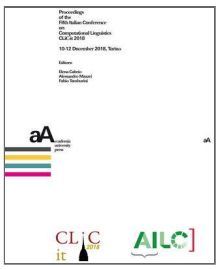

Elena Cabrio, Alessandro Mazzei and Fabio Tamburini (dir.)

Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018
10-12 December 2018, Torino

Accademia University Press

Tint 2.0: an All-inclusive Suite for NLP in Italian

Alessio Palmero Aprosio and Giovanni Moretti

DOI: 10.4000/books.aaccademia.3571
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2018
Published on OpenEdition Books: 8 April 2019
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788831978682

Tint 2.0: an All-inclusive Suite for NLP in Italian

Alessio Palmero Aprosio
Fondazione Bruno Kessler
Trento, Italy
aprosio@fbk.eu

Giovanni Moretti
Fondazione Bruno Kessler
Trento, Italy
moretti@fbk.eu

Abstract

English. In this we paper present Tint 2.0, an open-source, fast and extendable Natural Language Processing suite for Italian based on Stanford CoreNLP. The new release includes some improvements of the existing NLP modules, and a set of new text processing components for fine-grained linguistic analysis that were not available so far, including multi-word expression recognition, affix analysis, readability and classification of complex verb tenses.

Italiano. *In questo articolo presentiamo Tint 2.0, una collezione di moduli open-source veloci e personalizzabili per l'analisi automatica di testi in italiano basata su Stanford CoreNLP. La nuova versione comprende alcune migliorie relative ai moduli standard, e l'integrazione di componenti totalmente nuovi per l'analisi linguistica. Questi includono per esempio il riconoscimento di espressioni polirematiche, l'analisi degli affissi, il calcolo della leggibilità e il riconoscimento dei tempi verbali composti.*

1 Introduction

In recent years, Natural Language Processing (NLP) technologies have become fundamental to deal with complex tasks requiring text analysis, such as Question Answering, Topic Classification, Text Simplification, etc. Both research institutions and companies require accurate and reliable software for free and efficient linguistic analysis, allowing programmers to focus on the core of their business or research. While most of the open-source NLP tools freely available on the web (such as Stanford CoreNLP[1] and OpenNLP[2]) are designed for English and sometimes adapted to other languages, there is a lack of this kind of resources for Italian.

In this paper, we present a novel, extended release of Tint (Palmero Aprosio and Moretti, 2016), a suite of ready-to-use modules for Italian NLP. It is free to use, open source, and can be downloaded and used out-of-the-box (see Section 6). Compared to the previous version, the suite has been enriched with several modules for fine-grained linguistic analysis that were not available for Italian before.

2 Related work

There are plenty of linguistic pipelines available for download. Most of them (such as Stanford CoreNLP and OpenNLP) are language independent and, even if they are not available in Italian out-of-the-box, they could be trained in every existing language. A notable example in this direction is UDpipe (Straka and Straková, 2017), a trainable pipeline which performs most of the common NLP tasks and is available in more than 50 languages, and Freeling (Padró and Stanilovsky, 2012), a C++ library providing language analysis functionalities for a variety of languages. There are also some pipelines for Italian, such as TextPro (Emanuele Pianta and Zanoli, 2008), T2K (Dell'Orletta et al., 2014), and TaNL, but none of them are released as open source (and only TextPro can be downloaded and used for free for research purposes). Other single components are unfortunately available only upon request to the authors, for example the AnIta morphological analyser (Tamburini and Melandri, 2012).

In this respect, Tint represents an exception because not only it includes standard NLP modules, for example Named Entity Recognition and

[1] http://stanfordnlp.github.io/CoreNLP/
[2] https://opennlp.apache.org/

Lemmatization, but it also provides within a single framework additional components that are usually available as separate tools, such as the identification of multi-word expressions, the estimation of text complexity and the detection of text reuse.

Multi-word expression identification is a well studied problem, but most of the tools are available or optimized only for English. One of them, jMWE,[3] is written in Java and provides a parallel project[4] that adds compatibility to CoreNLP (Kulkarni and Finlayson, 2011). The mwetoolkit[5] is written in Python and uses a CRF classifier (Ramisch et al., 2010). The word2phrase module of word2vec attempts to learn phrases in a document of any language (Mikolov et al., 2013), but it is more a statistical tool for phrase extraction than for multi-word detection.

As for the assessment of text complexity, READ-IT (Dell'Orletta et al., 2011) is the only existing tool that gathers readability information for an Italian text. However, while the online demo can be used for free without registration, the tool is not available for offline use.

