The *PAISÀ* Corpus of Italian Web Texts

Verena Lyding [*] verena.lyding@eurac.edu
Egon Stemle[*] egon.stemle@eurac.edu
Claudia Borghetti[†] claudia.borghetti@unibo.it
Marco Brunello[‡] marcobrunello84@gmail.com
Sara Castagnoli[†] s.castagnoli@unibo.it
Felice Dell'Orletta[§] felice.dellorletta@ilc.cnr.it
Henrik Dittmann[¶] henrik.dittmann@bordet.be
Alessandro Lenci[‖] alessandro.lenci@ling.unipi.it
Vito Pirrelli[§] vito.pirrelli@ilc.cnr.it

Abstract

PAISÀ is a Creative Commons licensed, large web corpus of contemporary Italian. We describe the design, harvesting, and processing steps involved in its creation.

1 Introduction

This paper provides an overview of the *PAISÀ* corpus of Italian web texts and an introductory description of the motivation, procedures and facilities for its creation and delivery.

Developed within the *PAISÀ* project, the corpus is intended to meet the objective to help overcome the technological barriers that still prevent web users from making use of large quantities of contemporary Italian texts for language and cultural education, by creating a comprehensive and easily accessible corpus resource of Italian.

The initial motivation of the initiative stemmed from the awareness that any static repertoire of digital data, however carefully designed and developed, is doomed to fast obsolescence, if contents are not freely available for public usage, continuously updated and checked for quality, incrementally augmented with new texts and annotation metadata for intelligent indexing and browsing. These requirements brought us to design a resource that was (1) freely available and freely re-publishable, (2) comprehensively covering contemporary common language and cultural content and (3) enhanced with a rich set of automatically-annotated linguistic information to enable advanced querying and retrieving of data. On top of that, we set out to develop (4) a dedicated interface with a low entry barrier for different target groups. The end result of this original plan represents an unprecedented digital language resource in the Italian scenario.

The main novelty of the *PAISÀ* web corpus is that it exclusively draws on Creative Commons licensed data, provides advanced linguistic annotations with respect to corpora of comparable size and corpora of web data, and invests in a carefully designed query interface, targeted at different user groups. In particular, the integration of richly annotated language content with an easily accessible, user-oriented interface makes *PAISÀ* a unique and flexible resource for language teaching.

2 Related Work

The world wide web, with its inexhaustible amount of natural language data, has become an established source for efficiently building large corpora (Kilgarriff and Grefenstette, 2003). Tools are available that make it convenient to bootstrap corpora from the web based on mere seed term lists, such as the BootCaT toolkit (Baroni and Bernardini, 2004). The huge corpora created by the WaCky project (Baroni et al., 2009) are an example of such an approach.

A large number of papers have recently been published on the harvesting, cleaning and processing of web corpora.[1] However, *freely available*, *large*, *contemporary*, *linguistically annotated*, *easily accessible* web corpora are still missing for many languages; but cf. e.g. (Généreux et al., 2012) and the Common Crawl Foundations (CCF) web crawl[2].

[*] EURAC Research Bolzano/Bozen, IT
[†] University of Bologna, IT
[‡] University of Leeds, UK
[§] Institute of Computational Linguistics "Antonio Zampolli" - CNR, IT
[¶] Institut Jules Bordet, BE
[‖] University of Pisa, IT

[1] cf. the Special Interest Group of the Association for Computational Linguistics on Web as Corpus (SIGWAC) http://sigwac.org.uk/

[2] CCF produces and maintains a repository of web crawl data that is openly accessible: http://commoncrawl.org/

3 Corpus Composition

3.1 Corpus design

PAISÀ aimed at creating a comprehensive corpus resource of Italian web texts which adheres to the criteria laid out in section 1. For these criteria to be fully met, we had to address a wide variety of issues covering the entire life-cycle of a digital text resource, ranging from robust algorithms for web navigation and harvesting, to adaptive annotation tools for advanced text indexing and querying and user-friendly accessing and rendering online interfaces customisable for different target groups.

Initially, we targeted a size of 100M tokens, and planned to automatically annotate the data with lemma, part-of-speech, structural dependency, and advanced linguistic information, using and adapting standard annotation tools (cf. section 4). Integration into a querying environment and a dedicated online interface were planned.

