Introducing a Gold Standard Corpus from Young Multilinguals for the Evaluation of Automatic UD-PoS Taggers for Italian

Veronica Juliana Schmalz[1,4], **Jennifer-Carmen Frey**[2], **Egon W. Stemle** [2,3]
1. Free University of Bozen-Bolzano, Bozen, Italy
2. Institute of Applied Linguistics, Eurac Research, Bozen, Italy
3. Faculty of Informatics, Masaryk University, Brno, Czech Republic
4. KU Leuven, *imec* research group *itec*, Kortrijk, Belgium
veronicajuliana.schmalz@kuleuven.be,
jennifercarmen.frey@eurac.edu, egon.stemle@eurac.edu

Abstract

Part-of-speech (PoS) tagging constitutes a common task in Natural Language Processing (NLP) given its widespread applicability. However, with the advance of new information technologies and language variation, the contents and methods for PoS-tagging have changed. The majority of Italian existing data for this task originate from standard texts, where language use is far from multifaceted informal real-life situations. Automatic PoS-tagging models trained with such data do not perform reliably on non-standard language, like social media content or language learners' texts. Our aim is to provide additional training and evaluation data from language learners tagged in Universal Dependencies (UD), as well as testing current automatic PoS-tagging systems and evaluating their performance on such data. We use Italian texts from a multilingual corpus of young language learners, LEONIDE, to create a tagged gold standard for evaluating UD PoS-tagging performance on non-standard language. With the 3.7 version of Stanza, a Python NLP package, we apply available automatic PoS-taggers, namely ISDT, ParTUT, POSTWITA, TWITTIRÒ and VIT, trained with diversified data, on our dataset. Our results show that the above taggers, trained on non-standard data or multilingual treebanks, can achieve up to 95% of accuracy on young multilingual learner data, if combined.

1 Introduction

Part-of-Speech (PoS) tagging relates to the assignment of tags or labels to the words, punctuation marks and symbols of a text. It constitutes a basic task in NLP, with applications ranging from machine translation to speech recognition and beyond. PoS-tags usually correspond to the morphosyntactic word classes of a given language, i.e. nouns, verbs, conjunctions, etc. Since each language contains specific linguistic characteristics that distinguish itself from others, tagsets are usually language dependent. The first automatic tool for the assignment of PoS-tags in the Italian language was the TreeTagger built at the University of Stuttgart (Schmid, 1994) to perform lemmatization and PoS-tagging contemporarily. Another milestone in the history of Italian PoS-tagging is the so-called Baroni's TreeTagger tagset, released in 2003. It represents the initially most adopted tagset, containing no less than 50 labels, half exclusively dedicated to verbs (Baroni et al., 2004). Along with the latter, *Tanl* (Attardi and Simi, 2009) constitutes an additionally relevant and comprehensive tagset for Italian. It counts with numerous tags and includes morphological word features. Three subcategories with different numbers of elements can be found in it, namely 14 coarse-grained tags, 37 fine-grained tags and 336 morphed tags.

Originally, automatic tagging methods were mainly employed with standard texts, such as essays, literature, and newspaper articles (Del

Monte et al., 2007; Baroni et al., 2004). However, with the advent of new communication systems and the expansion of language studies to more informal and common areas, attention started to shift to non-standard texts. In this regard, in several of the EVALITA periodic evaluation campaings for Italian NLP and speech tools, PoS tagging non-standard language has been a topic of interest (cf. Tamburini, 2007; Attardi and Simi, 2009; Bosco et al., 2016, Bosco et al., 2020). These tasks proved that PoS-tagging still represents an unsolved issue when it comes to less widely used language from different domains. Therefore, more studies and investigations are needed on specific language varieties.

