A Diachronic Italian Corpus based on "L'Unità"

Pierpaolo Basile
Dept. of Computer Science
University of Bari, Italy
pierpaolo.basile@uniba.it

Annalina Caputo
ADAPT Centre
School of Computing, Dublin City University
annalina.caputo@dcu.ie

Tommaso Caselli
CLCG
University of Groningen, Netherlands
t.caselli@rug.nl

Pierluigi Cassotti
Dept. of Computer Science
University of Bari, Italy
pierluigi.cassotti@uniba.it

Rossella Varvara
DILEF
University of Florence, Italy
rossella.varvara@unifi.it

Abstract

English. In this paper, we describe the creation of a diachronic corpus for Italian by exploiting the digital archive of the newspaper "L'Unità". We automatically clean and annotate the corpus with PoS tags, lemmas, named entities and syntactic dependencies. Moreover, we compute frequency-based time series for tokens, lemmas and entities. We show some interesting corpus statistics taking into account the temporal dimension and describe some examples of usage of time series.

1 Motivation and Background

Diachronic linguistics is one of the two major temporal dimensions of language study proposed by de Saussure in his *Cours de languistique générale* and has a long tradition in Linguistics. Recently, the increasing availability of diachronic corpora as well as the development of new NLP techniques for representing word meanings has boosted the application of computational models to investigate historical language data (Hamilton et al., 2016; Tahmasebi et al., 2018; Tang, 2018). This culminated in SemEval-2020 Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020), the first attempt to systematically evaluate automatic methods for language change detection.

Italian is a Romance language which has undergone lots of changes in its history. Its official adoption as a national language occurred only after the Unification of Italy (1861), having previously been a literary language. Diachronic corpora of Italian are currently available and accessible to the public (e.g., DiaCORIS and MIDIA). Unfortunately, restricted access/distribution of these resources limits their utilisation. This actually prevents the investigation of more recent NLP methods to the diachronic dimensions.

To obviate this limit, we collect and make freely available[1] a new corpus based on the newspaper "L'Unità". Founded by Antonio Gramsci on February, 12^{th} 1924, "L'Unità" was the official newspaper of the Italian Communist Party (PCI [2], henceforth). The newspaper had a troubled history: with the dissolution of PCI in 1991, the newspaper continued to live as the official newspaper of the new Democratic Party of the Left (PDS/DS) until July, 31^{th} 2014. After that date, it ceased its publication until June, 30^{th} 2015, and it was definitely closed on June, 3^{rd} 2017.

Since 2017, the historical archive of "L'Unità" has been made again visible and available on the Web.[3] One of the main issues of this resource is the lack of information about who owns the rights of the original archive. To our knowledge, the online version of the archive was legally obtained by downloading the original archive before the closure of the newspaper. The current archive, available online, does not contain the local editions of the newspaper and the photographic archive.

The main contribution of this work lies in the

[1] https://github.com/swapUniba/unita/
[2] It is the acronym of *Partito Comunista Italiano*.
[3] https://archivio.unita.news/

resource itself and its accessibility to the research community at large. The corpus is distributed in two formats: raw text and pre-processed. The validity of the corpus for the automatic study of language change is currently tested as part of the DIACR-Ita task [4] at EVALITA 2020. However, we illustrate some further potential applications of the use of the corpus.

2 Italian diachronic corpora

Various Italian diachronic corpora are currently available and accessible to the public. DiaCORIS [5] (Onelli et al., 2006) comprises written Italian texts produced between 1861 and 1945, for a total of 100 million words, while MIDIA [6] (Gaeta et al., 2013) covers written documents in Italian between the beginning of the XIII century and the first half of the XX century, for a total of 7,5 million words over 800 texts belonging to different genres. The Corpus OVI dell'Italiano antico[7] consists of 1948 texts from the XII to the XIV centuries, for a total of 536.000 words. The LIZ[8] database comprehends 1,000 literary texts from the XIII to the XX century. Lastly, the *Corpus of Alcide de Gasperi*'s public documents (Tonelli et al., 2019) includes 1,762 documents (newspaper articles, propaganda documents, official letters, parliamentary speeches, for a total of 3.000.000 tokens) written from the Italian politician Alcide De Gasperi and published between 1901 and 1954.

