Dealing with Italian Adjectives in Noun Phrase: a study oriented to Natural Language Generation

Giorgia Conte
Dipartimento Studi Umanistici
Università di Torino
giorgiaconte.gc@gmail.com

Cristina Bosco
Dipartimento di Informatica
Università di Torino
boscodi.unito.it

Alessandro Mazzei
Dipartimento di Informatica
Università di Torino
mazzei@di.unito.it

Abstract

English. This paper describes a theoretical and empirical investigation about the position of adjectives in the Italian language. The long term goal which oriented the study is the formalization of this information into a natural language generation system. Providing that adjectives mainly occur within noun phrases, we focused on them and we collected data from corpora representing very different text genres, i.e. social media and standard ones, in order to compare the theoretical predictions with the real use of the adjective in Italian. The results obtained by confirm the previsions of the modern linguistic theories but also show the different behaviour of adjectives in the distinct analysed genres.

Italiano. *Questo lavoro presenta un'analisi teorica ed empirica sulla posizione degli aggettivi nella lingua Italiana. L'orientamento del lavoro è dato dalla necessità di formalizzare questa informazione nell'ambito di un sistema di generazione automatica della linguaggio. Poiché gli aggettivi si presentano principalmente nei sintagmi nominali, ci si è concentrati su questi, raccogliendo dati da corpora che rappresentano generi di testo diversi, ovvero social media e standard, al fine di confrontare le previsioni teoriche con l'uso reale dell'aggettivo in Italiano. I risultati ottenuti confermano le previsioni delle moderne teorie linguistiche ma mostrano anche il diverso comportamento degli aggettivi nei diversi generi analizzati.*

1 Introduction

Corpus linguistics is a methodological approach based on the extraction from a set of texts of data useful for the study of language. Even if in principle any collection of texts can be called *corpus*, the term assumes a more precise connotation in the context of modern linguistics, where a corpus is featured by sampling, representativeness, finite size, machine-readable form and a standard reference (McEnery and Wilson, 2001).

In this work we have applied a corpus-based approach and we considered two different corpora which represent two different text genres: one concerning social media language (PoSTWITA corpus) and one concerning balanced standard Italian (UD-it corpus). Indeed, while social media texts have recently gained great attention from the NLP community since they have many peculiar properties, standard texts can give a more accurate view on the status of some linguistic notions in "traditional" written text.

These above mentioned corpora allowed us an in depth investigation about the position of the adjective in the nominal phrase. Indeed, even if this grammatical category is described in several traditional Italian grammars (Renzi et al., 2001; Serianni, 2006; Patota, 2006), its theoretical status is not currently enough formalized to be used within the computational context. A more useful perspective on the behaviour of the adjective is proposed in a recent theoretical study which is focussed on the position of the adjective in Romance languages (Giusti, 2016).

This work aims at achieving two major goals. The first is to empirically confirm with the analysis of corpora the theoretical predictions given in (Giusti, 2016). The second goal is instead to provide a representation and classification of Italian adjective category that can be spent within the SimpleNLG-IT (Mazzei et al., 2016), a surface re-

alizer for Italian language.
The paper is organized as follows: in Section 2 we review the linguistic literature concerning the position of the adjective within the Italian noun phrase. In Section 3, we explain the details of our corpus linguistic investigation. In Section 4, we describe the use of the empirical data in the SimpleNLG-IT realizer. Finally, the Section 5 closes the paper with conclusions and some pointers to future work.

2 The Theoretical Status of the Adjective in the Nominal Phrase

We take into account the adjective in its primary use (Bhat, 1994), that is as modifier of a noun. In Italian, within the nominal phrase, the adjective can be positioned before or after the noun to which it refers. In accordance with the traditional grammar, e.g. (Serianni, 2006), these alternative positions are described as unmarked, when the adjective follows the noun, and marked, when it precedes the noun.

