Modelling and Linking an Old Latin-Portuguese Dictionary to the LiLa Knowledge Base

Lucas Dezotti[1], Marco Passarotti[2], Francesco Mambrini[2]
[1] Universidade Estadual Paulista, [2] Università Cattolica del Sacro Cuore
[1] Rod. Araraquara-Jaú km 1, 14800 Araraquara, Brazil, [2] Largo Gemelli 1, 20123 Milan, Italy
lucas.dezotti@unesp.br, {marco.passarotti, francesco.mambrini}@unicatt.it

Abstract

This paper describes the steps undertaken to include data from Antonio Velez's bilingual Latin-Portuguese dictionary (*Index Totius Artis*, 1744) into the LiLa Knowledge Base of interoperable linguistic resources for Latin. The paper focuses on how the lexical and lexicographic information of the source dictionary was modelled by using respectively the Lexicon Model for Ontologies (OntoLex-lemon) and its *lexicog* module. The linking process of the dictionary entries with those of the LiLa collection of Latin lemmas is detailed, discussing issues in dealing with ambiguities and typographical errors found in the source. The result is the first Latin-Portuguese lexical resource made interoperable with the (meta)data of the other linguistic resources for Latin interlinked in the LiLa Knowledge Base, providing new ways of assessing the dictionary information or using its content as starting point to explore the connections with other interlinked linguistic resources. A couple of use case scenarios illustrate those possibilities.

Keywords: Linguistic Linked Open Data, Computational Lexicography, OntoLex-lemon, Latin-Portuguese Bilingual Dictionaries, Index Totius Artis, Antonio Velez

1. Introduction

It is well-known that Latin played different roles in European cultural and linguistic history, first as the language of ancient Rome, then as the language of the Catholic church, and finally as the language of the Humanist and Modern scholarship and sciences. The two millennia of its sustained activity have generated an innumerable quantity of Latin written records in a variety of genres, a great amount of which have survived the test of time. Partly because of this, Latin was also the language of literacy and education in Western Europe for a long time. Hence, there have been many attempts to describe its linguistic system and usage, either for scientific or pedagogical purposes, resulting in a large set of grammars, teaching methods and vocabulary aids such as thesauri, lexica, glossaries and dictionaries. In particular, the Latin-vernacular bilingual dictionaries have played the dual role of functioning as a lexical reference for the emerging vernacular cultures in Europe as well as participating in the first experiences of printed lexicography.

Despite the presumed traditionalism of the field, part of the Classical Philology community has been interested in using computers to process this vast linguistic heritage since the avant-garde of Humanities Computing. Roberto Busa's *Index Thomisticus* project started in 1949 (Busa, 1974-1980), the indexing and statistical researches developed at the LASLA laboratory in Liège from the 1960's (Denooz, 2004), as well as David Packard's *Concordance to Livy* (Packard, 1968) are some of the pioneering examples. Moreover, the community has kept pace with the scientific technological and cultural transformations, from the first collections of Latin texts made available on the nascent World Wide Web in the 1990's (e.g. The Latin Library[1] and the Perseus Digital Library[2]) to the most recent practice of publishing linguistic resources according to the FAIR principles (findability, accessibility, interoperability and reusability, cf. Wilkinson et al., 2016), such as the LiLa Knowledge Base[3], in which the information from a set of Latin linguistic resources is described using common knowledge representation vocabularies to ensure interoperability between them and maximise their use (Passarotti et al., 2020).

The conversion of raw text into structured data has been a key moment in this process, since it has facilitated the circulation and reuse of linguistic resources. As regards Latin dictionaries, the digital edition of the Lewis and Short (1879) is a model case. Its TEI-XML edition, developed by the Perseus Project[4], has been used to integrate the dictionary entries into various websites and tools (e.g. Logeion[5], Latinitium[6], Collatinus[7],

[1] http://www.thelatinlibrary.com/

[2] http://www.perseus.tufts.edu/hopper/

[3] https://lila-erc.eu/

[4] Available at https://github.com/PerseusDL/lexica/. The Text Encoding Initiative (TEI) standards are available at https://tei-c.org/

[5] https://logeion.uchicago.edu/

[6] https://latinitium.com/

[7] http://outils.biblissima.fr/collatinus/.

LREC-COLING 2024, pages 11537–11547
20-25 May, 2024.

Diogénes[8], Scaife Viewer[9], Alpheios[10], among others). More recently, the same TEI-XML file was used as data source for a new digital edition (Mambrini et al., 2021) modelled using the Lexicon Model for Ontologies or simply OntoLex-lemon (Cimiano et al., 2016) and linked to the LiLa Knowledge Base, allowing the dictionary to be simultaneously queried with corpora documenting the attestations of any given lexical item.

Like Lewis and Short's dictionary, there are many lexical resources for Latin and beyond that are already available on the Web but not yet published in the Web, then lacking interoperability. The Latin-Portuguese dictionaries available in the Corpus Lexicográfico do Português website (CLP) (Verdelho and Silvestre, 2002), a collection of Portuguese lexicographic works compiled from the 16th to the 19th centuries, are among them. At present, the CLP provides an HTML-based interface to browse and concordance the dictionaries it contains, but its data is not structured in such a way as to be semantically interoperable with those of other lexical resources at the most granular level (i.e., that of the individual lexical entry).

The current state of the art for making shared resources interact with each other is modelling and publishing them according to the principles of the so-called Linked Data paradigm (Berners-Lee et al., 2001; Berners-Lee, 2006). The Linguistic Linked Open Data Cloud[11] provides an overview of the linguistic resources currently published as Linked Open Data. Despite the large number of resources included in the Cloud (most of which are published based on common vocabularies for knowledge description), in most cases they are interlinked at a still quite coarse-grained level, limited at descriptive metadata. One step forward towards a deep-level interoperability among distributed resources consists in interlinking them at the level of single word occurrences in corpora and entries in lexical resources. This is exactly what the LiLa Knowledge Base makes possible for Latin linguistic resources. As a consequence, a relevant task that is currently undertaken in LiLa is interlinking new textual and lexical resources in the Knowledge Base.

