

Identifying phrasal connectives in Italian using quantitative methods

1. Introduction

In recent decades, the analysis of phraseology has made use of the exploration of large corpora as a source of quantitative information about language. This paper intends to present the main lines of work in progress based on this empirical approach to linguistic analysis. In particular, we focus our attention on some problems relating to the morpho-syntactic annotation of corpora.

The CORIS/CODIS corpus of contemporary written Italian, developed at CILTA – University of Bologna (Rossini Favretti 2000; Rossini Favretti, Tamburini, De Santis *in press*), is a synchronic 100-million-word corpus and is being lemmatised and annotated with part-of-speech (POS) tags, in order to increase the quantity of information and improve data retrieval procedures (Tamburini 2000).

The aim of POS tagging is to assign each lexical unit to the appropriate word class. Usually the set of tags is pre-established by the linguist, who uses his/her competence to identify the different word classes. The very first experiments we made revealed how the traditional part-of-speech distinctions in Italian (generally based on morphological and semantic criteria) are often inadequate to represent the syntactic features of words in context. It is worth noting that the uncertainties in categorisation contained in Italian grammars and dictionaries reflect a growing difficulty as they move from fundamental linguistic classes, such as nouns and verbs, to more complex classes, such as adverbs, pronouns, prepositions and conjunctions. This latter class, that groups together elements traditionally used to express connections between sentences, appears inadequate when describing cohesive relations in Italian. This phenomenon actually seems to involve other elements traditionally

¹ C.I.L.T.A. – University of Bologna – Italy
{f.tamburini, c.desantis, e.zamuner}@cilta.unibo.it

The authors drafted this article together. As far as academic requirements are concerned, F. Tamburini takes official responsibility for sections 1 and 4, C. De Santis for section 2 and E. Zamuner for sections 3 and 5.

assigned to different classes, such as adverbs, pronouns and interjections. Recent studies proposed the class of 'connectives', grouping all words that, apart from their traditional word class, have the function of connecting phrases and contributing to textual cohesion. From this point of view, conjunctions can be considered as part of phrasal connectives, that can in turn be included in the wider category of textual connectives.

The aim of this study is to identify elements that can be included in the class of phrasal connectives, using quantitative methods. According to Shannon and Weaver's (1949) observation that words are linked by dependent probabilities, corroborated by Halliday's (1991) argument that the grammatical "system" (in Firth's sense of the term) is essentially probabilistic, quantitative data are introduced in order to provide evidence of relative frequencies.

Section 2 presents a description of word-class categorisation from the point of view of grammars and dictionaries arguing that the traditional category of conjunctions is inadequate for capturing the notion of phrasal connective. Section 3 examines the notion of 'connective' and suggests a truth-function interpretation of connective behaviour. Section 4 describes the quantitative methods proposed for analysing the distributional properties of lexical units, and section 5 comments on the results obtained by applying such methods drawing some provisional conclusions.

2. Categorisations in Italian grammars and dictionaries

The Italian reference grammars examined are: Serianni (1989), which can be considered an authoritative work among the traditional grammars of Italian, and Renzi et al. (1988, 1991, 1995), an innovative work within the framework of generative grammar. We will refer also to Sensini (1997) and Dardano / Trifone (1997).

The reference dictionaries are: De Mauro (1999), the most comprehensive lexicographic work on Italian in use, and Sabatini / Coletti (1997), which takes an innovative approach to the problem of categorisation. We will also refer to Zingarelli (2000) and Devoto / Oli (2001).

2.1 Italian grammars

Serianni, who adopts a traditional terminology in his *Grammar*, proposes a distinction into ten parts of speech (noun, article, adjective, numeral, pronoun, preposition, conjunction, interjection, verb, adverb). However, he admits the problematic nature of the class of conjunctions, which seems to be an open class, sharing features with other parts of speech (prepositions and adverbs). In particular, he underlines the uncertainties in the classification of elements such as *anche*, *pure*, *nemmeno*, *dunque*, *pertanto*, which are sometimes classified as conjunctions and sometimes as adverbs. He refers to recent studies that introduced the category of ‘connectives’ (defined as “words that, apart from their grammatical category, have the function of connecting the different parts of a text”). Together with ‘markers’ (defined as “signals of beginning and ending, placed in the boundaries of a text or a part of it”), they form the category of ‘discourse signals’ (“elements which have the function of organising the presentation of a communicative text”). Many of these elements tend to lose their original semantic value and assume the function of ‘fillers’ (Bustorf 1974).

Renzi et al., who assume the principle of the centrality of syntax in their description, starting from the phrase to go down to the parts of speech, adopt some traditional designations such as: name (the head of a noun phrase), article (determiner of a noun or a noun modifier), adjective (head of the adjectival phrase), verb (head of the verbal phrase), adverb (head of the adverbial phrase), preposition (head of a prepositional phrase), pronoun. This latter class includes: personal, reflexive, possessive and demonstrative. Indefinite pronouns (including also the definite *tutto*) are considered as quantifiers. The relative and interrogative pronouns are considered separately, as introducing, respectively, relative and interrogative phrases.

