The Italian Particle *ne*: Corpus Construction and Analysis

Malvina Nissim, Sara Perboni

Dipartimento di Studi Linguistici e Orientali
Università di Bologna
malvina.nissim@unibo.it saraaruspex@gmail.com

Abstract

The Italian particle *ne* exhibits interesting anaphoric properties that have not been yet explored in depth from a corpus and computational linguistic perspective. We provide: (i) an overview of the phenomenon; (ii) a set of annotation schemes for marking up occurrences of *ne*; (iii) the description of a corpus annotated for this phenomenon ; (iv) a first assessment of the resolution task. We show that the schemes we developed are reliable, and that the actual distribution of partitive and non-partitive uses of *ne* is inversely proportional to the amount of attention that the two different uses have received in the linguistic literature. As an assessment of the complexity of the resolution task, we find that a recency-based baseline yields an accuracy of less than 30% on both development and test data.

1. Introduction

Anaphora resolution in English has witnessed an unbroken string of interest since the 1970s, featuring both rule-based and statistical approaches. This is true not only for personal pronouns which are by far the most studied in the field, but also for less obvious anaphoric phenomena, such as other-anaphora, VP ellipsis, definite descriptions, and zero anaphora (Modjeska, 2002; Hardt, 1997; Vieira and Poesio, 2000; Nomoto and Nitta, 1993). In this paper we discuss the Italian particle *ne*, which exhibits interesting anaphoric properties and would certainly be a welcome addition to the study and modelling of non-standard anaphors.

The Italian particle *ne* has three main uses (Renzi et al., 2001). It can serve as a partitive pronoun, in which case it is often followed by a quantifier as in Example 1. It can be used purely anaphorically to refer to a previously introduced entity, such as "Altobelli" in Example 2. For both partitive and non-partitive uses, in order to interpret the *ne*, the antecedent must be identitified ("words" in Example 1 and "Altobelli" for Example 2). The third use is as a locative adverb, like in Example 3.[1]

(1) [...] oggi il gorilla usa regolarmente 500 parole e **ne** *conosce altre 500*.
Nowadays the gorilla regularly uses 500 words and *knows another 500* [**words**].

(2) Altobelli ha tuttavia insistito per scendere in campo. Le prestazioni e le lodi di questi mesi iniziali **ne** *hanno accentuato l'entusiasmo professionale.*
Nevertheless, Altobelli pressed to be let play. The performance and compliments of these first months *have enhanced* ***his*** *professional enthusiasm.*

(3) Me *ne* vado.
I'm leaving.

Note that *ne* is often used as a clitic, such as in Example 4 (a non-partitive occurrence). This can be the case with any of the three uses described above.[2]

(4) Giuseppe Liggio non era mai finito nei rapporti di polizia. Furono le confessioni di Totuccio Contorno, capomafia pentito, a *rivelar***ne** *il ruolo*.
Giuseppe Liggio never appeared in police records. It was the confession of Totuccio Contorno, a repentant mafia boss, that *uncovered* **his** *role*.

Although syntactic aspects of *ne* have been studied intensively (Belletti and Rizzi, 1981; Burzio, 1986; Sorace, 2000), this particle has received very little attention from a semantic and discourse perspective. Moreover, all the existing work is theoretical, and, to the best of our knowledge, no corpus-based or computational studies exist focusing on this phenomenon. However, annotated corpora are necessary not only for training and evaluating statistical models, but also for testing theories and possibly finding gaps therein.

The wider scope of this work is therefore twofold. On the one hand, we provide well-tested guidelines for the annotation of this anaphoric phenomenon. This can form the basis for large-scale annotation on the same or similar phenomena, useful for training and evaluating resolution models. On the other hand, we are after finding empirical evidence supporting (or discarding) theoretical claims and at the same time want to give a more comprehensive description of the behaviour of *ne* in real occurring discourse. In this paper we contribute to these aims by presenting the annotation scheme(s) developed for the markup of *ne*, a first annotated corpus and results on annotation agreement, add some observations on the general use of *ne*, and conclude with an assessment of the resolution task.

