Refined Tagging of Complex Verbal Phrases for the Italian Language*

Simone Faro[1] and Arianna Pavone[2]

[1] Dipartimento di Matematica e Informatica, Università di Catania,
Viale A. Doria 6, I-95125 Catania, Italy
[2] Dipartimento di Scienze Umanistiche, Università di Catania
Piazza Dante 32, I-95124 Catania, Italy
faro@dmi.unict.it, pavone@unict.it

Abstract. A verb phrase is a syntactic unit consisting of one verbal form, combined with any other elements, representing the verbal part of the speech. In Italian, as in many other languages, the verb phrase is the central element in a sentence. In this paper, we investigate the problem of the automatic recognition of complex verb phrases in the Italian language, where the wide variety of syntactic units and the complexity of morphology make the problem more difficult to solve than in English. In particular we propose an hybrid approach which faces the recognition and the disambiguation of Italian verb phrases by using language generation. We provide also a web tool[1] for testing and querying our method. The level of accuracy and the grade of detail reached by our solution is higher than any other known approach.

1 Introduction

The recognition of terms and phrases which compose a text is one of the main problems concerning with the automatic information extraction from natural language texts. This process is also at the base of a large area of applications, such as semantic analysis of natural language texts, automatic paraphrase, knowledge bases construction, automatic spelling and part of speech tagging. The process of recognition, analysis and paraphrase of the components of a natural language text is certainly more complex than the reverse process used in the automatic generation of the language itself. The complexity of such recognition process is due to the possible presence of a large number of variants, concerning the syntax and the grammar, that must be taken into account in the parsing process of the text. In addition it is also necessary to determine the appropriate syntactic and semantic features to be applied to it. Differently these details are prearranged in the process of automatic generation of natural language text.

However, there is considerable commercial interest in natural language recognition, mainly due to its numerous applications in various fields such as information extraction, categorization of texts, storage and analysis of large-scale content. The recognition of parts of speech (PoS) also finds application as a component in tools for grammatical spell-correction of texts. Such tools are currently unable to recognize the correctness of complex verb phrases as *se ne era accorta*⋆[2], but are limited to the

⋆ This work has been supported by project PRISMA PON04a2 A/F funded by the Italian Ministry of University and Research within the PON 2007-2013 framework.

[1] The web tool is provided to the reviewers in order to establish the effectiveness of our solution. It can be accessed at http://www.dmi.unict.it/~faro/tagger/voci_verbali.php .

[2] Along the paper we will present several examples of Italian verb phrases, together with the corresponding English translation, where necessary. Each example is also an anchor (identified by the symbol ⋆) which links to the recognition tool activated for the corresponding verb phrase.

Simone Faro, Arianna Pavone: Refined Tagging of Complex Verbal Phrases for the Italian Language, pp. 132–145.
Proceedings of PSC 2015, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-05787-2

correction of the terms composing the phrase. These instruments are not therefore able to recognize some types of grammatical errors, such as those which we can find in the sentence *l'aveva stato*⋆, in which the error is in the choice of the auxiliary relative to the main verb. Solutions to these problems would find application in many tools such as word processors, e-mail clients, electronic dictionaries, and search engines.

In this paper we address the problem of recognition and disambiguation of Italian phrases, with particular reference to the recognition of verb phrases. This work is part of a more complex project named A.R.I.ANN.A. (Automatic Refined Italian ANNotation Approach) whose aim is to produce a refined syntactic and logical tagger for the Italian language.

The paper is organized as follows. In Section 2 we will give a general view of the state of the art and we will briefly describe related works known in literature. Then we will introduce the new tool for the recognition of complex verb phrases for the Italian language. In particular we will discuss separately the features of the tool and the recognized verbal forms (Section 3) and the recognition scheme (Section 4). We will draw our conclusions in Section 5.

2 Related Work

The analysis of the parts of speech (PoS Tagging problem), with reference to the English language, is considered a simple problem today. The experimental results obtained by Tsuruoka and Tsujii [2] show that the PoS tagging solutions available for the English language can reach an accuracy up to 97%. Such solutions are generally based on machine learning techniques such as dependency networks [3], perceptrons [4], support vector machines (SVM, also known as support vector networks) [5] or hidden Markov models (HMM) [6]. Such problem consists in analyzing a natural language text and in associating each part of the speech to a tag, selected from a predetermined set of tags. Such tag set could be more or less refined.

The reference tag set used in PoS Tagging for the English language is the Penn Treebank tag set [7], which divides the parts of speech in 36 categories. The same problem, with reference to the Italian language, has been dealt with in the context of EVALITA (Evaluation of NLP and Speech Tools for Italian[3]) an initiative aimed at the evaluation of the tools for the analysis of natural language, with reference to the Italian language.