In particular *che*, by virtue of its properties, is not considered as a pronoun but as a conjunction which introduces all subordinate clauses except interrogative ones². Conjunctions are defined as lexical operators of coordination and subordination. Operators of coordination (divided into operators in the strict sense such as *e*, *o*, *ma*, and adverbial operators such as *perciò*, *tuttavia*, *quindi*) constitute the most substantial group of ‘discourse signals’ (defined as “elements

² Also Graffi (1994), discussing the traditional list of grammatical categories in the light of distributional criteria, suggests considering *che* as a relative pronoun (such as *cui* or *quale*), but prefers to consider it as an ‘operator of complementation’, which characterises the subordinate clauses with finite tenses that are not interrogative.

which lose their original meaning to assume additional values that highlight the nodes of the discourse”). Certain phrasal adverbs, interjections, and verbal or prepositional phrases can act also as discourse signals.

Sensini adopts a traditional classification with nine parts of speech, divided into variable (article, name, adjective, pronoun, verb) and invariable (adverb, preposition, conjunction, interjection). Dealing with conjunctions, he underlines the fluctuations, shown by many elements, between the value of conjunction and that of adverb (in particular the adjunctive *anche*, *inoltre*, *pure*, *altresì*, *per altro*, *nonché*, the closing ones *dunque*, *perciò*, *quindi* and elements such as *altrimenti*, *allora*, *ora*).

Dardano and Trifone propose the same classification, underlying the vague boundaries of the class of conjunctions, due to the mixing of the classes of preposition (such as *per* or *dopo*) and adverb (such as *anche*, *pure*, *dunque*, *allora*, *altrimenti*, *pertanto*). For elements of this kind, simple (*ebbene*, *eppure*, *infatti*, *inoltre*, *insomma*, *nondimeno*, *oltretutto*, *peraltro*, *perciò*, *sennò*, *tuttavia*) or complex (*a ogni modo*, *con ciò*, *d'altronde*, *del resto*, *in breve*, *in conclusione*, *in effetti*, *in realtà*, *in fin dei conti*, *tutt'al più*), they adopt the category of ‘textual connectives’. These elements appear to share two characteristics: the variety of functions and the tendency to determinate their function in relation to the context and the communicative situation. The notion of connective is resumed in the chapter devoted to text and discourse signals (of which the connectives are part). A distinction between ‘semantic connectives’ and ‘pragmatic connectives’ is introduced. The former underline the kind of relation (temporal, causal, logical) between two simple or complex phrases (and largely coincide with the traditional class of conjunctions); the latter express the attitude of the speaker towards the utterance (they are often at the beginning of a phrase).

2.2 Italian dictionaries

For the categorisation of lemmas, the GRADIT dictionary uses nine parts of speech (article, noun, adjective, pronoun, verb, adverb, preposition, conjunction, interjection). There are 53 adverbs also defined as prepositions, 52 adverbs also defined as conjunctions, 4 adverbs also defined as conjunctions or prepositions, and 12 prepositions also defined as conjunctions.

The same division into nine parts of speech can be found in other dictionaries, with some differences in listing, particularly for the classification of conjunctions and adverbs³.

Apart from the traditional categories, the DISC dictionary considers ‘textual conjunctions’ with their respective locutions. These are elements that cannot be easily included in the nine parts of discourse and are vaguely assigned to the category of conjunctions or adverbs (*dunque, ebbene, infatti, inoltre, insomma, oltretutto, peraltro, perciò, sennò, tuttavia, a ogni modo, con ciò, del resto, in realtà*). For some other elements that have a primary function in a phrase, the dictionary indicates any possible use as textual conjunction (for conjunctions as *benché, comunque, cosicché, quando, sebbene*, or adverbs as *allora, altrimenti, anche, ancora, anzi*). The phrasal value of some elements that, when prosodically isolated, concentrate a whole phrase, is also specified.

2.3 EAGLES standards

Mention should be made of the recommendations for the morpho-syntactic annotation of corpora drafted in 1996 by EAGLES (Monachini 1996), on the initiative of the European Commission, in order to define common methodologies and standards for the electronic processing of linguistic resources. In this document, parts of speech are considered as obligatory attributes to be included in a morpho-syntactic tagset. The recommended categories are 12: N (*noun*), P/D (*pronoun/determiner*), AP (*adposition*), I (*interjection*), PU (*punctuation*), V (*verb*), AT (*article*), C (*conjunction*), U (*unique/unassigned*, applied to classes with a unique or very small membership), AJ (*adjective*), NU (*numeral*), R (*residual*, assigned to classes which lie outside the traditional range of grammatical classes, such as foreign words or mathematical formulae).