This paper describes the steps undertaken for modelling and publishing as Linked Data in the Lila Knowledge Base one lexical resource, namely the Latin-Portuguese dictionary *Index Totius Artis* (Velez, 1744) provided by the CLP collection.

In the next section, we present the *Index* in its main characteristics and structural properties. Section 3 provides an overview of how the lexical resources are connected to LiLa and describes the strategies adopted for modelling the *Index* data and linking its entries to LiLa. Section 4 presents a couple of query examples that illustrate the new ways of assessing the dictionary information and exploring its connection with other interlinked linguistic resources provided by its publication as Linked Data. Finally, the concluding remarks outline some directions for future work.

2. The *Index Totius Artis*

The *Index Totius Artis* is a very small Latin-Portuguese dictionary produced in the course of the 17th century. As its name suggests, it was first conceived as an index of Latin words to the revised Portuguese edition of Manuel Alvares' famous Latin grammar *De institutiones grammaticae libri tres*, curated by Antonio Velez (then head of the University of Evora) and published in 1599 – the same year the *Ratio Studiorum* established Alvares' work as the official grammar book for all the Jesuit colleges over the world (Springhetti, 1961–1962; Kemmler, 2018; Salor and Gómez, 2020). Over the decades, what was a simple list of headings and associated references to the grammar pages was converted into a dictionary thanks to the addition of Latin definitions, Portuguese equivalents, information on verb complementation and phraseological examples[12]. The 1744 edition[13] we used as our data source is one of the latest: Velez's work was soon banned along with Alvares' grammar, following the Jesuit's expulsion from Portugal, in 1759.

This historical background helps to explain some of its main lexicographical features, as they have implications on the data structuring, especially the non-uniform way information is provided. Even though conveying all information types a lexicographic entry is expected to have (Hartmann, 2001) – namely, a headword, its formal characteristics (spelling, pronunciation, morphology) and its semantic properties (meaning and usage in particular contexts) –, none of these are provided regularly, nor are they even placed in the same position relatively to the entry structure. In addition, many dictionary entries provide information on multiple lemmas by means of text-delimited subentries[14].

[8] https://d.iogen.es/web

[9] https://scaife.perseus.org

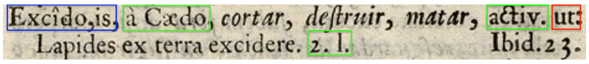
[10] https://alpheios.net

[11] https://linguistic-lod.org/llod-cloud

[12] The major improvements are found in the 1608 and the 1689 editions. On the *Index* creation process, refer to Iken, 2002.

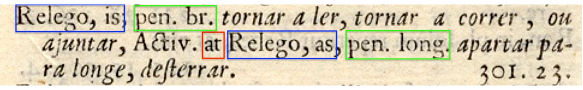
[13] Available on the CLP website (Velez, 2002). The CLP mentions "1599?" as the original source publication year; however, the 1744 edition was found to be the only one that completely matches with the content of the digital text.

[14] For the purposes of this article we adopt the Měchura (2023) definition of subentry as 'any element



Excîdo,is, à Cædo, *cortar, deſtruir, matar,* aƈtiv. ut:
Lapides ex terra excidere. 2. l. Ibid. 23.

Figure 1: Lexicographic entry for *excido*, with boxes highlighting, respectively, the lemma (blue line), the Latin connector 'ut' (red line) introducing a usage example, and other information (green line) surrounding the definition (in italics). Source: Velez (1744).



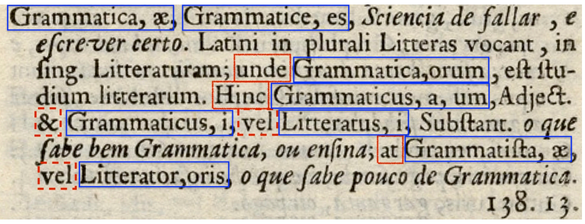
Relego, is; pen. br. *tornar a ler, tornar a correr, ou*
ajuntar, Aƈtiv. at Relego, as, pen. long. *apartar pa-*
ra longe, deſterrar. 301. 23.

Figure 2: Lexicographic entry for two homographs, with line boxes highlighting, respectively, the lemmas (blue line), the prosodic information (green line) preceding the definition (in italics), and the Latin connector 'at' (red line) delimiting the beginning of the subentry. Source: Velez (1744).

Actually, text delimiters are also extensively used for separating alternative forms, multiple senses, usage examples. Moreover, the language of description can be Latin or Portuguese, alternately. A couple of examples illustrate the point.

Figure 1 shows the dictionary entry for the Latin verb *excido* ('to cut'). As is costumary in traditional Latin lexicography, the lemma consists of a set of forms ('Excîdo, is'), providing both the canonical citation form of the word and its inflectional paradigm (*excido*, 3rd conjugation verb). The lexicographic description can be divided into five pieces of information, mostly in Latin: (a) information on its derivation base ('à Caedo', lit. 'from *Caedo*'); (b) information on its meaning through three Portuguese translation equivalents (lit. 'to cut, to destroy, to kill'), printed in italics; (c) information on verb complementation ('activ.' stands for *activum*, an ancient grammar label for verbs governing accusative[15]); (d) a usage example ('Lapides ex terra excidere') introduced by the Latin particle *ut* (lit. 'such as'); (e) information on prosody of the lemma ('2. l.' means the second syllable is long). The entry ends with a right-aligned locator, referring to Alvarez's grammar page ('Ibid. 23').

Figure 2 is an example of a dictionary entry embedding a subentry. It arranges the information on two homographs ('Relego, is', 3rd conjugation, and 'Relego, as', 1st conjugation verb) into two blocks similarly structured and delimited by the Latin connector *at* (lit. 'but'). Unlike the previous example, the first pieces of information, provided right after each lemma, are on prosody: 'pen. br.' and 'pen. long.' indicate, respectively, the length of the word's last syllable but one (lat. *paenultima*): short (*brevis*) for the former, long (*longa*) for the latter. Next, both entry and subentry are provided with sets of Portuguese translation equivalents (printed in italics), but only the main entry has its verb complementation indicated ('Activ.'). The last piece of information is the locator, set apart on the right, as usual.