Learner corpora exhibit a number of characteristics that differentiate them from the rest. In particular, numerous code-switching and code-mixing phenomena are common among them, as well as the presence of orthographical, syntactic and/or grammatical errors (Di Novo et al., 2019). More in detail, our data exhibited some peculiarities, for example the co-presence of variants for concepts ("Franco viene a casa e vede che *fuocare*/*brenn*") or new words combining different languages and morphologies ("Se sarò un giocatore famoso *richerò money*"). Given these distinctive aspects, analysing them in the context of PoS-tagging can offer interesting insights from the point of view of both the conception of these systems and their linguistic implications.

The rest of the paper is organized as follows. Section 2 provides relevant details concerning the Universal Dependencies (UD), as well as available Italian treebanks and taggers[1]. A brief overview about the differences in tagging standard and non-standard texts is presented in Section 3. Section 4 describes the methods and metrics commonly used for the evaluation of automatic taggers. We outline the tools and methodologies used for our experiments in Section 5 and the gold standard in Section 6. Next, in Section 7, we report the obtained results and in the subsequent section, namely 8, we discuss our findings, consider possible future works and draw our final conclusions.

[1] In this paper we use this term to refer to the Stanza models trained with the different available Italian Treebanks.

2 Universal Dependencies and Italian Treebanks

Over the years, alongside the different taggers and treebanks of each language, a new language-independent framework in PoS annotation has emerged, the Universal Dependencies (UD). UD is a cross-linguistic project with the aim of building common annotation frameworks for several world languages. Underlying the Universal Dependencies annotation scheme are universal Stanford dependencies (Marneffe et al., 2008), Google universal PoS-tags (Petrov et al., 2011) and the Interset interlingua for morphosyntactic tagsets (cf. McDonald et al., 2013). In particular, for the Italian language, the UD counts seven different Treebanks. These are VIT, or the Venice Italian Treebank (Delmonte et al., 2007), ISDT, Italian Stanford Dependency Treebank (Bosco et al., 2014), ParTUT, or the Parallel Text Universal Treebank (Sanguinetti et al. 2014), PoSTwita (Bosco et al., 2016), TWITTIRÒ (Cignarella et al., 2018), Valico-UD (Di Novo et al., 2019) and PUD, or the Parallel Universal Dependencies Treebank (Zeman et al., 2018). The UD universal Italian tagset counts a total of 17 different labels (Universal Dependencies, 2021).

3 Pos-Tagging Standard vs Nonstandard Language

Among the various available treebanks and taggers for Italian, most have been created using exclusively standard data, such as newspaper articles, non-fictional texts, talks and Wikipedia pages for training the models (as in the case of VIT, ISDT, ParTUT and PUD). However, recently more attention has been placed on the creation of linguistic resources for nonstandard language, as the quantity and dissemination of this type of content increases exponentially, so does the need for suitable tools for its analysis and exploitation. In this respect, PoSTwita (Bosco et al., 2016) and TWITTIRÒ (Cignarella et al., 2018) resorted to additional non-standard Italian linguistic data from Twitter, while Valico-UD (Di Novo et al., 2019) used texts from Italian learners for the creation of their treebanks. Some of the main reasons why the use of standard language data outweighs that of nonstandard data are

difficulties concerning the automatic processing and annotation of such texts. This applies especially when seeing the considerable amount of variation they contain, not only in the language itself, but also in the usage domains and among the individual language users (cf. Plank, 2016 and Sanguinetti et al. 2020). As a matter of fact, some distinctive features of non-standard texts are the broad variation in the structure and punctuation of utterances, namely in the syntax, but also at lexical level due to the use of abbreviations, domain-specific symbols or incorrect derivational forms, as well as code-switching for learners' language. The latter are likely to lead to issues regarding both automatic language processing, such as tokenization and lemmatization, and PoS-tagging, especially in the case of non-suitable or incomplete standard treebanks. For these reasons, the creation of resources from non-standard texts, like social media users or language learners, is crucial.