These different behaviour of the adjective also carry different semantic values: nominal phrases where the adjective precedes the noun indicate more subjectivity or more stylistic refinement if compared to the more neutral and objective expressions where the adjective follows the noun, as in the following examples (extracted from (Serianni, 2006)): *gli occhi neri* (the eyes black) and *gli alberi alti* (the trees high) vs. *i neri occhi* (the black eyes) and *gli alti alberi* (the high trees)[1]. In the left side of the versus, the adjectives *neri* (black) and *alti* (high) objectively qualify the nouns that they follow, and the information they carry is indeed verifiable by a true/false criterion; in the other side instead the same adjectives qualify the nouns but they also emphasize a desire for stylistic elaboration by those who write or speak (Serianni, 2006).

Moreover, a descriptive function is usually attributed in literature to pre-nominal adjectives, while a restrictive function is attributed to post-nominal ones, e.g. in (Serianni, 2006). This can be clearly exemplified by the difference between the following sentences: *le vecchie tubature hanno ceduto* (the old pipes has collapsed) and *le tubature vecchie hanno ceduto* (the pipes old has collapsed). In the first sentence, the pre-nominal adjective *vecchie* (old) has a descriptive function: it describes a quality of the related noun, i.e. *tubature* (pipes). Instead in the second sentence, the same adjective, in post-nominal position, has restrictive function with respect to the meaning of the related noun: it adds to the noun a distinctive qualification which identifies it as the only one with a certain quality (the *old* pipes, not the *new* ones) (Serianni, 2006). However the value of the adjective in the post-nominal position, being unmarked, may be ambiguous between these two functions, whereas an adjective in pre-nominal position can only have appositive (that is descriptive) function (Giusti, 2010).

2.1 A hierarchy of the Descriptive Adjectives

In (Giusti, 2010) a further distinction among the descriptive adjectives in sub-categories is provided. It is based on a cross-linguistically defined hierarchy where the rank that the adjective assumes is strictly related to the position that it can assume with respect to the noun. The categories are the following:

- evaluative, e.g. *bello* (beautiful)
- dimension, e.g. *alto* (high)
- age, e.g. *vecchio* (old)
- physical property, e.g. *duro* (hard)
- colour, e.g. *rosso* (red)
- relational, e.g. *nazionale* (national)

The adjectives collocated in the lower part of the hierarchy are more prone to assume post-nominal positions, where those in the higher part more frequently assume the pre-nominal ones. For instance, the relational adjectives, that are at the lower level of the hierarchy, are predominantly post-nominal. The others can be freely positioned before or after the noun, but those occupying lower positions within the hierarchy have a stronger tendency for post-nominal positions, while those in higher part of the hierarchy are more freely placed before or after noun (Giusti, 2016). For more details about the classification of the adjectives and how we applied it to those we extracted from corpora, see the following section.

3 Extracting Adjectives from Corpora

In order to validate the assumptions made in literature, and described in section 2 about the behaviour of the adjective, we selected corpora where Italian is annotated for what concerns morphology and syntax and representing also differ-

[1] The English glosses for the examples are literal and can not correspond to the correct English expressions.

ent text genres. We applied scripts in Python and SQL queries for detecting the presence of adjectives and noun phrases in both the reference corpora, but their classification is manually done, for carefully dealing with cases where ambiguity occurs.

We found a substantial help for finding a decision-making criterion for the classification of adjectives in the examples proposed in the Treccani online vocabulary. For instance, we tagged as evaluative the adjective *pericoloso* (dangerous), which is derived from the noun *pericolo* (danger), according to the vocabulary example *un viaggio pericoloso* (a dangerous journey). We tagged instead as relational the adjective *solare* (solar), like in the example *luce solare* (solar light), considering that the adjective is derived from the noun *sole* (sun), indicating an entity rather than a quality.

A particular attention must be paid to homonymous adjectives, like e.g. *reale* that may mean 'royal' or 'real'. In this case, two different entries in the vocabulary must be introduced, one for each meaning of the adjective: the first tagged as relational, for the meaning derived from the noun *re* (king), and the second tagged as evaluative, for indicating the meaning 'actually existing'.

In the rest of this section the resources we used in our investigation are described also showing the differences that make them especially interesting for validating our results in two different contexts and text domains.