Grammatica, æ, Grammatice, es, *Sciencia de fallar, e*
eſcre-ver certo. Latini in plurali Litteras vocant, in
ſing. Litteraturam; unde Grammatica,orum, eſt ſtu-
dium litterarum. Hinc Grammaticus, a, um, Adjeƈt.
& Grammaticus, i, vel Litteratus, i, Subſtant. *o que*
ſabe bem Grammatica, ou enſina; at Grammatiſta, æ,
vel Litterator,oris, *o que ſabe pouco de Grammatica.*
138. 13.

Figure 3: Lexicographic entry for *grammatica*, with boxes highlighting, respectively, the lemma (blue line), the Latin connectors delimiting subentries (solid red line), the Latin connectors delimiting the alternative lemmas that shares the same sense (dashed red line). Source: Velez (1744).

Figure 3 shows how intricate the entry structure can be when embedding multiple lemmas. The entry for the Greek loanword *grammatica* associates eight lemmas and four sense units by meaningfully using Latin connectors as delimiters. It can be divided into four blocks, based on the four senses it conveys. The first block describes the main entry, which refers to two alternative lemmas: the first ('Grammatica, ae') is the adapted form to Latin morphology of the second ('Grammatice, es'), which keeps the Greek inflection pattern; they are followed by a definition in Portuguese (in italics) and a note in Latin informing on Latin equivalents for the headword (namely, 'litterae' and 'litteratura'). The second block describes the lemma 'Grammatica, orum', whose Latin definition ('est studium litterarum') is associated with those Latin equivalents by means of the Latin adverb *unde* (lit. 'from which') used as a delimiter. Then, a third block is introduced by *hinc* (lit. 'hence') and consists of three quite different lemmas sharing the same Portuguese definition (lines 5-6, in italics, meaning 'who has a profound knowledge of grammar, or teaches it'): the first two are the homographs 'Grammaticus, a, um' (1st class adjective) and 'Grammaticus, i' (2nd declination noun), coordinated by the Latin conjunction *et* (in its ligature form '&'); the third ('Litteratus, i') is the Latin equivalent for the second, so they are coordinated by the Latin conjunction *vel* (lit. 'or'). A final block providing information on a pair of lemmas also coordinated by *vel* ('Grammatista, ae' and its Latin equivalent 'Litterator, oris') is introduced by the conjunction *at* (lit. 'but'), which is in line with its contrasting

inside a dictionary entry which has its own headword' and whose presence '*override* the entry's headword and provides its own'.

[15] Cf. Colombat (2003, p. 73)

definition to the previous one (lit. 'who has a superficial knowledge of grammar'). The entry is ended by the locator in the usual position.

The diversity of types and positioning of information, as well as the existence of multiple-lemma entries, pose some challenges to the task of data modelling, especially when seeking to preserve both the lexical description and structural information from the source. The strategies adopted to achieve this depend on the way the lexical resources are modelled and linked to the LiLa Knowledge Base, which is discussed in the next section.

3. Lexical resources in the LiLa Knowledge Base

The LiLa Knowledge Base is a Linked Open Data platform of linguistic resources for Latin. Its core consists of a Lemma Bank, a comprehensive collection of approximately 215,000 Latin headwords. Since lemmatisation is a common layer of annotation for both lexical and textual resources, this collection serves as a connection point for distributed linguistic resources in LiLa. Ultimately, the interoperability is achieved by linking all entries in lexical resources and tokens in corpora to their corresponding lemma in the Lemma Bank.

Resources are published as Linked Data according to the data model provided by the Resource Description Framework (RDF)[16], a W3C standard model for representing, describing and sharing data. In particular, RDF is based on the idea of making statements about resources on the Web in the form of subject-predicate-object expressions, known as triples; a triple is composed of: (1) a subject, generally represented by a Uniform Resource Identifier (URI) that uniquely identifies a resource on the Web,[17] (2) a predicate, representing the relationship or property asserted about the subject and also identified by a URI, (3) an object, which can be either another resource or a data property. RDF triples can also be conceptualised as directed, labelled graphs, with the subject and object as nodes connected by the predicate, and are searchable using a dedicated query language, called SPARQL[18].

The lexical resources linked to LiLa are modelled according to the OntoLex-lemon model (Cimiano et al., 2016), which provides the vocabulary to represent linguistic information related to ontology and vocabulary elements. It consists of a set of modules and vocabularies designed to address various types of content of lexical entries. The core vocabulary provided by the OntoLex-lemon model is *ontolex*. Its main class, `LexicalEntry`[19], represents a unit of analysis of the lexicon, usually consisting of a set of grammatically related forms and a set of base meanings that are associated with all of these forms. Each grammatical (inflected) form of a lexical entry is associated with its `LexicalEntry` by means of the property `lexicalForm`[20], but only one form can be linked by the property `canonicalForm`[21], which indicates the form that is used as the canonical form of citation for the lexical entry (in other words, the lemma). Given the central role played by lemmas in the LiLa Knowledge Base, `canonicalForm` is the property used to link the entries of the lexical resources for Latin to the LiLa Lemma Bank. In other words, lexical resources connected to LiLa are modelled as collections of lexical entries, each of which is linked to a lemma of the LiLa Knowledge Base via the `canonicalForm` property; different lexicons are made interoperable as entries pointing to the same words are linked to the same lemma.

As regards the meaning, the model follows the principle of semantics by reference – in the sense that the semantics of a lexical entry is expressed by reference to an individual, class or property defined in a given ontology. Nevertheless, it still allows the lexicon itself to add named concepts. Thus, a `LexicalEntry` can be associated with concepts either directly through a denotative property or by using the intermediate class `LexicalSense`,[22] which reifies the referential relationship of concept and lexical entry and helps to capture the particular lexicalisation of the ontology entity (see Figure 4).

Although the *ontolex* module has the ability to describe a natural language vocabulary from the lexicological perspective, it is not capable of representing structures typically used in lexicography such as nested subentries, sense clustering or translated usage examples. The OntoLex-lemon Lexicography Module *lexicog* (Bosque-Gil et al., 2019), which is also used for modelling the content of lexical resources in LiLa, was developed to overcome these limitations, enabling the existing lexicographic resources to be modelled as Linked Data. It is mainly based on the class `LexicographicComponent`[23], which can

[16] https://www.w3.org/RDF/

[17] The subject position may also be a blank node, whose identifier is not known or specified.

