The Lemma Bank of the LiITA Knowledge Base of Interoperable Resources for Italian

Eleonora Litta[1,*,†], Marco Passarotti[1,†], Paolo Brasolin[1,†], Giovanni Moretti[1,†], Francesco Mambrini[1,†], Valerio Basile[2,†], Andrea Di Fabio[2,†] and Cristina Bosco[2,†]

[1]CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milano, Italy
[2]Università degli Studi di Torino - Dipartimento di Informatica, Corso Svizzera 185, 10149 Torino, Italy

Abstract

The paper introduces the LiITA Knowledge Base of interoperable linguistic resources for Italian. After describing the principles of the Linked Data paradigm, on which LiITA is grounded, the paper presents the lemma-centred architecture of the Knowledge Base and details its core component, consisting of a large collection of Italian lemmas (called the Lemma Bank) used to interlink distributed lexical and textual resources.

Keywords

Linked Open Data, Linguistic Resources, Italian, Interoperability

1. Introduction

When considering the number of digital linguistic resources, either lexical or textual, Italian is among the richest languages: e.g., at the time of writing, a search on the CLARIN Virtual Language Observatory,[1] filtered for the Italian language, returns more than 8 000 results. Like other high-resource languages, Italian is provided with a large set of fundamental resources, including WordNets ([1] and [2]), a few treebanks available from the Universal Dependencies collection[2], historical corpora [3][4] and reference corpora of written (e.g., CORIS/CODIS [3]) and spoken language (e.g., KIParla [4]).

However, as is the case for many other languages, most linguistic resources for Italian vary in terms of data format, annotation criteria, and/or adopted tagsets. Such variation hinders full interaction between the (meta)data provided by the many available resources, with a negative impact on the empirical study of the language and resource usability. Indeed, different resources may provide different information or use different granularity of information about the same common object, namely words, which appear as occurrences in corpora and as entries in dictionaries or lexicons. Making this wealth of information interact represents one of today's main challenges, to best leverage the huge asset of (meta)data collected over decades of work.

As a consequence, a very active line of research currently focuses on the so-called Linguistic Linked Open Data (LLOD), aiming to define common practices for the representation and publication of linguistic resources according to the principles of the Linked Data paradigm, which underpins the Semantic Web[5].

A recently concluded COST Action (Nexus Linguarum[6]) resulted both in the creation of a large and cohesive scientific community and in the definition of a set of shared vocabularies for linguistic knowledge description. Some of these vocabularies have been widely applied in the LiLa Knowledge Base (KB), which is probably the main LLOD use case currently available. LiLa (Linking Latin) is a KB of Latin linguistic resources made interoperable through their representation and publication according to the Linked Data principles. Thanks to its streamlined and language-independent architecture, LiLa is today a reference model for projects aiming to achieve online interoperability between distributed linguistic resources.

Building on the experience of LiLa and reusing its ar-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†These authors contributed equally.
eleonoramaria.litta@unicatt.it (E. Litta); marco.passarotti@unicatt.it (M. Passarotti); paolo.brasolin@gmail.com (P. Brasolin); giovanni.moretti@unicatt.it (G. Moretti); francesco.mambrini@unicatt.it (F. Mambrini); valerio.basile@unito.it (V. Basile); andrea.difabio@unito.it (A. Di Fabio); cristina.bosco@unito.it (C. Bosco)
0000-0002-0499-997X (E. Litta); 0000-0002-9806-7187 (M. Passarotti); 0000-0003-2471-7797 (P. Brasolin); 0000-0001-7188-8172 (G. Moretti); 0000-0003-0834-7562 (F. Mambrini); 0000-0001-8110-6832 (V. Basile); 0000-0002-3290-8158 (A. Di Fabio); 0000-0002-8857-4484 (C. Bosco)

[1]https://vlo.clarin.eu
[2]https://universaldependencies.org
[3]https://www.corpusmidia.unito.it/
[4]http://www.ovi.cnr.it/

