Undefined X (2012) 1–X
IOS Press

A Curated and Evolving Linguistic Linked Dataset

Editor(s): Pascal Hitzler, Kno.e.sis Center, Wright State University, Dayton, OH, U.S.A.; Krzysztof Janowicz, University of California, Santa Barbara, U.S.A.
Solicited review(s): Ivan Herman, W3C; Marta Sabou, MODUL University Vienna, Austria; Jesse Weaver, Rensselaer Polytechnic Institute, U.S.A.

Emanuele Di Buccio, Giorgio Maria Di Nunzio and Gianmaria Silvello
University of Padua, Department of Information Engineering, Italy
E-mail: {dibuccio, dinunzio, silvello}@dei.unipd.it

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 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. In this context, ASIt is defined as a curated database which builds on 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 the 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.[1] 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 visualize particular linguistic features. These maps are usually produced either to study and safeguard the world's linguistic diversity or to display the geographic distribution of syntactic variables and their potential correlations. An example of the former is The National Geographic's Enduring

[1] http://www.wals.info/

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. The tool 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; and 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 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 developing and exploiting common analysis tools 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] and 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 (ASIt) enterprise [1], a scientific project carried on as a part of the Edisyn network. We define a mapping from a conceptual model of the ASIt 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 present the ASIt linguistic database as a curated and evolving linked dataset.

The paper is organized as follows: Section 2 reports on the issues that should be addressed to guarantee the quality of the data in the linguistic domain. Section 2.1 presents the ASIt 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 ASIt 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 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 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 an 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. In these cases, even an accuracy of 98% of the

[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/

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[7] [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 into 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 on 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 given rise to different efforts and projects over the years [3,1,2]. 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 [2].

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. The 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.

[7] A curated database is a database the content of which has been collected by a great deal of human effort.