In the course of the competition proposal in 2007 [8] the set of tags consisted in 32 lexical categories proposed by Treebank tag-set of the University of Pennsylvania, adapted to the Italian language, whose 6 categories were devoted to description of verb phrases. In that case the best solution achieved an accuracy of 98%.

In the course of the competition revived in 2009 [9] a set of lexical classes was used, widened to 37 elements with different morphological variants allowing the identification of 336 different elements. The set of reference tags is TANL (Text Analytics and Natural Language), made in accordance with the EAGLES guideline [10], a standard for English language, recognized by the community in natural language processing. The TANL tag set includes three levels of accuracy of tag, of which the highest level consists of 14 categories. those relating to the verb phrases are listed in Table 2, in

[3] Information related to different competitions proposals under EVALITA can be found at the web page www.evalita.it

Tag	Description	Examples (Italian)
VB	verb, lemma	*leggere, conoscere, andare*
VBD	verb, past	*leggevo, conobbi, andasti*
VBG	verb, gerund or present participle	*leggendo, conoscente, andando*
VBN	verb, past participle	*letto, conosciuta, andati*
VBP	verb, present, non-third singular person	*leggevamo, conosco, vai*
VBZ	verb, present, third singular person	*legge, conosce, va*

Table 1. The Treebank tag-set relative to verb phases.

Tag	Description	Examples
V	verb	*leggere, conosco, andato*
VA	auxiliary verb	*sono, eravamo, hanno*
VM	modal verb	*volevo, posso, dobbiamo*
Suffix	**Description**	**Examples**
-m	masculine	*letto, conosciuti, andato*
–f	feminine	*lette, conosciuta, andata*
–n	not specified	*leggo, conoscere, vanno*
–s	singular	*letto, conosci, va*
–p	plural	*lette, conoscevano, vanno*
–n	not specified	*leggere, conoscere, andare*
–1	first person	*leggevo, conosco, andammo*
–2	second person	*leggi, conoscevi, andrete*
–3	third person	*legge, conobbe, vanno*
–i	indicative	*leggo, conosceva, andavamo*
–m	imperative	*leggi, conosca, andate*
–c	subjective	*legga, conoscano, andassimo*
–d	conditional	*leggerei, conoscerebbe, andresti*
–g	gerund	*leggendo, conoscendo, andando*
–f	infinitive	*leggere, conoscere, andare*
–p	participle	*letto, conosciuta, andato*
–p	present	*leggo, conosco, vai*
–i	present perfect	*leggevo, conoscevi,*
–s	past	*lessi, conoscesti, andarono*
–f	future	*leggerà, conoscerete, andranno*
–c	clitics	*leggendocele, conoscilo*

Table 2. The TANL tag-set relative to verb phrases.

which are also shown suffixes that can be integrated to the main tag in order to describe form, tense, mood, number, person and also the possible presence of clitics.

Most of these solutions are able to recognize the parts of speech by associating the terms in the text with the entries in some lexical Knowledge Base (KB), as WordNet [12], Multi-WordNet [13], Euro-WordNet [14], BabelNet [15] or similar ones. Apart from WordNet, which contains only lemmas for the English language, other lexical

resources also contain lemmas of the Italian language, as well as those of many other languages. These lemmas include nouns, verbs, adjectives, adverbs etc. Each lemma or phrasal term in a KB, is associated to its sense, usually identified with one of the synsets related to the given term.

One of the most difficult challenges in the recognition of phrases in a natural language text is that these phrases are often composed of several terms. In WordNet 3.0, for instance, over 40% of the items are compound phrases, while the Italian version of MultiWordNet 1.5 the number of such phrases is 15%. The compound phrases are difficult to be accurately recognized for three main reasons:

a) In the first place, the terms which compose a compound phrase are themselves voices of the KB. For example, the verb phrase *essere caduto** (*to have fallen*, past infinitive) is composed by two separate verb phrases, *essere** (*to be*, present infinitive) and *caduto** (*fallen*, past participle). This is typically the output produced by the PoS tagging solutions described above, which ignore the issues of compound phrases splitting the entire term in the constituent subterms.
b) Secondly, the terms composing a compound phrase may not appear contiguously in the text. For example the verb phrase *essere improvvisamente caduto** (*to have suddenly fallen*) contains the verb phrase *essere caduto** (*to have fallen*) which is separated by the modal adverb *improvvisamente* (*suddenly*).
c) Finally, the conjugation of the terms contained in a compound verbal phrase may lead to a difficult recognition. For instance the verb phrase *esserle caduta addosso** (*to have fallen on top of her*) contains the verb phrase *esserle caduta** (*to have fallen on her*, past infinitive, clitic form, singular) with difficult recognition because of its pronominal form.

