Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus: C-Oral-Rom Italian

Alessandro Panunzi[1], Eugenio Picchi[2], Massimo Moneglia[1]

[1]Dipartimento di Italianistica, Università di Firenze
Piazza Savonarola, 1, 50123, Florence, Italy
alessandro.panunzi@email.it, moneglia@unif.it

[2]Istituto di Linguistica Computazionale, CNR Pisa
Via Moruzzi, 1, 50124, Pisa, Italy
picchi@ilc.cnr.it

Abstract

The automatic lemmatization and morpho-syntactic annotation of spoken language is a quite recent and complex task for Natural Language Processing. The state of the art on written corpora don't provide us with a satisfactory level of analysis regarding spontaneous spoken language (Uchimoto et al., 2002; Moreno & Guirao, 2003). The spontaneous speech corpus Italian C-ORAL-ROM has been tagged with Part of Speech (Pos) and morpho-syntactic information, using and adapting an already existing tool trained on Italian written resources (PiTagger, developed by Eugenio Picchi, ILC-CNR Pisa). The incidence of spoken domain on the performance is within a 10% of errors detected in the manual evaluation procedure. Some issues concerning spoken language emerged. The definition of significant contexts for PoS statistics is to be provided by utterance boundaries; moreover, the relevance of a series of phenomena related to the prosodic parsing has been highlighted: fragmentation phenomena, a relative lack of information for all word adjacent to utterance boundaries; under-specification of PoS for words in connection to secondary prosodic breaks and one word utterances.

1. Corpus and Tool

C-ORAL-ROM (IST 2000-26228) is a multilingual corpus of spontaneous speech for the main romance languages, French, Italian, Portoguese and Spanish (300,000 words for each language; see Cresti et al., 2002; Cresti & Moneglia, forthcoming).

The Italian resource has been tagged using PiSystem, an integrated procedure for textual and lexical analysis, which consists in: (i) DBT text encoding and analysis modules; (ii) a morpho-syntactic analyzer (PiMorpho); (iii) a Part of Speech tagger and lemmatizer (PiTagger).

To achieve automatic analysis and tagging, the PiMorpho and PiTagger components use a set of resources built on coherent bases:

- an electonic dictionary[1]
- a training corpus of 50,000 words of written language, manually tagged;
- the BDR, a database extracted from the training corpus.

The disambiguation of word-forms is processed by statistic measurements (on trigrams) extracted from the training corpus; the main program extimates the maximum likelihood pattern among the possible alternatives given by the morphological component, with a transitional probabilistic method (Picchi, 1994).

The tag-set used by these tools and resources is, for the greater part, in agreement with the EAGLES recommendations for morpho-syntactic annotation of Italian language (Monachini, 1996).

2. Extended tag set for spoken language

The spoken language transcripts of C-ORAL-ROM corpora contain elements of different kinds. More specifically, among the linguistic elements that belong to the dictionary, there is also a wide variety of non-standard linguistic forms in the corpora[2]. The table 1 shows the underlying structure of the tag sets, taking into account non linguistic (NL) and non standard (NS) elements.

ROOT classification	Secondary classification		Elements classified
linguistic elements	standard **(PoS tagset)**	compositional	PoS
		non-compositional	Interjection
	non-standard **(NS tagset)**	compositional	foreign and new forms
		non-compositional	onomatopoeia, acquisition
non-linguistic elements **(NL tagset)**	para-linguistic		fragments, supports and fillings
	extra-linguistic		coughs and laughs

Table 1. Structure of the tag set[3]

2.1. Non-Standard (NS) words tag set

Non-standard word-forms must be taken into consideration in order to give an automatic lemmatization as correct as possible. Excluding regional and dialectal forms (which are standard within the italian dialects), non-standard linguistic forms include the ones inserted in the table 2:

[1] The DMI, a morphologic dictionary of Italian language, developed within the ILC at CNR, Pisa.; it collects 106,090 lemmas encoded with PoS specifications and inflectional tags.

[2] See the tentative list of tags used in transcription presented in Furui, Maekawa & Isahara (2000). As far as we can see from the published material (Zavrel & Daelemans 2000; Van Eynde, Zavrel & Daelemans, 2000), the PoS tagging of the Spoken Dutch Corpus does not specify these phenomena in the tag set.

[3] The complete tag-sets, comprehending the PoS and morpho-synctactic traits annotated, is available at the webpage http://lablita.dit.unifi.it/coralrom/postag/italian.