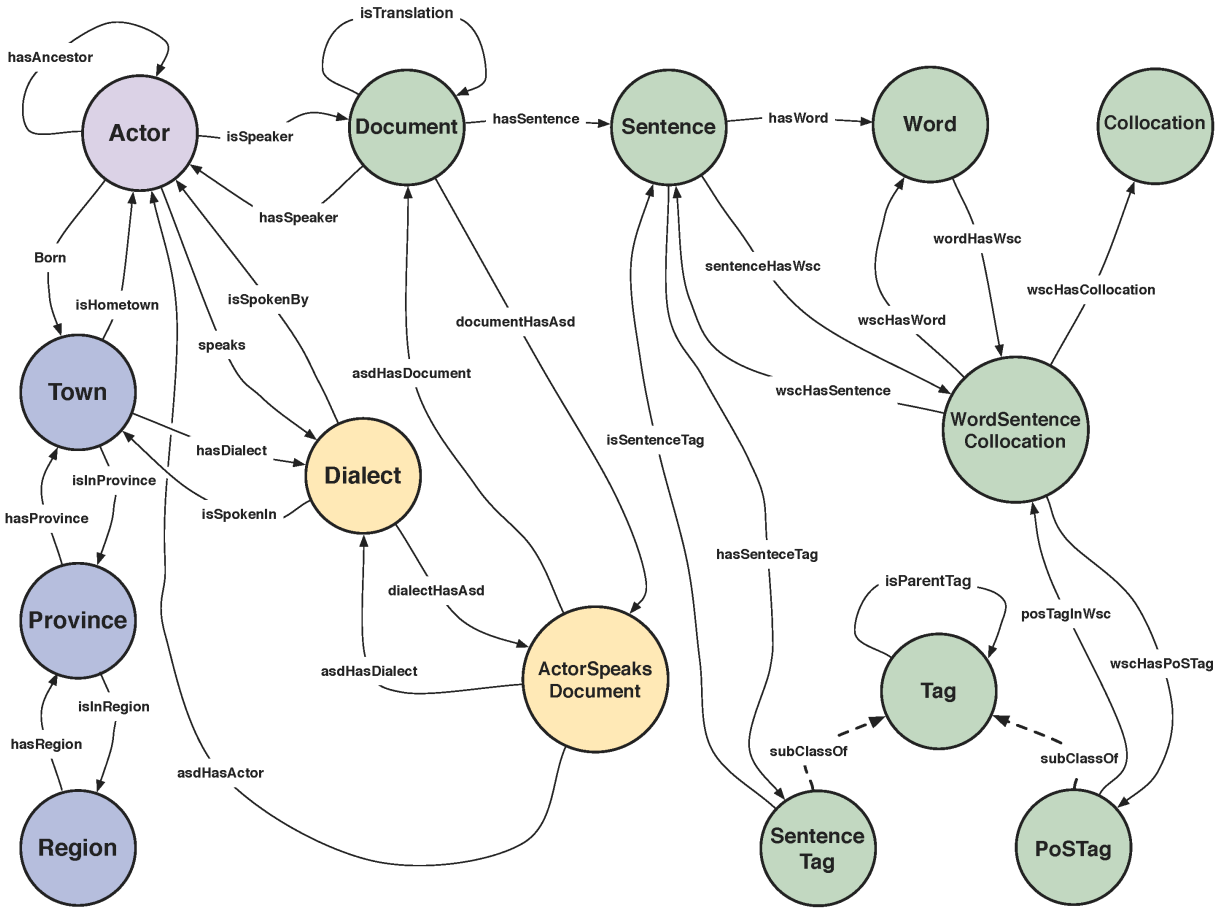

Fig. 1. Diagram representing the RDF/S defined for the ASIt enterprise.

Table 1
Namespaces and Prefixes adopted by the ASIt RDF Specification.

Prefix	Namespace	Description
asit	http://purl.org/asit/terms/	ASIt vocabulary terms
dcterms	http://purl.org/dc/terms/	Dublin Core terms
foaf	http://xmlns.com/foaf/0.1/	Friend of a friend
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	WGS84 Geo Positioning
gn	http://www.geonames.org/ontology#	GeoNames Ontology
owl	http://www.w3.org/2002/07/owl#	OWL vocabulary terms
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF vocabulary terms
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema

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

Table 2
Main data type properties of the classes of the schema of Fig. 1.

Area	Class	OWL Datatype Properties
geographical	Region	gn:officialName, asit:geographicPartition, asit:regionNotes
	Province	gn:officialName, gn:shortName, asit:provinceNotes
	Town	gn:officialName, geo:alt, geo:lat, geo:long, gn:population, asit:townNotes, asit:provinceCapital, asit:provinceLittoral, asit:altimetricArea, asit:mountainTown, asit:surface, asit:latitudine, asit:longitudine
	Dialect	asit:dialectName
derivation	Actor	foaf:firstName, foaf:lastName, foaf:name, foaf:birthday, foaf:gender, foaf:mailbox, asit:placeOfBirth, asit:education, asit:job, asit:country, asit:lang, asit:actorNotes, asit:affiliation
tagging	document	dcterms:title, dcterms:date
	sentence	asit:sentence, asit:transcription, asit:sentenceNotes
	word	asit:wordText, asit:transcription
	collocation	asit:position
	tag	asit:tagDescription, asit:mandatory

Table 3
Details of the ASIt Linguistic Linked Dataset.

Name	ASIt Linguistic Linked Dataset
URL	`http://purl.org/asit/alld`
Ver. No	1.02
Ver. Date	2012-08-03
Licensing	Creative Commons License Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)
Availability	Public
#Triples	421,948
Size	38.3 MB

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`

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 are 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 are not exposed because of privacy issues.

This dataset has been presented following the guidelines 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:

1. `http://purl.org/asit/resource/Region/Veneto`
2. `http://purl.org/asit/data/Region/Veneto`
3. `http://purl.org/asit/page/Region/Veneto`

Table 4
Statistics about the growth of main entities/relationships of the ASIt curated database

	Jan '11	Jul '11	Jan '12	Jul '12
tags	524	530	532	532
documents	462	468	512	540
sentences	47,973	48,575	51,256	54,091
tags/sentences	10,364	16,731	18,080	18,369
tags/words	0	5,411	18,509	27,046

The first of the three URIs identifies the non-information resource "Veneto"; dereferencing the first URI asking for `application/rdf+xml` gives the user, after a redirect, an RDF description of `http://purl.org/asit/resource/Region/Veneto` within the information resource `http://purl.org/asit/data/Region/Veneto`. In the event of a Web browser, the user is redirected to `http://purl.org/asit/page/Region/Veneto` information resource which is an HTML representation of "Veneto"; the HTML representations are made available through a browser embedded in the ASIt enterprise; the browser allows other resources within the ASIt dataset and resource of external datasets to be accessed through hyperlinks. Currently, the ASIt dataset is linked to DBpedia: the instances of the classes "Region", "Province" and "Town" are linked to the corresponding instances of the dbpedia.org class "Place" through the property `owl:sameAs`.

