Join Together? Combining Data to Parse Italian Texts

Claudia Corbetta[1,2,*,†], Giovanni Moretti[3,†] and Marco Passarotti[3,†]

[1]Università degli studi di Bergamo, via Salvecchio 19, 24129 Bergamo, Italy.
[2]Università di Pavia, corso Strada Nuova 65, 27100 Pavia, Italy.
[3]Università Cattolica del Sacro Cuore, largo A. Gemelli 1, 20123 Milan, Italy.

Abstract
In this paper, we create and evaluate non-combined and combined models using Old and Contemporary Italian data to determine whether increasing the size of the training data with a combined model could improve parsing accuracy to facilitate manual annotation. We find that, despite the increased size of the training data, in-domain parsing performs better. Additionally, we discover that models trained on Old Italian data perform better on Contemporary Italian data than the reverse. We attempt to explain this result in terms of syntactic complexity, finding that Old Italian text exhibits higher sentence length and non-projectivity rate.

Keywords
Parsing, Universal Dependencies, Combined Model, Old Italian, Contemporary Italian, Non-Projectivity

1. Introduction

High-quality textual data (semi-)manually enhanced with different layers of metalinguistic annotation are extremely valuable resources for conducting linguistic analysis. As for the syntactic layer of annotation, the de facto standard for dependency-based annotation is Universal Dependencies (UD),[1] an initiative that provides machine-readable annotations for a wide variety of languages, including historical languages [1]. At the current state of art,[2] Contemporary Italian is well-represented in UD, whereas Old Italian is only represented by one annotated text (a portion of the *Divine Comedy* of Dante Alighieri). The creation of additional Old Italian annotated data is therefore advisable.

Since a fully manual annotation process is time-consuming and requires significant effort, we aim to expedite it by using a parser that pre-parses the data, leaving the human annotator with only a manual revision task.

To address this, given the scarcity of Old Italian data, we create a combined parser using both Contemporary and Old Italian data. The objective is to determine whether a combined model with an expanded training dataset performs better compared to non-combined models (see [2] for Spanish language and [3] for Stanza combined models).

The paper is organised as follows: Section 2 provides a brief description of the Italian language, the syntactic resources and the Italian data available; Section 3 details the data used for the experiments, presents the performances of non-combined and combined models, and evaluates their performances; Section 4 analyzes the syntactic complexity of each test set (Old and Contemporary Italian) to address accuracy differences; and finally, Section 5 provides the conclusion.

2. Talking about Italian

Italian is a Romance language derived from Latin, and its development is closely connected with the political, cultural and economic system of Italy during the Late Middle Ages [4, 5, 6, 7]. Even though the evolution and history of the Italian language "can be properly understood only within the wider context of the evolution of the Italian dialects" [5, p. 3], the dialect spoken in Florence (Tuscany) in the thirteenth century, known as Florentine, played a pivotal role in establishing the foundation of the Italian language. The pre-eminence of Florentine over other Italian dialects was established due to the importance and prestige of Florentine literature. Its widespread success contributed to the codification of Florentine as the *lingua volgare* in the sixteenth century, distinguishing it as the spoken Italian language in contrast to Latin, which was still used for written cultural discourse [8].

Even though Florentine (and, more generally, Tuscan dialects) is considered conservative in its linguistic evolu-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
* Corresponding author.
† This paper is the result of the collaboration between the three authors. For academic purposes, all the authors are responsible of Sections 1,3,4,5; Claudia Corbetta is also responsible for Section 2, and Giovanni Moretti built the Stanza Models. Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
claudia.corbetta@unibg.it (C. Corbetta); giovanni.moretti@unicatt.it (G. Moretti); marco.passarotti@unicatt.it (M. Passarotti)
0009-0000-7425-196X (C. Corbetta); 0000-0001-7188-8172 (G. Moretti); 0000-0002-9806-7187 (M. Passarotti)

[1]See https://universaldependencies.org.
[2]We refer to version 2.14 of UD.

tion [5, p. 5], it is now widely recognized by most scholars as distinct from Contemporary Italian [9, p. 8]. Among the differences between Contemporary Italian and Florentine (henceforth referred to as Old Italian),[3] several syntactic distinctions have been noted [10, 11]. These include, among others, the position and order of clitics, the use of the marker *sì* 'that' as a thematic marker, and differences in the use of compound tenses [11, p. 425-444].