None of the above described problems are solved by state of the art PoS tagging solutions for the Italian language. Only very recently Del Corro *et al.* [1] addressed some of the above problems introducing a tool that allows to make jointly the recognition of the phrase and its disambiguation in Italian. The solution we describe in this paper solves a) and c) and could be adapted to solve b) as well.

Online there are many facilities for the generation of verb phrases, but limited only to the conjugation of verbs. Among the services for the Italian language the most used are Italian-Verbs[4], Coniugazione.it[5], it.bab.la[6], WordReference.com[7] e Virgilio.it[8]. Most of these services offer the possibility to conjugate verbs, not only in their active form, but also in the passive and reflective, if available.

3 A New Tool for the Recognition of Italian Verb Phrases

In linguistics a syntagma is a unit of varying syntactic complexity and autonomy, which is between the word and sentence. The verb phrase is a syntagma consisting of a verbal form together with any other elements, but it is still the verbal part of the speech. In this section we describe the features of our tool and the different verb forms which it is able to detected. We also focus our attention on some problems related to Italian syntax which make the recognition a difficult task.

[4] http://www.italian-verbs.com
[5] http://www.coniugazione.it
[6] http://it.bab.la/coniugazione/italiano/
[7] http://www.wordreference.com/conj/Itverbs.aspx
[8] http://parole.virgilio.it/parole/verbi_italiani/

In Italian, as in other languages, the verb phrase is the variable part of the speech and indicates an action, a state or a becoming in relation to a subject, expressed or implied, that does or undergoes an action. Some examples of verb phrases recognized by our tool are:

mangio⋆	(*I eat*)
sono andato⋆	(*I went*)
mi fu concesso⋆	(*I was allowed*)
le è stato mandato⋆	(*it was sent to her*)
mi pettino⋆	(*I comb my hair*)

The head of the verb phrase is the verb, the more complex part of speech under the grammatical aspect, which may vary according to different categories of reference. In Table 3 we show the tag set used in our solution. It reflects the level of detail of the recognition process. It includes 3 head tags and and 30 feature tags, beginning with a symbol ":", which can be added to any head tag in order to increase the level of detail. The tag set allows the identification of more than 10 000 different verb forms.

Regarding their value, verbs can be transitive (tag TR) or intransitive (tag IN); regarding their form or diathesis (describing the relationship between actor and action) a verb can be active (tag VSA), passive (tag VSP), reflexive (tag VPR). In addition, regarding the subject which they refers to, we can have verbs of first (tag 1), second (tag 2) or third person (tag 3). By number they can be singular (tag S) or plural (tag P). Regarding the performance of the action, verbal forms can vary according to a range of tenses and moods, as described below.

Our system is based on a manually annotated KB containing a set of more than 5.700 verb lemmas, including 151 intransitive reflexive forms. Verbal lemmas are also categorized according to their values. In particular about 180 verbal lemmas are recognized as intransitive verbs. The number of verbs accepting the auxiliary *avere*⋆ (*to have*) is about 3.650, while about 500 verbal entries accept the auxiliary *essere*⋆ (*to be*). There is also a third class of items that accepts both verbal auxiliaries, consisting of about 310 items in our KB.

3.1 Recognition of tenses and verbal moods

The verbal moods express attitudes that the speaker establishes against the interlocutor. In the Italian language, we distinguish the following moods: *indicative*, *subjunctive*, *conditional*, *imperative*, *gerund* and *participle*. Each of these verbal moods consist of some simple and some compound times. The latter ones are formed by combining the auxiliary verbs, *essere*⋆ or *avere*⋆, with the past participle of the verb itself.

The *indicative* (IND tag) shows the reality of a fact, which can be true or false. This verbal mode is very used in main clauses, i.e. independent grammatical clauses. For instance *mangio una mela*⋆ (*I eat an apple*) is an objective remark. For the conjugation of verbs, the indicative has four simple tenses (present, imperfect, far past and future) and four compound tenses (present perfect, past perfect, distant past perfect and future perfect).

The *subjunctive* (CNG tag) indicates a situation for which it is not possible to propose a real judgment of truth, because it concerns a desire, a possibility or a supposition. It consists of two simple tenses and two compound tenses.

The *conditional* (CND tag) indicates the presence of a real or unreal conditioning of the reality of facts, of an action or process. The conditional consists in a single simple tense (present) and a single compound tense (past).

The *imperative* tense (IMP tag) indicates an exhortation and a command. It has a single tense, the present, and only two forms: the second person (singular) and second person (plural). For other person, the imperative, borrows forms of the subjunctive, and in this case becomes exhortative subjunctive.