As for text reuse detection, i.e. when an author quotes (or borrows) another earlier or contemporary author, in the last years it has become easier thanks to new algorithms and high availability of texts (Mullen, 2016; Clough et al., 2002; Mihalcea et al., 2006). However, also in this case, no tools are available for Italian.

3 Tool description

The Tint pipeline is based on Stanford CoreNLP (Manning et al., 2014), an open-source framework written in Java, that provides most of the common Natural Language Processing tasks out-of-the-box in various languages. The framework provides also an easy interface to extend the annotation to new tasks and/or languages. Differently from some similar tools, such as UIMA (Ferrucci and Lally, 2004) and GATE (Cunningham et al., 2002), CoreNLP is easy to use and requires only basic object-oriented programming skills to extend it. In Tint, we adopt this framework to: (i) port the most common NLP tasks to Italian; (ii) make it easily extendable, both for writing new modules and replacing existing ones with more customized ones; and (iii) implement some new annotators as wrappers for external tools, such as entity linking, temporal expression identification, keyword extraction.

4 Modules

In this Section, we present a set of Tint modules, briefly describing those that were already included in the first release (Palmero Aprosio and Moretti, 2016) and focusing with more details on novel, more recent ones. While the old modules perform traditional NLP tasks (i.e. morphological analysis), we have recently integrated components for a more fine-grained linguistic analysis of specific phenomena, such as affixation, the identification of multi-word expressions, anglicisms and euphonic "d". These are the outcome of a larger project involving FBK and the Institute for Educational Research of the Province of Trento (Sprugnoli et al., 2018), aimed at studying with NLP tools the evolution of Italian texts towards the so-called neo-standard Italian (Berruto, 2012).

4.1 Already existing modules

As described in (Palmero Aprosio and Moretti, 2016), the Tint pipeline provides a set of pre-installed modules for basic linguistic annotation: tokenization, part-of-speech (POS) tagging, morphological analysis, lemmatization, named entity recognition and classification (NERC), dependency parsing.

Among the modules, two have been implemented from scratch and do not rely on the components available in Stanford CoreNLP: the tokenizer and the morphological analyser (see below). POS tagging, dependency parsing and NERC are performed using the existing modules in CoreNLP, trained on the Universal Dependencies[6] (UD) dataset in Italian (Bosco et al., 2013), and I-CAB (Magnini et al., 2006) respectively.

Additional modules include wrappers for temporal expression extraction and classification with HeidelTime (Strötgen and Gertz, 2013), keyword extraction with Keyphrase Digger (Moretti et al., 2015), and entity linking using DBpedia Spotlight[7] (Daiber et al., 2013) and The Wiki Machine[8] (Giuliano et al., 2009).

Tokenizer: This module provides text segmentation in tokens and sentences. At first, the text is grossly tokenized. Then, in a second step, tokens that need to be put together are merged us-

[3] http://projects.csail.mit.edu/jmwe/
[4] https://github.com/toliwa/CoreNLP-jMWE
[5] http://mwetoolkit.sourceforge.net/PHITE.php
[6] http://universaldependencies.org/
[7] http://bit.ly/dbpspotlight
[8] http://bit.ly/thewikimachine

ing two customizable lists of Italian non-breaking abbreviations (such as "dott." or "S.p.A.") and regular expressions (for e-mail addresses, web URIs, numbers, dates). This second phase uses (De La Briandais, 1959) to speedup the process.

Morphological Analyser: The morphological analyzer module provides the full list of morphological features for each annotated token. The current version of the module has been trained using the Morph-it lexicon (Zanchetta and Baroni, 2005), but it is possible to extend or retrain it with other Italian datasets. In order to grant fast performance, the model storage has been implemented with the mapDB Java library[9] that provides an excellent variation of Cassandras Sorted String Table. To extend the coverage of the results, especially for the complex forms, such as "porta-cene" or "bi-direzionale", the module tries to decompose the token into prefix-root-infix-suffix and tries to recognise the root form.

See Section 5 for an extensive evaluation of the modules.

4.2 New modules

Affixes annotation: This module provides a token-level annotation about word derivatives, based on *derIvaTario* (Talamo et al., 2016).[10] The resource was built segmenting into derivational cycles about 11,000 derivatives and annotating them with a wide array of features. The module uses this resource in input to segment a token into root and affixes, for example *visione* is analysed as *baseLemma=vedere*, *affix=zione* and *allomorph=ione*.

Classification of verbal tenses: Part-of speech tagger and morphological analyzer released with Tint can identify and classify verbs at token level, but sometimes the modality, form and tense of a verb is the result of a sequence of tokens, as in compound tenses such as participio passato, or passive verb forms. For this reason, we include in Tint a new tense module to provide a more complete annotation of multi-token verbal forms. The module supports also the analysis of discontinuous expressions, like for example *ho sempre mangiato*.