2.1. Syntactic resources

High-quality (semi-)manually annotated treebanks, i. e. corpora with annotations on various linguistic levels,[4] are indispensable tools for in-depth analysis of the syntax (and morphology) of languages. Treebanks not only facilitate faster, easier, and more precise querying of syntactic structures, but also aid in tracking the evolution of syntactic patterns in languages through time [13].

Among the dependency treebanks, UD is a pivotal initiative displaying cross-linguistically consistent treebanks for many languages [14]. As of the current version 2.14, UD includes 283 treebanks and 161 languages, encompassing historical languages such as Latin (e.g. *Index Thomisticus* Treebank, ITTB [15]), Old French (PROFITEROLE [16]) and Ancient Greek (e.g. PROIEL [17]), among others.

In Subsection 2.2, we describe UD treebanks of Italian language.

2.2. Italian data

Regarding Italian, UD includes 9 Contemporary Italian treebanks, spanning various genres, as reported in Table 1.

Table 1
Contemporary Italian UD treebanks (in UD 2.14).

Treebank	Syntactic words	Genre
ISDT	298K	legal,news, wiki
VIT	280K	news, non fiction
ParTUT	55K	legal, news, wiki
ParlaMint	20K	government legal
TWITTIRO	29K	social
Valico	6K	learner-essays
PoSTWITA	124K	social
MarkIT	40K	grammar-examples
PUD	23K	news, wiki

Concerning Old Italian, the only treebank present in UD is Italian-Old [18], encompassing the *Divine Comedy*, a poetic text written by Dante Alighieri (1 265-1 321). Currently, Italian-Old contains the first two *Cantiche* of the poem, namely *Inferno* and *Purgatorio*, amounting 80 694 tokens, 82 644 syntactic words[5] and 2 402 sentences.[6]

The divergence in annotated data available for Contemporary Italian (around 875K syntactic words) versus Old Italian (82K syntactic words) is considerable.

Considering that i) treebanks are essential for expanding the sample of comparable data and that ii) the manual annotation of data is an extremely time-consuming effort, the development of automatic parsers is crucial to expedite and assist the annotation process.

The shortage of gold-annotated data for Old Italian, compared to the large amount of data available for Contemporary Italian, led us to recognize the potential of testing combined models, i.e., models with a training set composed of both Old and Contemporary Italian data.

3. Combining Old Italian with Contemporary Italian data

Considering the aforementioned divergence in data, we create and evaluate the performance of a combined Contemporary-Old Italian model to understand whether joining datasets from different periods could improve parsing accuracy.

We train models using Stanza [19], a neural pipeline for natural language processing, with different training sets. Specifically, we train models based on Contemporary Italian data (henceforth CI), Old Italian data (henceforth OI), and a combination of Contemporary and Old Italian data (henceforth Combi).

In Subsection 3.1 we detail the selection and partitioning of the data. Subsection 3.2 outlines the creation of models and presents the resulting scores. Finally, Subsection 3.3 discusses the combined Contemporary-Old Italian model.

3.1. Selection and partitions of data

To build the model based on OI data, we use the only Old Italian treebank available, Italian-Old.

Among all the Contemporary Italian UD treebanks, we select two treebanks, ISDT (Italian Stanford Dependency Treebank) and VIT (Venice Italian Treebank). We select ISDT [20], as it is the Italian treebank with the highest

[3]We adhere to the definition of Salvi and Renzi [9], who use the term Old Italian to refer to the language spoken in Florence during the 13th and 14th centuries.

[4]Treebanks usually provide information on sentence tokenization, word lemmatization, and both morphological and syntactic details. Syntactic analysis is mandatory in a treebank, and can be encoded in either dependency syntax or constituency syntax [12].

[5]We use the term "syntactic words" and "tokens" following the UD definition (see https://universaldependencies.org/u/overview/tokenization.html).

[6]The numbers refer to UD version 2.14, see https://universaldependencies.org/treebanks/it_old/index.html.

Table 2
Number of sentences (sent) and tokens (tok) for the train/dev/test partitions of each dataset.

	VIT1	VIT2	VIT3	ISDT
train	1 697 sent - 53 662 tok	2 195 sent - 52 076 tok	2 189 sent - 52 016 tok	2 766 sent - 58 091 tok
dev	356 sent - 11 515 tok	317 sent - 11 168 tok	413 sent - 11 144 tok	591 sent - 12 465 tok
test	354 sent - 11 473 tok	318 sent - 11 136 tok	438 sent - 11 096 tok	606 sent - 12 402 tok

UD star ranking. This ranking, designed by the UD organizers, quantifies various qualities of the corpora, such as their usability and the variety of genres they encompass. Moreover, since Italian-Old is based on the poetry genre, to minimize a potential genre gap (the influence of genre on parsing has been addressed in [21]), we also select VIT [22], that includes, albeit with a limited number of words, literary texts.[7] We point out that, up to now, no CI treebanks contain poetry (see Table 1).