The *gerund* (tag GER) is a verbal mode which has just two tenses: simple and compound, present and past. It is used in subordinate clauses and establishes a relationship of contemporaneity to the action expressed by the verb in the main clause.

The *participle* (PAR tag) has two simple tenses: *present* and *past* participle. The participle is a mood participating in both the category names (from which it draws the conjugation, distinguishing between voice and aspect).

The *infinitive* (INF tag), denoted by the lemma of the verb, has a simple tense, the present, and a composed tense, the past.

All the grammatical forms of the verb relating to mood, tense and person, number and diathesis, constitute the conjugation. The Italian has three conjugations, distinguished by the infinite endings: *-are*, *-ere*, *-ire*: each conjugation has its paradigm, consisting in a series of endings and suffixes, by which, starting from the theme of the verb, the verbs are formed depending on different moods and tenses;

3.2 Recognition of pronominal verbal forms

In Italian there are particular verbal forms with particles, called *clitics*. These clitics attach themselves to a word and they form a single unit. For instance, *leggerla** (legger-la, *to read it*), *leggerne** (legger-ne, *to read some of them*) and *leggerci** (legger-ci, *to read to us*). Some of these verbs incorporate two clitics together, in these cases they are bi-pronominal verbs. Some examples are *leggersela** (legger-se-la, *to read it to himself*), *leggersene** (legger-se-ne, *to read some of them to himself*) and *leggerceli** (legger-ce-li, *to read them to ourselves*).

The pronominal verbs are divided into classes, distinguished by the clitic and the meaning. Specifically, we distinguish the following forms:

Verb forms including an direct object.
They are built with the particles *-mi -ti -lo -la -li -le -ci* and *-vi*, where the particle assumes the function of direct object (with the meaning, respectively, of *me*, *you*, *him her*, *us*, *you* and *them*). When the subject and the object are the same, the verbs in these forms indicate that the action expressed by the verb is reflexive, and it is related with the subject itself that performs the action. It is important to distinguish the different cases, for which we can not speak of reflexive constructions, as in the cases listed above. If the particles *-lo -la -li -le* are prefixed to the verb beginning with a vowel, the elision of the vowel is common: thus *l'amo** is equivalent to *la amo** (*I love her*).

Other examples are:

1. *lo porti** (*you bring it*)
2. *portarmi** (*to bring me*)
3. *se l'avessi portata** (*if you had brought it*)

Forms	**Description**	**Examples**
VSA	standard active	*capisco*⋆
VSP	standard passive	*sono capito*⋆
VPA	pronominal active	*avendolo capito*⋆
VPP	pronominal passive	*avendomi capito*⋆
VPR	pronominal reflexive	*essendomi capito*⋆
Values	**Description**	**Examples**
:TR	transitive	*capissi*⋆
:IN	intransitive	*andassi*⋆
Tenses	**Description**	**Examples**
:IND	indicative	*avevo capito*⋆
:CNG	subjective	*avessi capito*⋆
:CND	conditional	*avrei capito*⋆
:IMP	imperative	*capisci*⋆
:GER	gerund	*avendo capito*⋆
:PAR	participle	*capente*⋆
:INF	infinitive	*capire*⋆
Moods	**Description**	**Examples**
:PRE	present	*capisco*⋆
:PAS	past	*capivo*⋆
:FUT	future	*capirò*⋆
:IMP	present perfect	*avevo capito*⋆
:PRM	past perfect	*ebbi capito*⋆
:TRA	distant past perfect	*avessi capito*⋆
:FAN	future perfect	*avrò capito*⋆
Gender	**Description**	**Examples**
:M	male	*è stato capito*⋆
:F	female	*è stata capita*⋆
:N	neuter	*abbiamo capito*⋆
Number	**Description**	**Examples**
:S	singular	*capisci*⋆
:P	plural	*capiamo*⋆
:I	invariable	*capire*⋆
Person	**Description**	**Examples**
:P0	impersonal	*aver capito*⋆
:P1	first person	*abbiamo capito*⋆
:P2	second person	*avete capito*⋆
:P3	third person	*hanno capito*⋆
Clitic	**Description**	**Examples**
:COC	object complement	*avermi portato*⋆
:CTC	term complement	*avergli portata*⋆
:CPC	place complement	*averci portati*⋆
:CPF	partitive complement	*averne portate*⋆

Table 3. The new tag-set introduced in this paper. On top the list of head tags is listed. any of all other tags in the list, beginning with a symbol ":", can be added to the head tag in order to increas the level of detail. The set allows the identification of more than 10 000 different verb forms.

