A Curated and Evolving Linguistic Linked Dataset

Abstract. This paper describes the Atlante Sintattico d'Italia, Syntactic Atlas of Italy (ASIt) linguistic linked dataset. ASIt is a scientific project aiming to account for minimally different variants within a sample of closely related languages; it is part of the Edisyn network the goal of which to establish a European network of researchers in the area of language syntax that use similar standards with respect to methodology of data collection, data storage and annotation, data retrieval and cartography. In this context, ASIt is defined as a curated database which builds on a dialectal data gathered during a twenty-year-long survey investigating the distribution of several grammatical phenomena across the dialects of Italy.

Both the ASIt linguistic linked dataset and the Resource Description Framework Schema (RDF/S) on which it is based are publicly available and released with a Creative Commons license (CC BY-NC-SA 3.0). We report the characteristics of the data exposed by ASIt, the statistics about evolution of the data in the last two years, and the possible usages of the dataset, such as the generation of linguistic maps.

Keywords: Linguistic Data, Curated Database, Part-Of-Speech and Sentence Tagging, Interoperability

1. Introduction

Studying languages increases our understanding of how humans communicate and store knowledge. For over a century, linguists have produced atlases showing the geographical distribution of linguistic features in the dialects of a language [13]. In the last two decades, several large-scale databases of linguistic material of various types have been developed worldwide and have been offered on-line to be shared by any research community. The World Atlas of Languages Structures (WALS) [8] is the first linguistic feature atlas on a world-wide scale and one of the largest projects with 160 maps showing the geographical distribution of structural linguistic features. In Europe, the Common Language Resources and Technology Infrastructure project (CLARIN) [14] aims at creating an infrastructure which makes language resources (annotated recordings, texts, lexica, ontologies) and technology (speech recognisers, lemmatisers, parsers, summarisers, information extractors) available and readily usable to scholars of all disciplines, in particular the humanities and social sciences. One of the most important applications of linguistic databases is linguistic cartography, the goal of which is to create geographical maps which visualizes particular linguistic features. These maps are usually produced either to study and safeguard world’s linguistic diversity or to display geographic distribution of syntactic variables and their potential correlations. An example of the former is The

1http://www.wals.info/
National Geographic’s Enduring Voices Project\(^2\) the aim of which is to preserve endangered languages by identifying language hot spots and documenting the languages and cultures within them. Unesco\(^3\) as well has made available an online tool to assess the status of linguistic diversity in the world. It provides pieces of information for each language like: the name, the degree of endangerment, the countries where it is spoken, and the geographic coordinates. Other important online projects which refer to the problem of displaying correlations of linguistic features are: VIVALDI,\(^4\) DynaSAND,\(^5\) and the above mentioned WALS. The scientific value of these linguistic projects is undoubted; nevertheless, the use and the distribution of their data is very limited: users can only generate maps and save them as figures, and in a few cases export geographical XML files (as in the case of the Unesco project). The data of these systems are neither browsable nor exportable. Two recent international initiatives have started to tackle these issues: the Edisyn network\(^12\), the goal of which is to establish a European network of researchers in the area of language syntax that use similar standards with respect to methodology of data collection, data storage and annotation, data retrieval and cartography; the ISOcat\(^6\) linguistic concept database, developed by ISO Technical Committee 37, provides a reference to create a universally available resource for language-related metadata that can be used in a variety of applications and environments\([10]\). Furthermore, in recent years, the interoperability of linguistic resources has become a major topic in the several scientific fields, for instance computational linguistics and Natural Language Processing\([7]\). The different representation and management choices made by each linguistic project act as barriers toward the integration of all their linguistic resources. Furthermore, the lack of interoperability prevents the possibility of develop and exploit common analysis tool based on the linguistic data.

Exposing linguistic data as Linked Open Data enhances the interoperability between existing linguistic datasets and allows for their integration with other RDF resources such as lexical-semantic resources already available as Linked Data, e.g., a general knowledge base like DBpedia, or linguistic resources like WordNet or Wiktionary\([7]\). In this paper, we address the problem of the design and distribution of language resources by adopting an approach based on the Linked Open Data (LOD) paradigm\([9]\) exploiting its ability to enable interoperability at a data-level by overcoming the single collections characteristics and the particular system and its technological choices. We focus on the Atlante Sintattico d’Italia, Syntactic Atlas of Italy (ASI) enterprise\([2]\) a scientific project carried on as a part of the Edisyn network. We define a mapping from a conceptual model of the ASI linguistic curated database to a Resource Description Framework (RDF) schema, thus providing an instrument to expose linguistic data as LOD. This RDF schema defines a common layer allowing different linguistic projects to read, manipulate and re-use diversified linguistic data. Furthermore, the RDF schema allows us to expose the ASI linguistic database as a curated and evolving linked dataset.

