[6] LMF has been developed under the aegis of the ISO Committee TC37/SC4 (ISO-24613:2008)
[7] Here with a strong participation of ILC (one of the members of LaRI is the co-leader of the LMF working group).
[8] Here too with the participation of one of the members of LaRI who has a co-leader role.
[9] www.parthenos-project.eu
[10] ELEXIS is based on a previous Cost Action ENeL - aiming to establish a European network of lexicographers and a common approach to e-lexicography that forms the basis for a new type of lexicography (http://www.elexicography.eu/).

Data, developed under the banner of W3C (Ontolex-Lemon).

A strong impetus has been provided to the research directions mentioned above within ILC by the institute's official role as the leading Italian participant in the CLARIN-ERIC infrastructure[11]. Standardisation activities and the promotion of shared formats are crucial for CLARIN, and a Standard Committee is active within the infrastructure together with a task force dealing with Interoperability. While formats and best practices for corpora have been central till now, we foresee that the standardisation of lexical resources, and especially lexicographic resources, will become more and more important in the coming years. This is something that ILC, with its decades-long experience in standardisation, is well placed to make an significant contribution to.

3. Semantic Web Standards

In addition to activities described above ILC also has a strong commitment towards the adoption of semantic web technologies. In 2015, the semantic layer of PSC (Del Gratta et al. 2015) as well as ItalWordNet (Bartolini, Del Gratta, and Frontini 2013) were published as Linked Open Data (LOD), the former using the lemon model[12]. Other notable Semantic Web resources in whose creation ILC has been instrumental are the GeoDomain Wordnets (Frontini, Del Gratta, and Monachini 2016) – which connect the Geonames ontology with the ItalWordNet and Princeton Wordnets – and the sentiment lexicon for Italian (Maks et al. 2014).

Moreover ILC also participates in the W3C activities of Ontolex-lemon[13], with a particular focus on the modelling of dictionaries as well as the representation of etymology and language change. More broadly, ILC has carried out work on the creation of resources for historical languages (for instance the creation of Ancient Greek Wordnet[14] and the publication of the Intermediate Liddell Scott lexicon (1896) as LOD[15]). This interest for semantic web technologies extends towards other aspects such as the modelling of ontologies with OWL and the use of the semantic web rule language (SWRL).

With respect to the former the institute has published the OWL version of the SIMPLE ontology (Toral and Monachini 2007). As to the modelling of rules, an ongoing project aims at the translation of Italian inflexional morphology using SWRL (Khan et al. 2017).

[11] The Italian MIUR nominated the Department of Humanities and Social Sciences of CNR as the National Representative and gave ILC-CNR the role of building the national data center and the national repository (ILC4CLARIN, https:// ilc4clarin.ilc.cnr.it/). Monica Monachini was nominated National Coordinator of the CLARIN-IT Consortium.

[12] http://hdl.handle.net/20.500.11752/ILC-66, http//www.languagelibrary.eu/owl/simple/

[13] https://www.w3.org/community/ontolex

[14] See Bizzoni et al. (2014, 2015) and Del Gratta et al. (2015).

[15] Khan et al. (2016).

4. Towards an Ecosystem of Lexical Resources

These initiatives should be seen within the broader context of a new convergence of the once closely aligned but laterly somewhat estranged communities of language resources and digital humanities. In particular, we seem to be witnessing a new convergence between computational approaches to lexicography and the needs of e-lexicography. As more and more language technologists are collaborating on digital humanities projects the necessity of making the main formats (TEI, LMF, Ontolex-lemon) interoperable becomes more important and, at the same time, the encoding of levels of information that are of particular interest for DH - such as the representation of diachronic knowledge and language change - becomes essential.

The coexistence of various different competing standards is always a source of worry. Moreover, current lexicographic resources, both modern and historical, have different levels of structuring and are not equally suitable for application in other fields. However, we believe that current trends seem to be consistent with the idea of an ecosystem, where different standards can coexist and mutually enrich each other, with

i. TEI being a format for representing a digital edition of the lexical resource,
ii. LMF the basic tool for actionable lexicons within LT, as well as in contexts where an official ISO standard is required, and
iii. Ontolex-Lemon the standard format for interconnected lexical networks, in which individual datasets can refer back to TEI sources (when they exist).

This intuition underlies the current vision that ILC is promoting, i.e. developing strategies, tools and standards for extracting, structuring and linking lexicographic resources to unlock their full potential for LOD and the Semantic Web, as well as in the context of Digital Humanities.