3.2 Licenses

A crucial point when planning to compile a corpus that is free to redistribute without encountering legal copyright issues is to collect texts that are in the public domain or at least, have been made available in a copyleft regime. This is the case when the author of a certain document decided to share some rights (copy and/or distribute, adapt etc.) on her work with the public, in a way that end users do not need to ask permission to the creator/owner of the original work. This is possible by employing licenses other than the traditional "all right reserved" copyright, i.e. GNU, Creative Commons etc., which found a wide use especially on the web. Exploratory studies (Brunello, 2009) have shown that Creative Commons licenses are widely employed throughout the web (at least on the Italian webspace), enough to consider the possibility to build a large corpus from the web exclusively made of documents released under such licenses.

In particular, Creative Commons provides five basic "baseline rights": *Attribution (BY)*, *Share Alike (SA)*, *Non Commercial (NC)*, *No Derivative Works (ND)*. The licenses themselves are composed of at least *Attribution* (which can be used even alone) plus the other elements, allowing six different combinations:[3] (1) Attribution (CC BY), (2) Attribution-NonCommercial (CC BY-NC), (3) Attribution-ShareAlike (CC BY-SA), (4) Attribution-NoDerivs (CC BY-ND), (5) Attribution-NonCommercial-ShareAlike (CC BY-NC-SA), and (6) Attribution-NonCommercial-NoDerivs (CC BY-NC-ND).

Some combinations are not possible because certain elements are not compatible, e.g. *Share Alike* and *No Derivative Works*. For our purposes we decided to discard documents released with the two licenses containing the *No Derivative Works* option, because our corpus is in fact a derivative work of collected documents.

3.3 The final corpus

The corpus contains approximately 388,000 documents from 1,067 different websites, for a total of about 250M tokens. All documents contained in the *PAISÀ* corpus date back to Sept./Oct. 2010.

The documents come from several web sources which, at the time of corpus collection, provided their content under Creative Commons license (see section 3.2 for details). About 269,000 texts are from Wikimedia Foundation projects, with approximately 263,300 pages from Wikipedia, 2380 pages from Wikibooks, 1680 pages from Wikinews, 740 pages from Wikiversity, 410 pages from Wikisource, and 390 Wikivoyage pages.

The remaining 119,000 documents come from `guide.supereva.it` (ca. 19,000), `italy.indymedia.org` (ca. 10,000) and several blog services from more than another 1,000 different sites (e.g. `www.tvblog.it` (9,088 pages), `www.motoblog.it` (3,300), `www.ecowebnews.it` (3,220), and `www.webmasterpoint.org` (3,138).

Texts included in *PAISÀ* have an average length of 683 words, with the longest text[4] counting 66,380 running tokens. A non exhaustive list of average text lengths by source type is provided in table 1 by way of illustration.

The corpus has been annotated for lemma, part-of-speech and dependency information (see section 4.2 for details). At the document level, the corpus contains information on the URL of origin and a set of descriptive statistics of the text, including text length, rate of advanced vocabulary, readability parameters, etc. (see section 4.3). Also, each document is marked with a unique identifier.

[3] For detailed descriptions of each license see `http://creativecommons.org/licenses/`

[4] The *European Constitution* from wikisource.org: `http://it.wikisource.org/wiki/Trattato_che_adotta_una_Costituzione_per_l'Europa`

Document source	Avg text length
PAISÀ total	683 words
Wikipedia	693 words
Wikibooks	1844 words
guide.supereva.it	378 words
italy.indymedia.it	1147 words
tvblog.it	1472 words
motoblog.it	421 words
ecowebnews.it	347 words
webmasterpoint.org	332 words

Table 1: Average text length by source

The annotated corpus adheres to the standard CoNLL column-based format (Buchholz and Marsi, 2006), is encoded in UTF-8.

4 Corpus Creation

4.1 Collecting and cleaning web data

The web pages for *PAISÀ* were selected in two ways: part of the corpus collection was made through CC-focused web crawling, and another part through a targeted collection of documents from specific websites.

4.1.1 Seed-term based harvesting

At the time of corpus collection (2010), we used the BootCaT toolkit mainly because collecting URLs could be based on the public Yahoo! search API[5], including the option to restrict search to CC-licensed pages (including the possibility to specify even the particular licenses). Unfortunately, Yahoo! discontinued the free availability of this API, and BootCaT's remaining search engines do not provide this feature.