To avoid the CI data overwhelming the OI data due to their size disparity, we partition the CI data. The VIT treebank, consisting of 259.625 tokens, 280.153 syntactic words, and 10.087 sentences, allows us to partition the data into three parts, with each part closely matching the size of the Italian-Old dataset. Specifically, we divide the VIT dataset into three partitions of 34%, 33% and 33%, respectively named VIT1, VIT2 and VIT3. Additionally, we further divide each partition (VIT1, VIT2 and VIT3) into train, test, and dev sets with a split of 70%, 15%, and 15%, the same used in Italian-Old dataset. Unlike the VIT treebank, the ISDT is not directly partitionable, as it counts 278 461 tokens, 298 375 syntactic words, and 14 167 sentences. Therefore, we shuffled the data and extracted a total of 82 500 tokens (the same size of OI data), which were then partitioned into train, dev, and test sets with a ratio of 70%, 15%, and 15%, respectively.

We report in Table 2 the partition of each datasets in train/dev/test.

3.2. Creation of models and scores

With each partition (OI, VIT1, VIT2, VIT3 and ISDT), we train 5 models using Stanza, with the training and dev sets, and we evaluate them on the respective test sets. Within the CI-VIT datasets, we retain only the model that performs best, namely VIT1.

We then use the model built on OI data to parse the CI test sets, and vice versa.

In Table 3 and Table 4, we report the scores of both Label Attachment Score (LAS) and Unlabel Attachment Score (UAS)[8] of the OI model and the VIT1, and of the OI and the ISDT respectively.

For both VIT1 and ISDT scenarios, results show that using a model trained on in-domain data, namely data that pertain to the same textual domain as the test set (VIT1 on VIT1, OI on OI, and ISDT on ISDT), yields higher performance than using out-of-domain data (ISDT on OI, VIT1 on OI, and OI on VIT1 and ISDT). These results align with literature on in-domain testing [25].

[7]The VIT treebank contains 10 000 words of literally genre [22, 23]. Refer also to the read.me to further details (see A).
[8]Refer to [24] for an insight into the aforementioned metrics.

Table 3
Evalutation metrics with VIT1 and OI models (where "->" stands for "on").

	VIT1 -> VIT1	OI -> OI	VIT1 -> OI	OI -> VIT1
LAS	71.60	75.86	42.83	68.53
UAS	77.70	82.24	56.13	75.53

Table 4
Evalutation metrics with ISDT and OI models.

	ISDT -> ISDT	OI -> OI	ISDT -> OI	OI -> ISDT
LAS	88.55	75.86	51.62	74.83
UAS	91.41	82.24	63.03	80.93

While analyzing the scores of out-of-domain parsing (ISDT on OI, VIT on OI, and OI on ISDT and VIT), we notice that the model trained on OI data performs better on CI data in both scenarios, whereas CI models yield lower scores when applied to OI text. The differences in scores are approximately 20 points in favour of the OI model, specifically 25.7 (LAS) and 19.4 (UAS) compared to VIT1, and 23.21 (LAS) and 17.9 (UAS) compared to ISDT.

We attempt to explain the outperformance of the OI model in Section 4.

3.3. Joining model

To challenge the results obtained in 3.2, we build combined models with Stanza by merging OI data with CI data. Specifically, we create two models: CombiVIT, and CombiISDT. For each combined model, the test, dev, and train sets are created by merging the corresponding test, dev, and train sets of the VIT1 data and ISDT data with those of the OI data.

In Table 5, we report the UAS and LAS scores obtained.

We notice that in both scenarios the combined models perform better on CI data than on OI data, with the combined models outperforming by 13.74 (LAS) and 10.1 (UAS) for CI-VIT data and 12.58 (LAS) and 8.87 (UAS) for CI-ISDT data.

Table 5
Evaluation metrics with combined models.