Text reuse: Detecting text reuse is useful when, in a document, we want to measure the overlap with a given corpus. This is needed in a number of applications, for example for plagiarism detection, stylometry, authorship attribution, citation analysis, etc. Tint includes now a component to deal with this task, i.e. identifying parts of an input text that overlap with a given corpus. First of all, each sentence of the corpus is compared with the sentences in the processed text using the FuzzyWuzzy package[11], a Java fuzzy string matching implementation: this allows the system not to miss expressions that are slightly different with respect to the texts in the original corpus. In this phase, only long spans of text can be considered, as the probability of an incorrect match on fuzzy comparison grows as soon as the text length decreases. A second step checks whether the overlap involves the whole sentence and, if not, it analyzes the two texts and identifies the number of overlapping tokens. Finally, the Stanford CoreNLP quote annotator[12] is used to catch text reuse that is in between quotes, ignoring the length limitation of the fuzzy comparison.

Readability: In this module, we compute some metrics that can be useful to assess the readability of a text, partially inspired by Dell'Orletta et al. (2011) and Tonelli et al. (2012). In particular, we include the following indices:

- Number of content words, hyphens (using iText Java Library[13]), sentences having less than a fixed number of words, distribution of tokens based on part-of-speech.
- Type-token ratio (TTR), i.e. the ratio between the number of different lemmas and the number of tokens; high TTR indicates a high degree of lexical variation.
- Lexical density, i.e. the number of content words divided by the total number of words.
- Amount of coordinate and subordinate clauses, along with the ratio between them.
- Depth of the parse tree for each sentence: both average and max depth are calculated on the whole text.
- Gulpease formula (Lucisano and Piemontese, 1988) to measure the readability at document level.

[9] http://www.mapdb.org
[10] http://derivatario.sns.it/
[11] https://github.com/xdrop/fuzzywuzzy
[12] https://stanfordnlp.github.io/CoreNLP/quote.html
[13] https://github.com/itext/itextpdf

- Text difficulty based on word lists from De Mauro's Dictionary of Basic Italian[14].

Multi-word expressions: A specific multi-token annotator has been implemented to recognize more than 13,450 multi-word expressions, the so-called 'poliremathiche' (Voghera, 2004), manually collected from various online resources. The list includes verbal, nominal, adjectival and prepositional expressions (e.g. *lasciar perdere*, *società per azioni*, *nei confronti di*, *mezzo morto*). This annotator can identify also discontinuous multi-words. For example, in the expression *andare a genio* (Italian phrase that means "to like") an adverb can be included, as in *andare troppo a genio*. Similarly, in such phrases one can find nouns and adjectives (e.g. *lasciare Antonio a piedi*, where *lasciare a piedi* is an Italian multiword for *leave stranded*).

Anglicisms: A list of more than 2,500 anglicisms, collected from the web, is included in the last release of Tint, and a particular annotator identifies them in the text and distinguishes between adapted ("chattare", "skillato") and non-adapted anglicisms ("spread", "leadership"). This module can then be used to track the use of borrowings from English in Italian texts, a phenomenon much debated in the media and among scholars (Fanfani, 1996; Furiassi, 2008).

Euphonic "D": For euphonic reasons, the preposition *a*, and the conjunctions *e* and *o* usually become *ad*, *ed*, *od* when the subsequent word begins with *a*, *e*, *o* respectively. While traditionally this rule was applied to every vowel, a more recent grammatical rule has established that the euphonic 'd' should be limited to cases in which it is followed by the same vowel, for example *ed ecco* vs. *e ancora*[15]. Tint provides an annotator that identifies this phenomenon, and classifies each instance as correct, if it follows the aforementioned rule, or incorrect in all the other cases.

Corpus statistics: A collection of CoreNLP annotators have been developed to extract statistics that can be used, for instance, to analyse traits of interest in texts. More specifically, the provided modules can mark and compute words and sentences based on token, lemma, part-of-speech and word position in the sentence.

[14] http://bit.ly/nuovo-demauro

[15] http://bit.ly/crusca-d-eufonica

5 Evaluation

Tint includes a rich set of tools, evaluated separately. In some cases, an evaluation based on the accuracy is not possible, because of the lack of available gold standard or because the tool outcome is not comparable to other tools' ones.

When possible, Tint is compared with existing pipelines that work with the Italian language: Tanl (Attardi et al., 2010), TextPro (Pianta et al., 2008) and TreeTagger (Schmid, 1994).

