Reflexives, Impersonals and Their Kin: a Classification Problem

Kledia Topciu
Università degli Studi Di Siena
Via Roma 56
I-53100 Siena (Italy)
kledia.topciu@student.unisi.it

Cristiano Chesi
NETS - IUSS
P.zza Vittoria 15
I-27100 Pavia (Italy)
cristiano.chesi@iusspavia.it

Abstract

Despite the fact that true reflexives always require a local antecedent, attempting an automatic referential resolution is often far from trivial: in many languages, reflexives are morphologically indistinguishable from impersonals and both particles are sensitive to the syntactic structure in a non-trivial sense. Focusing on Italian, we annotated part of the Repubblica Corpus to attempt an automatic classification of the reflexive and impersonal *si* constructions. In this preliminary study we show that the accuracy of the automatic classification methods that do not use any relevant structural information are rather modest. A thoughtful discussion of the structural analysis required to distinguish among different contexts is provided, in the end suggesting that these structural configurations are not easily recoverable using a purely distributional approach.

1. Introduction

The non-triviality of reflexive/impersonal constructions in Italian is exemplified in (1):

(1) a. Ada$_i$ *si*$_i$ presentò.
A.$_i$ SI$_i$ introduced$_{3\text{-SG-PAST}}$
'A. introduced herself.'

b. *Si*$_{i/*j}$ presentò Ada$_i$.
SI$_{i/*j}$ introduced$_{3\text{-SG-PAST}}$ A.$_i$
'A. introduced herself.'

c. *Si*$_{*i/j}$ presentò ad Ada$_i$.
SI$_{*i/j}$ introduced$_{3\text{-SG-PAST}}$ to A.$_i$
'S/He introduced him/herself to A.'

d. *pro*$_i$ *Si*$_{i/*j}$ tolse la giacca$_i$.
pro$_i$ SI$_{i/*j}$ took$_{3\text{-SG-PAST}}$ off the jacket
'S/He took off the jacket.'

e. Il compagno$_j$ di Ada$_i$ *si*$_{*i/j}$ presentò.
The friend$_j$ of A.$_i$ SI$_{*i/j}$ introduced$_{3\text{-SG-PAST}}$
'A.'s friend introduced her/him-self.'

f. Riconosciuto il compagno$_j$ di Ada$_i$,
pro$_k$ *si*$_{*i/*j/k}$ presentò.
Recognized$_{3\text{-SG-P.PART}}$ the friend$_j$ of A.$_i$,
pro$_k$ SI$_{*i/*j/k}$ introduced$_{3\text{-SG-PAST}}$.
'Once s/he recognized A.'s friend,
s/he introduced her/him-self.'

g. *Si*$_{\text{generic}}$ pensa sempre a salvar*si* la pelle.
SI$_{\text{generic}}$ thinks always to save$_{\text{INF-REFL}}$the skin
'*We* always think about saving *our own* skin.'

Expecting the co-referential DP to be always "immediately to the left" of the reflexive form quickly leads to wrong predictions: if this generalization might seem sufficient in (1a) this is bluntly wrong in (1b), where we need to assume an empty referent (*pro*, Rizzi 1986) before the reflexive (see §1.1). Moreover, we should accept that the coreferential DP can be placed sometimes to the right of the predicate (structurally speaking, *pro* and post-verbal subject options are related, Belletti 2002); in this case, the (focalized/dislocated) post-verbal subject is a good candidate, (1b). Being "the closest DP" is however not a sufficient condition as suggested by the examples (1c-d). Hence, the null subject hypothesis as well as a structural analysis unravelling the role of each DP surrounding the predicate is requested, for the identification of the correct local binding domain (1e-f). Last but not least, a proper classification of the predicate admitting a reflexive or an impersonal pronoun is needed (1g). Under this perspective, we decided to run a little experiment to verify the consistency of a "usage-based" approach (Tomasello 2003) in this specific context and consider whether the "structural analysis" (Chomsky 1995; 2008) can be proved to be an outdated approach for the classification of the distinct kinds of *si*. In the remaining part of this introduction we will present the (possibly outdated) structural analyses proposed for reflexive (§1.1) and impersonal (§1.2) clitic *si*. We will then present our experiment consisting of the annotation of a small fragment of the Repubblica Corpus (Baroni et al. 2004) that we used to train and test a set of Machine Learning classification algorithms (§2). Results presentation (§3) and their discussion (§4) will follow.