These existing resources differ from each other and from the present corpus in different ways. First, the span of time the texts come from. The OVI Corpus considers texts from the early stages of the Italian language, with a time span of three centuries. The MIDIA corpus and the LIZ database cover 7 centuries, from the XIII to the first half of the XX century. DiaCORIS, the De Gasperi's corpus and L'Unità corpus contain texts from a shorter and more recent period of time. However, the time span considered in L'Unità corpus is interesting for the study of the Italian language because of the deep changes that occurred in that period. Indeed, the second half of the XX century has seen a wider spread and use of Italian among all the social classes.

Second, these corpora differ for the genres represented. The DiaCORIS and MIDIA corpora have been designed as representative and balanced samples of written Italian (considering, among other genres, academic prose, fiction, press, legal texts, etc). The OVI corpus and the LIZ database comprehend only literary texts. The De Gasperi's corpus is representative of political text from a single author. L'Unità corpus is representative only of press language, but this restriction may be an advantage in the study of diachronic lexical change. Indeed, observed semantic changes cannot be attributed to attestation from different genres in different periods, but can be interpreted as true semantic shifts.

Lastly, even if most of the corpora can be queried online (with the exception of the LIZ database), only the De Gasperi's corpus can be freely downloaded. This restriction affects the usability of these resources for the NLP community. With L'Unità corpus we aim at releasing a new diachronic resource that is freely available and that can be used in the theoretical and computational study of language change.

3 Corpus Creation

The corpus creation consists of several steps:

Downloading All PDF files are downloaded from the source site and stored into a folder structure that mimics the publication year of each article.

Text extraction The text is extracted from the PDF files by using the Apache Tika library.[9] First, the library tries to extract the embedded text if present in the PDF; otherwise the internal OCR is exploited. It is important to notice that during this step several OCR errors occur. In particular, during the processing of the early years, the newspaper has an unconventional format where a few large pages contain many articles split into several columns. Due to this format, the OCR is not able to correctly identify the column boundaries.

Cleaning In this step, we try to fix some text extraction issues. The previous step leaves an empty

[4] https://diacr-ita.github.io/DIACR-Ita/

[5] http://corpora.dslo.unibo.it/DiaCORIS/

[6] www.corpusmidia.unito.it

[7] http://gattoweb.ovi.cnr.it

[8] https://www.zanichelli.it/ricerca/prodotti/liz-4-0-letteratura-italiana-zanichelli

[9] https://tika.apache.org/

1	Ehud	Ehud	PROPN	SP	nsubj	3	B-PER	False	False	False	Xxxx
2	Barak	Barak	PROPN	SP	flat:name	1	I-PER	False	False	False	Xxxxx
3	scende	scendere	VERB	V	ROOT	0	O	False	False	False	xxxx
4	direttamente	direttamente	ADV	B	advmod	3	O	False	False	False	xxxx
5	in	in	ADP	E	case	6	O	False	False	True	xx
6	campo	campire	NOUN	S	obl	3	O	False	False	False	xxxx
7	per	per	ADP	E	mark	8	O	False	False	True	xxx
8	ufficializzare	ufficializzare	VERB	V	advcl	3	O	False	False	False	xxxx
9	la	la	DET	RD	det	10	O	False	False	True	xx
10	candidatura	candidatura	NOUN	S	obj	8	O	False	False	False	xxxx
11	dell'	dell'	DET	DD	det	13	O	False	False	False	xxxx'
12	ex	ex	ADJ	A	amod	13	O	False	False	True	xx
13	premier	premier	NOUN	S	obj	8	O	False	False	False	xxxx
14	laburista	laburista	PROPN	SP	amod	13	O	False	False	False	xxxx

