Automatic induction of a PoS tagset for Italian

R. Bernardi[1],
A. Bolognesi[2], C. Seidenari[2],
F. Tamburini[2]

[1]Free University of Bozen-Bolzano,
[2]CILTA –University of Bologna

1. Project: Italian Corpus Annotation

- **Project** carried out at the University of Bologna (CILTA);
- **Corpus** 100-million-words synchronic corpus of contemporary Italian (CORIS);
- **Deliverables** part-of-speech tagging for the complete corpus, and (possibly) in a later stage syntactic analysis for a subcorpus.

First question Which PoS classification should we use?

- **Other Projects**
  - Xerox, Grenoble (France)
  - Delmonte, Venezia (Italy)
  - TUT, Torino (Italy)
- **Standards** EAGLES project, guidelines by Monachini.

Second question How much do these classifications depend on linguistic-theories? Would the tagging satisfy the original purpose of Corpus annotation (to provide empirical support to NL applications)?

2. Comparison

- **Agreement** on the main PoS tags: nouns, verbs, adjectives, determiners, articles, adverbs, prepositions, conjunctions, numerals, interjections, punctuation and a class of residual items.

- **Disagreement** on the classification within the main PoS tags. For instance, ”molti luoghi diversi” - many different places- ”molti” (many) is considered

  - an Indefinite DETERMINER in Monachini
  - a Plural QUANTIFIER in Xerox, and
  - Indefinite ADJECTIVE in Delmonte and TUT.

- **Proposal** To follow a bottom-up approach and deduce the PoS classification from empirical data by considering the distributional behavior of words.

3. Distributional Method: Words

- **Aim** To examine the distributional behaviour of some target words we can compare the lexical distribution of their contexts [BM92]:

...	...	il	babbo	gioca	...	dad plays
...	macchina	del	babbo	...	...	car of dad
...	...	il	nonno	gioca	...	grandfather plays
...	macchina	del	nonno	...	...	car of grandfather

- **Result** Using this method on Italian, four different categories are obtained: Verbs (V), Nouns (N) and Grammatical Words (X). [TDSE02]

- **Drawback** sparse data problem which inflates the X category.

4. Distributional Method: Structures

- **First Solution** To solve this problem in [TDSE02] Tamburini et al. applied Brill's method on tags, obtaining a more fine-grained analysis of grammatical words.

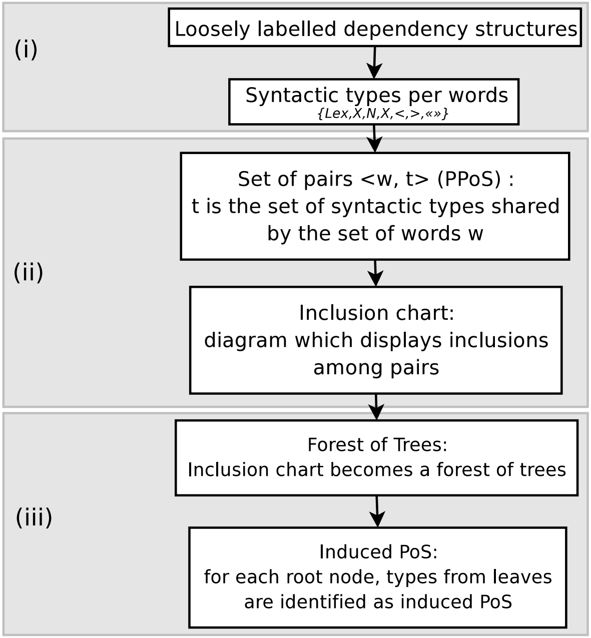
- Relying on limited distributional contexts (± 2 words), the method fails to manage linguistic phenomena involving larger chunks of language such as conjunctions.

			GW	N		GW	N			
la	mamma	incarta	il	regalo	per	il	babbo	. . .	. . .	. . .
(the)	mum	wraps	the	gift	for	(the)	dad			
la	mamma	incarta	il	regalo	e	il	babbo	scrive	il	biglietto
(the)	mum	wraps	the	gift	and	(the)	dad	writes the greetings card		

- Hence
  - With limited context "e" seems to act as "per"
  - Conjunctions may be clustered with prepositions.

- Tags carrying structural information could help overcome this problem.

