Auxiliary selection in Italian intransitive verbs: a computational investigation based on annotated corpora

Ilaria Ghezzi
Dipartimento di Lingue e Letterature Straniere e Culture Moderne
Università degli Studi di Torino
ghezzi.ila@gmail.com

Cristina Bosco
Alessandro Mazzei
Dipartimento di Informatica
Università degli Studi di Torino
{bosco,mazzei}@di.unito.it

Abstract

English. The purpose of this paper is the analysis of the auxiliary selection in intransitive verbs in Italian. The applied methodology consists in comparing the linguistic theory with the data extracted from two different annotated corpora: UD-IT and PoSTWITA-UD. The analyzed verbs have been classified in different semantic categories depending on the linguistic theory. The results confirm the theoretical assumptions and they could be considered as a starting point for many applicative tasks as Natural Language Generation.

Italiano. *Obiettivo di questo lavoro è l'analisi della selezione dell'ausiliare dei verbi intransitivi in italiano. La metodologia applicata consiste nel confrontare la teoria linguistica con dati estratti da due corpora annotati: UD-IT e PoSTWITA-UD. I verbi analizzati sono stati classificati nelle categorie semantiche individuate partendo dalla letteratura teorica. I risultati confermano con buona approssimazione gli assunti teorici e possono quindi essere il punto di partenza per l'implementazione di strumenti come sistemi di Natural Language Generation.*

1 Introduction

In this work we have applied a corpus-based approach to the investigation of the behavior of Italian intransitive verbs for what concerns the selection of the auxiliary verb. We considered two corpora, namely UD-IT[1] and PoSTWITA-UD (Sanguinetti et al., 2018), annotated following the Universal Dependencies standards. UD-IT and PoSTWITA-UD are treebanks (morphologically and syntactically annotated corpora) for the Italian language. UD-IT is made up of texts from various sources, namely the Italian Constitution, the Italian Civil Code, newspaper articles and Wikipedia. It is a balanced corpus and, therefore, a representative corpus for Italian standard language. On the other hand, PoSTWITA-UD contains tweets from the social media Twitter, and can therefore be considered a representative corpus for the Italian Language used in social media (non-standard Italian). This difference allows us to investigate verbs' behaviour in standard and non-standard Italian Language.

Intransitive verbs have been extensively studied in both traditional grammar and linguistics, since they do not always follow a standardized rule for the auxiliary selection (see examples Section 2). This fact could be the reason why their status is not currently formalized enough in NLP, as long as Italian is concerned. Among the most recent investigation which use a corpus linguistic methodology for the Italian language, we find (Amore, 2017).

Our analysis starts from traditional Italian grammars and then moves to the Auxiliary Selection Hierarchy by (Sorace, 2000), a syntactic and semantic perspective on the behaviour of intransitive verbs and auxiliary selection in Romance languages. That can be useful for formalizing the studied phenomenon and thus providing Natural Language Generation systems with the necessary information regarding the auxiliary selection, which is our final goal. Another contribute for the same systems but for what concerns adjectives has been published in (Conte et al., 2017).

2 Auxiliary Selection in Italian

As in several other languages, in Italian one among two auxiliary verbs can be used together

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

with the past participle verbal forms for compounding periphrastic tenses: *avere* (to have) and *essere* (to be), henceforth respectively indicated as A or E. When the verb is transitive, the auxiliary selection follows standard rules, depending on the diathesis: transitive verbs in active diathesis select A (e.g. *Luca ha mangiato la mela* – Luca ate the apple) while transitive verbs in passive diathesis select E (e.g. *La mela è mangiata da Luca* – The apple is eaten by Luca).

Problems in the auxiliary selection occur instead when the verb is intransitive. In fact, provided that the behaviour of intransitive verbs depends on both semantic and syntactic factors (Van Valin, 1990), a general rule for their auxiliary selection cannot always be formulated[2] (Patota, 2003). Some intransitive verbs can actually select both A or E depending on the semantics of the sentence, while others only admit E or A. See the examples[3] below:

1. *Maria ha corso alle olimpiadi / Maria è corsa a casa*
   (Maria has run at the Olympics / Maria is run home)

2. *Ieri ho camminato al parco / *Ieri sono camminato al parco*[4]
   (I walked in the park yesterday)

Even if all the verbs involved describe a form of movement and are semantically similar, in the first couple of examples the intransitive verb *correre* (to run) allows the selection of both E and A, while in the second one the intransitive verb *camminare* (to walk) only allows the selection of A, and the sentence generated by selecting E is indeed ungrammatical.