[18] https://www.w3.org/TR/rdf-sparql-query/

[19] http://www.w3.org/ns/lemon/ontolex#LexicalEntry

[20] http://www.w3.org/ns/lemon/ontolex#lexicalForm

[21] http://www.w3.org/ns/lemon/ontolex#canonicalForm

[22] http://www.w3.org/ns/lemon/ontolex#LexicalSense

[23] http://www.w3.org/ns/lemon/lexicog#

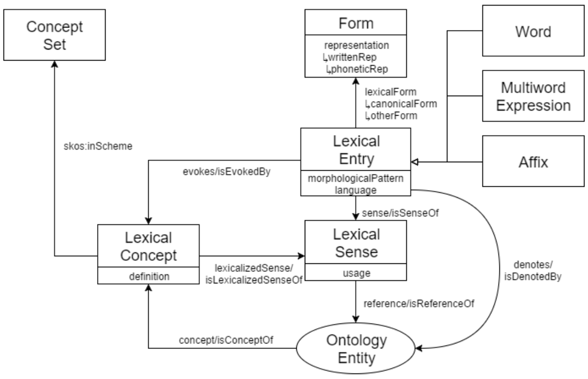


Figure 4: Diagram depicting the OntoLex-lemon core module. Source: Cimiano et al. (2016).

represent any (sub)structure of a lexicographic article that provides information about entries, senses or subentries; its subclass `Entry`[24] is used to indicate the root of the lexicographic article. The hierarchical relations between the lexicographic components are represented by means of the property `subComponent`[25], while the property `describes`[26] links each `LexicographicComponent` to the corresponding *ontolex* item that captures lexical data. Finally, the class `UsageExample`[27] represents the lexicographic examples, including when they comprise several string values such as translations.

It is important to emphasise the difference between a `LexicalEntry` (as defined in *ontolex*) and a `LexicographicEntry` (defined in *lexicog*). Whilst the former identifies the items in the lexicon of any given language and it is the only one that carries linguistic information, the latter is a purely structural unit, representing lexicographic records in a dictionary and their internal structure (e.g. the articulation in sense groups and subgroups in a dictionary entry). In simpler terms, whilst the English word *wolf* is a `LexicalEntry`, the article labelled *wolf* in a given dictionary is a `LexicographicEntry`. In most cases, a `LexicographicEntry` describes no more than a single lexical item (`LexicalEntry`), but it is also quite possible to have a single instance of the class `LexicographicEntry` linked to more lexical entries, for instance whenever a lexicon discusses multiple derived words (e.g. regular and substantivised adjectives) in the same record.

4. Modelling and Linking Data

The process of modelling the *Index Totius Artis* data is performed by describing its lexicological content (e.g. the definitions provided by the entries) through the classes and properties of *ontolex*, whilst the lexicographic structural elements of the dictionary entries (e.g. the hierarchical relations between entries and subentries) are modelled using the *lexicog* module.

To begin with, since the class `LexicalEntry` is intended to model single lexical items, all of the *Index*'s dictionary entries with multiple lemmas or subentries need to be separated into different lexical entries so that their content can be correctly linked. The data regarding their structural relationship is maintained through the use of the *lexicog* classes `Entry` and `LexicographicComponent`, respectively.

Each `LexicalEntry` is required to be linked to at least one lexical form, usually its lemma. But instead of the actual forms provided by the dictionary, the property `canonicalForm` is assigned the URI of its proper lemma in the Lemma Bank in order to link the *Index* lexical entries to the LiLa Knowledge Base. This is done by performing a string match between the lemmas of the LiLa Lemma Bank and those of the *Index*. To improve the matching process and reduce the number of ambiguous and false matches, each *Index* lemma is normalised and assigned its part-of-speech and inflectional category labels according to the LiLa tagset[28]. Normalisation consists of conversion to lowercase, substitution of *j* with *i* and *v* with *u*, suppression of diacritics and expansion of abbreviations and ligatures (e.g. from *cõpositè* to *composite*, from *Pompeij* to *pompeii*). The part-of-speech and inflectional category labels have been inferred from the set of principal forms provided by the entries for each lemma (e.g. 'Excîdo, is' becomes `excido_VERB_v3r`). The matching is done programmatically, in a three-step progressive approach: first, it tries to match the full strings, that is, the strings consisting of the lemma with part-of-speech and morphological labels; then, a second round is performed with the unmatched items, this time considering the pure lemma; finally, the edit distance between the remaining unmatched lemmas and the LiLa's lemma collection is calculated, resulting in a number of linking candidates for those lemmas.

The results are classified into single matches (1:1), ambiguous matches (1:N) or no matches (1:0) (see Table 1). The fact that more than 85%

`LexicographicComponent`

[24] `http://www.w3.org/ns/lemon/lexicog#Entry`

[25] `http://www.w3.org/ns/lemon/lexicog#subComponent`

[26] `http://www.w3.org/ns/lemon/lexicog#describes`

[27] `http://www.w3.org/ns/lemon/lexicog#UsageExample`

[28] As for PoS, LiLa adopts the Universal PoS tags (Petrov et al., 2011). As for inflectional categories, it makes use of a subset of the labels of the LEMLAT morphological analyser for Latin (Passarotti et al., 2017).

Table 1: Matching results, according to the number of candidates (single matches or 1:1, ambiguous or 1:N, no matches or 1:0).

Lexical entries	**n**	**%**
Total	4,723	100.0%
1:1 single matches	4,093	86.7%
1:N ambiguous matches	368	7.8%
1:0 no matches (total)	262	5.5%
1:0 with a single candidate	151	3.2%
1:0 with multiple candidates	104	2.2%
1:0 with no candidates	7	<0.1%

of the links are one-to-one matches reveals a good performance for the proposed method, due to its ability to distinguish homograph entries like the two *relego* as seen in Figure 2 (since they are converted into different strings, respectively, `relego_VERB_v3r` and `relego_VERB_v1r`).