5. Proposal: Architecture



(i)

Loosely labelled dependency structures

Syntactic types per words
{Lex,X,N,X,<,>,«»}

(ii)

Set of pairs <w, t> (PPoS) :
t is the set of syntactic types shared
by the set of words w

Inclusion chart:
diagram which displays inclusions
among pairs

(iii)

Forest of Trees:
Inclusion chart becomes a forest of trees

Induced PoS:
for each root node, types from leaves
are identified as induced PoS

6. (i): Explanation

DG structures 441 dependency trees with broadly accepted syntactic information:

- Head-Dependent relations ($H < D, D > H, H \ll D$ and $D \gg H$) and distinguishing each dependent either as:
  - an Argument ($H < D_{arg}$ and $D_{arg} > H$) or as
  - an Adjunct ($H \ll D_{adj}$ and $D_{adj} \gg H$).
- words are marked as $\mathtt{N}$ (nouns), $\mathtt{V}$ (verbs) or $\mathtt{X}$ (all others) according to the results obtained in [TDSE02].

From these dependency structures we extract syntactic type assignments by projecting dependency links onto formulas.

Types Formulas are built out of $\{<, >, \ll, \gg, \mathtt{N}, \mathtt{V}, \mathtt{X}, lex\}$ where the symbol lex stands for the word the formula has been assigned to.

7. Input of (i): Dependency Grammar structures

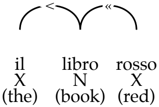
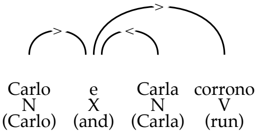
Initial dep. structure	Final type resolution
 il X (the) libro N (book) rosso X (red)	il: $lex{<}N$ libro: lex rosso: $N{\ll}lex$
 Carlo N (Carlo) e X (and) Carla N (Carla) corrono V (run)	Carlo: lex e: $N{>}lex{<}N$ Carla: lex corrono: $X{>}lex$

Figure 1: Type resolution example

8. Output of (i): Set of Types per word (example)

$$\mathsf{e} \;:\; \left\{ \begin{array}{l} X{>}lex{<}X \\ V{>}lex{<}V \\ N{>}lex{<}N \\ N{\ll}X{>}lex{<}X \\ V{\ll}X{>}lex{<}X \\ N{\ll}V{>}lex{<}V \\ N{>}lex{<}X \\ X{>}lex{<}N \\ X{>}lex{<}X \gg N \end{array} \right.$$

9. (ii): Explanation

1. Lexicon entries are gathered together by connecting words which have received the same types. This results in a **set of pairs** $\langle W, T \rangle$ consisting of a **set of words** W **and their shared set of types** T.

   - Sets of words are composed of at least two occurrence words.
   - From the given dependency structures we have obtained 215 pairs. They provide us with a first word class approximation with their associated syntactic behaviors.

   - We will refer to each pair $\langle W, T \rangle$ as Potential PoS (PPoS).

2. In order to interpret the classification obtained and to further refine it, we first organize the pairs into an **Inclusion chart** based on subset relations among the PPoS.

Basic Assumptions

1. a set of syntactic types represented by a single word does not have a linguistic significance.
2. type-set inclusions are due to syntactic similarities between words.

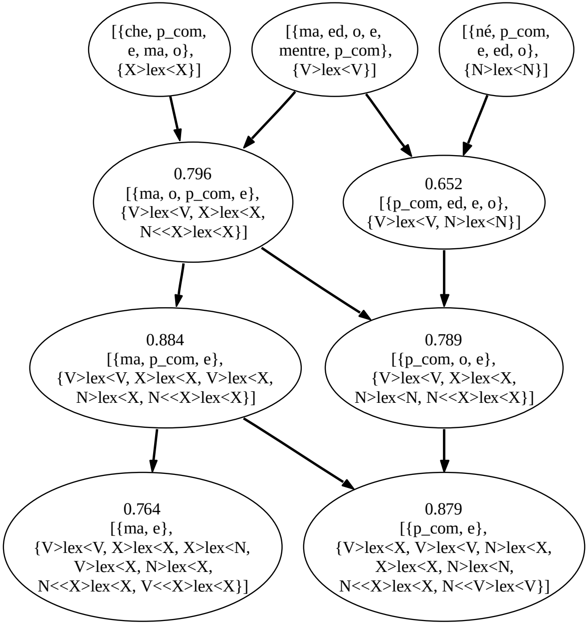
10. Input of (ii): Set of pairs (Examples)

Let us consider the lexicon entries “e” (**and**), “o” (**or**) and “p_com” (**comma separator**). The set of types assigned to “e” is shown above, those for “o” and “p_com” are as below.