The paper is organized as follows: Section 2 reports on the issues should be addressed to guarantee the quality of the data in the linguistic domain. Section 2.1 presents the ASI enterprise highlighting its main features and the method of creation and maintenance of the data. Section 3 reports the main characteristics and statistics about the ASI Linguistic Linked Dataset. Lastly, Section 4 draws some final remarks.

2. Linguistic Curated Data

Language resources that have been made publicly available can vary in the richness of the information they contain: on one hand, a corpus typically contains at least a sequence of words, sound or tags; on the other end, a corpus may contain a large amount of information about the syntactic structure, morphology, prosody, and semantic content of every sentence, plus annotation of discourse relations or dialogue acts\([5]\). However, the quality of such corpora may have been reduced by the intense, and often not well controlled, usage of automatic learning algorithms\([15]\). Depending on the type of analysis a researcher performs, linguistic datasets created by automatic Part-Of-Speech (POS) tagger can be either helpful or useless. For example, POS annotations are very important for performing particular linguistic analyses such as capturing fine-grained grammatical differences by comparing various dialectal translations of the same sentence.

\(^2\)http://travel.nationalgeographic.com/travel/enduring-voices/
\(^3\)http://www.unesco.org/new/en/culture/themes/endangered-languages/
\(^4\)http://www2.hu-berlin.de/vivaldi/
\(^5\)http://www.meertens.knaw.nl/sand/
\(^6\)http://www.isocat.org/
In these cases, even an accuracy of 98% of the best automatic POS taggers is not sufficient to pin down these subtle asymmetries. This specificity can only be reached manually [3].

The preparation of a linguistics resource of high quality requires several steps: crawling, downloading, cleaning, normalizing, and annotating the data are some of the actions that need to be performed to produce valuable content [11]. Some of these steps do require human intervention to achieve the highest quality possible for a resource of usable scientific data. Curated databases [6] are a possible solution for designing, controlling and maintaining collections that are consistent, integral and high quality. To this purpose, Bird et al. [5] discuss three important points about the design and distribution of language resources:

- How do we design a new language resource and ensure that its coverage, balance, and documentation support a wide range of uses?
- When existing data is in the wrong format for some analysis tool, how can we convert it to a suitable format?
- What is a good way to document the existence of a resource we have created so that others can easily find it?

In the context of the ASIt enterprise, these issues are addressed by adopting an approach based on the LOD paradigm with the aim of enabling interoperability at a data-level by overcoming the single collections characteristics depending by different methodological and technological choices.

2.1. The ASIt Curated Database

The ASIt enterprise builds on a long standing tradition of collecting and analyzing linguistic corpora, which has originated different efforts and projects over the years [3,2,1]. Dialectal data stored in the ASIt were gathered during a twenty-year-long survey investigating the distribution of several grammatical phenomena across the dialects of Italy [4]. Research on the syntax of Italian is of great interest to several important lines of research in linguistics: it allows comparison between closely related varieties (the dialects), hence the formation of hypotheses about the nature of cross-linguistic parametrization; it allows contact phenomena between Romance and Germanic varieties to be singled out, in those areas where Germanic dialects are spoken; it allows syntactic phenomena of Romance and Germanic dialects to be found, described and analyzed to a great level of detail [1].

At present, there are eight different questionnaires written in Italian and almost 500 questionnaires, corresponding to the eight Italian questionnaires, written in more than 240 different dialects, for a total of more than 50,000 sentences and more than 40,000 tags stored in the data resource managed by the ASIt digital library system.

In order to efficiently store and manage the amount of data recorded in the questionnaires, the interviews and the tagged sentences, ASIt has been realized as a linguistic curated database. ASIt curated database is organized in three main conceptual areas:

- The geographical area, which is the place where a given dialect is spoken and where a speaker is born;
- the derivation area, which focuses on the background of the speaker: the level of knowledge of the dialect, the particular variety of the dialect, the birthplace, the ancestors, the document that she/he translated;
- the tagging area, which is how the document is structured and how it has been tagged (at a sentence level and at a word level).