The disambiguation of the ambiguous matches and the linking candidates, in turn, must be done manually by comparing other properties available in the LiLa Knowledge Base. For instance, the string `excido_VERB_v3r` has two different matching verb possibilities, one with *caedo* ('to cut'), the other with *cado* ('to fall') as its lexical base[29]. In this case, the *Index* provides information about the base, enabling disambiguation by comparison with that provided by LiLa. Other ambiguous links require different strategies, especially when the *Index* does not provide any supplementary information besides the definition, as it is the case with the homographs `dissero_VERB_v3r` ('to sow' or 'to discuss'). To disambiguate such cases, we use information about the meaning of lexemes provided by other lexical resources already linked to the LiLa Knowledge Base, namely the Lewis and Short dictionary and the Latin WordNet. As regards the unmatched lemmas, those that are missing in the Lemma Bank are added to it as new lemmas (e.g., *quini*, lit. 'five each'), whilst graphical variations of already existing lemmas are added to it as new written representations (e.g. *ortographia* vs. *orthographia*); lemmas consisting of inflected forms of words already present in the Lemma Bank are linked to their respective lemma therein (e.g., *faxim*, form of *facio*), whilst typographical errors in the CLP source (e.g. *perper* instead of *perpes*, *cogniminis* instead of *cognominis*) are fixed and then equally linked. This process of data curation, although time-consuming, has led to improvements in both the dictionary data and the LiLa Lemma Bank: as for the former, 30 lemmas were cleared of typographical errors and 44 new lexical entries were identified amidst the data; as for the latter, the Lemma Bank was enhanced with 140 new lemmas and 60 new graphical variants for existing lemmas.

As regards the definitions provided by the *Index*, they are represented individually by the `LexicalSense` class through the `skos:definition`[30] property. Here the language alternation raises some data structuring issues, requiring a delimitation of the Portuguese and Latin definitions for the subsequent property value tagging. The separation must be done manually by observing common delimiters for Latin extended meanings (e.g. *pro*, 'used for'), as it reveals that, despite of making a considerable use of Latin, the *Index* is predominantly a bilingual dictionary: there are 4,577 senses exclusively conveyed in Portuguese, compared to 377 senses described in Latin solely and 100 senses conveyed in both languages simultaneously. The `LexicalSense` class is also the domain of the *lexicog* property `usageExample`, which links objects of the class `UsageExample` to their respective senses. Because of that, sometimes an "empty" `LexicalSense` must be created to serve as a container for blocks of the phraseological examples presented separately at the end of some dictionary entries.

Finally, the other pieces of information – that is, which are not a lemma, a definition, or a usage example – are considered as groups based on their position relative to the headword and the definition[31]. They are then distributed accordingly to the model's object classes they are associated with using two properties that can represent generic notes – namely, `lexinfo:note`[32] and `skos:note`[33]. For instance, the etymological and the prosodic information on the verb *excido*, seen above, is encoded as values of a `lexinfo:note` and a `skos:note`, respectively, and attributed to the `LexicalEntry`, whilst the information on verb complementation is represented by a `lexinfo:note` related to its respective `LexicalSense`. This strategy allows to preserve every textual content of the source, includ-

[29] The lexical base is defined as "the lexical morpheme of a word that is neither a prefix nor a suffix" (Passarotti et al., 2020). In the LiLa Lemma Bank, canonical forms for Classical Latin words are linked to their lexical base, which works as a connector between the words belonging to the same derivational family (Litta et al., 2019).

[30] http://www.w3.org/2004/02/skos#definition

[31] Accordingly to the Merrilees (1996)'s method of grouping what he calls 'supplementary information' into two categories: the 'post-lemmatic position' groups information placed between the lemma and the definition, and the 'post-definitional position' groups information placed after the definition.

[32] http://www.lexinfo.net/ontology/2.0/lexinfo#note

[33] http://www.w3.org/2004/02/skos#note

Table 2: Modelling results, grouped by OntoLex-lemon classes.

Class	Individuals
lexicog:Entry	3,619
lexicog:LexicographicComponent	1,056
ontolex:LexicalEntry	4,675
ontolex:LexicalSense	5,664
lexicog:UsageExample	2,428

ing lexicographic devices such as the Latin connectors and the page locators, which are encoded as `skos:note` of the respective *lexicog* module's classes. The quantitative results of data modelling are reported in Table 2.

A graph visualisation of the modelling result is shown in Figure 5. It represents the dictionary entry for *grammatica*, seen above, and illustrates the way the OntoLex-lemon modules interact to preserve the relationships between lexical entries as they are found within the dictionary entry structure. The upper part of the figure shows the nodes for the `lexicog` module elements, while the bottom part of the figure shows the nodes for the `ontolex` module elements. Thus at the top, the Lexicographic Resource (in bordeaux burgundy, at the upper centre) is connected to one of its entries (*grammatica*, in light violet purple, right below), which in turn is connected to six Lexicographic Components (in dark violet purple) representing each one a subentry. On the bottom, seven lexical entries (in yellow) are shown with their multiple connections: on one hand, they are linked as entries to the Lexical Resource they belong to (in green at the centre); on the other hand, each of them is linked both to the canonical form (in bordeaux, at the bottom) and to the lexical sense (in orange, at the bottom) they consist of. Last but not least, *lexicog* property `describe` acts as a bridge between the two structures, connecting lexical entries to lexicographic components.

5. Querying interlinked resources

The *Index totius artis* content is now part of the LiLa Knowledge Base (Velez, 2023). It can be accessed using a query interface for the Lemma Bank[34], one for the interlinked resources[35], or an SPARQL access point[36]. As the first Latin-Portuguese lexical resource made interoperable with the (meta)data of other linguistic resources, it provides different users with new ways of querying its dictionary data, from taking a deeper look at the kind of lexical information presented by the *Index* to exploring the connections with the other linguistic resources that form the LiLa Knowledge Base. Since the query possibilities depend on both the users' information needs and their programming abilities in SPARQL, we offer some examples that may be useful for users such as Latin students, scholars, historians of Lexicography, and classical philologists.