1.1 The reflexivization configuration

A popular structural analysis of reflexives is the unaccusative one: under this perspective, the subject of reflexives is an underlying object (just like the subject of unaccusatives) which has to raise to the subject position for Case reasons (reflexive morphology absorbs its Case). Two main variants of this approach are discussed in the literature: a lexical and a syntactic one. The lexical version predicts that the external argument is absorbed in the lexicon (Marantz 1984 and Grimshaw 1990), while the syntactic one proposes that the external argument is present in syntax via the reflexive clitic *se* (Kayne 1988, Pesetsky 1995, Sportiche 1998).

A different analysis is proposed by Reinhart & Siloni (1999, 2005): reflexives should be unergative entries since unaccusativity tests (e.g. *ne* cliticization, (2b)) fail with reflexive constructions:

(2) a. Ne sono arrivati tre.
of+them$_{\text{cl}}$ are arrived three
'Three of them arrived.'
b. *Se ne sono vestiti tre.
SI of+them$_{\text{cl}}$ are dressed three
'Three of them got dressed.'

Since the internal argument only can be cliticized and the reflexive verb fails the *ne* test, we conclude that the subject of the reflexives is an external argument, unlike the subject of unaccusatives. Another test helping us to tease apart external from internal argument structures is reduced relatives modification: when the modification is implemented via past participle, this does not allow for predicates with an external argument. The reduced relative in (3a) contains a reflexive predicate, while the one in (3b) is an impossible cliticization of a transitive reflexive past participle.

(3) a. Il bicchiere rottosi ieri apparteneva a mio nonno.
the glass broken-*him/herself* yesterday belonged to my grandfather
b. *L'uomo lavatosi ieri è mio nonno.
the man washed-*him/herself* yesterday is my grandfather

A robust evidence supports the idea that the subject of reflexive verbs patterns with the subject of unergatives, hence confirming its external argument nature (but see Pescarini 2015:42ff).

Kayne (1975) observes that reflexives occur in environments where transitive verbs are disallowed, e.g. in French causative constructions: when the verb embedded under the causative verb *faire* 'make' is a transitive verb (4a), its subject must be introduced by the preposition *a* 'to'; when the lower verb is intransitive or reflexive, its subject cannot be introduced by *a* (4b/c).

(4) a. Je ferai laver Jean *(a) Luc.
Io make$_{\text{FUT}}$ wash Jean to Luc.
'I will make Jean wash Luc'.
b. Je ferai courir (*a) Jean.
I make$_{\text{FUT}}$ Jean run.
'I will make Jean run.'
c. Je ferai se laver (*a) Jean.
I make$_{\text{FUT}}$ SE wash Jean.
'I will make Jean wash himself.'

When the lower verb is reflexive, its subject appears without the preposition, exactly like the subject of unergative verbs. Therefore, reflexive verbs are not transitive entries either.

Reinhart & Siloni (2005) suggest that these reflexive constructions are unergative entries derived from their transitive alternate by a reduction operation targeting the internal argument (identified with the external one). They take verbal reflexivization even further and propose a *lexicon-syntax parameter*: arity operations (on θ-roles) can apply either to the syntax or to the lexicon. Reflexivization is essentially the same phenomenon cross-linguistically, that is, two available θ-roles are assigned to the same syntactic argument, or, better said, the operation of reflexivization takes two θ-roles and forms one complex θ-role.