Dealing with the problem of ambiguity due to the phenomenon of homographs, ‘*port-manteau*’ tags are introduced, which retain more

³ For example, *dunque* in GRADIT and Zingarelli is classified as a conjunction, while Devoto-Oli specifies that “in interrogative phrases it has the value of adverb rather than of conjunction, seeing the equivalence with *insomma*”. *Anche*, that in GRADIT, Zanichelli and Devoto-Oli is classified as a conjunction, except for some literary uses with the adverbial value of *finora, ormai*, in DISC is qualified as an adverb when it modifies a preceding or a subsequent element, and as a conjunction when it has the function of linking phrases. *Allora*, that in Zanichelli, Devoto-Oli and GRADIT is classified as an adverb or conjunction, is classified in DISC as an adverb, save the mention of some uses with the function of ‘textual conjunction’ or ‘discourse signal’.

than one tag and signal the uncertainty in classification of the automatic program. Ambiguities due to human uncertainty are also signalled, particularly when dealing with categories that have fuzzy boundaries. Nevertheless, this problem is not considered a matter of great priority.

Among the codes proposed for Italian, we find traditional distinctions, such as the one between subordinating conjunctions (such as *perché*) and coordinating ones (such as *e*), not entirely adequate to describe the complexity of the phenomenon of connection in Italian.

3. Observations on the notion of 'connective'

Some descriptions of the notion of the phenomenon of connection have been examined so far. Even if we have considered only a restricted amount of the connectives bibliography, it may be suggested that the grammatical categorisations are not stable enough to provide an account of the results to be pointed out.

In recent decades, linguists such as Van Dijk (1977) have preferred the notion of 'connective' to that of 'conjunction', in order to highlight the cohesive function that such items develop. In this respect, we can consider the 'connective' as a term used in the grammatical classification of words or morphemes whose function is primarily to link linguistic units at any level. However, it is important to recall that the notion of 'connective' entails reference to truth-function semantics.

For the time being, our aim is to describe, based on the empirical data, the type of lexical units that can be traced to the category of 'connectives', as signals of inner links. In the discussion above of grammar categorisation, a range of definitions such as 'textual connectives', 'semantic connectives', 'pragmatic connectives' and 'discourse signals' were examined, and such definitions are strongly linked to the notions of text and context. Speaking of 'discourse signals', for example, implies a prior interpretation of the behaviour of these linguistic items, observed in their 'natural environment', namely propositions, portions of texts or whole texts. In contrast with this interpretation, a different point of view could be considered.

Distributional properties of lexical units may be represented by the quantitative method. In our opinion, distributional data may be assumed as a minimal syntactic description. A quantitative evaluation

of syntactic properties allows for no appeal to context or pragmatic indications. This means that the context of occurrence of the linguistic items is narrow, fixed within vectorial bounds. Based on this assumption, we tried to make a neat distinction between grammatical and lexical words. The result made it possible to pick out a class of linguistic items, among the grammatical words, in order to evaluate their distributional properties and trace them back to the category of 'connective'. (For further discussion of this point, see section 4).

3.1 Truth-function semantics

As we have seen, the syntactic notion of conjunction may be traced back to the general semantic concept of connective. There is a considerable resemblance between this concept and the formal-logic connective. This resemblance may be confirmed by the evidence that the function of natural connectives may be traced, at least in part, to that of logic connectives.

In this respect, we can refer to the conception of the meaning of connectives within the domain of truth-function semantics. The logical connectives (\neg , \wedge , \vee , \rightarrow) are functions that make it possible to calculate the truth-value of a molecular proposition from the truth-value of the atomic propositions; the connective is, at the same time, a constitutive part of the molecular proposition and rules the calculation.

This type of semantics has proved effective in the case of syntactically regimented artificial languages. In languages of this type, the biunivocal correspondence between grammatical and semantic categories is guaranteed; for example, an individual constant, a formal equivalent of the grammatical categories 'common noun' and 'proper noun' corresponds in the rules of the interpretation to an object (or a class of objects). This assumption underlies tarskian semantics and is shared by modellistic semantics. Some authors, especially Kaplan (1970) and Montague (1974a, 1974b, 1974c), have applied this to natural languages.

As shown in the following section, the meaning of the natural connective *e* is not merely a formal interpretation; the use of *e* can produce additional semantic effects, such as the temporal relation (Strawson 1952); effects which the representation of meaning, through the truth-functions, does not pick up. For this reason, truth-function semantics may be able to express only some aspects – mainly, the recurring ones – of the linguistic functioning of natural connectives. Formal semantics, such as the truth-function type, can interpret and

represent the logical properties of expressions but cannot represent the non-logical functioning (Van Dijk 1977). Nevertheless we think it possible to suggest a notion of linguistic functioning of natural connectives arising from the truth-function nucleus.

Undoubtedly, the linguistic meaning of a natural connective, such as *e*, cannot be completely carried out by the truth-table for the conjunction; however, in the case of *e* – as well as *o*, *non*, *se...allora* – we have the impression that the truth-functions capture part of our semantic intuitions.

4. Quantitative measures

Our approach is based on the hypothesis that if two words are syntactically and semantically different, then they will appear in different contexts, as suggested in Harris (1951). There are a number of studies that, starting from this hypothesis, have constructed automatic or semi-automatic procedures for clustering words (Brill *et al.* 1990, Brill 1993, Brown *et al.* 1992, Pereira *et al.* 1993, Martin *et al.* 1998). They examine the distributional behaviour of some target words, comparing the lexical distribution of their respective collocates using some quantitative measures of distributional similarity (Lee 1999). The work we present is based on a method first introduced by Brill and Marcus (1992), who set up a semi-automatic procedure that, starting from the lexical statistical data collected from a large corpus, aims to arrange some target words in a tree (more precisely a dendrogram), instead of clustering them automatically. This procedure requires a linguistic examination of the resulting tree, in order to identify the word class that is most appropriate to describe the phenomenon under investigation. In this sense they use a semi-automatic word-class generator method. Our work presents a similar procedure for clustering words based on their distributional behaviour, but some interesting differences have to be pointed out.