An earlier version of the corpus was collected using the tuple list originally employed to build itWaC[6]. As we noticed that the use of this list, in combination with the restriction to CC, biased the final results (i.e. specific websites occurred very often as top results) , we provided as input 50,000 medium frequent seed terms from a basic Italian vocabulary list[7], in order to get a wider distribution of search queries, and, ultimately, of texts.

As introduced in section 3.2, we restricted the selection not just to Creative Commons-licensed texts, but specifically to those licenses allowing redistribution: namely, CC BY, CC BY-SA, CC BY-NC-SA, and CC BY-NC.

Results were downloaded and automatically cleaned with the KrdWrd system, an environment for the unified processing of web content (Steger and Stemle, 2009).

Wrongly CC-tagged pages were eliminated using a black-list that had been manually populated following inspection of earlier corpus versions.

4.1.2 Targeted

In September 2009, the Wikimedia Foundation decided to release the content of their wikis under CC BY-SA[8], so we decided to download the large and varied amount of texts made available through the Italian versions of these websites. This was done using the Wikipedia Extractor[9] on official dumps[10] of Wikipedia, Wikinews, Wikisource, Wikibooks, Wikiversity and Wikivoyage.

4.2 Linguistic annotation and tools adaptation

The corpus was automatically annotated with lemma, part-of-speech and dependency information, using state-of-the-art annotation tools for Italian. Part-of-speech tagging was performed with the Part-Of-Speech tagger described in Dell'Orletta (2009) and dependency-parsed by the DeSR parser (Attardi et al., 2009), using *Multilayer Perceptron* as the learning algorithm. The systems used the ISST-TANL part-of-speech[11] and dependency tagsets[12]. In particular, the pos-tagger achieves a performance of 96.34% and DeSR, trained on the ISST-TANL treebank consisting of articles from newspapers and periodicals, achieves a performance of 83.38% and 87.71% in terms of LAS (*labelled attachment score*) and UAS (*unlabelled attachment score*) respectively, when tested on texts of the same type.

However, since Gildea (2001), it is widely acknowledged that statistical NLP tools have a drop of accuracy when tested against corpora differing from the typology of texts on which they were trained. This also holds true for *PAISÀ*: it contains

[5] http://developer.yahoo.com/boss/

[6] http://wacky.sslmit.unibo.it/doku.php?id=seed_words_and_tuples

[7] http://ppbm.paravia.it/dib_lemmario.php

[8] Previously under GNU Free Documentation License.

[9] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

[10] http://dumps.wikimedia.org/

[11] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

[12] http://www.italianlp.it/docs/ISST-TANL-DEPtagset.pdf

lexical and syntactic structures of non-canonical languages such as the language of social media, blogs, forum posts, consumer reviews, etc. As reported in Petrov and McDonald (2012), there are multiple reasons why parsing the web texts is difficult: punctuation and capitalization are often inconsistent, there is a lexical shift due to increased use of slang and technical jargon, some syntactic constructions are more frequent in web text than in newswire, etc.

In order to overcome this problem, two main typologies of methods and techniques have been developed: *Self-training* (McClosky et al., 2006) and *Active Learning* (Thompson et al., 1999).

For the specific purpose of the NLP tools adaptation to the Italian web texts, we adopted two different strategies for the pos-tagger and the parser. For what concerns pos-tagging, we used an active learning approach: given a subset of automatically pos-tagged sentences of *PAISÀ*, we selected the ones with the lowest likelihood, where the sentence likelihood was computed as the product of the probabilities of the assignments of the pos-tagger for all the tokens. These sentences were manually revised and added to the training corpus in order to build a new pos-tagger model incorporating some new knowledge from the target domain.

For what concerns parsing, we used a self-training approach to domain adaptation described in Dell'Orletta et al. (2013), based on ULISSE (Dell'Orletta et al., 2011). ULISSE is an unsupervised linguistically-driven algorithm to select reliable parses from a collection of dependency annotated texts. It assigns to each dependency tree a score quantifying its reliability based on a wide range of linguistic features. After collecting statistics about selected features from a corpus of automatically parsed sentences, for each newly parsed sentence ULISSE computes a reliability score using the previously extracted feature statistics. From the top of the parses (ranked according to their reliability score) different pools of parses were selected to be used for training. The new training contains the original training set as well as the new selected parses which include lexical and syntactic characteristics specific of the target domain (Italian web texts). The parser trained on this new training set improves its performance when tested on the target domain.