The data sets we used are respectively extracted from two different corpora: PoSTWITA (Bosco et al., 2016) and UD-it[2], both tagged in accordance with the Universal Dependencies annotation scheme [3]. While the PoSTWITA corpus is only morphologically tagged and it is taken from the social network Twitter, the other resource is a treebank which includes other variety of more standard texts.

3.1 PoSTWITA

PoSTWITA characterised by short texts (140 characters maximum) and a typical social media Italian jargon that is featured by a frequent use of creative expressions and incorrect words like in the following example:

ho un disparato bisogno di soffocati di coccole. <3 ti amo piccola mia. ([I] have a desperate need to suffocate you with pampering. <3 [I] love you my baby.)

where two incorrect words occur: *disparato* instead of *disperato* and *soffocati* instead of *soffocarti*.

Also distinctive graphic practices due to the particular medium are symbols are very frequent in Twitter posts, like e.g. acronyms and abbreviations and elements without a clearly defined syntactic function like hashtags, mentions and emoticons (Chiusaroli, 2016), whose presence is mainly motivated by communicative goals of the authors, like the following example shows: *"@pari_biosteria Alessandro #Bergonzoni Contro lo #stigma nei confronti della malattia mentale #passaparola http://t.co/daHsNTcBmh"* (@pari_biosteria Alessandro #Bergonzoni Against the #stigma towards the disease mental #passaparola http://t.co/daHsNTcBmh)

where some hashtag is exploited as common noun (*#stigma*), other as proper noun (*#Bergonzoni*) or with a proper communicative function *#passaparola*).

Each word of PoSTWITA is associated with a tag showing its grammatical category selected within the inventory of tags proposed for the part of speech tagging within the Universal Dependency project; only a few tags extends this inventory for better describe typical social media elements, like EMO for emoticons or URL for web addresses.

Within our corpora we focused only on the words tagged as ADJ (adjectives), NOUN (common nouns) and PROPN (proper nouns), that is those involved in the noun phrase structures. Nevertheless, it must be observed that since PoSTWITA corpus is only tagged morphologically, a proper notion of noun phrase is not marked in it. In order to detect adjectives that are syntactically linked to nouns within noun phrases, we considered the

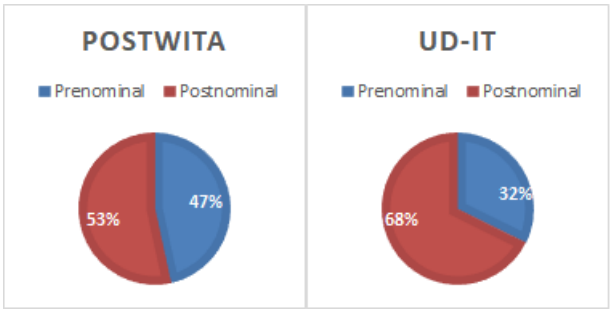


Figure 1: The percentage of pre-nominal and post-nominal adjectives in PoSTWITA and UD-it.

[2] http://universaldependencies.org/it/overview/introduction.html

[3] www.universaldependencies.org

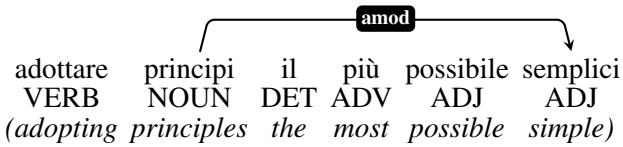
adjectives that were immediately before or after nouns or proper nouns. According to this strategy, the number of adjectives occurring in prenominal position is 1,519, while the number of those in postnominal position is 1,740.

3.2 UD-it

UD-it corpus is tagged both morphologically and syntactically. It is derived from the conversion of different resources developed by Turin and Pisa University's Computer Science Departments and Pisa CNR's Computational Linguistics Institute. This corpus is composed by legal texts (Italian Constitution and part of the Civil Code), Wikipedia and newspaper articles. We can therefore say that, unlike PoSTWITA corpus, UD-it corpus is representative of the so-called Standard Italian, that is encoded, over regional, elaborate, belonging to the upper classes, invariant and written (Berruto, 2010), like the following example shows: *"La prima attività ha lo scopo di creare e sviluppare una rete di ricognizione globale con l'intento di monitorare il rispetto dei trattati internazionali contro la proliferazione di armi di distruzione di massa e la definizione dei confini territoriali."* (The first activity has the objective of creating and developing a network of global reconnoiting with the goal of monitoring the respect of international treatises against the diffusion of the weapons of mass destruction and the definition of territorial borders.)