	CombiVIT -> VIT	CombiVIT -> OI	CombiISDT -> ISDT	CombiISDT -> OI
LAS	69.11	55.37	87.76	75.18
UAS	74.96	64.86	90.85	81.98

According to the results in Table 5, CI texts appear to be easier to parse, suggesting a simpler syntactic structure compared to OI text. To verify this claim and shed light on these results, in Section 4, we measure several syntactic parameters to gather information about the tree structures of both OI and CI tests.

4. An insight to OI and CI data

To analyze the complexity of tree structures in each test set (CI-ISDT, CI-VIT, and OI), we calculate:

- *type-token ratio* (TTR): the number of types divided by the number of tokens (excluding punctuation);
- *tree depth* (Depth): the longest path from the root of an oriented a-cyclic graph (i.e, the syntactic tree) to a leaf;
- *lexical density* (Lex. Den.): the number of content words, i.e. words that possess semantic content and contribute to the meaning of the sentence,[9] divided by the total number of syntactic words (excluding punctuation marks);
- *sentence length* (Length): the number of syntactic words (excluding punctuation marks) in each sentence.

Table 6 presents the average of the aforementioned measures. Additionally, we report for each test the minimum and maximum values of sentence length and tree depth.

Table 6
Average of type-token ratio, tree depth, lexical density, and sentence length of the OI, CI-ISDT and CI-VIT test sets.

	OI	CI-ISDT	CI-VIT
Avg. TTR	0.92	0.956	0.931
Avg. Depth	5.201	4.153	5.542
Avg. Lex. Den.	0.488	0.516	0.496
Avg. Length	30.095	16.873	26.636
Min - Max Length	7 - 112	1 - 92	2 - 100
Min - Max Depth	2 - 11	0 - 13	1 - 16

[9]We select as content words all words belonging to the following Universal parts of speech [26]: NOUN 'noun', VERB 'verb', ADJ ' adjective', ADV 'adverbs', and PROPN 'proper noun'.

Among the measures described, the OI test does not differ significantly from the CI values. The only measure in which the OI test differs from the CI tests is sentence length (Avg. Length): OI presents a higher average sentence length, surpassing the CI-ISDT average by 13 points and the CI-VIT average by 3.5.

Therefore, considering the parameters evaluated, only the sentence length could be considered to explain the possible overperformance of OI on CI data.

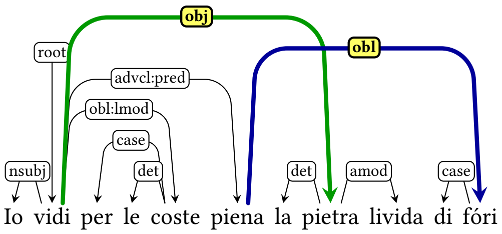
In Subsection 4.1, we evaluate another parameter that is related to the complexity of tree structure, namely non-projectivity (i.e., the number of structures where a head and its dependents form a discontinuous constituent). It has been demonstrated [27] that sentence length is interconnected with non-projectivity. Specifically, non-projective sentences exhibit greater sentence length compared to projective ones. By calculating non-projectivity, we aim to determine whether sentence length (which has been proven to be higher in OI test) and non-projectivity might indicate more complex structures in OI texts, thereby contributing to the overperformance of the OI model on CI data.

4.1. Non-projectivity

Non-projectivity arises when sentences exhibit non-local dependencies. While constituency approaches may handle similar structures using empty categories and coindexation [28], dependency-based approaches result in discontinuous dependencies that lead to non-projectivity.

We illustrate an example of non-projectivity, showing the non-local dependency relation of the oblique (`obl`) dependency relation of the node *fóri* 'holes', which is a dependent of the node *piena* 'full'. This relation causes non-projectivity with the node *pietra* 'rock', which is dependent on the root (`root`) of the sentence *vidi* 'saw' with an object (`obj`) dependency relation.

> *Inferno*, XIX, vv. 13–14:
>
> Io vidi per le coste (...) / piena la pietra livida di fóri
>
> 'Along the sides (...), / I saw that livid rock was perforated'



The non-projectivity of syntactic dependency trees presents a challenging task for parsing in natural language processing [29], with non-projective structures proving more difficult to parse. Concerning our task, we investigate the number of non-projective structures in each test set to determine whether the overperformance of OI on CIdata may be associated with a higher prevalence of non-projective structures, thereby confirming that having more non-projective structures in the training set is beneficial.

We calculate non-projectivity of the OI, CI-VIT, and CI-ISDT test sets. In Table 7 we report the total number of edges, the number of non-projective edges, and the ratio of non-projectivity expressed in percentage of each test set.