A SPARQL endpoint is provided at the URL: `http://purl.org/asit/rdf/sparql` and a GUI to submit queries to the ASIt Linguistic Dataset is available at the URL: `http://purl.org/asit/rdf/sparqlGui`.

4. Final Remarks

The Linguistic Linked Dataset we presented in this work aims to enable interoperability at a data-level by overcoming the single linguistic project boundaries depending on different methodological and technological choices. We imagine the use of the ASIt Linguistic Linked Dataset by third-party linguistic projects in order to enrich the data and build up new services over them. An example of a possible use of a new service is linguistic cartography.

ASIt aims to generate linguistic maps and also to expose the linguistic data by leaving users free to generate whatever type of map or analysis tool they like. In addition, the Linguistic Linked Dataset we presented allows for the investigation of correlation among geographical, linguistic and user data.

Acknowledgments

The authors wish to thanks 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

[1] Maristella Agosti, Birgit Alber, Giorgio Maria Di Nunzio, Marco Dussin, Diego Pescarini, Stefan Rabanus, and Alessandra Tomaselli. A Digital Library of Grammatical Resources for European Dialects. In Maristella Agosti, Floriana Esposito, Carlo Meghini, and Nicola Orio, editors, *IRCDL*, volume 249 of *Communications in Computer and Information Science*, pages 61–74. Springer, 2011.

[2] Maristella Agosti, Birgit Alber, Giorgio Maria Di Nunzio, Marco Dussin, Stefan Rabanus, and Alessandra Tomaselli. A Curated Database for Linguistic Research: The Test Case of Cimbrian Varieties. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[3] Maristella Agosti, Paola Benincà, Giorgio Maria Di Nunzio, Riccardo Miotto, and Diego Pescarini. A Digital Library Effort to Support the Building of Grammatical Resources for Italian Dialects. In Maristella Agosti, Floriana Esposito, and Costantino Thanos, editors, *IRCDL*, volume 91 of *Communications in Computer and Information Science*, pages 89–100. Springer, 2010.

[4] Paola Benincà and Cecilia Poletto. The ASIS Enterprise: A View on the Construction of a Syntactic Atlas for the Northern Italian Dialects. *Nordlyd. Monographic issue on Scandinavian Dialects Syntax*, 34(1), 2007.

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 1st edition, July 2009.

[6] Peter Buneman, James Cheney, Wang Chiew Tan, and Stijn Vansummeren. Curated Databases. In Maurizio Lenzerini and Domenico Lembo, editors, *Proc. of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2008*, pages 1–12. ACM Press, New York, USA, 2008.

[7] Christian Chiarcos. Interoperability of Corpora and Annotations. In Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff, editors, *Linguistic Linked Data*, pages 161–179. Springer, 2012.

[8] Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. *The World Atlas of Language Structures*. Oxford University Press, United Kingdom, 2005.

[9] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011.

[10] Marc Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. ISOcat: remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies*, 4(4):261–276, November 2009.

[11] Adam Kilgarriff. Googleology is Bad Science. *Computational Linguistics*, 33(1):147–151, March 2007.

[12] Jan Pieter Kunst and Franca Wesseling. The Edisyn Search Engine. *Language Variation Infrastructure*, 3(2):63–74, 2011.

[13] Alfred Lameli, Roland Kehrein, and Stefan Rabanus. *Language and space. Vol. 2: Language mapping*. De Gruyter Mouton, 2010.

[14] Jan Odijk. Recent Developments in CLARIN-NL. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[15] Karen Spärck Jones. Computational Linguistics: What About the Linguistics? *Computational Linguistics*, 33(3):437–441, September 2007.