We used this domain adaptation approach for the following three main reasons: a) it is unsupervised (i.e. no need for manually annotated training data); b) unlike the *Active Learning* approach used for pos-tagging, it does not need manual revision of the automatically parsed samples to be used for training; c) it was previously tested on Italian texts with good results (Dell'Orletta et al., 2013).

4.3 Readability analysis of corpus documents

For each corpus document, we calculated several text statistics indicative of the linguistic complexity, or 'readability' of a text.

The applied measures include, (1) *text length in tokens*, that is the number of tokens per text, (2) *sentences per text*, that is a sentence count, and (3) *type-token ratio* indicated as a percentage value. In addition, we calculated (4) the *advanced vocabulary per text*, that is a word count of the text vocabulary which is not part of the the basic Italian vocabulary ('vocabolario di base') for written texts, as defined by De Mauro (1991)[13], and (5) the *Gulpease Index* ('Indice Gulpease') (Lucisano and Piemontese, 1988), which is a measure for the readability of text that is based on frequency relations between the number of sentences, words and letters of a text.

All values are encoded as metadata for the corpus. Via the *PAISÀ* online interface, they can be employed for filtering documents and building subcorpora. This facility was implemented with the principal target group of *PAISÀ* users in mind, as the selection of language examples according to their readability level is particularly relevant for language learning and teaching.

4.4 Attempts at text classification for genre, topic, and function

Lack of information about the composition of corpora collected from the web using unsupervised methods is probably one of the major limitations of current web corpora vis-à-vis more traditional, carefully constructed corpora, most notably when applications to language teaching and learning are envisaged. This also holds true for *PAISÀ*, es-

[13]The advanced vocabulary was calculated on the basis of a word list consisting of De Mauro's 'vocabolario fondamentale' (`http://it.wikipedia.org/wiki/Vocabolario_fondamentale`) and 'vocabolario di alto uso' (`http://it.wikipedia.org/wiki/Vocabolario_di_alto_uso`), together with high frequent function words not contained in those two lists.

pecially for the harvested[14] subcorpus that was downloaded as described in section 4.1. We therefore carried out some experiments with the ultimate aim to enrich the corpus with metadata about text genre, topic and function, using automated techniques.

In order to gain some insights into the composition of *PAISÀ*, we first conducted some manual investigations. Drawing on existing literature on web genres (e.g. (Santini, 2005; Rehm et al., 2008; Santini et al., 2010)) and text classification according to text function and topic (e.g. (Sharoff, 2006)), we developed a tentative three-fold taxonomy to be used for text classification. Following four cycles of sample manual annotation by three annotators, categories were adjusted in order to better reflect the nature of *PAISÀ*'s web documents (cf. (Sharoff, 2010) about differences between domains covered in the BNC and in the web-derived ukWaC). Details about the taxonomy are provided in Borghetti et al. (2011). Then, we started to cross-check whether the devised taxonomy was indeed appropriate to describe *PAISÀ*'s composition by comparing its categories with data resulting from the application of unsupervised methods for text classification.

Interesting insights have emerged so far regarding the topic category. Following Sharoff (2010), we used topic modelling based on *Latent Dirichlet Allocation* for the detection of topics: 20 clusters/topics were identified on the basis of keywords (the number of clusters to retrieve is a user-defined parameter) and projected onto the manually defined taxonomy. This revealed that most of the 20 automatically identified topics could be reasonably matched to one of the 8 categories included in the taxonomy; exceptions were represented by clusters characterised by proper nouns and general language words such *bambino/uomo/famiglia* ('child'/'man'/'family') or *credere/sentire/sperare* ('to believe'/'feel'/'hope'), which may in fact be indicative of genres such as diary or personal comment (e.g. personal blog). Only one of the categories originally included in the taxonomy – natural sciences – was not represented in the clusters, which may indicate that there are few texts within *PAISÀ* belonging to this domain. One of the major advantages of topic models is that each corpus document can be associated – to varying degrees – to several topics/clusters: if encoded as metadata, this information makes it possible not only to filter texts according to their prevailing domain, but also to represent the heterogeneous nature of many web documents.