Traditional and normative Italian grammars do not provide an analysis of intransitive verbs and auxiliary selection which could be formalized and therefore usefully spent in NLP. In fact, they only suggest lists of verbs that select A or E as auxiliary, see e.g. (Moretti and Orvieto, 1979), (Patota, 2003), (Renzi et al., 1991), (Serianni, 1988), (Dardano and Trifone, 1997). For this reason, we decided to consider other theories too, starting from the *Unaccusative Hypothesis* discussed in (Perlmutter, 1978) and moving to the *Auxiliary Selection Hierarchy* proposed in (Sorace, 2000).
Moreover, we considered the application of a corpus-based approach, provided that corpora represent the way Italian native speakers use A or E together with intransitive verbs. We hypothesized that, this kind of probabilistic perspective can allow a reliable description of the phenomenon. In fact, when there is a lack of standard grammar rules, it is possible to determine certain linguistic aspects by extracting data from corpora. Doing so, we can compensate the lack of standard grammar rules with probabilistic and statistic data.

2.1 The theoretical status of intransitive verbs

For accounting for the behavior of intransitive verbs, in 1978, Perlmutter expressed the *Unaccusative Hypothesis*, which splits intransitive verbs in 2 subcategories: the *unaccusative verbs* and the *unergative verbs*. Perlmutter suggested that the unaccusative verbs are intransitive verbs whose grammatical subject is not an agent (e.g. *La nave è affondata* – The ship is sunk), while unergative verbs are intransitive verbs whose grammatical subject is an agent (e.g. *Giulia ha camminato* - Giulia has walked).

More recently other linguists and researchers analysed the topic, following two major lines: Rosen that suggested to follow a syntactic-only approach (Rosen, 1984), Van Valin and Dowty that suggested a semantic-only approach (Van Valin, 1990; Dowty, 1979).
A development of Perlmutter's hypothesis supported by experimental and psycho-linguistic results can be found in Sorace (2000) that proposed an interesting modelling of the behaviour of intransitive verbs with respect to the selection of auxiliary for Italian too. This theory especially inspired our current work.

2.2 A hierarchy for auxiliary selection

According to the theory proposed by Sorace, intransitive verbs can be hierarchically organized according to their different degree of telicity and agentivity. The more a verb is telic or agentive, the more it systematically selects the auxiliary verb E or A respectively.
This hierarchy of intransitive verbs, also known as *Auxiliary Selection Hierarchy* (ASH), includes categories defined on the basis of thematic and as-

[2]Flexibility in auxiliary selection can be accounted for a large number of cases if context is taken into account.

[3]The translation of the examples can be not correctly mapped on the English rules. When this happens the auxiliary is underlined.

[4]Sentences marked with * are ungrammatical.

ASH category	examples	auxiliary selection
Change of location (**maximum telicity**)	to go, to arrive	selects E
Change of state	to appear, to happen	
Continuation of pre-existing state	to stay, to last	
Existence of state	to exist, to seem	
Uncontrolled process	to sleep, to rain	
Controlled process - motional	to walk, to run	
Controlled process - non motional (**maximum agentivity**)	to act, to play	selects A

Table 1: Examples of verbs organized in the ASH: at the poles verbs that always select E or always select A, and between the verbs that alternatively select both.

pectual features. At one end of the ASH we find intransitive verbs which categorically select E as auxiliary, while at the other end we find intransitive verbs that always select A. The verbs between the two poles of the ASH can have an alternation in the auxiliary selection.
The ASH has been exploited in our work for classifying Italian intransitive verbs depending on its categories which are reported and exemplified in Table 1. This classification may seem wrong for verbs like "to go" (*andare*), which are both agentive and unaccusative, but, as Sorace (2000:863) points out, the verbs that express a change of location have the highest degree of dinamicity and telicity, and they always select E as auxiliary.

3 Intransitive verbs in the fundamental Italian vocabulary

3.1 Verbs selection

In order to focus our study on the intransitive verbs that are more commonly and competently used by Italian speakers, we decided to extract the intransitive verbs to be studied from the *Nuovo vocabolario di base della lingua italiana* (Chiari and De Mauro, 2016), a well known reference resource for Italian lexicography. The lexical entries are here organized in three basic vocabulary ranges according to their frequency of use and ease of recovery in speakers' brain: fundamental vocabulary (FO), high usage (AU) and high availability (AD).
For the present work, we considered only the verbs of the FO vocabulary, for a total of 51 intransitive verbs. But some of these verbs showed more than one single meaning and they could therefore be included in different categories of Sorace's ASH. In order to carry out a disambiguation process, we used Babelnet[5], a multilingual lexicalized semantic network and ontology. After the disambiguation process, the total number of verbs is 67.
For what concerns intransitive pronominal verbs (e.g.*rompersi, "to break"*), we decided not to take them into consideration for our research, since they always select the auxiliary E when constructed in compound tenses (eg. *Gli occhiali si sono rotti* (The glasses broke)). The choice to limit our research to the FO vocabulary is due to the fact that one should expect an expert usage of the verbs of this class also by an artificial speaker.