Brill and Marcus' method, in common with others in the bibliography above, simply collects lexical data, establishing relations among the collocates lexical distribution of the various target words. As usual, working in a strictly lexical environment leads to the well known sparse-data problem. A corpus, no matter how large it is, is not able to provide all the statistical information needed to analyse complex phenomena such as connection in great detail. In the early

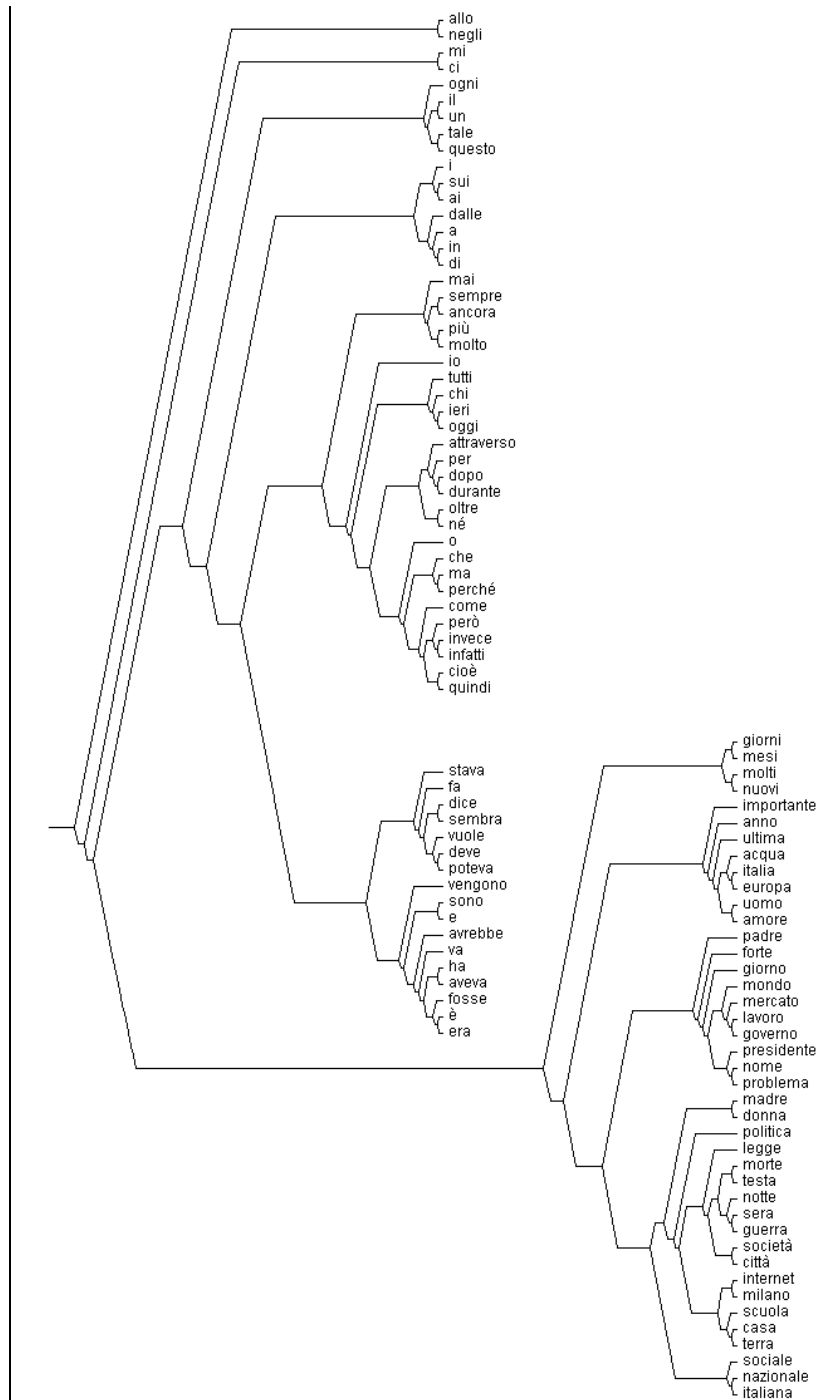


Figure 1. The distribution and grouping of lexical and grammatical words.

experiments, conducted on the CORIS corpus, mainly based on the collection of lexical co-occurrences of words, we obtained dendrograms that did not adequately represent the phenomenon under examination. Even considering that the CORIS is a large corpus, it did not provide enough data to perform such strictly lexical computations.