The distinctions follow from two different modes of operation: a lexical mode and a syntactic one. Languages such as Hebrew, English, Russian and Dutch have the parameter set to "lexicon", while in Romance languages, Greek and German the "syntax" value of the parameter is set. In the syntactic option (which is relevant here), what is to become a reflexive verb leaves the lexicon with the same number of θ-roles, which need to be assigned, as the basic verbal entry. Since the clitic itself cannot be viewed as an argument (the lack of Case blocks its merge), the "extra" θ-role has to be explained by an arity reduction operation.

In conclusion, an automatic classification algorithm, attempting at identifying the typology of the *si* reflexive pronoun, should necessarily have access to the subcategorization verbal frame and postulate an arity-reduction as suggested by (Reinhart & Siloni 2005). If this information is not available as lexical resource, we might try to rely on structural cues to infer the correct argument structure (as in Merlo & Stevenson 2001, Basili et al 1997 or Ienco et al. 2008). On the other hand, if statistical cues would be available, annotating them overtly would be unnecessary.

A further complication, however, is associated to the existence of a class of "reflexive" predicates (e.g. *alzar*si, 'to stand up') which are bona fide unaccusatives (inherent/lexical *si* constructions Pescarini 2015). In this case, the overlapping between the bare verbal root and a transitive form of some inherent *si* predicates does not help in automatic classification task (e.g. in "si lava la mano", *he/she wash his/her hand*, due to the transitive nature of lavare/*to wash*, the post-verbal DP "la mano" could be analyzed both as direct object or post-verbal subject).

1.2 Impersonal *si* constructions

The reflexive reading is not the only available option when the *si* pronoun is present: an impersonal reading is also possible. Impersonal *si* constructions are used to introduce a generic, unspecified subject and to make general statements about groups of people (Cinque 1988, Dobrovie-Sorin, C. 1998, 1999 a.o.). In Italian, *si* constructions are exemplified in (5a). The subject is unspecified and the sentence has a generic reading because of *si*, otherwise its absence would result in a sentence with a specific subject (5b) being Italian a *pro*-drop language (Rizzi 1986).

(5) a. In Italia *si* mangia troppo.
In Italy si eats$_{3rdSG}$ too much
'In Italy, people eat too much.'
b. In Italia pro mangia troppo.
In Italia *pro* reads$_{3rd\text{-}SG}$ a lot
'In Italia he/she reads a lot'

Notice that the adverbial modal modification "troppo" is coherent with the generic reading, while a punctual temporal adverbial modification would result inconsistent ("#In Italia si mangia domani" vs. "In Italia si mangia sempre").

As for the argumental status of *si*, there is a large disagreement in the linguistic community: Cinque (1988) proposes the existence of two different *si* items: the presence of *si* is usually restricted to finite clauses, however, it is also permitted in certain untensed clauses, namely in Aux-to-Comp (6) and Raising structures (7) with transitive and unergative verbs.

(6) Non essendo*si* ancora scoperto il colpevole…
not being$_{GERUND}$-SI yet discovered$_{P\text{-}PART\text{-}SG\text{-}MASC}$
the culprit$_{SG\text{-}MASC}$
'Not having yet discovered the culprit...'

(7) Sembra non esser*si* ancora scoperto il colpevole …
seems$_{3RD\text{-}SG}$ not being-SI yet
discovered $_{P\text{-}PART\text{-}SG\text{-}MASC}$ the culprit$_{SG\text{-}MASC}$
'It seems it hasn't yet been discovered
the culprit.'

Cinque considers these instances of *si* as argumental ones (+arg), which can be present in general only with verbs that project an external θ-role. The other *si* is a non-argumental one (-arg), which can be present with any verb class (therefore, also with verbs that do not assign an external θ-role).

Dobrovie-Sorin (1998, 1999) argues that it is not necessary to postulate this: according to her, what Cinque calls a +arg *si* is actually a middle passive Accusative *si*. The only Nominative *si* is Cinque's -arg *si*. She argues that *si* is not licensed in non-finite clause because it is a Nominative clitic and, in Italian, Nominative clitics are not allowed in non-finite clauses. Only transitive and unergative Aux-to-Comp and Raising structures allow *si* as Accusative. Dobrovie-Sorin tries to unify all the uses of SE in Romance languages and assumes that *si* is not a special lexical item that absorbs a θ-role or Case. Her analysis accounts for special cases, such as Romanian, which has *si* constructions but doesn't have Nominative clitics. Italian *si* constructions, on the other hand, rely either on Nominative (8) or Accusative (which also includes reflexive configurations) (9).