Table 7
Non-projectivity of OI, CI-VIT, and CI-ISDT test sets.

	OI	CI-VIT	CI-ISDT
Total edges	12 307	11 473	12 402
Non-projective edges	176	24	7
Non-projectivity ratio in %	1.43%	0.21%	0.06%

As shown in Table 7, OI shows a higher rate of non-projectivity compared to CI texts. In particular, the non-projectivity in OI is 7 times higher than in CI-VIT and 24 times higher than in CI-ISDT. The high rate of non-projective structures in OI could be related to the genre of the text, i.e., poetry, which reflects a more creative use of language and frequently employs inversions.

5. Conclusion

In this paper, we create and evaluate non-combined and combined models of Old Italian and Contemporary Italian data.[10] In light of the scarcity of manually annotated Old Italian data compared to the richness of Contemporary Italian data, the aim of this work is to determine whether combining data to train a combined model could lead to better accuracy in parsing, thereby facilitating the process for human annotators.

We observe that combining Contemporary Italian and Old Italian data, even though it increases the data size of the model, does not lead to better LAS and UAS accuracy scores. This confirms, in line with other studies [30, 31, 21, 32, 3], that having an in-domain training set is preferable.

Additionally, we notice that the model trained on OI data performs better on Contemporary Italian texts than the reverse (i.e. models trained on Contemporary data on OI texts). To explain these results, we investigate the syntactic complexity of each test set (OI, CI-ISDT, and CI-VIT). Specifically we evaluate sentence length, tree depth, lexical density and the type-token ratio. We notice that the tests differ only in the sentence length. We then proceed to calculate another parameter of syntactic complexity, namely non-projectivity.

We discover that OI texts present a higher number of non-projective sentences. We hypothesize that the high level of non-projectivity could be connected to the genre of OI text, namely poetry. Thus far, the lack of UD treebanks for OI prose texts and for CI poetry texts have prevented us from investigating whether the high degree of non-projectivity observed in OI test (based on the Italian-Old treebank) is characteristic of the poetry genre or specific to OI. Such question will be left for further studies.

Finally, we are currently working to increase the amount of manually annotated OI data, expanding both the range of authors and the genres of the texts considered. This will allow us to evaluate the model's performance both within and outside its domain (in terms of authorship and text typology), as well as to assess its potential applicability to other OI texts.[11]

References

[1] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308. URL: https://aclanthology.org/2021.cl-2.11. doi:10.1162/coli_a_00402.

[2] F. Sánchez-León, Combining different parsers and datasets for capitel ud parsing, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), 2020.

[3] A. Zeldes, N. Schneider, Are ud treebanks getting more consistent? a report card for english ud, 2023. URL: https://arxiv.org/abs/2302.00636. arXiv:2302.00636.

[4] B. Migliorini, Storia della lingua italiana, Bompiani, 2019.

[5] M. Maiden, Linguistic History of Italian, A, Routledge, 2014.

[6] A. Vàrvaro, La parola nel tempo. Lingua, società e storia, Bologna : Il Mulino, 1984.

[10] Models are available for public use at https://github.com/CIRCSE/Old_Italian_Model.

[11] For an overview of Old Italian resources, refer to [18].