Latin students can use a SPARQL query[37] to simply look up the information the *Index* provides for a given Latin word (e.g. the verb *do*, 'to give') or, inversely, a query[38] to pick up a set of Latin words associated with a given concept by means of its Portuguese lexicalisation (e.g. *fallar*, 'to speak', as written in the 18th century).

Latin scholars can use the semantic-related information (i.e. modelled as `lexinfo:note` of a Lexical Sense) to generate lists of words sharing the same properties. It is achieved by running a SPARQL query[39] that lists the frequency of `lexinfo:note` values associated with lexical senses. The result mostly consists of information on verb complementation diversely represented, whether by the ancient grammar verbal gender category (e.g. 'activum', 186 oc.; 'neutrum', 33 oc.), by case names (e.g. 'cum accusativo', 117 oc., 'dativo', 143 oc.) or even, to a lesser extent, by pronoun-based expressions (e.g. 'aliquid alicui', 65 oc.; 'aliquem aliqua re', 27 oc.). By taking one of these values (e.g. 'aliquem aliqua re', lit. 'someone from/with something') as input to another SPARQL query[40] that covers the links between the `lexinfo:note` value and the lemmas in the Lila Knowledge Base through the lexical entries, one can get a list of 29 verbs governing accusative of person and ablative of thing.

A similar query[41] may be of interest to historians of Lexicography as long as it groups the results by part-of-speech (as provided by the LiLa Knowledge Base), instead of displaying values for the `lexinfo:note`. It allows one to estimate the emphasis given to verbal entries in the *Index*, since they are mostly linked to those notes (830, or 65%), and helps one to understand the its ways of mirroring the grammar's syntactic chapters and reflecting the linguistic thought of the time by means of information on verb complementation[42]. This em-

[34] https://lila-erc.eu/query/

[35] https://lila-erc.eu/LiLaLisp/

[36] https://lila-erc.eu/sparql/

[37] https://github.com/lucascdz/psm/blob/main/Velez_lookup_Latin.rq

[38] https://github.com/lucascdz/psm/blob/main/Velez_lookup_Portug.rq

[39] https://github.com/lucascdz/psm/blob/main/Velez_count_notes_distinct.rq

[40] https://github.com/lucascdz/psm/blob/main/Velez_notes_to_lemmas.rq

[41] https://github.com/lucascdz/psm/blob/main/Velez_count_notes_byPos.rq

[42] See Colombat (2003) for an explanation of how the

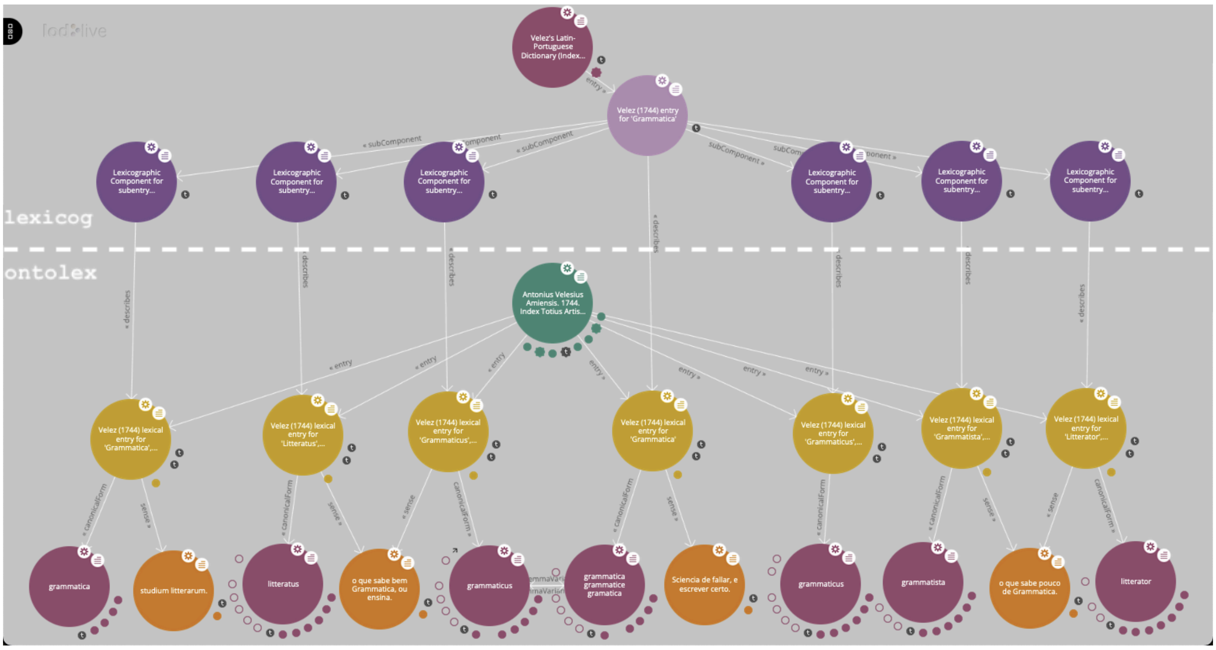

lod live
Velez's Latin-Portuguese Dictionary (Index...
entry »
Velez (1744) entry for 'Grammatica'
« subComponent
subComponent »
Lexicographic Component for subentry...
Lexicographic Component for subentry...
Lexicographic Component for subentry...
Lexicographic Component for subentry...
Lexicographic Component for subentry...
Lexicographic Component for subentry...
lexicog
ontolex
« describes
Antonius Velesius Amiensis. 1744. Index Totius Artis...
« entry
entry »
Velez (1744) lexical entry for 'Grammatica',...
Velez (1744) lexical entry for 'Litteratus',...
Velez (1744) lexical entry for 'Grammaticus',...
Velez (1744) lexical entry for 'Grammatica'
Velez (1744) lexical entry for 'Grammaticus',...
Velez (1744) lexical entry for 'Grammatista',...
Velez (1744) lexical entry for 'Litterator',...
« canonicalForm
sense »
grammatica
studium litterarum.
litteratus
o que sabe bem Grammatica, ou ensina.
grammaticus
grammatica grammatice gramatica
Sciencia de fallar, e escrever certo.
grammaticus
grammatista
o que sabe pouco de Grammatica.
litterator