A relevant aspect of the ASIt curated database is that it explicitly models sentence level tagging, which is not modeled by any other of the presented linguistic projects. Furthermore, we have developed a language-specific set of POS tags which is suitable for ASIt dialectal data but, at the same time, allows these data to be linked to other databases of dialect syntax. We can therefore imagine the creation of a language-specific tagset as starting from a universal core, shared by all languages, and subsequently developing a language-specific periphery, which is compatible with other databases and able to classify language-specific structures.

3. The Linguistic Linked Dataset

The ASIt curated database was the starting point for defining the RDF/S underlying the ASIt Linguistic Linked Dataset we present in this work. In Figure 1 we report the main classes and properties defining the RDF/S, whereas in Table 2 we report the Ontology Web
Language (OWL) data type properties of the presented classes.

As far as the vocabulary adopted in this specification is concerned, we use the namespaces and prefixes reported in Table 1; `asit` is the only vocabulary which is not inherited from other domains. RDF assumes that any instance of a class may have an arbitrary number (zero or more) of values for a particular property.

As an extension of RDF/S, OWL allows us to specify the maximum number of occurrences of a class within a property. Since in ASIt this number is either 1 or \(n\), we use the `owl:onProperty` from the OWL vocabulary to specify an `owl:cardinality` equal to 1. For instance an `Actor` can be born in one and only one `Town` or a `Province` can be in one and only one `Region`.

The complete RDF/XML serialization of the RDF/S specification is publicly available at the following URL:

http://purl.org/asit/rdf/asit-schema.rdf
Table 2
Main data type properties of the classes of the schema of Fig. 1.

<table>
<thead>
<tr>
<th>Area</th>
<th>Class</th>
<th>OWL Datatype Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>geographical</td>
<td>Region</td>
<td>gn:officialName, asit:GeographicPartition, asit:regionNotes</td>
</tr>
<tr>
<td></td>
<td>Province</td>
<td>gn:officialName, gn:shortName, asit:provinceNotes</td>
</tr>
<tr>
<td></td>
<td>Dialect</td>
<td>asit:dialectName</td>
</tr>
<tr>
<td>tagging</td>
<td>document</td>
<td>dcterms:title, dcterms:date</td>
</tr>
<tr>
<td></td>
<td>sentence</td>
<td>asit:sentence, asit:transcription, asit:sentenceNotes</td>
</tr>
<tr>
<td></td>
<td>word</td>
<td>asit:wordText, asit:transcription</td>
</tr>
<tr>
<td></td>
<td>tag</td>
<td>asit:tagDescription, asit:mandatory</td>
</tr>
</tbody>
</table>

We exploited this RDF/S to expose the linguistic data in the ASIt curated database as a Linked Dataset whose details are reported in Table 3.

The ASIt curated database is synchronized with the ASIt Linguistic Linked Dataset, where every entity in the database corresponds to a class in the linked dataset; therefore, the dataset is maintained following the same policies adopted for the database, ensuring the quality and the freshness of the exposed data. To this purpose the ASIt enterprise is provided with an RDF layer which is responsible for persistence and access to RDF triples. A synchronization service allows for the interaction with the RDF datastore which is responsible for the persistence of the RDF/S instantiation in the RDF Store. Therefore, the operations required by resource creation, deletion or modification are performed in parallel for each request to guarantee the synchronization between the database and the RDF store. As a consequence, the ASIt Linguistic Linked Dataset size grows proportionally to the size of the data in the curated database: the number of entries associated with a database entity is related to the number of instances of the RDF class we mapped from it. Since the research activities on the linguistic ASIt database is still ongoing, the number of documents and sentences is increasing over time as well as the tags the linguistic researchers associate with them. Table 4 reports the statistics about the evolution of the data in ASIt in the last two years. These statistics do not present data about the actors involved in the linguistic activities, which include dialectal speakers and data curators; these data is not exposed because of privacy issues.

This dataset has been exposed following the guidelines presented in [9]. As an example we report how it is possible to access and browse a resource referring to the resource named “Veneto” which is an instance of the class “Region”. It is possible to access the “Veneto” resource by means of three URIs, which are:
The authors wish to thank Maristella Agosti for her support and contribution in the design and development of the ASIt Digital Library. This work has been partially supported by the Project FIRB “Un’inchiesta grammaticale sui dialetti Italiani: ricerca sul campo, gestione dei dati, analisi linguistica” (cod. RBFR08KRA_003), and the PROMISE network of excellence (contract n. 258191).

References