[7] G. Rohlfs, Grammatica storica della lingua italiana e dei suoi dialetti, Torino : Einaudi, 1968.
[8] M. Vitale, La questione della lingua, Palermo : Palumbo, 1978.
[9] G. Salvi, L. Renzi (Eds.), Grammatica dell'italiano antico, il Mulino, Bologna, Italy, 2010. URL: https://www.mulino.it/isbn/9788815134585.
[10] M. Dardano, Sintassi dell'italiano antico. La prosa del Duecento e del Trecento, volume 1, Carocci, 2012.
[11] M. Dardano, G. Frenguelli, SintAnt. La sintassi dell'italiano antico, Roma, Aracne, 2004.
[12] A. Abeillé, Treebanks: Building and using parsed corpora, volume 20, Springer Science & Business Media, 2003.
[13] A. Taylor, Treebanks in historical syntax, Annual Review of Linguistics 6 (2020) 195–212.
[14] J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, D. Zeman, Universal Dependencies v2: An evergrowing multilingual treebank collection, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4034–4043. URL: https://aclanthology.org/2020.lrec-1.497.
[15] M. Passarotti, The project of the index thomisticus treebank, Digital classical philology. Ancient Greek and Latin in the digital revolution 10 (2019) 299–320. URL: https://doi.org/10.1515/9783110599572-017.
[16] S. Prévost, L. Grobol, M. Dehouck, A. Lavrentiev, S. Heiden, Profiterole: un corpus morpho-syntaxique et syntaxique de français médiéval, Corpus (2023).
[17] D. T. Haug, M. Jøhndal, Creating a parallel treebank of the old indo-european bible translations, in: Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008), 2008, pp. 27–34.
[18] C. Corbetta, M. Passarotti, F. M. Cecchini, G. Moretti, Highway to Hell. Towards a Universal Dependencies Treebank for Dante Alighieri's Comedy., in: CLiC-it, 2023.
[19] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, arXiv preprint arXiv:2003.07082 (2020).
[20] C. Bosco, F. Dell'Orletta, S. Montemagni, M. Sanguinetti, M. Simi, The evalita 2014 dependency parsing task, in: 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 University Press, 2014, pp. 1–8.
[21] F. Mambrini, M. C. Passarotti, Will a parser overtake achilles? first experiments on parsing the ancient greek dependency treebank, in: Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11). 30 November–1 December 2012, Lisbon, Portugal, Edições Colibri, 2012, pp. 133–144.
[22] L. Alfieri, F. Tamburini, (almost) automatic conversion of the venice italian treebank into the merged italian dependency treebank format., in: CEUR WORKSHOP PROCEEDINGS, volume 1749, Accademia University Press, 2016, pp. 19–23.
[23] R. Delmonte, A. Bristot, S. Tonelli, Vit-venice italian treebank: Syntactic and quantitative features., in: Sixth International Workshop on Treebanks and Linguistic Theories, volume 1, Northern European Association for Language Technol, 2007, pp. 43–54.
[24] S. Buchholz, E. Marsi, CoNLL-X Shared Task on Multilingual Dependency Parsing, in: L. Màrquez, D. Klein (Eds.), Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), Association for Computational Linguistics (ACL), New York City, NJ, USA, 2006, pp. 149–164. URL: https://aclanthology.org/W06-2920.
[25] M. Khan, M. Dickinson, S. Kübler, Towards domain adaptation for parsing web data, in: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, 2013, pp. 357–364.
[26] S. Petrov, D. Das, R. McDonald, A universal part-of-speech tagset, arXiv preprint arXiv:1104.2086 (2011).
[27] J. Macutek, R. Cech, J. Milicka, Length of non-projective sentences: A pilot study using a Czech UD treebank, in: X. Chen, R. Ferrer-i Cancho (Eds.), Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019), Association for Computational Linguistics, Paris, France, 2019, pp. 110–117. URL: https://aclanthology.org/W19-7913. doi:10.18653/v1/W19-7913.
[28] J. Nivre, Constraints on non-projective dependency parsing, in: D. McCarthy, S. Wintner (Eds.), 11th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Trento, Italy, 2006, pp. 73–80. URL: https://aclanthology.org/E06-1010.
[29] J. Nivre, J. Nilsson, Pseudo-projective dependency parsing, in: K. Knight, H. T. Ng, K. Oflazer (Eds.), Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 99–106. URL: https://aclanthology.org/P05-1013. doi:10.3115/

1219840.1219853.
[30] M. Khan, M. Dickinson, S. Kuebler, Does size matter? text and grammar revision for parsing social media data, in: C. Danescu-Niculescu-Mizil, A. Farzindar, M. Gamon, D. Inkpen, M. Nagarajan (Eds.), Proceedings of the Workshop on Language Analysis in Social Media, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 1–10. URL: https://aclanthology.org/W13-1101.
[31] M. Khan, M. Dickinson, S. Kübler, Towards domain adaptation for parsing web data, in: R. Mitkov, G. Angelova, K. Bontcheva (Eds.), Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2013, pp. 357–364. URL: https://aclanthology.org/R13-1046.
[32] C. Corbetta, M. Passarotti, G. Moretti, The Rise and Fall of Dependency Parsing in Dante Alighieri's Divine Comedy, in: R. Sprugnoli, M. Passarotti (Eds.), Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, ELRA and ICCL, Torino, Italia, 2024, pp. 50–56. URL: https://aclanthology.org/2024.lt4hala-1.7.

A. Online Resources

- Italian-Old,
- Italian-ISDT,
- Italian-VIT,
- ITTB,
- PROFITEROLE,
- PROIEL,
- Stanza,
- Old-Italian-Model.