[14] In fact, even the nature of the targeted texts is not precisely defined: for instance, Wikipedia articles can actually encompass a variety of text types such as biographies, introductions to academic theories etc. (Santini et al., 2010, p. 15)

5 Corpus Access and Usage

5.1 Corpus distribution

The *PAISÀ* corpus is distributed in two ways: it is made available for download and it can be queried via its online interface. For both cases, no restrictions on its usage apply other than those defined by the Creative Commons BY-NC-SA license. For corpus download, both the raw text version and the annotated corpus in CoNLL format are provided.

The *PAISÀ* corpus together with all project-related information is accessible via the project web site at `http://www.corpusitaliano.it`

5.2 Corpus interface

The creation of a dedicated open online interface for the *PAISÀ* corpus has been a declared primary objective of the project.

The interface is aimed at providing a powerful, effective and easy-to-employ tool for making full use of the resource, without having to go through downloading, installation or registration procedures. It is targeted at different user groups, particularly language learners, teachers, and linguists. As users of *PAISÀ* are expected to show varying levels of proficiency in terms of language competence, linguistic knowledge, and concerning the use of online search tools, the interface has been designed to provide four separate search components, implementing different query modes.

Initially, the user is directed to a basic keyword search that adopts a 'Google-style' search box. Single search terms, as well as multi-word combinations or sequences can be searched by inserting them in a simple text box.

The second component is an advanced graphical search form. It provides elaborated search options for querying linguistic annotation layers and allows for defining distances between search terms as well as repetitions or optionally occurring terms. Furthermore, the advanced search supports regular expressions.

The third component emulates a command-line search via the powerful CQP query language of

the Open Corpus Workbench (Evert and Hardie, 2011). It allows for complex search queries in CQP syntax that rely on linguistic annotation layers as well as on metadata information.

Finally, a filter interface is presented in a fourth component. It serves the purpose of retrieving full-text corpus documents based on keyword searches as well as text statistics (see section 4.3). Like the CQP interface, the filter interface is also supporting the building of temporary subcorpora for subsequent querying.

By default, search results are displayed as KWIC (KeyWord In Context) lines, centred around the search expression. Each search hit can be expanded to its full sentence view. In addition, the originating full text document can be accessed and its source URL is provided.

Based on an interactive visualisation for dependency graphs (Culy et al., 2011) for each search result a graphical representations of dependency relations together with the sentence and associated lemma and part-of-speech information can be generated (see Figure 1).

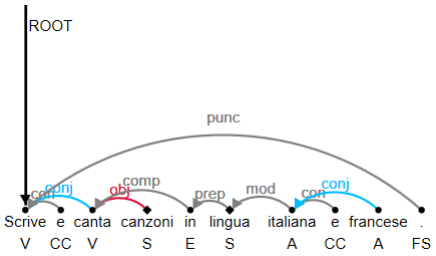

Figure 1: Dependency diagram

Targeted at novice language learners of Italian, a filter for automatically restricting search results to sentences of limited complexity has been integrated into each search component. When activated, search results are automatically filtered based on a combination of the complexity measures introduced in section 4.3.

5.3 Technical details

The *PAISÀ* online interface has been developed in several layers: in essence, it provides a front-end to the corpus as indexed in Open Corpus Workbench (Evert and Hardie, 2011). This corpus query engine provides the fundamental search capabilities through the CQP language. Based on the CWB/Perl API that is part of the Open Corpus Workbench package, a web service has been developed at EURAC which exposes a large part of the CQP language[15] through a RESTful API.[16]

The four types of searches provided by the online interface are developed on top of this web service. The user queries are translated into CQP queries and passed to the web service. In many cases, such as the free word order queries in the simple and advanced search forms, more than one CQP query is necessary to produce the desired result. Other functionalities implemented in this layer are the management of subcorpora and the filtering by complexity. The results returned by the web service are then formatted and presented to the user.

The user interface as well as the mechanisms for translation of queries from the web forms into CQP have been developed server-side in PHP. The visualizations are implemented client-side in JavaScript and jQuery, the dependency graphs based on the xLDD framework (Culy et al., 2011).

5.4 Extraction of lexico-syntactic information

PAISÀ is currently used in the CombiNet project "Word Combinations in Italian – Theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary".[17] The project goal is to study the combinatory properties of Italian words by developing advanced computational linguistics methods for extracting distributional information from *PAISÀ*.