(8) Non si_i e_i è mai contenti.
not SI is$_{3RD\text{-}SG}$ ever satisfied
'One is never satisfied.'

(9) Il greco$_i$ si_i traduce e_i facilmente.
the Greek SI translates$_{3RD\text{-}SG}$ easily
'Greek translates easily.'

In (8), s*i* is an anaphor and if we assume a restricted theory of binding, the anaphoric status of the clitic is transferred to its trace. The indexing configuration corresponds to a single argument, the Theme. On the other hand, the *si* in (9) is not an anaphor and therefore imposes no relation between the subject and object positions; it binds an empty category in the subject A-position.

A rephrase of Dobrovie-Sorin's proposal is formulated by Salvi (2018), who argues that in modern Italian there are two reflexive *si* constructions: a *passive* one and an *impersonal one* (the reader should refer to Pescarini 2015 for a more detailed discussion of a richer classification). The first one, exemplified in (10b), is characterized by the cancelation of the subject (10a) and the transformation of the direct object into the grammatical subject (triggering agreement); the derived grammatical subject can occur also in the canonical preverbal position (10c):

(10) a. Il preside ha consegnato i diplomi.
The dean has awarded the diplomas
b. *Si* sono consegnati i diplomi.
SI$_{generic}$ are awarded the diplomas
'Diplomas got awarded'
c. I diplomi *si* consegnano (agli studenti).
the diplomas *SI*$_{generic}$ awarded
(to the students)
'Diplomas are getting awarded
(to the students)'

This construction is only possible with (di)transitive predicates, since the promotion of the object to the grammatical subject role is only available when a direct object is available.

On the other hand, the impersonal version of *si* does not induce the promotion of the internal argument to the grammatical subject role and in fact this construction is available without any verbal class restriction:

(11) a. *Si* guarda la partita
SI$_{generic}$ watches the game
'We watch the game'
b. *Si* dorme
SI$_{generic}$ sleeps
'We sleep'
c. *Si* cade
SI$_{generic}$ falls
'We fall'

In sum, with the impersonal *si* construction, the subcategorization verbal frame (i.e. the verbal argumental structure) could help in isolating the passive *si* construction, but not the impersonal one. As for reflexive *si*, the full argument structure must be identified and then either the passive strategy (deletion and promotion) or the impersonal one (simple deletion) considered. As a consequence of the null subject option in Italian, the difference between impersonal and passive *si* is often blurred.

2. Materials and methods

From Repubblica Corpus (Baroni et al 2004), we extracted all contexts in which the "si" lemma was present: 2.737.558 contexts are returned by the simple query including a left and right context of maximum 8 words around the *si* + predicate cluster; each left and right context was cut at full stops, colons, semi colons, exclamative and question marks, whenever those were found within the 8 tokens context. The tagset used in the Repubblica Corpus neither distinguishes among reflexive and various types of impersonal forms ("CLI/si" is the generic tag used) nor among different verbal classes with respect to their argumental structure (only VB for "be", VH for "have", and VV for other verbs are included). We then decided to annotate manually the first 2.000 contexts returned by our query (0,07% of the total) using the following scheme much simplified with respect to the structural asymmetries revealed by the discussion in §1: **I** (impersonal), **L** (local, DP immediately preceding "si" is the correct one), **PV** (post-verbal: the first DP after the predicate following "si" is the correct co-referent) and **LM** (the DP immediately preceding, in the hierarchical sense, the reflexive "si" is the correct one, but such DP is "modified" by a PP or a relative clause) and **A** (the referent is not present/retrievable in the extracted context; these are in the great majority pro-drop cases, in just two cases the referent was lexically realized outside the context isolated). Both authors annotated independently the corpus and discussed about the disagreement cases (less than 1% of the sample) in order to find an agreement in the annotation. Table 1 indicates the distribution of the classes across the annotated corpus fragment, while Table 2 exemplifies the classification. Due to the simplicity of this classification (that essentially focus on the identification of the reflexive antecedent, if present/necessary), we would expect a better performance compared to any richer classification, which is apparently necessary according to the structural analysis previously discussed.