Table 1: An example of generated token features for the sentence: "*Ehud Barak scende direttamente in campo per ufficializzare la candidatura dell'ex premier laburista.*" [Ehud Barak takes the field to announce the candidacy of the former labour leader.]

line when the end of a paragraph is reached. However, a paragraph can be composed of multiple lines which sometimes contain a word break at the end of the line. We manage word breaks in order to obtain a paragraph on a single text line; we still retain the empty line for delimiting paragraphs. Moreover, we remove noisy text by adopting two heuristics: (1) paragraphs must contain at least five tokens composed by only letter characters; (2) 60% of the paragraph must contain words that belong to a dictionary. The dictionary is built by extracting words that occur into the Paisà corpus (Lyding et al., 2014) taking into account only words composed by letters. The output of this process is a plain text file for each year where each paragraph is separated by an empty line.

Processing All plain text files produced by the cleaning step are processed by a Python script that splits each paragraph into sentences and analyses each sentence by performing several natural language processing tasks. We rely on the spaCy[10] Python library for performing: tokenization, PoS-tagging, lemmatization, named entity recognition and dependency parsing. The spaCy library provides performance comparable to the state-of-the-art approaches with a good processing speed when compared to other NLP tools.[11] We also provide the plain text in order to allow the processing with other tools. Each plain text file is analysed and transformed in vertical format adding two tags: `<p>...</p>` for the begin and the end of a paragraph, and `<s>...</s>` for delimiting sentences. The vertical format is compliant to the CONLL representation standard and the tagset for the Italian[12] is automatically mapped to the Universal Dependencies scheme[13].

Feature	Description
Position	The token position in the sentence starting from 1
Token	The token
Lemma	The lemma
PoS-tag	The PoS tag
Tag	Additional tags, such as morphological tags
Dependency	Dependency type
Head position	Head position of the dependency
IOB2 NE	IOB2 tag of the named entity
Punctuation	Boolean indicating if punctuation
Space	Boolean indicating if space character
Stop word	Boolean indicating if stop word
Shape	The word shape – capitalisation, punctuation, digits

Table 2: Description of token features.

The corpus spans 67 years from 1948 to 2014. For each year, we provide two files: (1) the plain text file containing the cleaned text extracted from PDF where each paragraph is delimited by an empty line; (2) a vertical file. In the vertical file format, exemplified in Table 1, each paragraph is split in sentences and tokens occurring in each sentence are annotated with 12 features, whose symbols and descriptions are reported in Table 2.

4 Corpus Statistics

In this section, we report some corpus statistics. Table 3 illustrates the total number of occurrences and the dictionary size for each feature (token, lemma, and named entity, respectively).

	dict. size	occurrences
token	4,177,128	425,833,098
lemma	4,053,561	425,833,098
named entity	5,429,470	26,330,273

Table 3: Dictionary size and total number of occurrences.

[10] https://spacy.io/

[11] https://spacy.io/usage/facts-figures

[12] https://spacy.io/api/annotation

[13] http://universaldependencies.org/u/pos/

The corpus contains more than 400 million occurrences and more than 25 million named entities occurrences. The most frequent entities are *Italia*, *Roma* and *PCI*. This result is expected since "L'Unità" was the newspaper of the Italian Communist Party.

Figure 1 shows the PoS-tags[14] frequency over time for open-class tags: NOUN, VERB, ADJective, ADVerb and PROPer Noun. The most frequent tag is NOUN followed by VERB, PROPN, ADJ and ADV. We observe that the frequency of PoS-tags is almost constant over time (excluding PROPN) underlying a stable language style that is typical for the news domain. We observe a variable usage of proper nouns, that may be related to the different types of events narrated over time that do not depend on a particular language style. Moreover, after the 1976, we observe a complementary trend between the adjectives and adverbs frequencies: the former slightly increase over time, while the latter decrease. This may denote a change in the language style that has varied to prefer the usage of adjectives over adverbs in more contemporary writing.

