10.12871/CLICIT2014175

The Italian Module for NooJ

Simonetta Vietri
Department of Political, Social and
Communication Sciences
University of Salerno, Italy
vietri@unisa.it

Abstract

English. This paper presents the Italian module for NooJ. First, we will show the basic linguistic resources: dictionaries, inflectional and derivational grammars, syntactic grammars. Secondly, we will show some results of the application of such linguistic resources: the annotation of date/time patterns, the processing of idioms, the extraction and the annotation of transfer predicates.

Italiano. *In questo articolo si presenta il modulo italiano per NooJ. In un primo momento si descrivono le risorse lessicali di base: i dizionari, le grammatiche flessive, derivazionali e sintattiche. Si presentano poi i risultati relativi all'applicazione di tali risorse: l'annotazione dei pattern temporali, il parsing delle frasi idiomatiche, l'estrazione e l'annotazione dei predicati di trasferimento.*

1 Introduction

NooJ is a development environment used to construct large-coverage formalized descriptions of natural languages, and apply them to corpora, in real time. NooJ, whose author is Max Silberztein (Silberztein 2003-), is a knowledge-based system that makes use of huge linguistic resources.

Dictionaries, combined with morpho-syntactic grammars, are the basic linguistic resources without which it would be impossible to perform a text analysis. The system includes various modules for more than twenty languages, among them Italian (.nooj4nlp.net). Most of the Italian linguistic resources are completely new.

The goal of the NooJ project is twofold: to provide tools allowing linguists to implement exhaustive descriptions of languages, and to design a system which processes texts in natural language (see Silberztein 2014).

NooJ consists of higher and higher linguistics levels: tokenization, morphological analysis, disambiguation, named entity recognition, syntactic parsing[1].

Unlike other systems, for example TreeTagger, developed by Helmut Schmidt (1995)[2], NooJ is not a tagger, but the user can freely build disambiguation grammars and apply them to texts.

Section 2 describes the Italian dictionary and the inflectional/derivational grammars associated with it. Section 3 shows the extraction of date/time patterns, section 4 the parsing of idioms. Section 5 describes the XML annotation and extraction of transfer predicates.

2 The dictionaries and the inflectional grammars

The first Machine Italian dictionary was built at the Institute for Computational Linguistics, C.N.R, directed by Antonio Zampolli (see Bortolini et al (1971), Gruppo di Pisa (1979)). More than a decade later a group of researchers of the Linguistics Institute at the University of Salerno, directed by Annibale Elia, started to implement an electronic Italian dictionary on the principles of the Lexicon-Grammar framework (Gross 1968, 1979, Elia et al 1981)[3].

More recently Baroni and Zanchetta (2005) developed *Morph-it!*, that contains more than 505,000 entries and about 35,000 lemmas [4].

[1] See textpro.fbk.eu/docs.html for **TextPro**, an NLP system implemented at FBK. It is a suite of modules performing various tasks. **Unitex** is a system developed by Sébastien Paumier, see igm.univ-mlv.fr/~unitex/index.php?page=1.

[2] See cis.uni-muenchen.de/~schmid/tools/TreeTagger/ and elearning.unistrapg.it/TreeTaggerWeb/TreeTagger.html. See also the Venice Italian Treebank (**VIT**), the Turin University Treebank (**TUT**), the Italian Syntactic Semantic Treebank (**ISST**).

[3] For the literature on Lexicon-Grammar, see infolingu.univ-mlv.fr/english/. A very first version of the Italian dictionary was built for Intex. See De Bueriis and Montemagno (1995).

[4] See dev.sslmit.unibo.it/linguistics/morph-it.php. As concerns the corpus utilized, see Baroni et al. (2004).

The Italian dictionary of simple words (S_dic) for NooJ contains 129,000+ lemmas, whereas the dictionary of compounds includes 127,000+ nouns and 2,900+ adverbs Elia (1995). Furthermore, the Italian module consists of a number of satellite dictionaries including toponyms (1,000+), first and last names (2,000+), acronyms (200+). Some dictionaries are richer than others which are still under construction. The canonical forms of dictionary entries, either simple or compound, are of the following type:

```
americano,A+FLX=N88
il,DET+FLX=D301
su,PREP
surfista,N+FLX=N70
tavola a vela,N+FLX=C41
tavola,N+FLX=N41
volare,V+FLX=V3
```