In modern linguistics there is a wide acceptance of the distinction, mainly based on the concept of open and closed set of words, between lexical words (content words) and grammatical words (or functional words) (Halliday 1985). Accepting such a distinction, it is possible to postulate four main categories of words, three belonging to the set of lexical words (nouns, verbs, qualitative adjectives) and one large class that collects all the grammatical words. We also had to include the set of mood adverbs in the lexical classes, but the complex behaviour of adverbs and the different positions of various studies led us to include them in the class of grammatical words, meaning that we do not attribute any kind of specific function to them. Thus, the class of grammatical words becomes a global class that contains all those word classes not widely or universally accepted as lexical words. In support of this idea, we applied the lexical Brill and Marcus method to the whole CORIS corpus, considering as target words some high frequency words. Figure 1 shows a clear distinction among lexical words (at the bottom of the dendrogram) and grammatical words (at the top of the dendrogram). Verbs, nouns and adjectives are grouped together and are clearly divided from grammatical words.

If we tag a corpus using only four part-of-speech classes, nouns, verbs, qualitative adjectives and grammatical words (actually we have to add some more categories for punctuation marks, but they are not problematic or controversial), we can apply a method similar to that proposed by Brill and Marcus and analyse the distributional behaviour of the target words among the word classes that appear in their context. Having only a small set of word classes, it is possible to collect the required information, while totally avoiding the sparse-data problem. The class of grammatical words is the category which our work mainly focuses on and is not further divided into subclasses, allowing for an unbiased analysis using the methods described above.

The corpus used to derive the statistical data consist of 25 million tokens, automatically tagged using the word classes described above and the tagger designed by Tamburini (2000). The texts are taken from the fiction subcorpus of CORIS.

In our work, *target words* are represented by the most frequent grammatical elements (including multiword units) among those traditionally assigned to the sets of conjunctions, interjections, adverbs

and relative or interrogative pronouns. For each target word (**tw**), the tags of the words appearing in its context are collected, and their probability of appearance in the corpus, in terms of probability distribution, is computed and stored to form a *distributional fingerprint* of the target word (figure 2).

The probability distributions computed considering the tag distributions in the context of each target word (maintaining the distributional position of the various tags) are concatenated to build one single data vector forming the fingerprint of the examined word and then compared with all the fingerprints of the other target words, by pairs, using the distance:

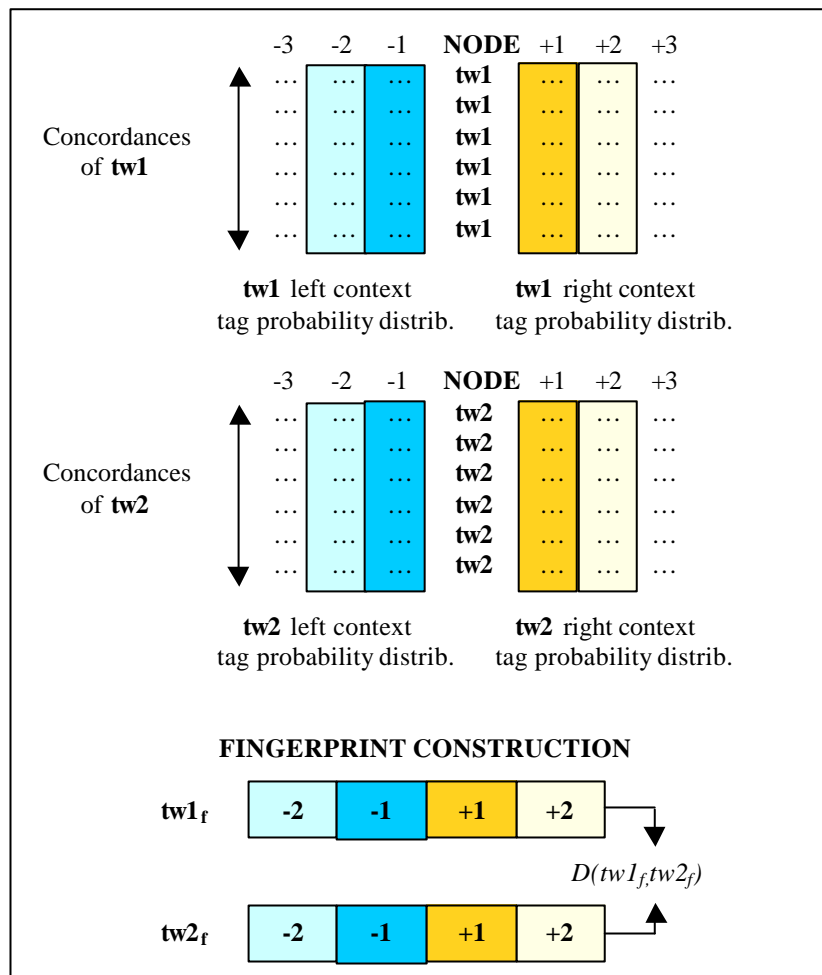


Figure 2. Word distributional fingerprint construction.

$$D(tw1_f, tw2_f) = \sum_d \left(\sum_{x=1}^{NTAGS} tw1_d(x) \ln \left(\frac{tw1_d(x)}{tw2_d(x)} \right) + \sum_{x=1}^{NTAGS} tw2_d(x) \ln \left(\frac{tw2_d(x)}{tw1_d(x)} \right) \right),$$

where d is the distance from the node $(-2, -1, +1, +2)$ and x spans over all the tags. The distance is a measure of the similarity of the two side contexts, here globally represented by the distributional fingerprint; it is zero if they are equal, and greater than zero if they are different. A great distance between two fingerprints means that they are very different and the target words from which they derive cannot be considered as distributionally similar.