Verb forms including an indirect object.
Some pronominal forms use the particles *-mi* and its conjugations in gender and number, *-ti -gli -le -ci -vi*. In this case the pronominal particle is used as an indirect object (with the meaning of *to me*, *to you*, *to him*, *to her*, etc). This form is used with both transitive and intransitive verbs.

Other examples are:

1. *gli porti** (*you bring to him*)
2. *portarmi** (*to bring to me*)
3. *le avessi portata** (*you had brought to her*)

Verb forms including an adverb of place.
They are built by using the pronominal particle *-ci* or *-ne*, which have the function of adverb of place. The particle *-ci* is used with the meaning of *in that/this place* while the particle *-ci* is used with the meaning of *from that/this place*. In this context, the verb phrase *andarci** (*to go there**) can be paraphrased as *andare in quel luogo** (*to go in that place**). Other examples are:

1. *arrivarci** (*to reach that place*)
2. *ne vengo ora** (*I came now from there*)
3. *lui ci viene** (*he came here*)

Verb forms including a partitive complement.
The particle *-ne** can be used also with the meaning *of that/this/them* with a partitive function. It can be applied to transitive and to intransitive verbs as well. Example of these verb phrases are:

1. *parlarne** (*to speak about that*)
2. *ne avevamo spesi** (*we spent some of them*)
3. *ne porterò due** (*I will bring two of them*)

Bi-pronominal verb forms.
Many of the particles used for the composition of pronominal verb forms listed above can, in general, composed to create bi-pronominal verb forms. The particles obtained in this way, can be listed in the following forms:

1. adverbof place + direct object *ci + (lo/la/li/le)*
2. direct object + adverbof place *(me/te/se/ve) + ci*
3. adverb of place + partitive complement *ci + ne*
4. indirect object + partitive complement *(me/te/se/ce/ve) + ne*
5. indirect object + direct object *(me/te/se/ce/ve) + (lo/la/li/le)*

Example of these verb phrases are, respectively:

1. *portarcelo** (*to bring it in that place*)
2. *portarmici** (*to bring me in that place*)
3. *portarcene** (*to bring there some of them*)
4. *portarmene** (*to bring some of them to me*)
5. *portarvelo** (*to bring it to you*)

3.3 Recognition of irregular verbal forms

In the Italian language there are many verbs that do not follow the whole regular paradigm related with their conjugation and for this reason they are called irregular; these verbs are used quite common, for instance, the verbs *essere** (*to be*), *avere** (*to have*), *andare** (*to go*), *fare** (*to do*), etc. In relation to the conjugation, we distinguish defective verbs, i.e. without some forms, such as *vertere** and overabundant verbs, which follow different conjugations in all or in some tenses, such as *starnutare** – *starnutire** (*to sneeze*), *adempiere** – *adempire** (*to fulfill*). In our system verbal entries are divided according to three main conjugations. Irregular forms are listed based on the class of regularity.

For instance, the verbs *abbassare** (*to decrease*) and *appellare** (*to appeal*) belong to the same class because they have the same irregularities in their conjugation. In particular belong to the first conjugation more than 4 200 entries, of which about 1 200 are irregular forms, divided into 6 classes of irregularities; belong to the second conjugation approximately 500 entries, of which 440 are irregular forms, divided into 19 classes; the voices belonging to the third conjugation are about 520, of which about 480 irregular forms, divided into 12 classes of irregularities.

3.4 Ambiguity of the recognition of compound tenses

The compound tenses consist in (at least) two terms: an auxiliary verb, *essere** (*to be*) or *avere** (*to have*), conjugates in a simple tense, and a main verb conjugated in the past participle. In this context the past participle can be composed depending on the number or on the gender. The correct recognition (and the consequent tagging) of this verbal form creates some problems since in the Italian language the compound verbs can be composed in different ways.

In particular the question of the correspondence of the past participle is one of the most difficult chapter of Italian syntax. The main errors that usually arise in the correspondence, and that we addressed in our solution, can be summarized as follows: **a)** if the verb is transitive and accepts the auxiliary *avere**, then it is possible to accord the participle of the verb, both in masculine or feminine, and also with the object complement, even if the first form is more used than the second. Thus the sentence *I chose the best solutions* can be translated as:

a1. *ho scelto le migliori soluzioni**
a2. *ho scelte le migliori soluzioni**

b) If the verb is transitive and accepts the auxiliary *avere** (*to have*) and the compound verb is preceded by a personal or a relative then it is possible to accord the participle of the verb with the prefixed object. Thus the sentence *He has cheated us* can be translated as:

b1. *ci ha ingannato**
b2. *ci ha ingannati**

c) If the verb accepts the auxiliary *essere** (*to be*) then it is possible to accord the participle of the verb with the subject or with the predicate complement. Thus the sentence *it was a news* can be translated as:

c1. *lo è stato una novità**
c2. *lo è stata una novità**

d) When the verb phrase is in pronominal transitive form, the participle of the verb can be accorded with the subject or with the object complement, even if it is prefixed to the verb. Thus, the sentence *since we set ourselves that goal* can be translated as:

d1. *essendocelo prefissati*⋆
d2. *essendocelo prefissato*⋆

The possibilities of choice among the points reported above have always existed in the Italian language and the restrictions indicated by some grammarian are considered to be unfounded. Our system recognizes as accurate all previous combinations, providing the correct interpretation of the correspondence of participle.