Each entry is associated to an alphanumeric code that refers to an inflectional grammar, as the following example[5]:

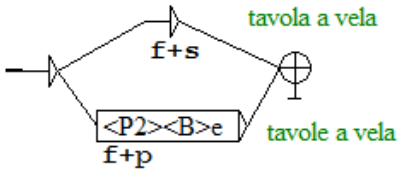

Fig. 1. Sample of an Inflectional Grammar

On the basis of the entries and the inflectional codes, NooJ generates the dictionaries of inflected forms (more than one million of simple forms, and 260.000+ of noun compounds) in a few seconds. By applying these resources, NooJ will annotate a sentence such as *Le surfiste volavano sulle tavole a vela* as follows:

```
le,il,DET+f+p
le,PRON
surfiste,surfista,N+f+p+Um
volavano,volare,V+IM+3+p+i+a+e
su,PREP
tavole,tavola,N+f+p
tavole a vela,tavola a vela,N+f+p
a,PREP
vela,N+f+s
vela,velare,V+PR+3+s+t
```

Each form is associated with morpho-syntactic information. Since NooJ is not a tagger, the annotations show the ambiguities (unless the user applies disambiguation grammars)[6]. For example, *vela* may be not only a feminine (+f) singular (+s) noun (N), but also the Present Indicative (+PR) form of the transitive (+t) verb *volare*, in the 3rd person (+3), singular (+s).

[5] For the FSA/FST grammars, see Silberztein (2003-).
[6] For reason of space, some annotations are not shown.

2.1 Proper Names and derivation

The dictionary of proper names is built according to the same criteria used for the main dictionary. Although proper names do not inflect, they are linked to derived forms. Such forms like *renzismo*, *antirenziano*, *renzista* are relatively new and are not included in the S_dic. The dictionary of proper names and a derivational grammar associated with it allow NooJ to annotate these very productive forms, as in the following:

```
renzismo,Matteo Renzi,N+Npr...
antirenziano,Matteo Renzi,A+Npr
```

2.2 The Annotation of Pronominal forms

Italian is particularly rich of agglutinated forms such as *vederti*, *mandandogliela*, *dimmi*, *comprata*, etc. which are constituted of a verb (infinitive, gerund, imperative, past participle) and one or more clitics. Although these forms are formally single words, they are analyzed by means of a morpho-syntactic grammar which separates the verb form from the pronoun. Therefore, the forms above will be annotated as follows:

```
vedere,V+t+a+INF
ti,PRON+Persona=2+s
mandando,mandare,V+G
gli,PRON+Persona=3+m+s
la,PRON+Persona=3+f+s
dì,dire,V+IMP+2+s+t+a
mi,PRON+Persona=1+s
comprata,comprare,V+PP+f+s
la,PRON+Persona=3+f+s
```

3 The extraction of date/time patterns

Among the syntactic resources, the Italian module includes a grammar for the extraction and annotation of date and time sequences. It's a complex net of local grammars which, applied to a text of 1MB (129,000+ word forms), extracts and annotates sequences like the following:

Nell'arco di tre mesi/<DURATA>
fino al giugno 2006/<DURATA>
intorno alle 23,20/<DATA>
Dal 1987 al 2004/<DURATA>
Per la fine di gennaio/<DATA>
Nel novembre del 2001/<DATA>
Il 18 e 19 dicembre scorsi/<DATA>
un mese dopo/<DATA>
in due giorni/<DURATA>
dieci anni fa/<DATA>
il prossimo 9 gennaio/<DATA>
dal 18 al 21 gennaio prossimi/<DURATA>
per 30 mesi/<DURATA>
nell'ottobre del 2004/<DATA>
fino al dicembre 2005/<DURATA>

4 The Annotation of Idioms

The formal representation and processing of idioms has always been a very debated issue (Abeillé 1995, Sag et al 2001, Fothergill et al 2012). In the NooJ dictionaries, Italian idioms (Vietri 2014a, 2014c) are represented as strings formed by a verb that requires one or more fixed elements as in the following (simplified) example:

```
alzare,V+C1+FLX=V3+DET=<il,DET+m+s>
+N=<gomito,N+m+s>
```

The verb *alzare* is associated with the determiner *il* and the fixed noun *gomito*. The idiom *alzare il gomito* ('lift one's elbow') belongs to class **C1** (+C1), the verb inflects (+FLX) according to the code **V3**, and the **DET**erminer has to be masculine singular (+m+s) because the noun *gomito* is obligatory masculine singular. NooJ is an "open" system, and the user can choose to assign a property like +Passive only to those idioms that accept this construction. In such a case, the property ±Passive can be recalled in the grammar which is associated with the dictionary of idioms.