[5]A few resources for Italian are available as Linked Open Data, namely the CompL-it lexicon (http://hdl.handle.net/20.500.11752/ILC-1007), the ItalWordNet v.2 (http://hdl.handle.net/20.500.11752/ILC-66), and a collection of names from the PAROLE SIMPLE CLIPS (PSC) lexicon (http://hdl.handle.net/20.500.11752/ILC-558).
[6]https://nexuslinguarum.eu

chitecture, the LiITA (Linking Italian)[7] project has started the creation of a KB of interoperable linguistic resources for Italian published as Linked Data. This paper describes the development of the fundamental component of the LiITA KB, which consists of a collection of Italian lemmas (called the Lemma Bank) that serves as the connection point between word occurrences and their entries in the corpora and lexical resources that will be published in the KB.

2. Linguistic Linked Data

Introduced by Tim Berners-Lee et alii [5], the concept of the Semantic Web is based on the assumption that documents published on the World Wide Web are associated with information and metadata structured in such a way as to allow their querying and semantic interpretation not only by humans but also by automated agents.

This structuring is implemented in the form of Linked Data, which are the pillars of the Semantic Web. Unlike a web made of hypertexts, where links are not semantically interpretable, the Semantic Web consists of links between "objects" associated with a unique and persistent identifier (URI: Uniform Resource Identifier). The links between objects are semantically interpretable as they are represented through vocabularies for knowledge description recorded in the form of ontologies.

The Linked Data paradigm is founded on four principles defined by Berners-Lee himself[8]:

1. Use URIs as "names for things" to identify them uniquely and persistently. The "things" dealt with when handling linguistic (meta)data in Linked Data are linguistic objects, such as occurrences of words in texts, lexical entries in dictionaries, or sets of parts of speech;
2. Use HTTP URIs to allow people (and machines) to look up things on the Web;
3. Use standards such as RDF and SPARQL to provide useful information about what is identified by a URI, for the purpose of representation and retrieval of (meta)data. RDF (Resource Description Framework) [6] is the data model that underlies the Semantic Web. According to this model, information in the Semantic Web is organised and represented in terms of triples, i.e., relationships between a Subject and an Object through a Property. The classes to which Subjects and Objects belong, as well as the semantics of Properties, are established by ontologies shared by the different communities that enrich and use the Semantic Web. SPARQL (SPARQL Protocol And RDF Query Language)[9] is a query language for (meta)data represented in RDF;
4. Include links to other URIs to allow people (and machines) to discover more things.

Applying the principles of the Linked Data paradigm to (meta)data derived from linguistic resources and publishing them on the Web offers several benefits [7]. Firstly, as for representation and modelling of (meta)data, RDF is a very versatile model, suitable for representing metadata such as those conveyed by the various levels of annotation available in linguistic resources (morphology, syntax, lemmatisation, etc.). Moreover, the adoption of a common data model (RDF) enables both structural (or syntactic) interoperability, which is the ability of different systems to process exchanged data using shared protocols and formats (such as HTTP and URI), and conceptual (or semantic) interoperability, which is the ability of a system to automatically and semantically interpret the exchanged information using a common set of classes and data categories defined in ontologies and vocabularies [8]. The Italian language is no stranger to this paradigm[10][11][12]. But this is the first attempt to create such a kind of resource in the form of a lemma bank in Italian.

3. The LiITA Knowledge Base

This Section introduces the fundamental architecture of the LiITA KB and details its core component, i.e., a collection of canonical forms of citations (lemmas) for the Italian language[13]. The base URI of the resource is `http://www.liita.it/data/`, a namespace we reserved by buying the domain from a registrar to use also as a URL, e.g., for the project website.

3.1. The Architecture of LiITA

The architecture of the LiITA KB resembles that of the LiLa KB for Latin[14], which is based on the assumption that the sources of the (meta)data that the KB makes interoperable are all related to words. These sources are linguistic resources and specifically:

- lexical resources, such as dictionaries or lexicons, which describe the properties of words and consist of lexical entries;
- textual resources, such as corpora and digital libraries, which provide texts and are made of occurrences of words (tokens).