Using the distance measurement, and comparing each pair of target words it is possible to build up a tree using hierarchical clustering techniques: the procedure is to link the two words (or clusters) that are most similar forming a new cluster and reiterate the operation until all the target words are clustered. The resulting tree is the basis for linguistic analysis. The words exhibiting similar behaviour from a distributional point of view should lie in the same area of the tree. Figure 3 shows the results of our computations based on the corpus described.

5. Observations on dendrogram results

The dendrogram shows a wide range of items, which may lead back to traditional categories: conjunctions, conjunctive locutions, adverbs, adverbial locutions and pronouns. Such a wide range of grammatical words was chosen in order to provide a larger means of comparison. Many grammatical words were shown with an exclusive adverbial value in order to obtain clearer results in the connective placing. Those items having no connective function work as a contrast medium.

The dendrogram is divided into two sections. The first from *tale che* to *sempre*, contains items which are part of different traditional categories (pronouns, adverbs, prepositions). The main property of this section is the presence of several items with an adverbial meaning. The presence of item *ieri*, in its adverbial meaning, is probably implied here. Besides the proximity of *ancora* and *sempre* marks out the adverbial value of *ancora*, that may also have a function of connection.

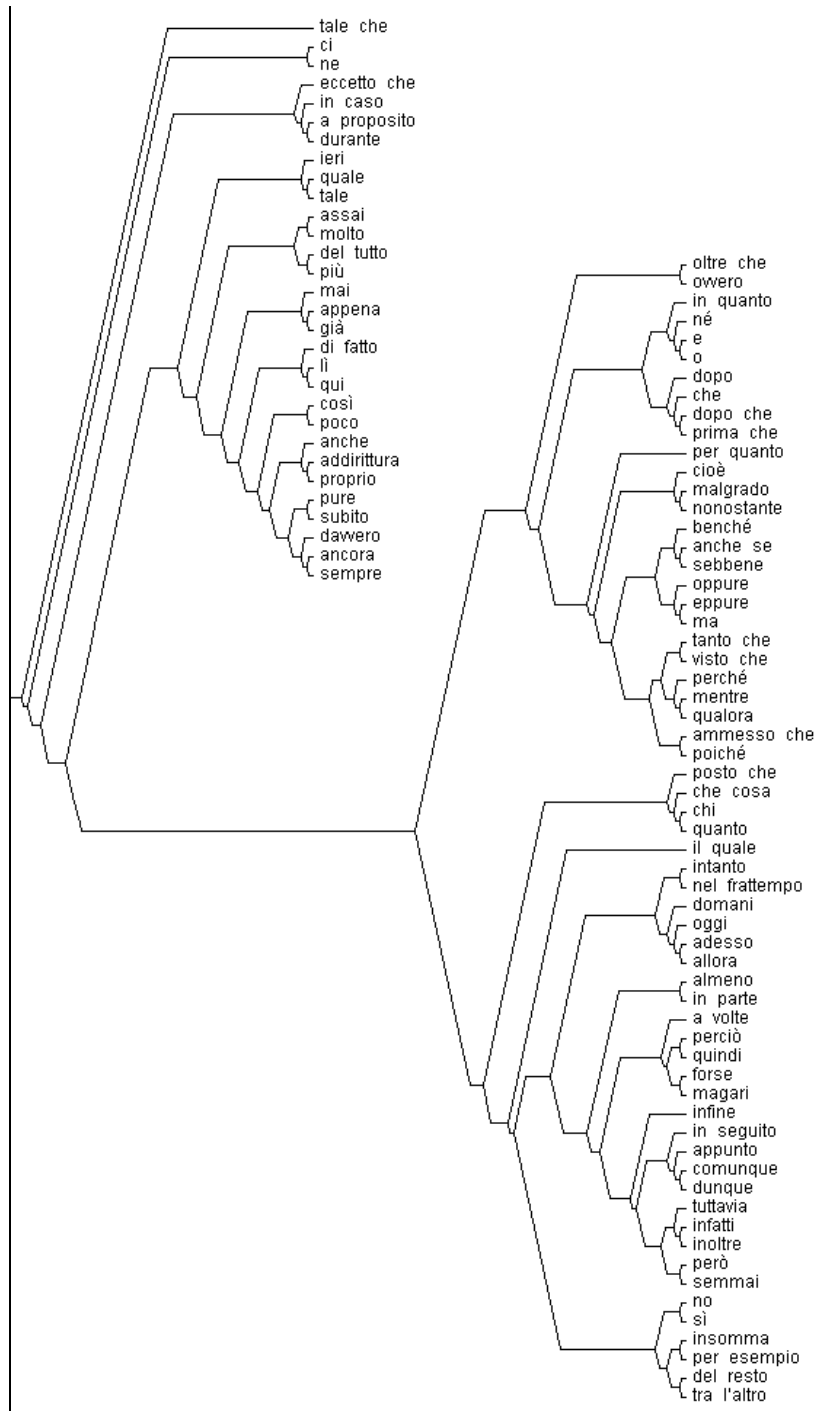


Figure 3. The connectives grouping among other grammatical words.