Providing that UD-it corpus is fully annotated according to the dependency grammar framework of the Universal Dependencies, a notion of noun phrase can be derived from its structures, even if it is not properly annotated, as usual in dependency formats. We considered in this corpus all the adjectives that are related with a noun or a proper noun with the dependency relation *amod*, that is the dependency featuring the adjectival modifiers. Taking into account this relation, we collected 4,469 adjectives occurring in pre-nominal position and 9,362 in the post-nominal one.

It must be observed that the availability of the syntactic annotation of the UD-it corpus has allowed more reliable results with respect to that obtained from PoSTWITA. Indeed we can not be sure that an adjective is related to a specific noun just because it is near that noun, providing that an adjective can refer to a noun even if distant from it, as the following example shows, where an adverbial modifier is collocated between the noun and the adjective that modifies it:



3.3 Discussion of Results

The pie charts (Fig. 1) show the data extraction results. The largest percentage of the post-nominal adjectives provides some hints about the markedness of the pre-nominal position for both PoSTWITA and UD-it.

For what concerns the distribution in pre- and post-nominal position of the categories of adjectives described in sec. 2.1, it is represented in the histograms as detected in Figure 3 (PoSTWITA) and Figure 2 (UD-it). We collected these data by applying to our datasets scripts in Python and SQL queries running on a database version of the resources.

The diagrams show how the adjectives in the lower portion of the hierarchy (relational, colour and physical property) are predominantly in post-nominal position within the noun phrase, whereas the adjectives in the higher portion of the hierarchy (age and dimension) are in majority in the pre-nominal one. Evaluative adjectives are the most equally distributed. These results confirm the theoretical tenets presented in the previous part of the paper and collocate the behaviour of the adjective within a context that can be used for modelling in a computational perspective this grammatical category.

4 Ordering adjectives in SimpleNLG-IT

The formalization of linguistic properties is a fundamental process both for NL processing as well as for NL generation systems. In particular, a widespread architecture for NLG assumes a specific module for the linguistic *realization*, that is essentially an algorithmic implementation of a formal grammar (Reiter and Dale, 2000). Recently, as can be read in (Mazzei et al., 2016), a common set of API for the linguistic realization has been adapted also for Italian language. A key component of SimpleNLG-IT is the reference lexicon, i.e. the computational dictionary specifying the computational properties of the words that the realizer can generate (Mazzei et al., 2016). The de-

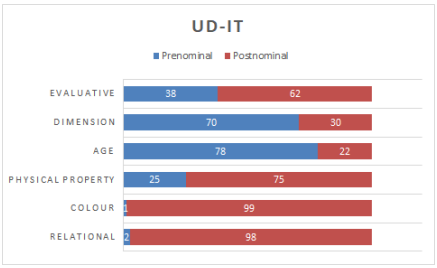


Figure 2: The distribution of the classes of the descriptive adjective in UD-it.

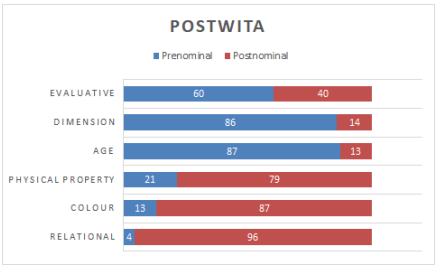


Figure 3: The distribution of the classes of the descriptive adjective in PoSTWITA.

Category	NVdB/UD-it
dimension	15/16
age	7/7
physical property	4/4
colour	10/11
relational	111/121
evalutative^{pre}	33/35
evalutative^{post}	61/68

Table 1: The adjectives occurrences in NVdB/UD-it respectively.

fault position for adjective which is assumed in SimpleNLG-IT is the post-nominal one, with the only exception of ordinals adjectives.