4 Evaluation of Automatic PoS-Taggers

When it comes to evaluating the performance of a PoS-tagger, generally an annotated gold standard reference corpus is used. The latter requires a distribution of the particular linguistic phenomena that is representative of the PoS-tagger's target application. Additionally, since a PoS-tagger combines several functions, like tokenization, word/sentence segmentation, and PoS-tag disambiguation, one of these parts must be firstly chosen as the test object. After selecting the aspect under analysis, it is necessary to choose which metrics to use to compare the results. The metrics commonly adopted for the evaluation of the tags assigned to a linguistic corpus are accuracy, precision, recall, F1-scores and Cohen's *K* (cf. Arstein and Poesio, 2008). These metrics vary not only in terms of the aspects they measure but also according to the type of data that constitute the corpus and its size.

Although various available UD taggers for Italian exist, little is known about how these perform on non-standard data. Some evaluations have been done on user-generated texts in social media (Bosco et al. 2016; Cignarella et al. 2018) and recently also on spoken language (Bosco et al. 2020) and adult learners of Italian with English, French, German and Spanish as first languages (Di Novo et al. 2019). However, this is still a nascent process, and the number of studies and analysed varieties are limited. Therefore, a closer examination and evaluation of an automatic tagger on an additional non-standard resource from a different domain promises to enhance our knowledge about PoS-tagging.

5 Methodology

In this study, we evaluate automatic PoS-tagging on the LEONIDE corpus (Glaznieks et al., 2020) to investigate how existing tagging models trained with the already available Italian treebanks perform with data from young language learners.

Given the inaccessibility of an evaluation sample for UD PoS-tagging on Italian learner language, we built our own pre-tokenized gold standard sample (see Section 6). Once we had our gold standard, we created a processing pipeline to test available tagging models for Italian on our data. For this, we used Stanza, a Python natural language analysis package designed using the UD formalism, as it offered easy access to a number of pre-trained models for PoS-tagging UD in Italian. The following models have been used in our evaluation: ISDT, ParTUT, POSTWITA, TWITTIRÒ and VIT. In order to evaluate only the PoS-tag disambiguation step of the PoS-taggers, regardless from other steps such as tokenization, we tagged the pre-tokenized texts using Stanza but deactivated the tokenizer (*tokenize_pretokenized=True*) and selected a different model as parameter each time. With the results obtained from each model, we resorted to *sklearn.metrics.classification_report* and *sklearn.metrics.cohen_kappa_score* to evaluate the total number of tags assigned to the more than 7,000 gold standard tokens according to accuracy and Cohen's *K*. In this way, the use of the exact same tokens and comparison metrics would have allowed an equal and meaningful comparison.

We closely focused on the accuracy and Cohen's *K* values (cf. Artstein and Poesio, 2008) because the first allowed us to check the overall performance of the tagger as well as the results on each tag's class, and the second to evaluate the similarity between the gold-standard and the automatically assigned tags.

As the available models had been trained on different data, both in quantity and type compared to each other but also compared to our corpus, it was particularly interesting to consider how they would deal with the young language learner data at hand, but also which type of errors they would make. We thus investigate common misclassifications for taggers and human annotators, discussing possible improvements and considerations to bear in mind when using these automatic PoS-tagging systems. For the latter, we

used confusion matrices, so that we could check the types of errors made, and which were the most correctly assigned tags out of the total.

6 Gold Standard

For the creation of our gold standard, we used a subset of the Longitudinal lEarner cOrpus iN Italian, Deutsch, English (LEONIDE) (Glaznieks et al., 2020), a collection of 2,512 texts from 163 trilingual pupils attending lower secondary school (*scuola media*) in South Tyrol. The corpus contains texts in three languages, namely English, German, and Italian, and in two text genres, meaning *narrative* in the form of a picture-inspired story and *argumentative* in the form of a simple opinion text. Over the span of three years, the pupils were asked to write one text for each of the three languages and each of the text genres per year. The portion of Italian data in the corpus amounts to 844 texts counting 93,378 tokens. For our gold standard[2], we randomly selected a sample of 10% of the total available Italian texts, i.e. 84 texts with 7,665 tokens. We pre-tokenized and pre-tagged the texts in the sample using Stanza with the combined PoS-tagging model[3] in order to present our annotators with vertical files with one token per line and a PoS-tag to be eventually corrected. Once this step was completed, two language experts, native speakers of Italian, independently annotated the texts, correcting and adjusting the automatically pre-tagged version using the guidelines and documentation for the UD PoS tags and making use of the whole UD tagset. Their inter-annotator agreement in the independent tagging was relatively high, achieving a Cohen's Kappa of 0.98. In order to investigate a possible effect given by the use of a pre-tagged corpus version by the annotators, we also tested tagging the texts from scratch, meaning without any pre-assigned labels in the tokenized texts. For this purpose, we selected a random sample of ten texts extracted from the original corpus. Once again, to compare the two tagged versions we calculated the Cohen's *K* value, which resulted in 0.95. Hence, we can conclude that the pre-tagged version had no particular effect on the annotators and did not significantly affect their annotation.