Non-standard element	Tag	Example
Compositional		
Foreign words	(PoS+)K	they\PERK
New formations	(PoS+)Z	torniante\SZ
Non-compositional		
Acquisition	ACQ	cutta\ACQ
Onomatopeic	ONO	zun\ONO

Table 2. NS tag set

Although these forms are not widely present in the corpus, their treatment is important when the quality level of the tagging results is taken into consideration. The main feature which marks the distinction between these forms is the syntactic value of these elements within the linguistic structures: while foreign and new forms are compositional elements (which follows the syntactic criterion), onomatopoeia and language acquisition forms are non-compositional. The following examples show the different behavior of compositional and non-compositional non-standard elements in speech contexts. The foreign word in the example (a) and the new formation in the example (b) produce complex Noun Phrases (bracketed in the text) and preserve the agreement features. On the contrary, the onomatopoeic element in example (c) is not compositional; i.e. it lacks both syntactic and argumental relationships with respect to the other linguistic elements of the utterance:

(a) le raccomandazioni del gruppo per la prevenzione / [gruppo **spread\AZ**] /
[recommendations for the prevention group / spread group /]

(b) con [quelli **babbussi\SZ** brutti] che arrivavono ...
[with all those ugly thugs coming towards]

(c) prese lo sportello / **bum\ONO** //
[he took the door / bum (he banged it)]

The distinction between compositional and non compositional elements is supported by the analogue behaviour of all these elements with respect to the prosodic structure of the utterance. Non-compositional elements (which include interjections as standard elements) are always isolated by primary or secondary prosodic boundaries and are frequently the only element of the utterance.

2.2. Non Linguistic (NL) Elements

Non linguistic elements present in the transcript are tagged with special codes[4]. These graphic elements represent two orders of phenomena:
- *para-linguistic elements*, which include:
  a) phonetic support elements (mostly pause-fillers);
  b) word-fragments;
- *extra-linguistic elements* (laughs and coughs).

Moreover, all words (or word chain) that the transcribers have not understood are reported as "xxx" in the texts, and receive the special tag X in the lemmatised resource.

[4] The weight of all these elements in the corpus amount in 7,237 over 306,638 tagged tokens (2,36%).

Non-linguistic element	Tag	Example
Paralinguistic	PLG	&he\PLG
Extralinguistic	XLG	hhh\XLG
Not understandable words	X	xxx\X

Table 3. NL tag set

3. Evaluation of Pos tagging and lemmatization

The C-ORAL-ROM Italian resource comprises 306,638 tagged tokens. Since the non-standard and regional forms were inserted in a special pre-dictionary, the PiTagger system, reached a 100% recall of the number of tokens.

The evaluation of the precision of the automatic PoS-tagging procedure is based on a random sampling of 1/100 tokens picked out of the whole C-ORAL-ROM Italian resource, and evaluated in their utterance contexts. Each token is extracted from a different utterance, also randomly selected.

The manual revision of the tagged samples has evaluated the automatic procedure. Only the PoS tag errors are reported below[5]:

Total Evaluated	3069	100%
Correct PoS tag	2773	90,36%
PoS tag errors	296	9,64%

Table 4. Precision

The results of the evaluation are reported in a *confusion matrix* (available at the web page mentioned in note 3). From the confusion matrix, it is possible to obtain the data on precision, recall and f-measure for each category (Table 5); the following table details these measurements, which give an overall estimate of the automatic tagging procedure:

PoS	precision	recall	f-measure
Demonstratives	100,00%	100,00%	1
Prepositions	97,31%	95,83%	0,9656
Verbs	96,05%	96,55%	0,963
Possessives	90,91%	100,00%	0,9524
Articles	94,27%	92,82%	0,9354
Interjections	97,01%	89,45%	0,9308
Adverbs	95,26%	85,52%	0,9013
Personals	86,29%	93,92%	0,8995
Nouns	81,03%	95,86%	0,8782
Number	97,56%	78,43%	0,8696
Conjunctions	86,21%	83,68%	0,8493
Indefinites	70,42%	96,15%	0,813
Adjectives	76,92%	73,17%	0,75
Relatives/Interrog.	66,67%	51,06%	0,5783
Numeral Adjectives	100,00%	28,57%	0,4444

Table 5. Precision, recall and f-measure by PoS

[5] Tested on a corpus of official documents of the UE Commission (500,000 tokens, reviewed by Enrica Calchini), PiTagger reached a 97% of correctness. The same recognition rate was reached on the LABLITA literary sampling corpus.

4. Spoken-specific tasks

4.1. Definition of the relevant context

Lemmatization and PoS tagging need a detailed and coherent definition of the relevant context within which the statistics on disambiguation have to operate. Therefore resources must include annotations regarding context boundaries. As for written language, this is defined by period boundaries, which are given by punctuation signs; on the contrary, the minimal relevant unit for spoken language is defined by the *utterance*, which must be detected and identified within the speech flow. Although the utterance boundaries have been defined, in previous experiences (Uchimoto et alii, 2002), by the automatic detection of pauses (longer than 200 ms), in the C-ORAL-ROM corpora such boundaries are provided by the marking of terminal prosodic breaks ("//", "?", "…"), manually detected in the whole resource by expert operators. PiTagger is sensible to these boundaries.