$$\text{o} \;:\; \begin{cases} X{>}lex{<}X \\ X{>}lex{<}X{\gg}V \\ N{>}lex{<}N \\ V{>}lex{<}V \\ N{\ll}X{>}lex{<}X \\ N{\ll}N{>}lex{<}N \end{cases} \qquad \text{p\_com} \;:\; \begin{cases} X{>}lex{<}X \\ V{>}lex{<}V \\ N{>}lex{<}N \\ N{\ll}X{>}lex{<}X \\ N{>}lex{<}X \\ N{\ll}V{>}lex{<}V \\ N{>}lex{<}X \\ V{>}lex{<}X \end{cases}$$

The set of words $W_1 = \{$ p_com, e, o$\}$ with the shared set of types $T_1 = \{V{>}lex{<}V, X{>}lex{<}X,$
constitute the pair $\langle W_1, T_1 \rangle$.

11. Output of (ii): Inclusion chart (a fragment)


[{che, p_com, e, ma, o}, {X>lex<X}]
[{ma, ed, o, e, mentre, p_com}, {V>lex<V}]
[{né, p_com, e, ed, o}, {N>lex<N}]
0.796 [{ma, o, p_com, e}, {V>lex<V, X>lex<X, N<<X>lex<X}]
0.652 [{p_com, ed, e, o}, {V>lex<V, N>lex<N}]
0.884 [{ma, p_com, e}, {V>lex<V, X>lex<X, V>lex<X, N>lex<X, N<<X>lex<X}]
0.789 [{p_com, o, e}, {V>lex<V, X>lex<X, N>lex<N, N<<X>lex<X}]
0.764 [{ma, e}, {V>lex<V, X>lex<X, X>lex<N, V>lex<X, N>lex<X, N<<X>lex<X, V<<X>lex<X}]
0.879 [{p_com, e}, {V>lex<X, V>lex<V, N>lex<X, X>lex<X, N>lex<N, N<<X>lex<X, N<<V>lex<V}]

12. (iii): Explanation

1. **From inclusion chart to forest of trees**: In order to extract a suitable PoS classification from the inclusion chart, this must be pruned by discarding less relevant nodes; hence, we need to introduce a relevance criterion to highlight the closest pairs.

   Word Frequency focuses on the similarity between words in W by rating how far words agree in their syntactic behaviour. Roughly, if the word frequency returns a high value for a pair then we can conclude that words within that pair have a close syntactic resemblance.

   Type Frequency rates the similarity between types in T according to the number of times the words to which they have been assigned in the lexicon have shown that syntactic behavior in the dependency structures.

   Pair Frequency is the average of the two cohesion evaluations.

Basic Assumption

1. the relevance of a PPoS depends on how representative its members are with respect to each other: suitable PoS are the closest ones in the inclusion chart

13. (iii): Explanation (Cont'd)

2. **From forest of trees to induced types** Each tree in the Forest marks off complex groups of syntactic types. However, the same types occur in more than one tree, therefore we need to identify all and only those belonging to a given tree.

   Syntactic core Leaves of each tree are grouped together; such groups constitute the whole type set partition. Clearly each group corresponds to a unique root node. Syntactic types from leaf nodes encode few specialized syntactic patterns, i.e. the relevant syntactic component of the corresponding PPoS root node.

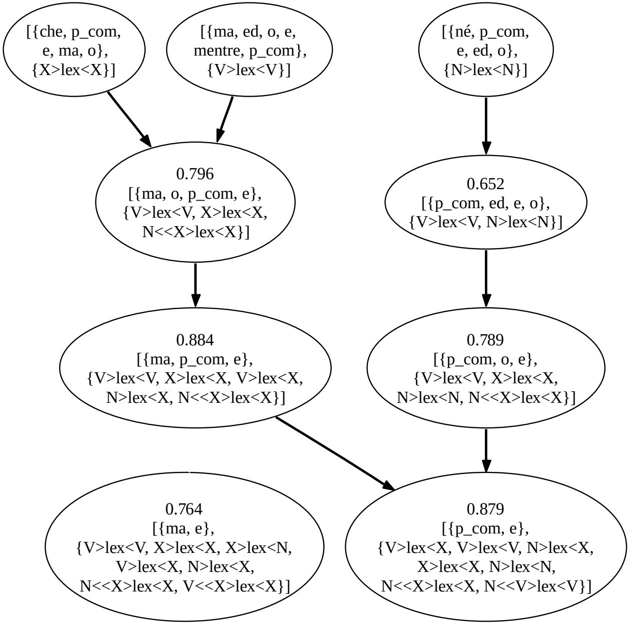
   Lexical core Once a syntactic core is defined, the corresponding lexical core is automatically derived by identifying word sets showing exclusively sets of types belonging to that syntactic core.