We aim to create a unified platform for interlinked lexical resources with a focus on Italian and on classical languages, where language resources are distributed in different formats for different purposes and are:

- accessible by web based query interface for linguists, lexicographers, students and the general public;
- downloadable in various formats (via the ILC4CLARIN repository[16]);
- exposed as LOD, browsable through a SPARQL query interface (as a service of CLARIN-IT) for lexicographic linked open data.

[16] The list of ILC-CNR lexical conceptual resources (mentioned here) is available in the ILC4CLARIN repo: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/discover?filtertype=type&filter_relational_operator=equals&filter=lexicalConceptualResource

The Linguistic LOD paradigm provides a suitable approach for the development of such an ecosystem.

5. References

Bartolini, Roberto, Riccardo Del Gratta, and Francesca Frontini. 2013. "Towards the Establishment of a Linguistic Linked Data Network for Italian." In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and Other Language Data*, 76–81. Pisa, Italy: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-5512.

Bizzoni, Yuri, Federico Boschetti, Riccardo Del Gratta, Harry Diakoff, Monica Monachini, and Gregory Crane. 2014. "The Making of Ancient Greek WordNet." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, 1140–47. Reykjavik, Iceland: ELRA. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1071_Paper.pdf.

Bizzoni, Yuri, Riccardo Del Gratta, Federico Boschetti, and Marianne Reboul. 2015. "Enhancing the Accuracy of Ancient Greek WordNet by Multilingual Distributional Semantics." In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-It 2015*, 47. Accademia University Press.

Calzolari, Nicoletta, Monica Monachini, and Claudia Soria. 2013. "LMF – Historical Context and Perspectives." In *LMF Lexical Markup Framework*, edited by Gil Francopoulo and Patrick Paroubek, 1–18. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118712696.ch1.

Del Gratta, Riccardo, Francesca Frontini, Fahad Khan, and Monica Monachini. 2015. "Converting the Parole Simple Clips Lexicon into Rdf with Lemon." *Semantic Web Journal* 6 (4):387–92.

Del Gratta, Riccardo, Federico Boschetti, Angelo Del Grosso, Fahad Khan, and Monica Monachini. 2015. "Cooperative Philology on the Way to Web Services: The Case of the CoPhiWordNet Platform." In *Worldwide Language Service Infrastructure*, edited by Yohei Murakami and Donghui Lin, 173–87. Lecture Notes in Computer Science 9442. Springer International Publishing. https://doi.org/10.1007/978-3-319-31468-6_13.

Francopoulo, Gil, ed. 2013. *LMF Lexical Markup Framework*. John Wiley & Sons.

Frontini, Francesca, Riccardo Del Gratta, and Monica Monachini. 2016. "GeoDomainWordNet: Linking the Geonames Ontology to WordNet." In *Human Language Technology. Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers*, edited by Zygmunt Vetulani, Hans Uszkoreit, and Marek Kubis, 9561:229–42. Lecture Notes in Computer Science (LNCS). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-43808-5_18.

Khan, Fahad, Andrea Bellandi, Francesca Frontini, and Monica Monachini. 2017. "Using SWRL Rules to Model Noun Behaviour in Italian." In *LDK 2017: Language, Data, and Knowledge*, 134–42. Lecture Notes in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-319-59888-8_11.

Khan, Fahad, Francesca Frontini, Federico Boschetti, and Monica Monachini. 2016. "Converting the Liddell Scott Greek-English Lexicon into Linked Open Data Using Lemon." In *Digital Humanities 2016: Conference Abstracts*, 593–96. Kraków: Jagiellonian University & Pedagogical University. http://dh2016.adho.org/abstracts/236.

Liddell, Henry George, and Robert Scott. 1896. *An Intermediate Greek-English Lexicon: Founded upon the Seventh Edition of Liddell and Scott's Greek-English Lexicon*. Harper & Brothers.

Maks, Isa, Ruben Izquierdo, Francesca Frontini, Rodrigo Agerri, Piek Vossen, and andoni Azpeitia. 2014. "Generating Polarity Lexicons with WordNet Propagation in 5 Languages." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA).

Toral, Antonio, and Monica Monachini. 2007. "Formalising and Bottom-up Enriching the Ontology of a Generative Lexicon." In *Proceedings of RANLP07-Recent Advances in Natural Language Processing*.