3.5 Actives, passives and reflexives forms

In the Italian language, and also in other languages, verbal entries can take various forms: *active*, *passive*, *reflexive* and *pronominal*.

A verb is in active form, when the subject performs the action:

lei guarda⋆ (*she looks*)

A verb is in passive form when it is the subject who undergoes the action:

lei è guardata⋆ (*she is looked*)

The passive form is characterized by the auxiliary *essere*⋆ (*to be*) followed by the past participle of the verb. Only transitive verbs can take the passive form. Reflexive verbs are accompanied by a reflexive pronoun (*mi*⋆, *ti*⋆, *si*⋆, *ci*⋆, *vi*⋆) which comply with the subject. The presence of the reflexive pronoun, which can be prefixed or postponed to the verb, makes a phrase in pronominal form.

Examples are:

mi guardo⋆ (*I look at me*)
guardatevi⋆ (*look at yourselves*)

There are different types of reflexive verbs. The wider class of reflexive verbs is obtained by entries that admit both transitive reflexive forms, and active (*io lavo*, *io mi lavo*). There are also reflexive verbs that are used with reciprocal value, that allow a reading for which an event, which has at least two promoters subjects, is realized when the effects produced by the first fall on the second, and the effects produced by the second fall on the first, as in:

amarsi⋆ (*to love each other*)
sposarsi⋆ (*to marry*)
spingersi⋆ (*to push each other*)

4 The recognition process

In this section we briefly describe the recognition process on which our solution is based. As we noticed above, in general, the process of recognition of the components of a natural language text is more complex than the reverse process of language generation. This is particularly true in the case of Italian verb phrases, where the grammar is hardly structured and allows phrases as complex compound sequences of terms.

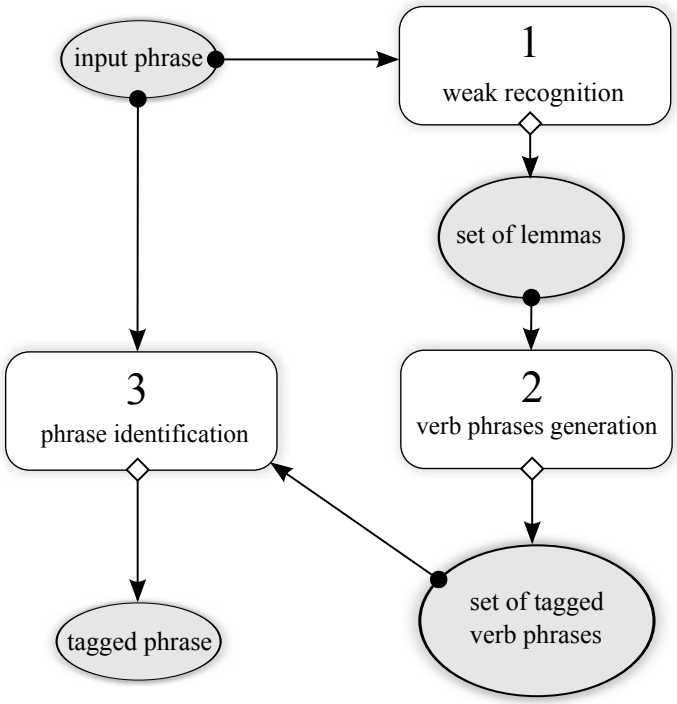

Figure 1. The scheme which describes the recognition process. Data are represented by grey circles, while recognition steps are represented by white rectangles. Input data starts with a black circle, while output data come from white square.

The tool takes as input a text, which consists in a sequence of terms. It identifies all the verb phrases contained in such a text, assuming that each term of the phrases can participate only of a single verb phrase.

In order to simplify the correct and detailed recognition of verb phrases we mix a ruled based weak recognition approach with a robust finite model approach for language generation. The process is then finalized by using a fast string matching subroutine. Specifically the recognition process is divided in three steps.

- *weak recognition:* the input text is tokenized and each term is associated with a (possibly empty) set of verb lemmas;
- verb phrases generation: each lemma is processed and a set of possible (already tagged) matching verb phrases is generated;
- final identification: the input phrases are matched against the candidate verb phrases in order to perform the correct association.