annotation	# of contexts	%
I	332	16.6
L	994	49.7
LM	417	20.8
PV	183	9.15
A	74	3.7

Table 1. Distribution of the annotated categories across the sample.

annotation	example
I	*si* è deciso di ridurre il deficit *we decided to reduce the deficit*
L	[i fedeli]$_i$ si$_i$ sono tuttavia sciolti *the faithfulls, nevertheless, split up*
LM	[il vertice di Dublino]$_i$ si$_i$ è dimostrato *the Dublin summit proved to be …*
PV	nel cortile si$_i$ stendono [le stuoie]$_i$ *in the courtyard the mats unfolded*
A	per 16 anni si$_i$ è occupato dei processi *for 16 years [he] took care of the trials*

Table 2. Sample annotation using 5 categories.

2.1 Classifiers descriptions

Under the "usage-based" approach the disambiguation (i.e. the interpretation of the correct referent, if necessary) of the distinct *si* constructions should be possible on the basis of the purely statistical distribution of the (implicit) features across the corpus (Tomasello 2003 and related works). To test this hypothesis we created a set of classifiers using the Weka environment (Frank et al 2016). 4 different classifiers are used including the original extracted context of maximum 8 words before and after the clitic *si* + predicate cluster (Table 3): pure Bag-of-Words (BoW) approach was used for the first two classifiers, one with only the left context included, the other with both left and right context; then we manipulated the left context classifier substituting the words with their POS (classifier C3-POS-L) and with a more coarse set of POS tags (C4-CPOS-L). POS and CPOS annotation are obtained using a free online tool (*ItaliaNLP REST API*, Cimino & Dell'Orletta 2016).

Class. ID	Approach	Context
C1-BOW-L	**BoW**	Left context
C2-BOW-LR		Left & Right context
C3-POS-L	**POS**	Left context
C4-CPOS-L	**CPOS**	Left context

Table 3. Classifier description

2.2 Classification algorithms

Given the baseline classification of 49.7% of accuracy, obtained by choosing always the reflexive local class (L classification), we compared Naïve Bayesian algorithms (i.e. NaïveBayes, *n.bayes* in table 4, and NaïveBayesMultimodal, *n.bayes.mul.* in table 4) with a decision tree-based algorithm (i.e. *J48*) and then with both 3 layers convoluted (with LSTM layer; *conv.net* in table 4) and simple recurrent neural networks using Weka wrappers for Deeplearning4j 1.5.13 (*srnn.net* in table 4) for a total of 5 classifiers. We run our experiments within Weka 3.8.3 environment with CUDA 10.1 GPU nVIDIA support. Word embeddings are built using a larger fragment of left and right contexts (+/-10 words at most, breaking the left/right context at full stops) extracted from Repubblica corpus including the "si" seed (first 1.000.000 sentences returned using the publicly available Sketch Engine search interface).

3. Results

The results of the classification tests are reported in table 4. The accuracy indicates the rate of correct classifications and the standard deviation running 10 experiments with cross-fold validation (standard deviation is indicated) and the significance is expressed with respect to the baseline: 👍 indicates that the accuracy is significantly better than baseline, 👎 significantly worse and no sign means no significant difference (pair-wise comparison using corrected resampled T-Test, Witten & Frank 2005).