Despite the generally good agreement between the annotators, some difficulties emerged. These mainly concerned cases of German code-switching, particles, clitic pronouns and auxiliary verbs (see Discussion), and occasionally orthographical or overgeneralization errors (ex. Da grande *facherò* [X/VERB] il calciatore). For the gold standard these issues were unanimously resolved in accordance with the Italian UD guidelines[4].

7 Results

Table 1 displays the obtained results in terms of tagging models' accuracy and Cohen's *K*, this time comparing the gold standard and the taggers' assigned tags, along with the accuracy scores reported in Stanza for the CoNLL 2018 Shared Task[5] on UD v2.5 Treebanks evaluation.

Tagger	Training data (in tokens)	Accuracy (Stanza)	Accuracy on learner data	Cohens' K (learner data)
Combined	Pre-trained	-	**0.95**	0.94
TWITTIRÒ	28,387 (ironic tweets)	0.94	0.86	0.84
ParTUT	Multilingual parallel treebank	0.98	0.84	0.82
PoSTWITA	119,238 (Tweets)	0.96	0.79	0.77
ISDT	278,429 (articles, newspapers, legal texts, Wikipedia)	0.98	0.76	0.73
VIT	272,000 (news, bureaucracy, finance, science, literature texts)	0.95	0.75	0.72

Table 1. Comparison of taggers' results on the LEONIDE's dataset (with additional training information) in terms of accuracy and Cohens' K values.

The highest accuracy on our gold standard for learner data has been achieved by the combination of models chosen by Stanza per default. We

[2]Available at http://hdl.handle.net/20.500.12124/34

[3]This indicates the Stanza model which originates from a combination of the existing taggers given by the Treebanks for the Italian language https://stanfordnlp.github.io/stanza/combined_models.

[4]https://universaldependencies.org/it/

[5]https://universaldependencies.org/conll18/evaluation.html

would have expected better results from ISDT, considering the high accuracy values on the standard data used to train it, and PoSTWITA for non-standard texts. However, regardless of this, in respect to our gold standard, the best models for accuracy value and Cohen's K are TWITTIRÒ and ParTUT. These latter performed well despite the fact that their tagsets did not contain all the tags used in our gold standard. In fact, both TWITTIRÒ and ParTUT, as well as PoSTWITA, did not include the PART tag (contrary to the other treebanks such as VIT and ISDT) , and thus did not assign it to particles. However, our human annotators referred to this tag to mark the common use of pronominal, reflexive and adverbial particles, such as 'mi' and 'si' in the corpus (ex. Più lingue *ci* sono; *Si* deve studiare molto). Furthermore, the parTUT treebank also lacked the tag for interjections, INTJ, as opposed to other treebanks that did make use of this category. Nevertheless, the training data for TWITTIRÒ was the treebank provided in Stanza that was closest to our data in type. It was created using data from social networks, therefore far from the scientific, nonfictional, or journalistic canon. On the other hand, ParTUT had been designed using standard texts but in Italian, English, and French in parallel.

8 Discussion

The results show that the performance of the models was significantly influenced by the particular type of data in our gold standard corpus, which presented incorrect orthographical or morphological tokens, but also contained numerous foreign words and abnormally disposed parts-of-speech within the sentence.