[7] http://www.liita.it/
[8] https://www.w3.org/DesignIssues/LinkedData
[9] https://www.w3.org/TR/rdf-sparql-query/
[10] http://hdl.handle.net/20.500.11752/ILC-1007
[11] http://hdl.handle.net/20.500.11752/ILC-66
[12] http://hdl.handle.net/20.500.11752/ILC-558
[13] https://github.com/LiITA-LOD
[14] https://lila-erc.eu/data-page/

Lexical entries and word occurrences coming from distributed resources are made interoperable in LiITA by linking them to their respective lemmas. This makes it possible to perform federated searches on the different linguistic resources that LiITA makes interoperable. For example, one can search for all occurrences (tokens) of the same lemma in multiple textual corpora; or extract from multiple corpora all those tokens that have certain lexical properties provided by one or more lexical resources.

Given the central role played by lemmas in the architecture of LiITA, the core component of the KB is a collection of conventional citation forms (lemmas) of Italian words, called the Lemma Bank.

In the LiLa KB lemmas are described with the help of custom ontology.[15] This ontology, on the one hand, provides detailed information on some morphological and linguistic features of the lemmas (e.g. the part of speech, the gramatical gender for nouns and the inflectional class) relying on the OLiA annotation model [9, 151-155]. On the other hand, the LiLa ontology defines classes and properties to model the task of lemmatization, such as the property `lila:hasLemma`[16] which links lemmas to corpus tokens. The class of `lila:hasLemma`[17] is defined as a subclass of `ontolex:Form` (on which, see sec. 3.2), so that the LiLa KB is not a lexical resource in itself, but rather a collection of canonical forms that can be either used to lemmatize texts or to index lexical entries.

3.2. The LiITA Lemma Bank

Data modelling

The Lemma Bank of LiITA consists of a collection of lemmas of the Italian language, i.e., lexical citation forms adopted (more or less conventionally) in linguistic resources. These lemmas are the names of entries in (most) lexical resources and the forms chosen to gather all occurrences of a particular word in (lemmatised) textual resources. As mentioned above, the Lemma Bank plays a fundamental role in the LiITA KB, acting as the connection point between entries in various lexical resources and word occurrences in textual resources.

Following the principles of the Linked Data paradigm, conceptual interoperability among the distributed resources connected in LiITA is achieved by applying a vocabulary for knowledge description commonly used in the world of Linguistic Linked Open Data. In the specific case of the Lemma Bank, this means adopting the vocabulary defined by OntoLex-Lemon [10], one of the most widely used models for representing and publishing lexical resources as Linked Data. Figure 1 shows the

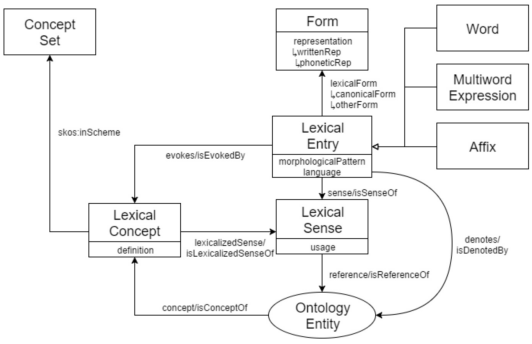

Figure 1: The OntoLex-Lemon model.

OntoLex-Lemon model.

In Figure 1, the Classes of OntoLex-Lemon are graphically represented within rectangles. The relationships between Classes are shown as arrows associated with the name of the Property that connects two Classes.

The main Class of OntoLex-Lemon is `ontolex:LexicalEntry`[18], understood as the unit of lexicon analysis that gathers one or more forms (`ontolex:Form`[19]) and one or more lexical senses (`ontolex:LexicalSense`[20]), lexical concepts (`ontolex:LexicalConcept`[21]) or entities from ontologies.

Lexical senses are lexicalised senses: a sense belongs exactly to one lexical entry. Semantic aspects that can be expressed by multiple words are represented through lexical concepts, which can therefore have more than one lexicalisation. A typical example of a lexical concept is the synset in a resource like WordNet, which groups multiple words related by a conceptual synonymy relationship.