Figure 5: Visualisation graph of the *Index* entry for *grammatica* as represented in LiLa Knowledge Base LOD Viewer. Source: https://lila-erc.eu/lodlive/.

phasis on verbs can be confirmed by running a query[43] which takes the number of individuals of the class LexicalEntry and the morphological information provided by LiLa to estimate the ratio of lexicographic entries by part-of-speech in both the *Index* and a reference dictionary like the Lewis and Short. The results show that the *Index* has a higher proportion of entries for verbs (30% against 14%), whereas the opposite occurs with adjectives (15% against 28%, respectively). Another way of assessing the *Index* wordlist has particular interest, since its very small macrostructure could give the impression of a sort of 'essential Latin dictionary'[44]. Actually, when measured[45] against the top one thousand most commonly used lemmas in a considerable corpus such as the LASLA corpus[46], the *Index* wordlist is evidently far from covering the essential Latin vocabulary: it contains only 60% of the most commonly used Latin lemmas, with the verbs corresponding to the largest part (75%), the nouns to the smallest (40%).

Finally, classical philologists can explore the *Index* connections with other interlinked resources in new and significant ways, such as using the Portuguese translation equivalents as starting points to investigate the text corpora linked to LiLa. For instance, one can be interested in the distribution of the Latin verbs that potentially mean 'to speak' in a set of narrative texts (namely, the writings of five Roman historians plus two epic poems[47]). So a SPARQL query[48] can (1) take a given Portuguese word that represents the chosen concept as input (e.g. 'fallar'), (2) go through the links from the *Index* definition to the LiLa Lemma Bank, where it gets a list of lexical bases[49], (3) select all the verbal lemmas linked to those bases (i.e. belong to the same family), and (4) count the number of tokens that are linked to the selected lemmas. Four lexical bases were selected, namely those of *dico*, *loquor*, *for* and *taceo*. While the first three do mean 'to speak', the last one actually means the opposite ('não fallar', lit. 'not to speak'). Table 3 shows the results, grouped by author and lexical base. Although an extended interpretation of the results is beyond the scope of this paper, some patterns in the authors' usages are noteworthy. To begin with, *dico* is the most frequently used lexical base for referring to

syntax is described in terms of complementation patterns by the Humanist grammarians, in particular the assessment of the Alvares' practice inherited by Velez.

[43] https://github.com/lucascdz/psm/blob/main/Velez_assess_wordlist_vs_LS.rq

[44] It is stated by Verdelho (1995, 461) and restated by Iken (2002, 58).

[45] https://github.com/lucascdz/psm/blob/main/Velez_assess_wordlist_vs_lasla.rq

[46] As linked to the LiLa KB, the size of the LASLA corpus is estimated to be 1.8 million tokens according to Fantoli et al. (2022).

[47] All the texts of Caesar, Hirtius, Sallustius, Curtius, and Tacitus, as well as Virgil's *Aeneid*, are part of the LASLA Latin corpus, except for the *Pharsalia*, which belongs to the CIRCSE Latin Library.

[48] https://github.com/lucascdz/psm/blob/main/Velez_definition_base_corpus.rq

[49] In the LiLa Lemma Bank, lemmas are linked to their lexical bases by means of the property http://lila-erc.eu/ontologies/lila/hasBase

Table 3: Normalised frequency (by 10k) of verbal tokens linked to the lexical bases that mean *to speak*.

	Base of DICO		Base of LOQUOR		Base of FOR		Base of TACEO	
author	n	n/10k	n	n/10k	n	n/10k	n	n/10k
Caesar	172	21.8	45	5.7	0	0.0	2	0.3
Hirtius	10	15.1	0	0.0	0	0.0	0	0.0
Sallustius	92	28.5	15	4.7	0	0.0	5	1.6
Curtius	179	24.7	41	5.7	4	0.6	12	1.7
Tacitus	198	13.8	59	4.1	7	0.5	16	1.1
Vergilius	163	24.3	43	6.4	125	**18.6**	8	1.2
Lucanus	14	**2.1**	24	3.6	51	7.6	35	**5.2**

the act of speaking in all authors' texts, except in Lucanus, where its frequency is remarkably low. In turn, the verbs linked to the base *loquor* are moderately used by all authors, except for Hirtius, who does not use it at all. As regards the base *for*, it is mostly used in the two poetic texts, as expected; nonetheless, the existence of some occurrences in Curtius and Tacitus could be worth investigating. Finally, the base of *taceo* has low frequencies in all authors, but Lucanus stands out again, this time for showing the higher frequency of the group; the association between the low frequency of *dico* and the high frequency of *taceo* suggests a potential relevance of the theme of silence in Lucanus' epics that could be explored by further and properly designed researches.

6. Conclusions and future work

Old lexical resources concerning Latin account for a significant part of our knowledge of historical languages and cultural heritage. In this paper we described a way to represent this wealth of information as Linked Data, according to practices widely adopted in the Semantic Web. To do so, we coped with the difficulties in modelling loosely structured lexical/lexicographic entries according to the OntoLex-lemon model. Since the *Index Totius Artis* micro-structure is typical of the early modern Latin lexicography, the success of the method used here makes it suitable for reuse in further bilingual or monolingual Latin dictionaries from that period, if not beyond.

Moreover, by linking the dictionary lemmas to a Linked Data Knowledge Base like LiLa we have improved the potential of researching, assessing and (re)using that data in many ways. Among the advantages made possible by this work we can mention:

- improvements in dictionary consultation, as it can now be accessed by means of the lexicographic categories provided by the *Index* itself (lemma, definition, usage example, etc.) as well as by means of those provided by the other resources linked to the LiLa Knowledge Base (such as lexical bases, WordNet synsets, valency patterns, and corpora);
- improvements in the dictionary data description and assessment, providing the possibility of performing lexicographic-related research – such as describing its wordlist characteristics or evaluating the lexical and lexicographic information provided in relation to the entries, senses, and examples;
- improvements in the dictionary data availability and reuse, given the possibility of automatically producing new resources (e.g. a Latin-Portuguese vocabulary for a selected corpus for educational purposes).