In particular, CombiNet uses a pattern-based approach to extract a wide range of multiword expressions, such as phrasal lexemes, collocations, and usual combinations. POS n-grams are automatically extracted from *PAISÀ*, and then ranked according to different types of association measures (e.g., pointwise mutual information, log-likelihood ratios, etc.). Extending the LexIt methodology (Lenci et al., 2012), CombiNet also extracts distributional profiles from the parsed layer of *PAISÀ*, including the following types of information:

1. syntactic slots (subject, complements, modi-

[15] To safeguard the system against malicious attacks, security measures had to be taken at several of the layers, which unfortunately also make some of the more advanced CQP features inaccessible to the user.

[16] Web services based on REST (Representational State Transfer) principles employ standard concepts such as a URI and standard HTTP methods to provide an interface to functionalities on a remote host.

[17] 3-year PRIN(2010/2011)-project, coordination by Raffaele Simone – University of Rome Tre

fiers, etc.) and subcategorization frames;

2. lexical sets filling syntactic slots (e.g. prototypical subjects of a target verb);

3. semantic classes describing selectional preferences of syntactic slots (e.g. the direct obj. of *mangiare*/'to eat' typically selects nouns referring to food, while its subject selects animate nouns); semantic roles of predicates.

The saliency and typicality of combinatory patterns are weighted by means of different statistical indexes and the resulting profiles will be used to define a distributional semantic classification of Italian verbs, comparable to the one elaborated in the VerbNet project (Kipper et al., 2008).

6 Evaluation

We performed post-crawl evaluations on the data. For licensing, we analysed 200,534 pages that were originally collected for the *PAISÀ* corpus, and only 1,060 were identified as containing no CC license link (99.95% with CC mark-up). Then, from 10,000 randomly selected non-CC-licensed Italian pages 15 were wrongly identified as CC licensed containing CC mark-up (0.15% error). For language identification we checked the harvested corpus part with the CLD2 toolkit[18], and $> 99\%$ of the data was identified as Italian.

The pos-tagger has been adapted to peculiarities of the *PAISÀ* web texts, by manually correcting sample annotation output and re-training the tagger accordingly. Following the active learning approach as described in section 4.2 we built a new pos-tagger model based on 40.000 manually revised tokens. With the new model, we obtained an improvement in accuracy of 1% on a test-set of 5000 tokens extracted from *PAISÀ*. Final tagger accuracy reached 96.03%.

7 Conclusion / Future Work

In this paper we showed how a contemporary and free language resource of Italian with linguistic annotations can be designed, implemented and developed from the web and made available for different types of language users.

Future work will focus on enriching the corpus with metadata by means of automatic classification techniques, so as to make a better assessment of corpus composition. A multi-faceted approach combining linguistic features extracted from texts (content/function words ratio, sentence length, word frequency, etc.) and information extracted from document URLs (e.g., tags like "wiki", "blog") might be particularly suitable for genre and function annotation.

Metadata annotation will enable more advanced applications of the corpus for language teaching and learning purposes. In this respect, existing exemplifications of the use of the *PAISÀ* interface for language learning and teaching (Lyding et al., 2013) could be followed by further pedagogical proposals as well as empowered by dedicated teaching guidelines for the exploitation of the corpus and its web interface in the class of Italian as a second language.

In a more general perspective, we envisage a tighter integration between acquisition of new texts, automated text annotation and development of lexical and language learning resources allowing even non-specialised users to carve out and develop their own language data. This ambitious goal points in the direction of a fully-automatised control of the entire life-cycle of open-access Italian language resources with a view to address an increasingly wider range of potential demands.

Acknowledgements

The three years *PAISÀ* project[19], concluded in January 2013, received funding from the Italian Ministry of Education, Universities and Research (MIUR)[20], by the FIRB program (Fondo per gli Investimenti della Ricerca di Base)[21].

References

G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proc. of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia.

M. Baroni and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proc. of LREC 2004*, pages 1313–1316. ELDA.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed

[18] Compact Language Detection 2, `http://code.google.com/p/cld2/`

[19] An effort of four Italian research units: University of Bologna, CNR Pisa, University of Trento and European Academy of Bolzano/Bozen.

[20] `http://www.istruzione.it/`

[21] `http://hubmiur.pubblica.istruzione.it/web/ricerca/firb`

web-crawled corpora. *Journal of LRE*, 43(3):209–226.