Forms can have one or more graphical variants (written representations), represented through the Data Property `ontolex:writtenRep`[22], and possibly one or more phonetic variants (Property `ontolex:phoneticRep`[23]). One of these forms, the object of the `ontolex:canonicalForm` Property[24], is the form that is conventionally chosen to represent the entire set of inflected forms of a lexical entry. The Lemma Bank of LiITA is a collection of such forms, modelled as individuals of the Class `lila:Lemma`[25], which is a subclass of `ontolex:Form`, originally created for the LiLa project, and adopted in the LiITA Lemma Bank accordingly. The lemmas of the LiITA Lemma

[15] http://lila-erc.eu/ontologies/lila/.
[16] http://lila-erc.eu/ontologies/lila/hasLemma
[17] http://lila-erc.eu/ontologies/lila/hasLemma

[18] http://www.w3.org/ns/lemon/ontolex#LexicalEntry
[19] http://www.w3.org/ns/lemon/ontolex#Form
[20] http://www.w3.org/ns/lemon/ontolex#LexicalSense
[21] http://www.w3.org/ns/lemon/ontolex#LexicalConcept
[22] http://www.w3.org/ns/lemon/ontolex#writtenRep
[23] http://www.w3.org/ns/lemon/ontolex#phoneticRep
[24] http://www.w3.org/ns/lemon/ontolex#canonicalForm
[25] http://lila-erc.eu/ontologies/lila/Lemma

Bank are unbound by any relationship with a lexical entry, as the Lemma Bank is not a lexical resource consisting of lexical entries but a set of canonical forms of citation. This reflects the role of the Lemma Bank in LiITA as a collection of lemmas used to make resources interoperable.

The LiITA Lemma Bank makes textual resources for Italian interoperable through the `lila:hasLemma` Property[26], which links a token in a corpus with its lemma in the Lemma Bank. Lexical resources, on the other hand, are connected to the Lemma Bank through the `ontolex:canonicalForm` Property, which links a lexical entry in the resource to its corresponding lemma in the Lemma Bank.

By using the Property `lila:hasPos`[27], each lemma in the Lemma Bank is assigned one part of speech, following the Universal PoS tagset [11].

In the case of words that are assigned multiple PoS tags in lexical resources, multiple lemmas are created in the Lemma Bank. For instance, the word *sopra* 'over' is usually assigned four PoS: preposition, adverb, adjective and noun. Thus, four distinct lemmas are created in the Lemma Bank with four different PoS represented via the `lila:hasPos` Property.

Data harmonisation

To harmonise different lemmatisation criteria that may be found in linguistic resources, the Lemma Bank of LiITA includes two specific Properties. The symmetric Property `lila:lemmaVariant`[28] connects different forms of the inflectional paradigm of a word that can be used as lemmas. A typical case is that of *pluralia tantum*, which can be lemmatised either in the plural form or in the singular form. This model allows, for example, for both the `lila:Lemma` *pantaloni* and *pantalone*, which are linked to each other by the `lila:lemmaVariant` Property.

While `lila:lemmaVariant` links lemmas that are assigned the same part of speech, the Property `lila:hasHypolemma`[29] (and its inverse property`lila:isHypolemma`[30]) connects lemmas that can be used for the same word but have different parts of speech. This is the case for the adjectives used as adverbs, e.g. *veloce* which can be interpreted (and lemmatised) either as a form of adjective (hence modelled as a `lila:Lemma`) or as an adverb (hence modelled as a `lila:Hypolemma`[31], a subclass of `lila:Lemma`).

Past participles are another kind of hypolemma (e.g. *caduto* 'fallen'), which in the Lemma Bank are assigned the part of speech Adjective. Participles are modelled as individuals of the `lila:Hypolemma` Class and are connected to their verbal lemma (*cadere* 'to fall') through the `lila:isHypolemma` Property.
Regardless of whether two resources lemmatise participles according to different criteria (namely, one under the participial lemma and the other under the verbal lemma), the two different lemmatisations are harmonised in the Lemma Bank.