An interesting analysis concerns the tokens occurrences per year, whose result is plotted in Figure 2. We observe a low number of occurrences in the period (1948-1970), probably due to two factors: (1) the first period contains many OCR errors and noise removed during the cleaning step; (2) the number of pages of the newspaper increases over time. The latter may also explain the lower number of tokens for some of the years, such as 1981, 1995, 2000, 2007-2008, 2014. In particular, the latest years are characterised by management issues (e.g. the newspaper liquidation in July 2000) that were reflected in the newspaper format.

We also compute the time series of normalised occurrences (frequency) for each token, lemma, and named entity. All the aforementioned statistics are distributed in separate files together with the corpus.

As an illustrative example of the potential use of the corpus, in Figure 3 we plot the time series for two keywords. The first, *comunismo* [comunism], is assumed to be pivotal to this corpus due to the specific role played by the newspaper in relation to the PCI. The second keyword, *antipolitica* [anti-politics], is particularly interesting as it is a term used to describe the current state of the political life in Italy, characterised by a high level of distrusts in parties and, more generally, in politics. The lifespan of *comunismo* [comunism] appears to be extremely influenced and characterised by history. We observe two big spikes in the time series. The first is around 1962, one of the harshest year of the Cold War, witnessing the Cuban missile crisis. The second spike is between 1989 and 1991, corresponding to the beginning of the worldwide crisis of the communist movement and the dissolution of PCI. After 1991, the frequency of the term constantly decreases. Interestingly, the frequency for *comunismo* [comunism] is low between 1968 and 1988, a period of time that witnessed a cultural hegemony of leftist movements and strong criticism against the U.S.S.R. On the other hand, we observe that *antipolitica* [anti-politics] is a recent term whose first appearance dates back to 1977. The word frequency starts to increase slowly from 1999 and it reaches its peak in 2012 with the unexpected electoral success of the populist 5 Star Movement at the local elections in May.

Using the same approach, we plot the time series for two named entities: *PCI* and *Berlusconi*. We notice that the frequency of *PCI* start dropping in 1986, few years before its dissolution in 1991, while the name *Berlusconi* has a substantial increase in 1994 when he became the Italian Prime Minister.

Finally, we investigate how the vocabulary changes between two periods: $T_1 = [1948-1958]$ and $T_2 = [2004-2014]$. For each period we build the vocabulary V_i taking into account only words that occur at least 10 times. Then, we compute the differences between the two dictionaries, $V_1 \setminus V_2$ and $V_2 \setminus V_1$, and sort the words in descending order by occurrences. We observe that the words *agrari, imperialisti, mezzadri, monarchici*[15] appear frequently in T_1 and never appear in T_2, conversely the words *euro, centrosinistra, centrodestra, immigrati*[16] appear only in T_2. A similar analysis was executed on named entities[17] and shows that *Scelba, D.C., PSI, U.R.S.S.* are specific to T_1, while *Berlusconi, PD, Bush, Obama* to T_2, revealing differences in topics and people covered

[14]The used tag-set is described here `https://universaldependencies.org/u/pos/`

[15]In English: *agrarians, imperialists, sharecroppers, monarchists*.

[16]In English: *euro, centre-left politics, centre-right politics, immigrants*.

[17]In this case we consider only entities that appear at least 5 times.