   Induced types The syntactic and lexical core is the output of our algorithm. They are the syntactic (and lexical) prototype to be used for PoS classification.

Basic Assumption

1. Syntactic types from leaf nodes encode few specialized syntactic patterns. We assume those patterns to be the syntactic core of a given tree, i.e. the relevant syntactic component of the corresponding PPoS root node.

14. Input of (iii): Forest of trees (a fragment)


[{che, p_com, e, ma, o}, {X>lex<X}]
[{ma, ed, o, e, mentre, p_com}, {V>lex<V}]
[{né, p_com, e, ed, o}, {N>lex<N}]
0.796 [{ma, o, p_com, e}, {V>lex<V, X>lex<X, N<<X>lex<X}]
0.652 [{p_com, ed, e, o}, {V>lex<V, N>lex<N}]
0.884 [{ma, p_com, e}, {V>lex<V, X>lex<X, V>lex<X, N>lex<X, N<<X>lex<X}]
0.789 [{p_com, o, e}, {V>lex<V, X>lex<X, N>lex<N, N<<X>lex<X}]
0.764 [{ma, e}, {V>lex<V, X>lex<X, X>lex<N, V>lex<X, N>lex<X, N<<X>lex<X, V<<X>lex<X}]
0.879 [{p_com, e}, {V>lex<X, V>lex<V, N>lex<X, X>lex<X, N>lex<N, N<<X>lex<X, N<<V>lex<V}]

15. Output of (iii): Induced PoS

The first root node in the fragment of the forest tree has no leaf, being a root without branches, so it contains no syntactic core. On the other hand, the second has the following three leaves:

$$\langle\ \{\text{che},\ \text{p\_com},\ \text{e},\ \text{ma},\ \text{o}\}\,,\{X{>}Lex{<}X\}\ \rangle$$
$$\langle\ \{\text{ma},\ \text{ed},\ \text{o},\ \text{e},\text{mentre},\ \text{p\_com}\}\,,\{V{>}Lex{<}V\}\ \rangle$$
$$\langle\ \{\text{nep\_apo},\ \text{p\_com},\ \text{e},\ \text{ed},\ \text{o}\}\,,\{N{>}Lex{<}N\}\ \rangle$$

Thus its type set contains the **syntactic core** $\{X{>}Lex{<}X, V{>}Lex{<}V, N{>}Lex{<}N\}$

The word "o" has shown $X{>}Lex{<}X$, $V{>}Lex{<}V$, $N{>}Lex{<}N$, but also $N{\ll}X{>}Lex{<}X$ which belongs to both root nodes so the word "o" cannot be part of the lexical entries the syntactic core is represented by.

The second root node is associated with the **lexical core** consisting of {ed, mentre, né, che} Hence the algorithm concludes the existence of the following **PoS prototype**:

$$\langle\{\text{ed},\ \text{mentre},\ \text{né},\ \text{che}\},\{X{>}Lex{<}X, V{>}Lex{<}V, N{>}Lex{<}N\}\rangle$$

16. Classification of PoS tags

PoS Label		Associated types	Prototypical words
Nouns		N	nuvola, finestra, tv
Verbs		V	stupire, raggiunto, concludendo, abbiamo
X	Prepositionals & Determiners	Lex<N, Lex<X, N≪Lex<N, N≪Lex<X, N≪Lex<V, X≪Lex<N, X≪Lex<V, X≪Lex<X	alcuna, della, dieci, diversi, le, molti, negli, numerose, quegli, questi, sei, sull’
	Verb-Modif. Prepositionals	V≪Lex<N, Lex<N≫V, V≪Lex<X, Lex<X≫V	a_causa_del, attraverso, contro, davanti_al, secondo, senza
	Left Adjectivals	Lex≫N	forti, giovane, grande, nuove, piccolo, suo,
	Right Adjectivals	N≪Lex, X≪Lex	economici, elettorale, idrica, importanti, positiva, ufficiale
	Adverbials	V≪Lex, Lex≫V, Lex≫X	allora, appena, decisamente, ieri, mai, molto, persino, rapidamente, presto, troppo
	Coordinators	V>Lex<V, N>Lex<N, X>Lex<X, N>Lex<X, X>Lex<N, V>Lex<X, V≪X>Lex<X, N≪V>Lex<V, N≪X>Lex<X	e, ed, ma, mentre, o, sia
	Subordinators	Lex<V, Lex<V≫V, V≪Lex<V	in_modo_da, oltre_a, quando, perché, se
	Relatives	N>Lex	che, cui, dove, quale
	Entities	Lex	ci, di_più, in_salvo, io, inferocito, noi, ti, sprovveduto, una

Table 1: Resulting PoS classification

Note: the *Coordinators* PoS in the Table above correspond to the one of the previous example, but there it was simpler because of the simplification of the *Inclusion chart* taken by means of example.