Data acquisition

The lemmas and PoS that constitute the Lemma Bank is based on the lexical base of an online version of the dictionary Nuovo De Mauro[32], which amounts to about 145 000 entries; out of these, 13 000 multi-word expressions were excluded because they were deemed unnecessary, as lemmatisers usually deal with single tokens. About 94 000 lemmas were derived from the remaining 131 000 entries. The most numerically abundant PoS with which the Lemma Bank was populated are listed in Table 1.

Table 1
Distribution of lemmas across different parts of speech

Lemmas	Part of Speech
56 575	Nouns
19 912	Adjectives
15 885	Verbs
359	Proper Nouns
311	Adverbs
112	Pronouns
106	Conjunctions
40	Prepositions
58	Articles

This population process was not an easy task for two main reasons. Firstly, the online version of *Nuovo De Mauro* is tailored for visualisation: data is mixed with graphical information. Secondly, *Nuovo De Mauro* stems from one of the greatest efforts in Italian lexicographic history, namely GRADIT (*Grande dizionario italiano dell'uso* [12]). The resource includes information especially hard to handle computationally: De Mauro and colleagues described for every lemma not only each of its usual lexicographic metadata (meaning, PoS, examples, etc.) but also frequency, semantic domain, grouping of senses, multi-word expressions and more. The extraction of data is in practice hindered by information that must be filtered out because it is not relevant for our purposes of building a lemma bank or is provided in some non-homogeneous forms. Therefore, in order to ease this

[26] http://lila-erc.eu/ontologies/lila/hasLemma
[27] http://lila-erc.eu/ontologies/lila/hasPOS
[28] http://lila-erc.eu/ontologies/lila/lemmaVariant
[29] http://lila-erc.eu/ontologies/lila/hasHypolemma
[30] http://lila-erc.eu/ontologies/lila/isHypolemma
[31] http://lila-erc.eu/ontologies/lila/Hypolemma

[32] https://dizionario.internazionale.it/. PoS tags were converted automatically into the Universal tagset, adopted in the Lemma Bank.

initial work, we decided to preliminary extract the aforementioned PoS, leaving out a part of the minor lexical categories like acronyms (e.g. *NASA*, *FBI*), exclamation marks, or unit symbols (e.g. *cm*, *kg*) setting them aside for future developments of LiITA.

For the time being, the Nuovo De Mauro's PoS categorisation rationale was adopted with some in-house adjustment. In fact, the Nuovo De Mauro's PoS categorisation rationale was mapped to the UPOS tagset. The original tagging was that of the Italian grammarian tradition, hence we had to adapt some tags, for example conjunctions. As a matter of fact, De Mauro's conjunctions didn't distinguish between subordinate and coordinate, so, we aligned manually each of the dictionary's conjunctions to the UPOS tags. For the rest of De Mauro's PoS we have manually found the correspondence with UPOS tagset.

4. Conclusion and Future Work

In this paper we presented the first steps towards the publication as LLOD of a collection of canonical forms of citation (lemmas) for Italian. Such Lemma Bank is the core component of LiITA, a knowledge base of interoperable linguistic resources for Italian inspired by the LiLa knowledge base for Latin. LiITA aims to compensate the current lack of interoperability between Italian resources, as well as to become the pivot to interlink all the present and future lexicons and corpora for Italian. To this aim, the Lemma Bank is modelled such that it can harmonise different lemmatisation criteria found in lexical and textual resources, following a bottom-up approach rather that a top-down one.

Building a Lemma Bank to make distributed resources interoperable in Linked Data is an open-ended process. As the linking of more and more resources to the KB might require the inclusion of new lemmas, the LiITA Lemma Bank will keep on growing, both through the extraction of lemmas from other lexical sources and in a resource-driven fashion.

Beside extending the Lemma Bank and linking the first resources, the LiITA project will develop online services, following what has been done for LiLa [13]. The process of linking a text or corpus in the KB must be supported by an accessible tool performing automatic lemmatisation, PoS-tagging and linking. Currently, a new Stanza model [14] has been trained combining all the existing Italian treebanks. This model will serve as the foundation for the linkage process of textual resources to be included in the LiITA KB.[33] The advanced interrogation of data offered by all the resources interlinked in LiITA will be eased by a graphical interface which will help with the task of writing complex SPARQL queries.