In fact, when inspecting the tags incorrectly assigned by the different models with confusion matrices (see Figure 1, 2 and 3 below), we noticed that:

- The foreign or misspelled words, which according to the UD rules had to be assigned the X tag, proved to be those with the highest number of errors. In fact, they were often confused with proper nouns, PROPNs, especially in the case of code-switching with the German language, where nouns are spelled with initial capital letters (ex. Dopo la scuola media voglio fare la *Hotelfachschule* [~~PROPN~~-X]). This was particularly evident with the ParTUT model that did not assign the X tag at all (see Figure 3);
- The second most incorrectly tagged words were particles, PART, which are not included in the tagsets of all models although they could have been assigned to pronominal, reflexive and adverbial particles (see section 7). Instead, these words were usually assigned the PRON tag for pronouns (ex. *Si* [~~PRON~~-PART] deve parlare questa lingua);
- The third most inaccurate group of tagged words was that of interjections, INTJ, which were also not included in all treebanks. These were often confused with particles or foreign words, PART or X (ex. *Ehm* [~~X~~-INTJ] ciao! fece Alessandra) as it is visible from Figure 2 in the case of the TWITTIRÒ model.

On the other hand, regarding discrepancies between the tags assigned by the human annotators, we found that:

- The groups on which there was most disagreement between the two annotators concerned particles, PART, and auxiliary verbs, AUX[6] (ex. Le strategie che funzionano peggio *sono* [~~AUX~~-VERB] studiare con il computer). Concerning the first, the models did not always include the PART tag in their employed tagsets. Auxiliary verbs, additionally, were also at times abnormally positioned within the sentence and were often automatically annotated incorrectly.
- Foreign words were often not annotated according to the X tag, probably because the annotators also had knowledge of the German language and therefore tended to assign the corresponding tag in the other language (ex. Faccio la *Landesberufschule* [~~NOUN~~-X]).

We can therefore argue that there were errors common to both automatic models and annotators, although the reasons for the errors were evidently different.

[6] This might be due to the fact that annotators could be influenced by the presence next to each token of a tag automatically assigned by Stanza, that had performed the tokenization of the texts.

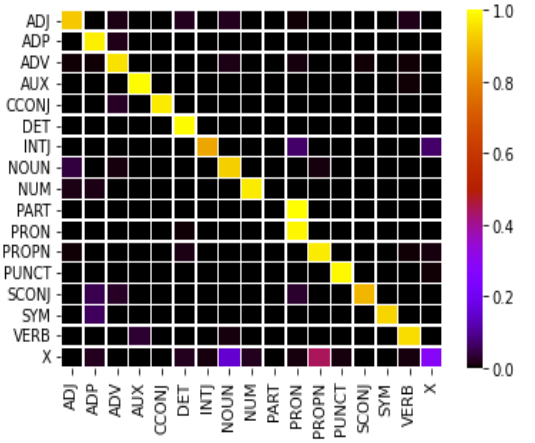

Figure 1. Confusion matrix related to the *combined* tagger (95% accurate)

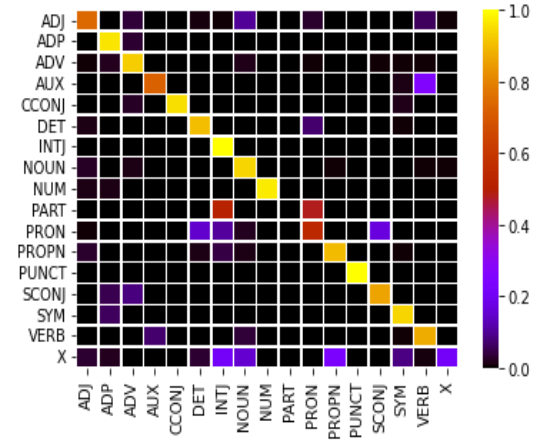

Figure 2. Confusion matrix related to the *TWITTIRÒ* tagger (86% accurate)