Based on this datum, the dendrogram places items with a clear adverbial value. This value may be that of an intensifier and, less frequently, a modifier (*ieri*). Intensifiers have a heightening or lowering effect on the meaning of other elements, in the sentence. This datum probably derives from the narrow context of evaluation. The typical position of an intensifier is $-1/+1$, according to the position of the term to be modified.

The second section of the dendrogram, from *oltre che* to *tra l'altro*, shows the items under discussion: the connectives. A scale within the class of connectives may be supposed. It can be argued that the section from *oltre che* to *poiché* shows the connectives which exhibit a strong connective function, while the section from *posto che* to *tra l'altro* exhibits a weaker one.

As stated above, some natural connectives may be considered to be the lexical realisation of truth-functional logic operators. This applies to *né, e, o*. This is, however, a theoretical assumption, which is supported by empirical evidence. The fact that the dendrogram shows the natural connectives, which are nearer to the truth-functions, close to each other, is significant; in a certain sense, confirming the presence of a hard connective nucleus – from a logical point of view. The proximity is a consequence of the considerable syntactic similarities between *né, e, o*.

In this respect, the lexical realisations of the logical operators are numerous; natural connectives such as *ma, eppure* lead back to the truth-functions of the logical connective \wedge . In spite of this, the conceptual affinities between logical operators and lexical realisations cannot be completely shown by the distribution of the lexical units in the dendrogram; this confirms the difference between the syntactic properties of the operators and the distributional ones of the natural connectives.

The dendrogram shows certain semantic similarities. For example, items such as *malgrado/nonostante, benché/anche se/sebbene, ma/eppure, dopo che/prima che*. The branch from *dopo* to *prima che* is interesting from a morphological point of view. The different functions of connectives *dopo* and *che* can be observed; the locution *dopo che* derives from them, merging the value of both the connectives and showing similar distributional properties.

The section from *posto che* to *tra l'altro* is formed by numerous and heterogeneous items. Those items may carry out a weak connective function. As shown in the case of *né, e, o* we can state the existence of a class of connectives having a neat linguistic meaning.

On the other hand, in the case of *posto che*, *nel frattempo*, *per esempio* the items are ingredients of a function of connection and cohesion but not connectives in the proper sense. They have no linguistic meaning such as the former: their meaning derives from cooperation. They give rise to connection interacting with other items in the text. The dendrogram shows how the connection can be a function of lexical units, having no truth-functional meaning, or standard connective value (such as the value expressed by the conjunction). In cases like this, the connection derives from the identity of reference between two lexical units. Some words performs a phrasal or inter-phrasal connective function, by depending on an antecedent (point of link): a lexical unit (simple or complex) which fixes the value of the following connective unit. Their value is established by coreference.

This is the case of items having 'phoric' properties; they may be defined 'substitute words'; they may have their linking point either in a phrase or outside of it, at an extra-phrase level. The dendrogram shows a sequence of *substitute words* deriving from the same node: *che cosa*, *chi*, *quanto*. Some substitute words may have their linking point in an extra-phrasal context, as in the case of deictics. Their function typically refers to the linking points outside the sentence. The dendrogram shows the syntactic-semantic proximity of the deictics *domani*, *oggi*, *adesso*, *allora*.

A function of substitution may be also attributed to the couple *sì/no*, when interpreted as specialised; in this case the linking point is formed by a whole verbal phrase.

This account confirms our starting point. At the beginning of this paper we argued that traditional part-of-speech distinctions are inadequate to represent the syntactic properties of some grammatical words. We focused on a wide class of items and tried to revalue their distributional properties, in consideration of quantitative analysis. We obtained results which were quite stable and significant for our purposes: we can demonstrate, based on distributional data, that there is a class of grammatical words which function as connectives. As we argued, this is not a solid class, in which every item has the unique function of connective, but a shaded class within which we may recognise a hard nucleus that leads back to logical operators and a peripheral one, in which connective function works by merging several values and variables. The dendrogram shows some distributional results that make it possible to speak of a semantic-syntactic phenomenon we named 'function of connection'. We

consider this data as a starting point for the definition of some categories useful for creating a part-of-speech tag set.