17. Evaluation of PoS Classification

- **Preposition and Determiners** the overlapping of determiners and prepositions within the same PoS is noteworthy. The lack of accuracy this classification results in is due, on the one hand, to the wide range of highly specific syntactic constructions involving determiners and prepositions that share the same loosely labeled dependency structures.

Monosyllabic preposition Moreover, Italian monosyllabic (or 'proper') prepositions may be morphologically joined with the definite article (for example *di* ('of') + *il* ('the') = *del* ('of the')), performing sintactically both as a preposition and a determiner. Clearly this class will be further specialized by exploiting morphological information.

Polysyllabic preposition (or 'not proper') prepositions, as opposed to monosyllabic ones, tend to occur in a lower number of syntactic patterns and, more crucially, cannot be fused with the article. In this case our system performs more accurately as it is able to correctly detect the syntactic similarities between such prepositions. As they typically tend to carry the function of the head (together with prepositional locutions) in verb-modifying structures they have been classified as 'Verb-Modifying Prepositionals' as shown in Table 1.

18. Evaluation PoS Classification (Cont'd)

- **Adjectives and Conjunction** The 4 word classes grouping words commonly classified as adjectives and conjunctions may be considered an interesting result of the syntactically motivated induction algorithm presented here.

  Adjectives they have been divided into 2 separate classes depending on predicative or attributive distribution with respect to the noun they modify ('Left/Right Adjectivals' in Table 1).

  Conjunction As far as conjunctions (and conjunctional locutions) are concerned, again, their syntactic patterning enforced a very clear split between 'Coordinators' and 'Subordinators'.

- **Adverbs** By contrast a relatively strong syntactic resemblance has been automatically recognised between words (and locutions) traditionally described as adverbs (and adverbial locutions): hence, the single 'Adverbials' word class is derived. Again, further anlysis exploiting distributional and morphological data may be useful in obtaining a finer-grained classification if necessary.

- **Copulative structures** A final point to make is about copulative structures: our system proved not to properly process them in general, as shown by the fact that their predicative components ended up classified under either 'Entities' or 'Prepositionals & Determiners'.

19. Evaluation of Data and Results

- The sets of automatically extracted syntactic types represent the prototypical syntactic behaviors of the corresponding words summarized by the explanatory PoS labels.

- This classification is not fine-grained enough to be used by a tagger to reach an informative and useful annotation and should be intended as a first step through the empirical construction of a hierarchical tagset, e.g. following the parameters for taxonomic classification shown in [Kaw05]. Further analysis for each class must be carried out to increase the granularity of the tagset, for instance by exploiting morphological information.

- The present study was carried out on a limited quantity of data; the sparseness of primary information we used to derive the proposed tagset might affect the conclusions we have drawn. The results will need to be checked with more data and with different treebanks to avoid biases introduced by the treebank used (TUT) from which the initial dependency structures were extracted.

20. Further Research

The final output of the three phase system will be a **hierarchy** of PoS tags. Such structured organization is expected to help the linguist during the annotation phase as well as when searching the annotated corpus.

Annotate the linguist can browse the graph for a given word to get a sense of its syntactic distribution or to improve the proposed classification (e.g. by splitting an induced category that is too coarse.)

Search since the resulted PoS classification is organized as a hierarchy with inclusion relations, a more intelligent search interface can be constructed to help the user extract the relevant information from the annotated corpus.

References

[BM92] E. Brill and M. Marcus. Tagging an unfamiliar text with minimal human supervision. In *Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language*, pages 10–16, Cambridge, 1992.

[CORIS] `http://corpus.cilta.unibo.it:8080`

[Kaw05] Y. Kawata. *Tagsets for Morphosyntactic Corpus Annotation: the idea of a 'reference tagset' for Japanese*. PhD thesis, University of Essex, Colchester, UK, 2005.

[TDSE02] F. Tamburini, C. De Santis, and Zamuner E. Identifying phrasal connectives in Italian using quantitative methods. In S. Nuccorini, editor, *Phrases and Phraseology -Data and Description*. Berlin: Peter Lang, 2002.