In calculating speed, we run each experiment 10 times and consider the average execution time. When available, multi-thread capabilities have been disabled. All experiments have been executed on a 2,3 GHz Intel Core i7 with 16 GB of memory.

The Tanl API is not available as a downloadable package, but it's only usable online through a REST API, therefore the speed may be influenced by the network connection.

No evaluation is performed for the Tint annotators that act as wrappers for an external tools (temporal expression tagging, entity linking, keyword extraction).

5.1 Tokenization and sentence splitting

For the task of tokenization and sentence splitting, Tint outperforms in speed both TextPro and Tanl (see Table 1).

System	Speed (tok/sec)
Tint	**80,000**
Tanl API	30,000
TextPro 2.0	35,000

Table 1: Tokenization and sentence splitting speed.

5.2 Part-of-speech tagging

The evaluation of the part-of-speech tagging is performed against the test set included in the UD dataset, containing 10K tokens. As the tagset used is different for different tools, the accuracy is calculated only on five coarse-grained types: nouns (N), verbs (V), adverbs (B), adjectives (A) and other (O). Table 2 shows the results.

5.3 Lemmatization

Like part-of-speech tagging, lemmatization is evaluated, both in terms of accuracy and execu-

[16] The (considerable) speed of TreeTagger includes both lemmatization and part-of-speech tagging.

System	Speed (tok/sec)	Accuracy
Tint	28,000	**98%**
Tanl API	20,000	n.a.
TextPro 2.0	20,000	96%
TreeTagger	**190,000**[16]	92%

Table 2: Evaluation of part-of-speech tagging.

tion time, on the UD test set. When the lemma is guessed starting form a morphological analysis (such as in Tint and TextPro), the speed is calculated by including both tasks. Table 3 shows the results. All the tools reach the same accuracy of 96% (with minor differences that are not statistically significant).

System	Speed (tok/sec)	Accuracy
Tint	97,000	96%
TextPro 2.0	9,000	96%
TreeTagger	**190,000**[16]	96%

Table 3: Evaluation of lemmatization.

5.4 Named Entity Recognition

For Named Entity Recognition, we evaluate and compare our system with the test set available on the I-CAB dataset. We consider three classes: PER, ORG, LOC. In training Tint, we extracted a list of persons, locations and organizations by querying the Airpedia database (Palmero Aprosio et al., 2013) for Wikipedia pages classified as `Person`, `Place` and `Organisation`, respectively. Table 4 shows the results of the named entity recognition task.

System	Speed	P	R	F_1
Tint	**30,000**	**84.37**	79.97	**82.11**
TextPro 2.0	4,000	81.78	**80.78**	81.28
Tanl API	16,000	72.89	52.50	61.04

Table 4: Evaluation of the NER.

5.5 Dependency parsing

The evaluation of the dependency parser is performed against Tanl (Attardi et al., 2013) and TextPro (Lavelli, 2013) w.r.t the usual metrics Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS). Table 5 shows the results: the Tint evaluation has been performed on the UD test data; LAS and UAS for TextPro and Tanl is taken directly from the Evalita 2011 proceedings (Magnini et al., 2013).

System	Speed	LAS	UAS
Tint	**9,000**	84.67	87.05
TextPro 2.0	1,300	87.30	91.47
Tanl (DeSR)	900	**89.88**	**93.73**

Table 5: Evaluation of the dependency parsing.

6 Tint distribution

The Tint pipeline is released as an open source software under the GNU General Public License (GPL), version 3. It can be download from the Tint website[17] as a standalone package, or it can be integrated into an existing application as a Maven dependency. The source code is available on Github.[18]

The tool is written using the Stanford CoreNLP paradigm, therefore a third part software can be integrated easily into the pipeline.

7 Conclusions and Future Works

In this paper, we presented the new release of Tint, a simple, fast and accurate NLP pipeline for Italian, based on Stanford CoreNLP. In the new version, we have fixed some bugs and improved some of the existing modules. We have also added a set of components for fine-grained linguistics analysis that were not available so far.

In the future, we plan to improve the suite and extend it with additional modules, also based on the feedback from the users through the github project page. We are currently working on new modules, in particular Word Sense Disambiguation (WSD) based on linguistic resources such as MultiWordNet (Pianta et al., 2002) and Semantic Role Labelling, by porting to Italian resources such as FrameNet (Baker et al., 1998), now available only in English.

The Tint pipeline will also be integrated in PIKES (Corcoglioniti et al., 2016), a tool that extracts knowledge from English texts using NLP and outputs it in a queryable form (such RDF triples), so to extend it to Italian.