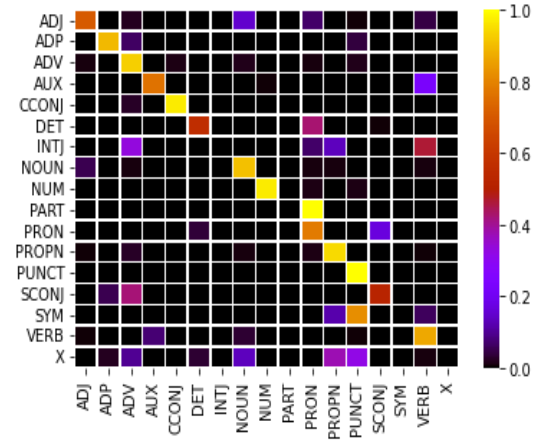

Figure 3. Confusion matrix related to the *ParTUT* tagger (84% accurate)

9 Conclusion

Although all taggers managed to execute the task of automatically PoS-tagging pre-tokenized Italian non-standard language with an accuracy of at least 75% (with the combined model offered by Stanza showing the best performance with 95% accuracy and 0.94 Cohen's *K*), there were differences in the performance shown by the individual models. The best performing two individual models were TWITTIRÒ (86%) and ParTUT (84%), while ISDT and PoSTWITA, that performed better in other evaluation tasks (Bosco et al. 2014, Cignarella et al. 2018) had a lower accuracy on our data. These results hint towards the fact that in order to automatically tag non-standard texts relating to language learners, the use of high-performance systems in the generic task is not sufficient, but the characteristics of the actual texts must also be taken into account.

Improvements could be made in the future regarding the adaptation of the models to the particular type of data used here. They could be, indeed, re-trained again in case a complete Treebank with Italian non-standard data becomes available. In addition, further attempts could be made to adapt or add the missing tags to the tagsets of all models so as not to have results biased by the lack of matching tags. Finally, as far as the annotators are concerned, they could be provided with the automatically pre-tokenized texts from the models, but in order to avoid pre-assigned tags influencing their annotation process, it would be preferable to omit these. Thus, human annotators would only get the taggers' tokenized text versions, so that the same tokens will be available for everyone, while the assignment of PoS would be completely up to them.

References

Artstein, Ron and Poesio, Massimo (2008). Inter-Coder Agreement for Computational Linguistics. Association for Computational Linguistics.

Attardi, Giuseppe and Simi, Maria (2009). Overview of the EVALITA 2009 Part-of-Speech tagging task. *Poster and Workshop proceedings of the 11th Conference of the Italian Association for Artificial Intelligence,* 12 December 2009, Reggio Emilia.

Baroni, Marco, Bernardini, Silvia, Comastri, Federica, Piccioni, Lorenzo, Volpi, Alessandra, Aston, Guy and Mazzoleni, Marco (2004). Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04).* Lisbon: ELDA, 1771-1774.

Bosco, Cristina, Dell'Orletta, Felice, Montemagni, Simonetta, Sanguinetti, Marco and Simi, Maria (2014). The Evalita 2014 Dependency Parsing task. *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of*

the Fourth International Workshop EVALITA 2014, 9-11 December 2014, Pisa, 1-8.

Bosco, Cristina, Tamburini, Fabio, Bolioli, Andrea and Mazzei, Alessandro (2016). Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian task. In: Tamburini, F. (Ed.), *EVALITA Evaluation of NLP and Speech Tools for Italian. Proceedings of the Final Workshop*, Torino: Accademia University Press, 78-84.

Bosco, Cristina, Ballaré, Silvia, Cerruti, Massimo, Goria, Eugenio, & Caterina, Mauri (2020). KIPoS@ EVALITA2020: overview of the task on Kiparla part of speech tagging. In EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (pp. 1-8). CEUR.