Class. ID	Algorithm	Accuracy (SD)	Sign.
baseline		*49.70%*	
C1-BOW-L	n.bayes	56.95% (2.79)	👍
	n.bayes.mul.	54.28% (2.03)	👍
	J48	58.34% (2.48)	👍
	conv.net	51.88% (1.44)	👍
	srnn.net	39.63% (11.79)	👎
C2-BOW-LR	n.bayes	49.21% (3.40)	
	n.bayes.mul.	51.61% (1.17)	👍
	J48	48.66% (2.53)	
	conv.net	49.77% (0.41)	
	srnn.net	39.05% (12.77)	👎
C3-POS-L	n.bayes	54.49% (2.35)	👍
	n.bayes.mul.	53.26% (1.99)	👍
	J48	60.76% (2.97)	👍
	conv.net	57.58% (1.98)	👍
	srnn.net	43.52% (7.17)	👎
C4-CPOS-L	n.bayes	59.96% (2.85)	👍
	n.bayes.mul.	50.89% (1.03)	👍
	J48	61.49% (3.08)	👍
	conv.net	49.70% (0.25)	
	srnn.net	44.20% (6.17)	👎

Table 4. Classification accuracy results

In both left and left-right context classifiers, BoW approach (C1-BOW-L and C2-BOW-LR) is clearly not sufficient to solve the classification problem; the introduction of a right context (C2-BOW-LR) significantly reduces the performance of the classifier. Notice that in almost 10% of the cases the availability of the referent is post-verbal (PV classification). Decision trees (J48), overall, perform better (M=58.34% SD=2.48) but this performance represents a significant improvement only with C1-BOW-L and C4-CPOS-L classifiers. None of the deep learning approaches (conv.net and srnn.net) are significantly better than decision trees (in some cases SRNs perform significantly worse). The best absolute performance in obtained substituting words with coarse POS (C4-CPOS-L). In this case J48 obtains the best accuracy (M=61.49% SD=3.08).

4. Discussion

In this paper, we discussed the nature of some *si* constructions in Italian, suggesting that, despite their apparent simplicity, their structural intricacies require a deep syntactic analysis for identifying correctly the typology of the clitic in various contexts and retrieve, when necessary, a proper referent. Also using a simplified set of five classes (I = impersonal; L = local immediately preceding coreferential DP; PV = local, immediately post-verbal coreferential DP; LM = local preceding coreferential DP but with prepositional phrase or relative clause modification; A = absent referent), we demonstrated that, using an annotated sample of the Repubblica corpus, no classifier has exceeded the performance of 61.49% of accuracy. This is well below any human reasonable performance (as suggested by the 99% agreement in classification between annotators). These results, even though still based on a small fragment of the Repubblica Corpus, extend Chesi & Moro (2018) original considerations using a wider dataset and more advanced ML algorithms.

These results showed that neither the algorithms used nor the extension of the context (both left and right) helped in classifying correctly the instances of “si” when the referent had to be retrieved non-locally or in impersonal “si” cases. Replacing the words with their POS mildly helped in improving the performance of some classifiers (especially using the coarse tagset), with decision tree classifier (J48) obtaining the best performance (on average) across the tests.

Given the poor performance of the classifiers tested, we concluded that the “usage-based” intuition is not sufficient here to account for the acquisition of the discriminative capabilities any Italian native speaker owns and that enable her/him to identify correctly the relevant referent both pre- and post-verbally, even in the case of complex subjects (referent DPs modified by prepositional phrases or relative clauses), as well as its unnecessity (in generic/impersonal readings) or its recovery in case of pro-drop. We might expect then that a richer syntactic annotation could help to boost the automatic classification results in accordance with the structural analysis summarized in §1.1 and §1.2: first, a verbal subcategorization specification properly describing the predicate argument structure could be useful, then a correct analysis of the subject phrase structure, including agreement cues should be used, as well as a richer classification of temporal/modal adverbials/modifiers.

As suggested by an anonymous reviewer, information structure, which is largely obliterated in written texts, is expected to disambiguate between reflexive and impersonal constructions: for instance, non-dislocated preverbal subjects (L(M) in our classification) should be ruled out in impersonal constructions (see Raposo & Uriagereka 1996); moreover, non-focalized (or right-dislocated) postverbal subjects (PV in our classification) should be ruled out in reflexive constructions. Then, despite the fact that prosody/information structure cannot be assessed within a corpus-based study, we might expect an improvement of the classifiers performance considering some relevant features associated to these configurations: e.g. post-verbal subject annotation in connection with the verbal class and adverbials placement between the subject and verb indicating a dislocated subject.