In the following examples, two possible transcripts of the same dialogic turn are presented: the first one without prosodic boundaries (followed by a second row bearing the possible PoS tags related to ambiguoous word), the second one according to the Italian C-ORAL-ROM corpus (words in italic are ambiguous).

sì dice *che taglia porta* io lì per lì un' elle penso
Conj/Rel Noun/V Noun/V

sì // dice / *che taglia porta* ? io / lì per lì / un' elle / penso //
[yes // he says / which size do you take? There and then / I (say) / an L / I think]

The PoS of ambiguous elements can be decided only in connection with the context boundaries (represented by utterance boundaries). If the disambiguation process worked on the simple nude transcripts, it would also operate on a non-coherent word pattern.

Without prosodic boundaries, the disambiguation of the PoS tags belonging to the words in the dialogic turns would be arbitrary. The presence of significant boundaries annotation is therefore necessary to give the automatic lemmatizer the information on the relevant linguistic units.

4.2. Disambiguation and prosodic boundaries

A second level of evaluation provides us with an estimate of the incidence of these phenomena on the number of errors. As a result, it becomes possible to identify the contexts in which the system encounters problems and to select which of them are caused by specific features of spoken language.

In detail, in spontaneous speech, the automatic PoS tagging procedure is made complex at three levels, in connection with: (a) words adjacent to utterance boundaries; (b) fragmentation phenomena (retracting, interruptions); (c) secondary prosodic boundaries.

4.2.1. Words adjacent to utterance boundaries

a. One word utterances

On the set of total errors in PoS tag (296), 26% are one word utterances where an ambiguous word occurs. In this case, which is typical of spontaneous speech, the algorithms of the disambiguation system, that are based on PoS order statistics, are radically under-determined.

The following is an example of one word utterance. The automatic PoS tagging cannot find contextual evidence to provide a disambiguation.

(d) esatto**Adj?Adv?** //[6]
[exactly]

b. Words in peripheral position

Apart from the words that constitute an utterance, from the collected data it can be observed that around 22% of the tagged tokens with a PoS error occur in the first position of the utterance, and that roughly 13% occur in the last position. The following examples show two real error evaluated in the :

(e) **che\Conj*** devo dire / sono nelle mani del mio compagno //
*[**what** can I say, I'm in my partner's hands]*

(f) e vai sempre **diritto\Noun*** //
*[and keep going **straight on**]*

These data, summed to single word utterances, tell us that in 61% of cases the errors in PoS tagging occur with a peripheral position of the ambiguous words.

Indeed in such position, immediately before and immediately after the end of an utterance, the system is faced with little contextual information. Given the low average utterance length which characterizes spontaneous speech, the percentage of such contexts appears to be quite relevant (roughly 30% of words). To improve the results, the tagging system should be able to take into account the main prosodic breaks as relevant positions for the disambiguation procedure, and to be trained on a training corpus comprising such information..

4.2.2. Fragmentation phenomena

On the total number of utterances within which the selected word was wrongly tagged (296), interruptions and retracting appear in around 10% of cases.

a. Interruptions

Interruptions constitute a peculiar trait of spoken language. An interruption involves the ending of the utterance, and it may create both an anomalous syntactic configuration (which is impossible to find in the written language) and a lack of information necessary to make a decision for disambiguation:

(g) eh / e beh / lui pensa **che\?** +
Conj/Rel?
[hm/well/he thinks **that**]

The lemmatizers do not have either rules or statistics about these kinds of endings which, by definition, are not regular. As disambiguation techniques cannot be applied, in the Italian C-ORAL-ROM annotated output, these contexts do not receive any particular treatment: the choice made was merely that of the automatic tagger.

[6] In the reported examples, an interrogative point (**?**) after the Pos tag means that there is a doubt on categorization, while an asterisk (*****) marks an error in categorizing done by the tagger.

b. Retracting

The retracting phenomenon is often considered under the generic label of the "disfluency phenomena" which occurs in the speech flow. Retracting phenomena are relevant with respect to the morphosyntactic disambiguation process, because they produce an irregular PoS sequence. Taking the utterance as the relevant context in which to apply the disambiguation rules, the retracting phenomenon (marked with "[/]" in the transcripts) gives us unpredictable linear patterns, as:

(h) che volevano l'\Art [/] **l'\Personal*** [/] l'\Art esclusività / più che tutto / ecco //
[well, that they wanted (**the**) exclusivity above all]

In this utterance an ungrammatical pattern made up of a linear sequence of 3 articles is produced. The tagger made an error in tagging the second token, labeling it as a personal pronoun. For a statistic-based lemmatizer, this irregular sequence upsets the n-grams that are the objects of the application of statistics we assumed for disambiguation. For a rule-based lemmatizer, the presence of such an ungrammatical pattern makes the text impossible to analyze by grammatical rules.