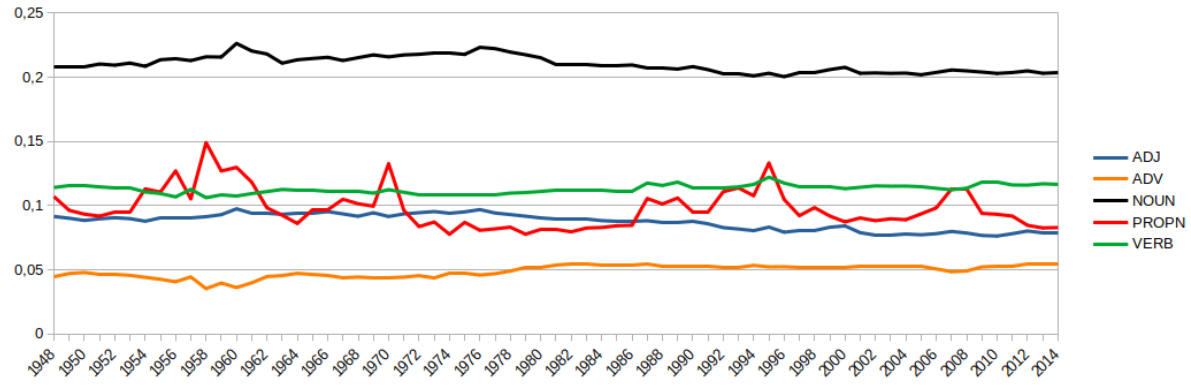

Figure 1: PoS tags frequency over time for: NOUN, VERB, ADJective, ADVerb

Figure 2: The plot of token occurrences per year.

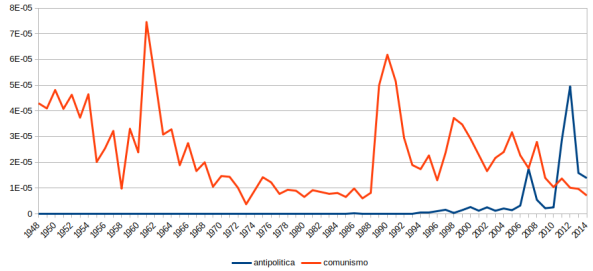

Figure 3: Plot of the time series for the words *comunismo* [comunism] and *antipolitica* [anti-politics].

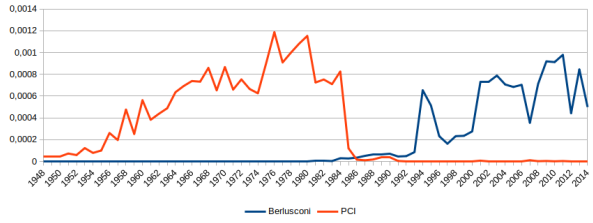

Figure 4: Plot of the time series for the entities *PCI* and *Berlusconi*.

by the newspaper.

5 Conclusions

In this paper, we describe an Italian diachronic corpus based on the newspaper "L'Unità". The corpus spans 67 years (1948-2014) and is provided both in plain text and in an annotated format that includes PoS-tags, lemmas, named entities, and syntactic dependencies. We compute some statistics and time series for each token, lemma and named entity. We think that the corpus and the precomputed data are a valuable source of information both for linguists and researchers interested in diachronic analysis of the Italian language, and for historians, political scientists, and journalists as a digital resource enriched with automatic text analysis technologies.

However, the corpus has some issues that we plan to fix in the future, such as OCR errors and logical document structure recognition. We also plan to process the corpus by exploiting other Italian NLP pipelines in order to understand the differences between the output of different tools. Finally, we are working on generating and making available temporal word embeddings for each year.

References

Livio Gaeta, Iacobini Claudio, Ricca Davide, Angster Marco, De Rosa Aurelio, and Schirato Giovanna. 2013. Midia: a balanced diachronic corpus of italian. In *21st International Conference on Historical Linguistics*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 3, pages 1489–1501, may.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The paisa'corpus of italian web texts. In *9th Web as Corpus Workshop (WaC-9)@ EACL 2014*, pages 36–43. EACL (European chapter of the Association for Computational Linguistics).

Corinna Onelli, Domenico Proietti, Corrado Seidenari, and Fabio Tamburini. 2006. The DiaCORIS project: a diachronic corpus of written Italian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Lexical Semantic Change. *1st International Workshop on Computational Approaches to Historical Language Change 2019*.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676, sep.

Sara Tonelli, Rachele Sprugnoli, Giovanni Moretti, and Fondazione Bruno Kessler. 2019. Prendo la parola in questo consesso mondiale: A multi-genre 20th century corpus in the political domain. In *CLiC-it*.