Figure 1 shows the scheme which describes the recognition process. Data are represented by grey circles, while recognition steps are represented by white rectangles. Input data starts with a black circle, while output data come from white square.

In what follows we briefly describe each step of the process of Figure 1. We suppose that the input phrase t is a sequence of n terms $t = \langle x_1 x_2 \cdots x_n \rangle$, with $n \geq 1$.

I. Weak recognition step.
During the first step the input phrase is tokenized and each term x_i is analyzed, for $i = 1, \ldots, n$. In particular the algorithm uses a weak recognition process in order to estabilish whether a term x_i is a verb or not. At this level the tool is unable to identify the correct features of the term, like tense, mood, person and gender. Such a process allows only the identification of the lemma (or the list of lemmas) associated with the term, if the latter is a verb phrase. For instance if x_i is the term*andassimo* the tool will associate it with the lemma *andare*.

Specifically, each term x_i is decomposed in two substrings p_i (a prefix) and s_i (a suffix) such that $x_i = p_i \cdot s_i$. Any possible decomposition of the type $x_i = p_i \cdot s_i$ is taken into account, with $|p_i| > 0$ and $|s_i| > 0$. If we find a prefix p_i which is equal to the radix of a verb v in our KB then we investigate if the corresponding suffix s_i could be a desinence of v. In such a case the verb v is returned as a lemma of x_i.

Observe that in some cases two or more lemmas can be associated to a single term. For instance the term *stato* can be associated to both lemmas *essere* (*to be*) and *stare* (*to stay*). If the input phrase is *ce lo avevano portato** (*they had brought it to us*) then the tool recognizes the following set of lemmas:

1. *ce*	$\emptyset$
2. *lo*	$\emptyset$
3. *avevano*	$\{avere^\star\}$
4. *portato*	$\{portare^\star\}$

while the set of lemmas relative to the input phrase *le era stato dato** (*it has been given to her*) are:

1. *le*	$\emptyset$
2. *era*	$\{essere^\star\}$
3. *stato*	$\{essere^\star, stare^\star\}$
4. *dato*	$\{dare^\star\}$

II. Verb phrases generation step.
During the second step of the recognition process the algorithm generates all possible verb phrases which are connected to the lemmas which have been identified at the previous step. Specifically, let x_i a term of the input text t, and let $\{\ell_1, \ell_2, \ldots, \ell_m\}$ the set of lemmas associated to x_i. The algorithm generates all possible verb phrases which are licensed by lemma ℓ_j, for $j = 1, \ldots, m$, by using a finite state model based on conjugation details stored in our KB.

As stated above, the set of verb phrases generated from a single lemma ℓ_i could contain more than 10 000 elements, even if, in most practical cases it is not larger than 9 000 elements, including active, passive, pronominal, simple and compound verb forms.

In addition, during the generation process, each produced verb phrase is associated with high precision to the correct tag. This can be done since form features are stored in the KB together with the conjugation details.

For example, some of the tagged verb phrases generated from the lemma *portare** (*to bring*) are

$portare^\star \rightarrow \{$ *porto**, (VSA:TR:IND:PRE:N:S:P1)
porti* (VSA:TR:IND:PRE:N:S:P2)
porta* (VSA:TR:IND:PRE:N:S:P3)
...
avessi portati* (VSA:TR:CNG:TRA:N:S:P2)
avesse portati* (VSA:TR:CNG:TRA:N:S:P3)
...
eravate state portate* (VSP:TR:IND:IMP:F:P:P2)
erano state portate* (VSP:TR:IND:IMP:F:P:P3)
...
ce lo avessi portato* (VSA:TR:CNG:TRA:N:S:P2:COC:CTC)
ce lo avesse portato* (VSA:TR:CNG:TRA:N:S:P3:COC:CTC)
$\ldots\}$

III. Final identification step.
During the final step of the process the algorithm identifies any possible verb phrase in the input text t by using information generated at the previous step. Let x_i be a term in t and let ℓ_j a lemma associated to x_i during the first step. Moreover let V_j be the set of all possible verb phrases which are licensed by lemma ℓ_j, generated at the previous step. Notice that each verb phrase $v \in V_j$ is a sequence of terms $v = \langle y_1 y_2 \cdots y_k \rangle$, with $k \geq 1$.

In order to identify all verb phrases the algorithm checks whenever each sequence $v \in V$ is equal to any subsequence of length k in t which involves the term x_i. More formally the sequence p is compared with the subsequence $\langle x_h x_{h+1} \cdots x_{h+k} \rangle$, for $h = \max(1, i-k) \cdots \min(n, i+k)$.