A follow up of this study should test these predictions and, possibly, extend the study to the whole Repubblica corpus, confirming (or disconfirming) our preliminary results that suggest we cannot avoid a deep structural analysis of these constructions to classify (and interpret) them correctly.

References

Baroni, Marco, 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 International Conference on Language Resources and Evaluation (LREC 2004).*

Basili, Roberto, Maria Teresa Pazienza, and Michele Vindigni. 1997. Corpus-driven unsupervised learning of verb subcategorization frames. *Congress of the Italian Association for Artificial Intelligence*. Springer, Berlin, Heidelberg.

Belletti, Adriana. 2002. Aspects of the low IP area. Forthcoming in The structure of IP and CP. *The Cartography of Syntactic Structures*, vol. 2, L. Rizzi (ed.). New York: Oxford University Press.

Burzio, Luigi 1992. On the morphology of reflexives and impersonals. *Theoretical analyses in Romance linguistics*. Amsterdam: Benjamins, 399-414.

Chesi, Cristiano, & Moro, Andrea 2018. Il divario (apparente) tra gerarchia e tempo. *Sistemi intelligenti*, 30(1), 11-32.

Chomsky, Noam 1995. *The minimalist program. Cambridge*, MA: MIT press.

Cimino, Andrea, Dell'Orletta, Felice. 2016. "Building the state-of-the-art in POS tagging of Italian Tweets". In Proceedings of EVALITA '16, Evaluation of NLP and Speech Tools for Italian, 7 December, Napoli, Italy.

Cinque, Guglielmo 1988. On si constructions and the theory of arb. *Linguistic inquiry*, 19(4), 521-581.

Dillon, Brian, Alan Mishler, Shayne Sloggett, and Colin Phillips. 2013. Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. J. *Mem. Lang*. 69, 85–103.

Dobrovie-Sorin, Carmen. 1998. Impersonal se constructions in Romance and the passivization of unergatives. *Linguistic Inquiry*, 29(3), 399-437.

Frank, Eibe, Mark A. Hall, and Ian H. Witten. 2016. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Grimshaw, Jane. 1990. Argument Structure. MIT Press, Cambridge, MA.

Ienco, Dino, Serena Villata, and Cristina Bosco. 2008. Automatic extraction of sub-categorization frames for Italian. In *LREC08*, pp. 2094-2100. European Language Resources Association (ELRA)

Marantz, Alec. 1984. On the Nature of Grammatical Relations. MIT Press, Cambridge.

Merlo, Paola and Stevenson, S Suzanne, 2001. Automatic verb classification based on statistical distributions of argument structure. Computational Linguistics, 27(3), pp.373-408.

Pescarini, Diego. 2015. Le costruzioni con si. Italiano, dialetti, lingue romanze. Roma: Carocci.

Pesetsky, David. 1995. Zero Syntax. MIT Press, Cambridge, MA

Raposo, Eduardo & Juan Uriagereka. 1996. Indefinite SE. Natural Language and Linguistic Theory 14: 749—810.

Reinhart, Tania, & Siloni, Tal. 2005. The lexicon-syntax parameter: Reflexivization and other arity operations. Linguistic inquiry, 36(3), 389-436.

Rizzi, Luigi. (1986). Null objects in Italian and the theory of 'pro'. *Linguistic inquiry*, 17(3), 501-558.

Salvi, Giampaolo 2018. La formazione della costruzione impersonale in italiano. Linguística: Revista de Estudos Linguísticos da Universidade do Porto, 3, 13-37.

Sportiche, Dominique. 1998. Partitions and atoms of clause structure: Subjects, agreement, Case and clitics. New York: Routledge.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University press.

Witten, Ian, H. and Eibe Frank 2005. Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.