6. References

- Brill, Eric / Magerman, David / Marcus, Mitchell / Santorini, Beatrice 1990. Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the DARPA Speech and Natural Language Workshop*, San Mateo, CA: Morgan-Kaufmann, 275-282.
- Brill, Eric / Marcus, Mitchell 1992. Tagging an unfamiliar text with minimal human supervision. In *Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language*, Cambridge, MA: American Association for Artificial Intelligence, 10-16.
- Brill, Eric 1993. *A corpus-based approach to language learning*. PhD thesis, Department of Computer and Information Science, Philadelphia, PA: University of Pennsylvania.
- Brown, Peter F. / Della Pietra, Vincent J. / de Souza, Peter V. / Lai, Jenifer C. / Mercer, Robert L. 1992. Class based n-gram models of natural language. *Computational Linguistics*, 18/4, 467-479.
- Bustorf, Wolfgang 1974. Riflessione sui cosiddetti "riempitivi" in italiano. In Mario Medici and Antonella Sangregorio (eds.), *Fenomeni morfologici e sintattici nell'italiano contemporaneo*, Roma: Bulzoni, 21-25.
- Chierchia, Gennaro 1997. *Semantica*, Bologna: Il Mulino.
- Chierchia, Gennaro / McConnell-Ginet, Sally 1993. *Significato e grammatica*, Padova: Murzio.
- Dardano, Maurizio / Trifone, Pietro 1997. *La nuova grammatica della lingua italiana*, Bologna: Zanichelli.
- Devoto, Giacomo / Oli, Giancarlo 2001. *Il dizionario della lingua italiana*, Firenze: Le Monnier.
- De Mauro, Tullio 1999. *Grande dizionario dell'uso*, Torino: UTET.
- Dijk van, Teun A. 1977. *Testo e contesto*, Bologna: Il Mulino.
- Graffi, Giorgio 1994. *Le strutture del linguaggio. Sintassi*, Bologna: Il Mulino.
- Halliday, Michael A.K. 1985. *Spoken and Written Language*, Deakin University: Victoria.

- Halliday, Michael A.K. 1991. Corpus studies and probabilistic grammar. In Karin Aijmer and Bengt Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London: Longman, 30-43.
- Harris, Zellig 1951. *Structural Linguistics*, Chicago: University of Chicago Press.
- Kaplan, David 1970. What is Russell's Theory of Descriptions? In Wolfgang Yourgrau and Allen D. Breck (eds.), *Physics, Logic and History*, New York: Plenum Press, 277-288.
- Lee, Lillian 1999. Measures of distributional similarity. In *Proceedings of the 37th ACL*, College Park, MD, 25-32.
- Lyons, John 1977. *Semantics*, Cambridge: Cambridge University Press.
- Martin, Sven / Liermann, Jörg / Ney, Hermann 1998. Algorithms for bigram and trigram word clustering. *Speech Communication*, 24, 19-37.
- Monachini, Monica 1996. ELM-IT: EAGLES Specification for Italian Morphosyntax Lexicon Specification and Classification Guidelines. *EAGLES Document EAG CLWG ELM IT/F*.
- Montague, Richard 1974a. English as a Formal Language. In Richmond H. Thomason (ed.) *Formal Philosophy. Selected Papers of Richard Montague*, New Haven-London: Yale University Press, 188-221.
- Montague, Richard 1974b. *Formal Philosophy. Selected Papers of Richard Montague*, in Richmond H. Thomason (ed.), New Haven-London: Yale University Press.
- Montague, Richard 1974c. The Proper Treatment of Quantification in Ordinary English. In Richmond H. Thomason (ed.) *Formal Philosophy. Selected Papers of Richard Montague*, New Haven-London: Yale University Press, 221-242.
- Pereira, Fernando / Tishby, Tali / Lee, Lillian 1993. Distributional clustering of English words. In *Proceedings of the 31st ACL*, Columbus, Ohio, 183-190 .
- Renzi, Lorenzo / Salvi, Gianpaolo / Cardinaletti, Anna (eds.) 1988, 1991, 1995. *Grande grammatica italiana di consultazione*. Bologna: Il Mulino.
- Rossini Favretti, Rema 2000. Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS. In Rema Rossini Favretti (ed.), *Linguistica e informatica. Multimedialità, corpora, percorsi di apprendimento*, Roma: Bulzoni, 39-56.
- Rossini Favretti, Rema / Tamburini, Fabio / De Santis, Cristiana CORIS/CODIS: A corpus of written Italian based on a defined

- and a dynamic model. In Andrew Wilson, Paul Rayson and Tony McEnery (eds.) *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Munich: Lincom-Europa. *In press*.
- Sabatini, Francesco / Coletti, Vittorio 1997. *Disc: dizionario italiano Sabatini Coletti*. Firenze: Giunti.
- Sensini, Marcello 1997. *Grammatica italiana*, Milano: Mondadori.
- Serianni, Luca 1989. *Grammatica italiana. Italiano comune e lingua letteraria*, Torino: UTET.
- Shannon, Claude / Weaver, Warren 1949. *The Mathematical Theory of Communication*, Urbana: University of Illinois Press.
- Strawson, Peter F. 1952. *Introduction to Logical Theory*, London: Methuen.
- Tamburini, Fabio 2000. *Annotazione grammaticale e lemmatizzazione di corpora in italiano*. In Rema Rossini Favretti (ed.) *Linguistica e informatica. Multimedialità, corpora, percorsi di apprendimento*, Roma: Bulzoni, 57-74.
- Zingarelli, Nicola 2000. *Lo Zingarelli 2001: vocabolario della lingua italiana*, Bologna: Zanichelli.