C. Borghetti, S. Castagnoli, and M. Brunello. 2011. I testi del web: una proposta di classificazione sulla base del corpus paisà. In M. Cerruti, E. Corino, and C. Onesti, editors, *Formale e informale. La variazione di registro nella comunicazione elettronica.*, pages 147–170. Carocci, Roma.

M. Brunello. 2009. The creation of free linguistic corpora from the web. In I. Alegria, I. Leturia, and S. Sharoff, editors, *Proc. of the Fifth Web as Corpus Workshop (WAC5)*, pages 9–16. Elhuyar Fundazioa.

S. Buchholz and E. Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proc. Tenth Conf. Comput. Nat. Lang. Learn.*, number June in CoNLL-X '06, pages 149–164. Association for Computational Linguistics.

C. Culy, V. Lyding, and H. Dittmann. 2011. xldd: Extended linguistic dependency diagrams. In *Proc. of the 15th International Conference on Information Visualisation IV2011*, pages 164–169, London, UK.

T. De Mauro. 1991. *Guida all'uso delle parole*. Editori Riuniti, Roma.

F. Dell'Orletta, G. Venturi, and S. Montemagni. 2011. Ulisse: an unsupervised algorithm for detecting reliable dependency parses. In *Proc. of CoNLL 2011, Conferences on Natural Language Learning*, Portland, Oregon.

F. Dell'Orletta, G. Venturi, and S. Montemagni. 2013. Unsupervised linguistically-driven reliable dependency parses detection and self-training for adaptation to the biomedical domain. In *Proc. of BioNLP 2013, Workshop on Biomedical NLP*, Sofia.

F. Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia.

S. Evert and A. Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proc. of the Corpus Linguistics 2011*, Birmingham, UK.

M. Généreux, I. Hendrickx, and A. Mendes. 2012. A large portuguese corpus on-line: Cleaning and preprocessing. In *PROPOR*, volume 7243 of *Lecture Notes in Computer Science*, pages 113–120. Springer.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. 2008. A large-scale classification of english verbs. *Journal of LRE*, 42:21–40.

A. Lenci, G. Lapesa, and G. Bonansinga. 2012. Lexit: A computational resource on italian argument structure. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proc. of LREC 2012*, pages 3712–3718, Istanbul, Turkey, May. ELRA.

P. Lucisano and M. E. Piemontese. 1988. Gulpease: una formula per la predizione della difficolt dei testi in lingua italiana. *Scuola e città*, 39(3):110–124.

V. Lyding, C. Borghetti, H. Dittmann, L. Nicolas, and E. Stemle. 2013. Open corpus interface for italian language learning. In *Proc. of the ICT for Language Learning Conference, 6th Edition*, Florence, Italy.

D. McClosky, E. Charniak, and M. Johnson. 2006. Reranking and self-training for parser adaptation. In *Proc. of ACL 2006, ACL*, Sydney.

S. Petrov and R. McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Proc. of SANCL 2012, First Workshop on Syntactic Analysis of Non-Canonical Language*, Montreal.

G. Rehm, M. Santini, A. Mehler, P. Braslavski, R. Gleim, A. Stubbe, S. Symonenko, M. Tavosanis, and V. Vidulin. 2008. Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proc. of LREC 2008*, pages 351–358, Marrakech, Morocco.

M. Santini, A. Mehler, and S. Sharoff. 2010. Riding the Rough Waves of Genre on the Web. Concepts and Research Questions. In A. Mehler, S. Sharoff, and M. Santini, editors, *Genres on the Web: Computational Models and Empirical Studies.*, pages 3–33. Springer, Dordrecht.

M. Santini. 2005. Genres in formation? an exploratory study of web pages using cluster analysis. In *Proc. of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK05)*, Manchester, UK.

S. Sharoff. 2006. Creating General-Purpose Corpora Using Automated Search Engine Queries. In M. Baroni and S. Bernardini, editors, *Wacky! Working Papers on the Web as Corpus*, pages 63–98. Gedit, Bologna.

S. Sharoff. 2010. Analysing similarities and differences between corpora. In *7th Language Technologies Conference*, Ljubljana.

J. M. Steger and E. W. Stemle. 2009. KrdWrd – The Architecture for Unified Processing of Web Content. In *Proc. Fifth Web as Corpus Work.*, Donostia-San Sebastian, Basque Country.

C. A. Thompson, M. E. Califf, and R. J. Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proc. of ICML99, the Sixteenth International Conference on Machine Learning*, San Francisco, CA.