Given the *Index* intricate way of structuring the entry elements, we have chosen to focus on distributing their content through the OntoLex model classes of data, whilst keeping it as freetext in order to maximally preserve the linguistic and structural information as conveyed in the source. Undoubtedly, the actual Linked Data resource would benefit from the interlinking with a Portuguese lexical database, a conceptual ontology, or an ontology for linguistic description. These are all potential improvements for future releases. Notwithstanding, in the face of the extreme lack of Latin-Portuguese bilingual resources on the Web, it will also be worth replicating this experiment on Velez's dictionary to model and publish as Linked Data in the LiLa Knowledge Base the other dictionaries available from CLP. This will provide LiLa and the entire community with a good set of bilingual dictionaries made interoperable with each other as well as with the several textual corpora for Latin currently published in LiLa.

7. Acknowledgements

This publication is part of a doctoral research developed at the São Paulo State University (UNESP/FCL-Ar, Brazil) and was financed in part by the CAPES Foundation (Ministry of Education of Brazil – Finance Code 001 – Grant Agreement No. 88887.716744/2022-00).

The "LiLa - Linking Latin" project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

We thank all the LiLa Team for their help with this work, especially Giovanni Moretti, Federica Iurescia, Flavio Cecchini and Paolo Ruffolo.

8. Bibliographical References

Tim Berners-Lee. 2006. Linked data.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American*, 284(5).

Julia Bosque-Gil, Jorge Gracia, John McCrae, Fahad Khan, Katrien Depuydt, Jesse de Does, Francesca Frontini, and Ilan Kernerman. 2019. The Ontolex Lemon Lexicography Module: Final community group report, 17 September 2019. Technical report, Ontology Lexica under the W3C Community.

Roberto Busa. 1974-1980. *Index Thomisticus*. Frommann-Holzboog, Stuttgart-Bad Cannstatt.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon model for ontologies: Final community group report, 10 may 2016. Technical report, Ontology-Lexicon Community Group under the W3C Community Final Specification Agreement (FSA).

Bernard Colombat. 2003. Le traitement de la construction verbale dans la grammaire latine humaniste. In Sylvain Auroux, editor, *History of Linguistics 1999*, volume 99, pages 63–81. John Benjamins, Amsterdam/Philadelphia.

Joseph Denooz. 2004. Opera latina: une base de données sur internet. *Euphrosyne*, 32:79–88.

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the lasla corpus in the lila knowledge base of interoperable linguistic resources for latin. In *Proceedings of The 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 26–34, Marseille, France. European Language Resources Association.

Reinhard R. K. Hartmann. 2001. *Teaching and Researching Lexicography*, 2nd. edition, pages 57–79. Routledge, London/New York.

Sebastião Iken. 2002. *Index totius artis* (1599-1755): algumas reflexões sobre o índice lexicográfico latino-português da gramática de Manuel Álvares, elaborado por António Velez. In Rolf Kemmler, Barbara Schäfer-Prieß, and Axel Schönberger, editors, *Estudos de história da gramaticografia e lexicografia portuguesas*, pages 53–83. Domus Editoria Europaea, Frankfurt am Main.

Rolf Kemmler. 2018. Handwritten annotations in the early editions of Manuel Alvares' *De institvtione grammatica libri tres*. *Philologica Jassyensia*, 28:57–69.

Charlton T. Lewis and Charles Short. 1879. *A Latin Dictionary*. Clarendon Press, Oxford.

Eleonora Litta, Marco Passarotti, and Francesco Mambrini. 2019. The treatment of word formation in the LiLa knowledge base of linguistic resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 35–43, Prague, Czechia. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Francesco Mambrini, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2021. Linking the Lewis & Short dictionary to the LiLa Knowledge Base of interoperable linguistic resources for Latin. In *Proceedings of the Eighth Italian Conference on Computational Linguistics*, Aachen, Germany. Italian Conference on Computational Linguistics, CEUR Workshop Proceedings.

Brian Merrilees. 1996. The shape of the medieval dictionary entry. *Digital Studies/le Champ Numérique*, 4.

Michal Měchura. 2023. Avoiding Recursion in the Representation of Subsenses and Subentries in Dictionaries. *International Journal of Lexicography*, 36(3):260–278.

David W. Packard. 1968. *A Concordance to Livy*. Harvard University Press, Cambridge, MA.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The lemlat 3.0 package for morphological analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31, Gothenburg. Linköping University Electronic Press.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the

LiLa Knowledge Base of linguistic resources for Latin. *Studi e Saggi Linguistici (SSL)*, 58(1):177–212.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset.

Eustaquio Sánchez Salor and Juan María Gómez Gómez. 2020. Introdução. In *Manuel Álvares: Instituição da Gramática, ampliada e explicada por António Velez*, pages 7–32. Imprensa da Universidade de Coimbra, Coimbra.

Emilio Springhetti. 1961–1962. Storia e fortuna della grammatica di Emmanuele Alvares. *Humanitas (Coimbra)*, 13–14:283–304.

Antonio Velez. 1744. Index totius artis. In *Emmanuelis Alvari S. J. De Institutione Grammatica Libri Tres*, pages 366–661. Typographia Academiae, Eborae.

Antonio Velez. 2002. Index totius artis. In Telmo dos Santos Verdelho and João Paulo Martins Silvestre, editors, *Corpus Lexicográfico do Português*. Universidade de Aveiro and Centro de Linguística da Universidade de Lisboa, Portugal.

Telmo dos Santos Verdelho. 1995. *As origens da Gramaticografia e da Lexicografia latino-portuguesas*. Instituto Nacional de Investigação Científica, Aveiro, Portugal.

Telmo dos Santos Verdelho and João Paulo Martins Silvestre, editors. 2002. *Corpus Lexicográfico do Português*. Universidade de Aveiro and Centro de Linguística da Universidade de Lisboa, Portugal.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

9. Language Resource References

Antonio Velez. 2023. *Index Totius Artis*. CIRCSE: Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione. LiLa Knowledge Base, Latin-Portuguese Dictionaries, 1.1.