Acknowledgments

The research leading to this paper was partially supported by the EU Horizon 2020 Programme via the SIMPATICO Project (H2020-EURO-6-2015, n. 692819).

[17] `http://tint.fbk.eu/`
[18] `https://github.com/dhfbk/tint/`

References

G. Attardi, S. Dei Rossi, and M. Simi. 2010. The Tanl Pipeline. In *Proc. of LREC Workshop on WSPP*.

Giuseppe Attardi, Maria Simi, and Andrea Zanelli. 2013. Tuning desr for dependency parsing of italian. In *Evaluation of Natural Language and Speech Tools for Italian*, pages 37–45. Springer.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Gateano Berruto. 2012. *Sociolinguistica dell'italiano contemporaneo*. Carocci.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank.

Paul Clough, Robert Gaizauskas, Scott SL Piao, and Yorick Wilks. 2002. Meter: Measuring text reuse. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 152–159. Association for Computational Linguistics.

Francesco Corcoglioniti, Marco Rospocher, and Alessio Palmero Aprosio. 2016. A 2-phase frame-based knowledge extraction framework. In *Proc. of ACM Symposium on Applied Computing (SAC'16)*.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. Gate: An architecture for development of robust hlt applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 168–175, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Rene De La Briandais. 1959. File searching using variable length keys. In *Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference*, IRE-AIEE-ACM '59 (Western), pages 295–298, New York, NY, USA. ACM.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '11, pages 73–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. T2kˆ 2: a system for automatically extracting and organizing knowledge from texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Christian Girardi Emanuele Pianta and Roberto Zanoli. 2008. The textpro tool suite. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Massimo Fanfani. 1996. Sugli-anglicismi nell"italiano contemporaneo (xiv). *Lingua nostra*, 57(2):72–91.

David Ferrucci and Adam Lally. 2004. Uima: An architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348, September.

Cristiano Furiassi. 2008. *Non-adapted Anglicisms in Italian: Attitudes, frequency counts, and lexicographic implications*. Cambridge Scholars Publishing.

Claudio Giuliano, Alfio Massimiliano Gliozzo, and Carlo Strapparava. 2009. Kernel methods for minimally supervised wsd. *Comput. Linguist.*, 35(4):513–528, December.

Nidhi Kulkarni and Mark Alan Finlayson. 2011. jmwe: A java toolkit for detecting multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 122–124. Association for Computational Linguistics.

Alberto Lavelli. 2013. An ensemble model for the evalita 2011 dependency parsing task. In *Evaluation of Natural Language and Speech Tools for Italian*, pages 30–36. Springer.

Pietro Lucisano and Maria Emanuela Piemontese. 1988. GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, 3(31):110–124.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-cab: the italian content annotation bank. In *Proceedings of LREC*, pages 963–968. Citeseer.

Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta. 2013. *Evaluation of Natural Language and Speech Tool for Italian: International Workshop, EVALITA 2011, Rome, January 24-25, 2012, Revised Selected Papers*, volume 7689. Springer.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the dirt: Extracting keyphrases from texts with kd. *CLiC it*, page 198.

Lincoln Mullen, 2016. *textreuse: Detect Text Reuse and Document Similarity*. R package version 0.1.4.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *LREC2012*.

A. Palmero Aprosio and G. Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *ArXiv e-prints*, September.

Alessio Palmero Aprosio, Claudio Giuliano, and Alberto Lavelli. 2013. Automatic expansion of DBpedia exploiting Wikipedia cross-language information. In *Proceedings of the 10th Extended Semantic Web Conference*.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Developing an aligned multilingual database. In *Proc. 1st Intl Conference on Global WordNet*. Citeseer.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The textpro tool suite. In *LREC*. Citeseer.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Multiword expressions in the wild?: the mwetoolkit comes in handy. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 57–60. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees.

Rachele Sprugnoli, Sara Tonelli, Alessio Palmero Aprosio, and Giovanni Moretti. 2018. Analysing the evolution of students' writing skills and the impact of neo-standard italian with the help of computational linguistics. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Luigi Talamo, Chiara Celata, and Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, 9(1):72–102.

Fabio Tamburini and Matias Melandri. 2012. Anita: a powerful morphological analyser for italian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Sara Tonelli, Ke Tran Manh, and Emanuele Pianta. 2012. Making readability indices readable. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 40–48, Montréal, Canada, June. Association for Computational Linguistics.

Miriam Voghera. 2004. Polirematiche. *La formazione delle parole in italiano*, pages 56–69.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the italian language. *Corpus Linguistics 2005*, 1(1).