Since each term can be involved in a single verb phrase, if two overlapping subsequences of t are recognized as verb phrases, then only the longest one is taken into account. For instance, in the sentence *ce lo avevano portato** ($t = \langle x_1 \cdots x_5 \rangle$) the tool will recognize the following verb phrases:

verb phrase	lemma	tag	position
1. *ce lo avevano**	*avere**	VSA:TR:IND:PAS:N:P:P3:CPC:COC	$\langle x_2 \ldots x_4 \rangle$
2. *lo avevano**	*avere**	VSA:TR:IND:PAS:N:P:P3:COC	$\langle x_3 \ldots x_4 \rangle$
3. *ce lo avevano portato**	*portare**	VSA:TR:IND:IMP:N:P:P3:CTC:COC	$\langle x_2 \ldots x_5 \rangle$
4. *ce lo avevano portato**	*portare**	VSA:TR:IND:IMP:N:P:P3:CPC:COC	$\langle x_2 \ldots x_5 \rangle$
5. *avevano**	*avere**	VSA:TR:IND:PAS:N:P:P3	$\langle x_3 \ldots x_3 \rangle$
6. *avevano portato**	*portare**	VSA:TR:IND:IMP:N:P:P3	$\langle x_4 \ldots x_5 \rangle$
7. *portato**	*portare**	VSA:TR:PAR:PAS:M:S:P1	$\langle x_5 \ldots x_5 \rangle$
8. *portato**	*portare**	VSA:TR:PAR:PAS:M:S:P2	$\langle x_5 \ldots x_5 \rangle$
9. *portato**	*portare**	VSA:TR:PAR:PAS:M:S:P3	$\langle x_5 \ldots x_5 \rangle$

However only the verb phrases n.3 and n.4 are returned as output since they overlaps all other choices.

5 Conclusions and Future Works

In this paper we presented a web tool for the recognition of complex verb phrases in the Italian language. Our solution is able to recognize a refined set of verbal forms, including passive, reflexive and pronominal forms. In addition the new proposed tool

is able to recognize compound verbs and to associate such forms to their right features, which include mood, tense, person, number, gender and type of the clitics, if present.

Future work will be devoted to increase the number of details recognized by our solution, allowing the identification of features related to the direct object or indirect object referred by the verb phrase, as number and gender. Moreover, the tool framework is general enough to be adapted to other languages like English, French or Spanish. In addition the recognition process could be also integrated in a more general solution to the PoS tagging problem for the Italian language.

References

1. L. Del Corro, R. Gemulla, and G. Weikum: *Werdy: Recognition and Disambiguation of Verbs and Verb Phrases with Syntactic and Semantic Pruning*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, 2014, pp. 374–385.
2. Y. Tsuruoka and J. Tsujii: *Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data*, in Proceedings of HLT-EMNLP, 2005, pp. 467–474.
3. K. Toutanova, D. Klein, C. Manning, and Y. Singer: *Feature-rich part-of-speech tagging with a cyclic dependency network*, in Proceedings of HLT-NAACL, 2003, pp. 505–512.
4. M. Collins: *Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms*, in Proceedings of EMNLP, 2002, pp. 1–8.
5. J. Gimenez and L. Marquez: *Fast and accurate part-of-speech tagging: the SVM approach revisited*, in Proceedings of RANLP, 2003, pp. 158–165.
6. T. Brants: *TnT: a statistical part-of-speech tagger*, in Proceedings of the 6th Applied NLP Conference, 2000, pp. 224–231.
7. M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz: *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational Linguistics, vol. 19, issue 2, 1993, pp. 313–330.
8. F. Tamburini: *EVALITA 2007: The Part-of-speech Tagging Task*. IA-Intelligenza Artificiale, Anno IV, issue 2, 2007, pp. 4–7.
9. G. Attardi and M. Simi: *EVALITA 2009: The Part-of-speech Tagging Task*, 2009.
10. M. Monachini: *ELM-IT: An Italian Incarnation of the EAGLES-TS. Definition of Lexicon Specification and Classification Guidelines*. Technical report, Pisa, 1995.
11. S. Montemagni et al.: *Building the Italian Syntactic-Semantic Treebank*, in Abeillé, ed., Building and using Parsed Corpora, Language and Speech series. Kluwer, Dordrecht, 2003, pp. 189–210.
12. G. A. Miller: *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, No. 11, 1995, pp. 39–41.
13. E. Pianta, L. Bentivogli, and C. Girardi: *MultiWordNet: developing an aligned multilingual database*, in Proceedings of the First International Conference on Global WordNet, 2002, pp. 21–25.
14. P. Vossen: *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.
15. R. Navigli, S. P. Ponzetto: *BabelNet: Building a Very Large Multilingual Semantic Network*, in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 216–225.