Nevertheless, providing that a more correct modelling of the behaviour of words has a positive impact on the human-machine interaction, in SimpleNLG-IT we devised a new version of the lexicon by following the procedure described in (Mazzei, 2016). We started from the newly released *Vocabolario di base della lingua italiana*[4] (NVdB) (Chiari and De Mauro, 2014) which represent the basic lexicon typical of a standard Italian speaker. Moreover, according to (Giusti, 2016), we classified the adjectives as: relational, colour, physical property, age, dimension, evalutative^{pre} and evalutative^{post}. Indeed, following the data reported in the Figure 2, we formalized that adjective belonging to the relation, colour, physical property sets are generated in pre-nominal position. In contrast, adjectives belonging to age and dimension classes are generated in post-nominal position. Since evaluative adjectives do not show a clear default position, we further split the set in two different subsets that are generated in pre-/pos-tnominal position respectively. Note that not all the adjectives used for UD-it analysis belong to NVdB, e.g. *maggiore* (greater) or *agrario* (agrarian). Table 1 reports the occurrences of the adjectives in NVdB/UD-it respectively.

All the resource developed are made available on a free access repository[5].

5 Conclusion and future work

The paper presents a study about the behaviour of the adjective within the noun phrase. Providing that the qualitative description given by traditional grammars does not allow the definition of a formal model, we considered a recent study that classifies the descriptive adjectives. The long term goal which oriented this study is to contribute to the development of a natural language generation system for Italian featured by a more careful modelling of the behaviour of words within sentence structures.

Assuming a corpus-based perspective we tested on two corpora for Italian the tenets of this study. The results confirm and validate the theory thus opening the window for a definition of a formal model that can be exploited in our computational framework.

Future work is planned to extend the validation of our model on larger datasets, where a wider variety of adjectives is used and also more complex noun phrase structures are taken into account with respect to the simple <*adjective - noun*>or <*noun - adjective*>associations here considered. In particular, providing that more than one adjective can occurs within a noun phrase and can be syntactically linked to a single noun, we intend to investigate on the preference order also in these cases.

[4] https://dizionario.internazionale.it/nuovovocabolariodibase

[5] https://github.com/alexmazzei/SimpleNLG-IT

References

G. Berruto. 2010. Italiano standard. www.treccani.it.

D.N.S. Bhat. 1994. *The adjectival category. Criteria for differentiation and identification*. John Benjamins Publishing Company.

Cristina Bosco, Fabio Tamburini, Andrea Bolioli, and Alessandro Mazzei. 2016. Overview of the EVALITA 2016 Part Of Speech on Twitter for ITAlian task. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Isabella Chiari and Tullio De Mauro. 2014. The New Basic Vocabulary of Italian as a linguistic resource. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *1th Italian Conference on Computational Linguistics (CLiC-it)*, volume 1, pages 93–97. Pisa University Press, December.

F. Chiusaroli. 2016. Scritture brevi e tendenze della scrittura nella comunicazione di Twitter. In *Linguaggio e apprendimento linguistico. Metodi e strumenti tecnologici*. Officinaventuno.

G. Giusti. 2010. Il sintagma aggettivale. In Giampaolo Salvi and Lorenzo Renzi, editors, *Grammatica dell'italiano antico*. Il Mulino.

G. Giusti. 2016. The structure of the nominal group. In *The Oxford guide to the Romance Languages*. Oxford University Press.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. SimpleNLG-IT: adapting SimpleNLG to Italian. In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

Alessandro Mazzei. 2016. Building a computational lexicon by using SQL. In Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*, volume 1749, pages 1–5. CEUR-WS.org.

T. McEnery and A. Wilson. 2001. *Corpus linguistics. An introduction.* Edimburgh University Press.

G. Patota. 2006. *Grammatica di riferimento dell'italiano contemporaneo.* Garzanti Linguistica.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

L. Renzi, G. Salvi, and A. Cardinaletti. 2001. *Grande grammatica italiana di consultazione.* Il Mulino.

L. Serianni. 2006. *Grammatica italiana. Italiano comune e lingua letteraria*. Utet università.