The dictionary is associated with a grammar, since the fixed lexical elements have to be linked to each other. Figure 2 shows a simplified example of grammar where the variable (indicated by the rounded parentheses) containing the verb is directly linked to the determiner (**V$DET**) and to the noun (**V$N**). This formalism keeps the fixed elements linked together also in case of modifiers or adverbs insertion, or in case of discontinuous idioms such as *prendere qc. per la gola*.

The dictionary/grammar pair, whose formalism is explained in details in Silberztein (2012), allows NooJ to automatically annotate sequences like *alzare il gomito*. Since this construction is ambiguous, NooJ produces both the idiomatic annotation, signaled by the little curve, and the literal one, as shown in Figure 3.

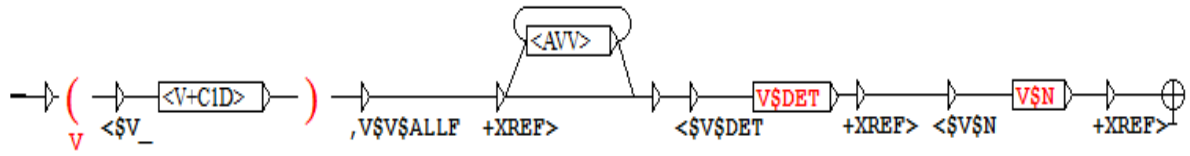

Fig. 2. The 'Active' Grammar

Maria alzò il gomito

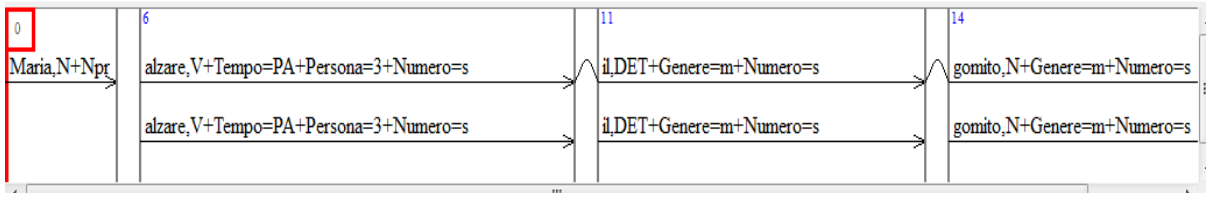

Fig. 3. Text Annotation

4.1 Parsing Idioms

Once NooJ has annotated idioms, it is possible to syntactically parse the sentence in question by applying an appropriate syntactic grammar. However, a sentence such as *Maria alzò il gomito* is ambiguous, therefore it has to be assigned a double representation. The representations in Figures 4 and 5 are flat trees which can be (re)designed according to the user's choice. Figure 4 represents the idiomatic construction: the blue boxes indicate that the lexical entries are linked.

The tree in Figure 5 represents instead the non-idiomatic construction, where the lexical entries are not linked.

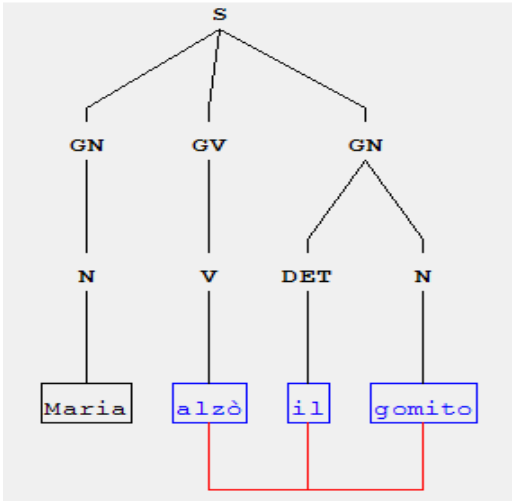

Fig. 4. Idiomatic Representation

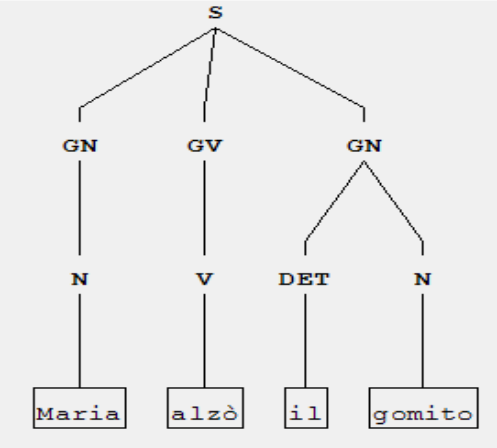

Fig. 5. Non-idiomatic Representation

Furthermore, the user can freely decide to assign only the idiomatic representation by means of the property +UNAMB.