3.2 Verbs classification

After having selected the verbs, we proceeded to their classification, following the theory proposed by (Sorace, 2000). The intransitive verbs belonging to the FO Italian vocabulary have therefore been included in different categories, depending both on the semantics and the syntax.

Table 2 shows some examples of Italian intransitive verbs belonging to the FO class, classified depending on the ASH by Sorace (2000).

ASH	FO verbs
Change of location	andare (to go)
Change of state	apparire (to appear)
Contin. pre-existing state	rimanere (to last)
Existence of state	esistere (to exist)
Uncontrolled process	dormire (to sleep)
Control. proc. (motion)	camminare (to walk)
Control. proc. (nonmotion)	agire (to act)

Table 2: Examples of intransitive verbs belonging to FO and classified according to ASH.

[5]https://babelnet.org/

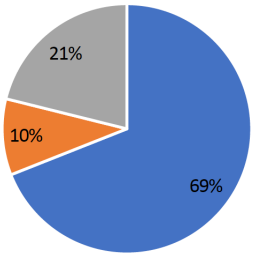

Figure 1: The percentage of intransitive verbs selecting E (in blue), A (in orange) or not detected (in grey) in UD-it.

4 Reference corpora

As mentioned above, the reference corpora for this work are the treebanks UD-IT and PoSTWITA-UD, both annotated according to the Universal Dependencies (UD) format for what concerns morphology and syntax. Provided that UD is currently a standard de facto, the exploitation of this format allows us the application of the same methodology on other resources or languages.

The exploitation of both the data set is motivated by the need to extend our research on the larger available amount of data, and by the fact that UD-IT is representative of the standard Italian language, while PoSTWITA-UD represents the Italian language used in social media. This allows us to obtain a comprehensive set of results.

4.1 Data extraction

To extract the data concerning the auxiliary selection on UD-it and PoSWITA we used the Sets Treebank Search provided by the University of Turku, available for free at `http://bionlp-www.utu.fi/dep_search/`.

We formulated an expression that allowed us to extract data related only to intransitive verbs that appear in the reference corpora at the past participle form together with an auxiliary verb (A or E). We then compared the data from the corpora against the classification based on the linguistic theory.

5 Results

After the data extraction from UD-IT and PoSTWITA-UD, a first consideration is to be made about the percentages of intransitive verbs that select A or E in the two corpora.

As figure 1 shows, in UD-IT the auxiliary A is selected by 10% of the verbs and the auxiliary E by 69%. As long as PoSTWITA-UD is concerned (see fig.2), 49% of verbs select E and 9% select A in this corpus. The remaining percentages (in grey) are made up by the verbs that do not appear in compound tenses in the corpus and did not provide useful result for our study; they must be studied in larger corpora.

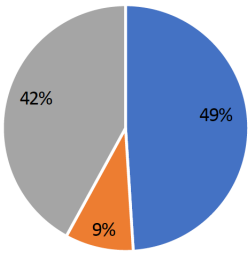

Figure 2: The distribution of verbs selecting E (in blue) and A (in orange) in postwita-UD.

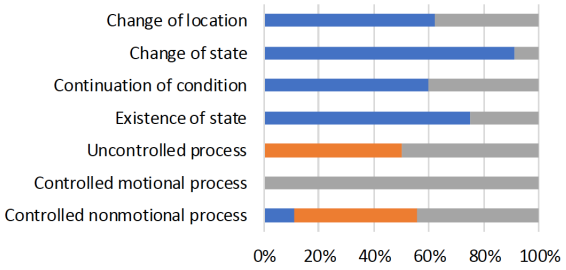

Figure 3: The distribution of verbs selecting E (in blue) and A (in orange) across Sorace's verbal classes in postwita-UD.

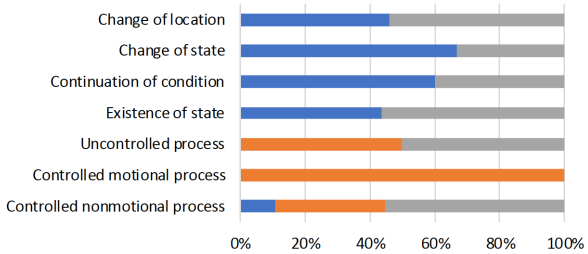

Figure 4: The distribution of verbs selecting E (in blue) and A (in orange) across Sorace's verbal classes in it-UD.