Cignarella, Alessandra Teresa, Bosco, Cristina, Patti, Viviana, & Lai, Mirko (2018). Application and analysis of a multi-layered scheme for irony on the Italian Twitter Corpus TWITTIRÒ. In: Calzolari, Nicoletta, Choukri, Khalid, Cieri, Christopher, Declerck, Thierry, Goggi, Sara, Hasida, Koiti, Isahara, Hitoshi, Maegaard, Bente, Mariani, Joseph, Mazo, Hélène, Moreno, Asuncion, Odijk, Jan, Piperidis, Stelios and Tokunaga, Takenobu (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki*:* European Language Resources Association, 4204-4211.

De Marneffe, Marie Catherine and Manning, Christopher D. (2008). Stanford typed dependencies manual. Technical report, Stanford University, 338-345.

Delmonte, Rodolfo, Bristot, Antonella and Tonelli, Sare (2007). VIT - Venice Italian Treebank: Syntactic and Quantitative Features. In: De Smedt, Koenraad, Hajic, Jan and Kübler, Sandra (Eds.), *Proceedings Sixth International Workshop on Treebanks and Linguistic Theories*, Bergen: Northern European Association for Language Technology (NEALT) Proceedings Series Vol.1, 43-54.

Di Novo, Elisa, Bosco, Cristina, Mazzei, Alessandro and Sanguinetti, Manuela (2019). Towards an Italian Learner Treebank in Universal Dependencies. In: Bernardi, Raffaella, Navigli Roberto and Semeraro, Giovanni (Eds.), *Proceedings of the 6th Italian Conference on Computational Linguistics, CliC-it 2019* (Vol. 2481), Bari, Italy, CEUR WS: 1-6.

Glaznieks, Aivars, Frey, Jennifer-Carmen, Stopfner, Maria, Zanasi, Lorenzo and Nicolas, Lionel (2020): LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research* (IJLCR).

McDonald, Ryan, Nivre, Joakim, Quirmbach-Brundage, Yvonne, Goldberg, Yoav, Das, Dipanjan, Ganchev, Kuzman, Hall, Keith, Petrov, Slav, Zhang, Hao, Täckström, Oscar, Bedini, Claudia, Bertomeu Castelló, Nuria and Lee, Jungmee (2013). Universal Dependency Annotation for Multilingual Parsing. In: Schütze, Henrich, Fung, Pascale, Poesio, Massimo (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia: Association for Computational Linguistics, 92-97.

Petrov, Slav, Das, Dipanjan and McDonald, Ryan (2011). A universal part-of-speech tagset. In: Calzolari, Nicoletta, Choukri, Khalid, Declerk, Thierry, Uğur Doğan, Mehmet, Maegaard, Bente, Mariani, Joseph, Moreno, Asuncion, Odijk, Jan and Piperidis, Stelios (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Instanbul: European Language Resources Association (ELRA)*, 2089-2096.

Plank, Barbara (2016). What to do about non-standard (or non-canonical) language in NLP. In: Sharma Misra, Dipti, Sangal, Rajeev, Singh Kumar, Anil (Eds.), *Proceedings of the 13th Conference on Natural Language Processing*, Varanasi: NLP Association of India.

Sanguinetti, Manuela and Bosco, Cristina (2014). Converting the parallel treebank ParTUT in Universal Stanford Dependencies. In: Basili, Roberto, Lenci, Alessandro and Magnigni, Bernardo (Eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa*, Pisa: Pisa University Press, 316-321.

Sanguinetti, Manuela, Cassidy, Lauren, Bosco, Cristina, Çetinoğlu, Özlem, Cignarella, Alessandra Teresa, Lynn, Teresa, Rehbein, Ines, Ruppenhofer, Josef, Seddah Djamé & Zeldes, Amir (2020). Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. arXiv preprint arXiv:2011.02063.

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Tamburini, Fabio (2007). EVALITA 2007: The Part-of-Speech Tagging Task. *Contributi scientifici Associazione Italiana per l' Intelligenza Artificiale.* Anno IV, Giugno 2007.

Universal Dependencies (February 2021). UD for Italian. https://universaldependencies.org/it/.