5 Annotation of Transfer Predicates

The annotation of the Predicate-Argument structure of Transfer Predicates is described in details in Vietri (2014b). In the following examples, the transfer predicate is *consegnare* (= to deliver) in (1), *effettuare la consegna* (= make delivery) in (2), and *consegna* (= delivery) in (3):

(1) *Il fornitore consegna la merce al cliente*
The supplier delivers the goods to the customer

(2) *Il fornitore effettua la consegna della merce al cliente*
The supplier makes delivery of the goods to the customer

(3) *La consegna della merce al cliente dal fornitore*
The delivery of the goods to the customer by the supplier

They are all transfer predicates with three arguments: the Giver (*il fornitore* = the supplier), the Receiver (*il cliente* = the customer), and the Object (*la merce* = the goods) that is transferred from the Giver to the Receiver. Therefore, the Predicate-Argument structure is a function of the type **T (Giver, Object, Receiver)**. NooJ can build a concordance and annotate sequences such as (1)-(3), according to their Transfer Predicate-Argument Structure. This can be done by applying to a text/corpus a complex grammar that contains more than 70 sub-graphs. The annotated text can be exported as an XML document. Here is the XML text referring to the examples (1)-(3):

```
<G> Il fornitore </G> <T> consegna </T>
<O> la merce </O> al <R> cliente </R>,
ma prima di <T> effettuare la consegna
<\T> della <O> merce <\O> ...
<T> La consegna </T> della <O> merce
</O> al <R> cliente </R>.
```

The Transfer Grammar applied to the Italian Civil and Commercial Codes produce more than 2,600 occurrences. The most frequent Predicate-Argument structure is formed of the Transfer predicate **T** and the Object **O** (1,200 occurrences), immediately followed by the passive constructions where the Object **O** precedes the predicate **T** (387 occurrences).[7]

6 Conclusion

The application of the Italian module to a corpus of 100MB (La Stampa 1998) produced the following results: 33,866.028 tokens, 26,785.331 word forms. The unknown tokens are loan words, typos, acronyms, alterates[8].

The Italian module consists of exhaustive dictionaries/grammars formally coded and manually built on those distributional and morpho-syntactic principles as defined within the Lexicon-Grammar framework. Such a lingware (a) constitutes an invaluable linguistic resource because of the linguistic precision and complexity of dictionaries/grammars, (b) can be exploited by the symbolic as well as the hybrid approach to Natural Language Processing. The linguistic approach to NLP still constitutes a valid alternative to the statistical method that requires the (not always reliable) annotation of large corpora. If the annotated data contain errors, those systems based on them will produce inaccurate results. Moreover, corpora are never exhaustive descriptions of any language.

On the other hand, formalized dictionaries/grammars can be enriched, corrected and maintained very easily. Silberztein (2014) contains a detailed discussion on the limits, errors and naïveté of the statistical approach to NLP. The Italian module for NooJ constituted the basis of several research projects such as Elia et al. (2013), Monti et al. (2013), di Buono et al. (2014), Maisto et al. (2014). Therefore, it has been tested, verified and validated. The results constitute the basis for the updating of the module itself. Ultimately, the lexical resources of the Italian module can be easily exported into any format usable by other systems.

[7] In a different perspective, the *Lexit* project, directed by Alessandro Lenci, explores the distributional/semantic profiles of Italian nouns, verbs, and adjectives.
[8] The grammar that annotates alterates is under construction.

References

Anne Abeillé. 1995. The Flexibility of French Idioms: a Representation with Lexicalized Tree Adjoining Grammar. In M. Everaert, E-J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: structural and psychological perspectives*. Hillsdale, NJ: Lawrence Erlbaum Associates: 15-41.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the "la Repubblica" corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian, in *Proceedings of the Fourth Language Resources and Evaluation Conference*, (Lisbon: ELDA): 1771-1774.