Finally, given its language-independent architecture and the use of common vocabularies for knowledge description, LiITA promises to have a substantial methodological impact on how linguistic resources are published and made interoperable as Linked Data.

[33]The current model's performances are presented in Table 2 in Appendix. The model can be found at https://github.com/LiITA-LOD/LiITA$_N LP_M odels$

Acknowledgments

This contribution is funded by the European Union - Next Generation EU, Mission 4 Component 1 CUP J53D2301727OOO1. The PRIN 2022 PNRR project **"LiITA: Interlinking Linguistic Resources for Italian via Linked Data"** is carried out jointly by the Università Cattolica del Sacro Cuore, Milano and the Università di Torino.

References

[1] E. Pianta, L. Bentivogli, C. Girardi, Multiwordnet: developing an aligned multilingual database, in: First international conference on global WordNet, 2002, pp. 293–302.

[2] A. Roventini, R. Marinelli, F. Bertagna, ItalWordNet v.2, 2016. URL: http://hdl.handle.net/20.500.11752/ILC-62, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

[3] R. R. Favretti, F. Tamburini, C. De Santis, Coris/-codis: A corpus of written italian based on a defined and a dynamic model, A rainbow of corpora: Corpus linguistics and the languages of the world (2002) 27–38.

[4] C. Mauri, S. Ballarè, E. Goria, M. Cerruti, F. Suriano, et al., Kiparla corpus: a new resource for spoken italian, in: CEUR WORKSHOP PROCEEDINGS, SunSITE Central Europe, 2019, pp. 1–7.

[5] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Scientific american 284 (2001) 34–43.

[6] E. J. Miller, An introduction to the resource description framework, Journal of library administration 34 (2001) 245–255.

[7] C. Chiarcos, S. Moran, P. N. Mendes, S. Nordhoff, R. Littauer, Building a linked open data cloud of linguistic resources: Motivations and developments, The People's Web Meets NLP: Collaboratively Constructed Language Resources (2013) 315–348.

[8] N. Ide, J. Pustejovsky, What does interoperability mean, anyway? toward an operational definition of interoperability for language technology, in: Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China, 2010.

[9] P. Cimiano, C. Chiarcos, J. P. McCrae, J. Gracia, Linguistic Linked Data: Representation, Generation and Applications, Springer, Cham, 2020. URL: https://www.springer.com/gp/book/9783030302245. doi:10.1007/978-3-030-30225-2.
[10] J. P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, P. Cimiano, The ontolex-lemon model: development and applications, in: Proceedings of eLex 2017 conference, 2017, pp. 19–21.
[11] S. Petrov, D. Das, R. McDonald, A Universal Part-of-Speech Tagset, in: N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2089–2096. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.
[12] T. De Mauro, Grande dizionario italiano dell'uso-Gradit, UTET, 1999.
[13] M. Passarotti, F. Mambrini, G. Moretti, The services of the lila knowledge base of interoperable linguistic resources for latin, in: Proceedings of the 9th Workshop on Linked Data in Linguistics@ LREC-COLING 2024, 2024, pp. 75–83.
[14] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Appendix

Table 2
Performance the current LiITA model.

Metric	Prec.	Recall	F1 Score	Al.Acc.
Tokens	99.81	99.77	99.79	
Sentences	89.26	89.66	89.46	
Words	99.62	99.61	99.61	
UPOS	97.03	97.02	97.03	97.41
XPOS	92.69	92.68	92.68	93.04
UFeats	94.66	94.65	94.65	95.02
AllTags	90.61	90.60	90.60	90.96
Lemmas	97.39	97.38	97.39	97.77
UAS	86.49	86.48	86.48	86.82
LAS	82.31	82.30	82.31	82.63
CLAS	75.90	75.61	75.76	76.00
MLAS	69.37	69.09	69.23	69.45
BLEX	73.89	73.60	73.75	73.99