The overall results confirm the linguistic theory for what concerns the distribution in semantic classes organized by Sorace in hierarchy. In fact, as Sorace affirms in (Sorace, 2000), the auxiliary E is selected by intransitive verbs belonging

to the categories of Change of location, Change of state, Continuation of condition and Existence of state as shown in figure 3 and 4 with respect to our two reference corpora. Figure 5 shows an example with the verb "to go" taken from UD-it.

On the other hand, the auxiliary A is selected

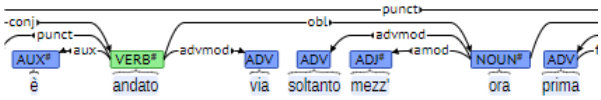

Figure 5: Example taken from UD-IT. In English: "He has gone away only half an hour before the end".

by verbs belonging to the categories of Uncontrolled process, Controlled motional Process and Controlled nonmotional process. This is an example taken from the corpus UD-It, for the verb "to act", *agire* in Italian: *Se, a richiesta del mittente, il vettore emette la lettera di trasporto aereo, si considera, sino a prova contraria, che egli abbia agito in nome del mittente* [6].

As fig. 4 shows, the results related to the category of "controlled nonmotional process" show that both auxiliary A and E can be admitted. This fact is also mentioned by (Sorace, 2000), when she says that some Italian native speakers may accept the auxiliary verb E for this category of verb (e.g. *Il cibo dell'ONU ha / è funzionato solo come palliativo*).

6 Conclusion and future work

The paper presents a study about the auxiliary selection in intransitive verbs in Italian. Providing that the qualitative description given by traditional grammars does not allow the definition of a formal model for the auxiliary selection, we considered a study (Sorace, 2000) that classifies the intransitive verbs taking into account both semantic and syntactic features and behaviors. The long-term goal of this study is to contribute to the development of a natural language generation system for Italian (Mazzei et al., 2016; Mazzei, 2016; Conte et al., 2017). In particular, the facilities of a fluent automatic selection of the auxiliary can be an important feature also in context where the realizer module of the system is used for extracting suggestions for non-native speakers learning Italian as L2.

We adopted in this study a corpus-based perspective and we tested our assumption on two treebanks for Italian respectively representig standard and social media language. The results confirm and validate the theory and they could be used to develop a formal model that can be exploited in a computational context.

[6] English translation: If, under request of the sender, the carrier issues the airway bill, it is considered, if not proven otherwise, that he has acted in the name of the sender.

References

M. Amore. 2017. I verbi neologici nell'italiano del web: Comportamento sintattico e selezione dell'ausiliare. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017.*

I. Chiari and T. De Mauro. 2016. *Nuovo vocabolario di base della lingua italiana.*

G. Conte, C. Bosco, and A. Mazzei. 2017. Dealing with Italian adjectives in noun phrase: a study oriented to natural language generation. In *Proceedings of 4th Italian Conference on Computational Linguistics (Clic-it 2017)*, Rome, Italy.

M. Dardano and P. Trifone. 1997. *La nuova grammatica della lingua italiana*. Zanichelli, Bologna.

D. Dowty. 1979. *Word Meaning and Montague Grammar,*. D. Reidel, Dordrecht.

A. Mazzei, C. Battaglino, and C. 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, December.

G.B Moretti and G.R. Orvieto. 1979. *Grammatica italiana*. Benucci, Perugia.

G. Patota. 2003. *Grammatica di riferimento della lingua italiana per stranieri*. Le Monnier, Firenze.

D. M. Perlmutter. 1978. Impersonal passives and the unaccusative hypothesis. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society 38*. Linguistic Society of America.

L. Renzi, G. Salvi, and A. Cardinaletti. 1991. *Grande grammatica italiana di consultazione*. Il Mulino, Bologna.

C. Rosen. 1984. The interface between semantic roles and initial grammatical relations. In D.M. Perlmutter and C. Rosen, editors, *Studies in Relational Grammar 2*, pages 38–77. University of Chicago Press.

M. Sanguinetti, C. Bosco, A. Lavelli, A. Mazzei, and F. Tamburini. 2018. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of 11th International Conference on Language Resources and Evaluation - LREC 2018*, Miyazaki, Japan, 7-12 May.

L. Serianni. 1988. *Grammatica italiana. Italiano comune e lingua letteraria. Suoni, forme e costrutti*. UTET, Torino.

A. Sorace. 2000. Gradients in auxiliary selection with intransitive verbs. *Language*, 76(4):859–890.

R. D. Van Valin. 1990. Semantic parameters of split intransitivity. *Language*, 66(2):221–260.