In any case, as the retracting patterns are non grammatical ones, the lemmatization and tagging procedure would not treat them within the statistical patterns to disambiguate. The information on retracting, since its idiosyncratic occurrence within the speech flow, would not constitute a relevant information for statistics. In principle the expression(s) which are the object of retracting are redundant from the point of view of a correct PoS chain.

4.2.3. Secondary prosodic boundaries.

The data collected in the evaluation phase of the automatic tagging show that 13,2% of the errors are represented by words in a single tone unit (i.e. isolated by secondary prosodic breaks), mostly standard words with uncertain classification. Indeed, in connection to secondary prosodic boundaries, the statistics extracted from the written training corpus may be hardly recognized and properly applied. In such positions a word-form may have a value that doesn't match with its typical PoS, even if it is not ambiguous. E.g. (contexts in which the tagger has made errors):

a) Standard forms with LEMMA and PoS assignment problems (fix verb-forms used as Discourse Markers, isolated in the speech flow within non-terminal prosodic breaks):

(i) **scusa\Noun*** / cinquantatré / costì è +
[**excuse me** fifty-three\ there\ is]

(l) **ascolta\Noun*** / ci si sente dopo //
[**listen**I'll call you\ later]

b) Standard forms with PoS assignment problems (mostly conjunctions and adverbs that appear both in the first and in the last position of utterances, with special functional values)

(m) dopo_di_che / **appunto\Noun*** +
[after that**exactly**]

(n) di ricaricare le pile / **ecco\Interj*** //
[**well**to recharge the batteries]

A tagging system that is able to assume the information about both the primary and the secondary prosodic boundaries as contextual inputs is needed to achieve an adequate treatment of these forms in the relevant contexts specified by prosodic cues. The training of tools on corpora that comprehend such an annotation level will be highly relevant for the improvement of the results on the automatic tagging for spoken corpora, as the PoS tagging experience of the Dutch Corpus shows (see the tag set in Van Eynde, Zavrel & Daelemans, 2000).

However, the focus on the context information must be correlated to the detection of the relevant boundaries that play a role in the spoken structures.

Spoken language shows peculiar characters that are necessary to be evaluated and adequately treated. All the phenomena highlighted in the paper point out to a correlation between the information provided by the prosodic breaks and the PoS assignment.

5. References

Cresti, E. et alii., (2002). The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus. In Rodriguez, M. C., Suarez Araujo, C. (Eds.), Proceedings of LREC 2002, vol. 1 (pp. 2--10). Paris: ELRA.

Cresti, E., Moneglia, M. (Eds.), (forthcoming). C-ORAL-ROM. Amsterdam: John Benjamins.

Furui, S., Maekawa, K., Isahara, H. (2000). A Japanese national project on spontaneous speech corpus and processing technology, in Proceedings of ASR2000 (pp.244--248). Paris.

Mendes, A., Amaro, R., Bacelar, F., (2003) Reusing Available Resources for Tagging a Spoken Portuguese Corpus. In Tagging and Shallow Processing of Portuguese: workshop notes of TASHA 2003. Lisboa.

Monachini, M. (1996). ELM-IT: EAGLES specifications for Italian morphosyntax. Lexicon Specification and Classification Guidelines. Eagles Document EAG-CLWG-ELM-IT/F, ILC-CNR, Pisa. http://www.ilc.cnr.it/EAGLES96/browse.html

Moreno, A., Guirao, J. M. (2003). Tagging a spontaneous speech corpus of Spanish. To appear in Proceedings of RANPL 2003. Borovets.

Picchi, E. (1994). Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer of Italian. In Proceedings of EURALEX 1994. Amsterdam.

PiSystem, http://www.ilc.cnr.it/pisystem/

Uchimoto, K. et alii, (2002). Morphological Analisis Of The Spontaneous Speech Corpus. In Proceedings of COLING 2002, Taipei.

Van Eynde, F., Zavrel, J., Daelemans, W. (2000). Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus. In Proceedings of LREC 2000, vol. 3 (pp. 1427--1433). Paris: ELRA.

Zavrel, J and Daelemans, W. (2000). Bootstrapping a Tagged Corpus through Combination of Existing Heterogeneous Taggers. In Proceedings of LREC 2000 (pp. 17--20) Paris: ELRA.