Ugo Bortolini, Carlo Tagliavini, and Antonio Zampolli. 1971. *Lessico di Frequenza della Lingua Italiana*. Milano: Garzanti.

Giustino De Bueriis and Mario Monteleone. 1995. *Dizionario elettronico DELAS_I - DELAF_I ver. 1.0*, Dipartimento di Scienze della Comunicazione dell'Università degli Studi di Salerno.

Maria Pia di Buono, Mario Monteleone, and Annibale Elia. 2014. How to populate ontology. Computational linguistics applied to the Cultural Heritage Domain. In E. Métais, M. Roche, and M. Teisseire (Eds.): *NLDB 2014 - 19th International Conference on Application of Natural Language to Information Systems*, 18-20 June 2014 - Montpellier, France: 55-58.

Annibale Elia, Daniela Guglielmo, Alessandro Maisto, and Serena Pelosi. 2013. A Linguistic-Based Method for Automatically Extracting Spatial Relations from Large Non-Structured Data. In *Algorithms and Architectures for Parallel Processing*. Springer International Publishing: 193-200.

Annibale Elia, Maurizio Martinelli, and Emilio D'Agostino. 1981. *Lessico e strutture sintattiche. Introduzione alla sintassi del verbo italiano*, Napoli: Liguori.

Annibale Elia. 1995. Chiaro e tondo, in *Tra sintassi e semantica. Descrizioni e metodi di elaborazione automatica della lingua d'uso*, E. D'Agostino (ed.), ESI: Salerno.

Richard Fothergill and Timothy Baldwin. 2012. Combining resources for MWE-token classification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*: 100-104.

Maurice Gross. 1968. *Syntaxe du verbe*. Paris: Larousse.

Maurice Gross. 1975. *Mèthodes en syntaxe*. Paris: Hermann.

Gruppo di Pisa. 1979. Il dizionario di macchina dell'italiano. In Daniele Gambarara, Franco Lo Piparo, Giulianella Ruggiero (eds), *Linguaggi e formalizzazioni*, Atti del Convegno internazionale di studi, Catania, 17-19 settembre 1976. Bulzoni, Roma: 683-707.

Alessandro Maisto and Serena Pelosi. 2014. A Lexicon-Based Approach to Sentiment Analysis. The Italian Module for Nooj. *Proceedings of the International Nooj 2014 Conference*, University of Sassari, Italy (forthcoming).

Johanna Monti, Mario Monteleone, Maria Pia di Buono, and Federica Marano. 2013. Natural Language Processing and Big Data. An Ontology-Based Approach for Cross-Lingual Information Retrieval**.** *Proceedings of the Social Computing (SocialCom)* - 2013 ASE/IEEE International Conference: 725-731.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword Expressions: A Pain in the Neck for NLP. *Computational Linguistics and Intelligent Text Processing*. Berlin Heidelberg: Springer: 1-15.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland: 172-176.

Max Silberztein. 2003. *NooJ Manual*. Available for download at: www.nooj4nlp.net .

Max Silberztein. 2012. Variable Unification in NooJ v3. In K. Vučković, B. Bekavac, & M. Silberztein (Eds.), *Automatic Processing of Various Levels of Linguistic Phenomena*. Newcastle upon Tyne: Cambridge Scholars Publishing: 1-13.

Max Silberztein. 2014. *Formaliser les langues: l'approche de NooJ*. London: ISTE eds.(forthcoming).

Simonetta Vietri. 2014a. The Lexicon-Grammar of Italian Idioms. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, Coling 2014, Dublin: 137-146.

Simonetta Vietri. 2014b. The Construction of an Annotated Corpus for the Analysis of Italian Transfer Predicates, *Lingvisticae Investigationes*, 37-1, Amsterdam & Philadelphia*:* John Benjamins: 69-105.

Simonetta Vietri. 2014c. *Idiomatic Constructions in Italian. A Lexicon-Grammar Approch.* Linguisticae Investigationes Supplementa, 31. Amsterdam & Philadelphia*:* John Benjamins (forthcoming).

Eros Zanchetta and Marco Baroni. 2006. Morph-it! A free corpus-based morphological resource for the Italian language. *Proceedings of Corpus Linguistics 2005*, online at corpus.bham.ac.uk/PCLC/.