2. Annotation Schemes

The development of the annotation guidelines was based on the existing theoretical work (in particular (Renzi et al., 2001)) and on the direct analysis of 150 occurrences of *ne* in newswire data (Baroni et al., 2004). In our scheme we introduce three main annotation categories: *ne*, *antecedent*, and *predicate*.

Anaphoric uses of *ne* can fall in one of four classes: *partitive*, *non-partitive*, *vague*, and *cataphoric*. The first two apply to cases such as Examples 1 and 2, respectively. The tag *vague* is used when the antecedent is not clearly identifiable

[1] All examples are from the "la Repubblica Corpus" (Baroni et al., 2004), a corpus of about 380M tokens of Italian newswire data. The *ne* is bold-faced, the antecedent is underlined, and the predicate is in italics.

[2] Clitics are not considered in this study, but will be included in future extensions. In the "la Repubblica Corpus" they are one fourth (83,655) of the total occurrences of *ne* (315,467).

or out of context (farther back than two sentences). The last category was introduced simply to exclude cataphoric cases in this pilot study. We cover for adverbial or idiomatic occurrences with the category *non-anaphoric*. Such uses are left out of any further annotation. The tree in Figure 1 gives an overview.

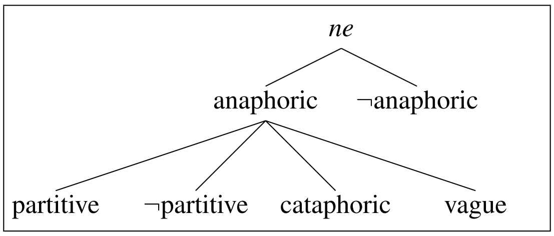

Figure 1: Annotation tree for the particle *ne*

For the annotation of the antecedent we are mainly concerned with its syntactic behaviour. Thus, we specify whether it is an NP, a VP, or a full S. In the case of NPs, we annotate the grammatical function by assigning one of three categories (subject, object, or other), and whether modification is present (yes/no). The annotation tree for antecedents is given in Figure 2.

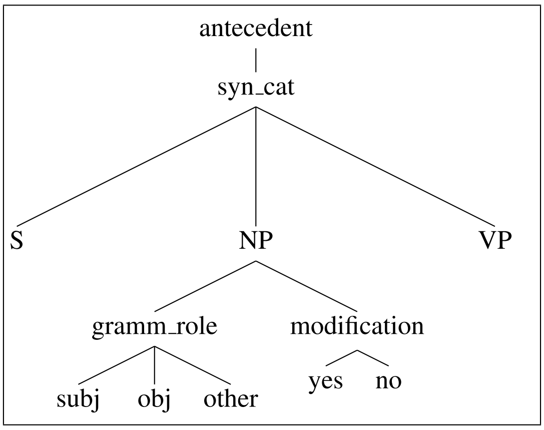

Figure 2: Annotation tree for the antecedent

The predicate is the VP that would take the antecedent of *ne* for saturation. For instance, in Example 1, the predicate is "conosce altre 500" (knows another 500) which is completed by the antecedent "parole" (words). As shown in Figure 3, the only feature we mark up is whether there is a parallelism between the predicate of *ne* and that of the antecedent. This choice was motivated by the observation that parallelism seems to be a characterising aspect of partitive uses.

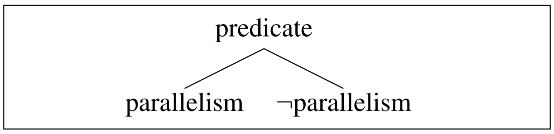

Figure 3: Annotation tree for the predicate

Table 1: Inter-annotator agreement (f-score and kappa) for the three annotation classes.

class	feature	f-score	kappa
ne		0.993	
	type		0.869
antecedent		0.878	
	syn_cat		0.955
	gramm_func		0.977
	modif		0.774
predicate		0.806	
	parallelism		1.000

3. Scheme Validation

3.1. Data Collection

We extracted 300 occurrences of non-clitic *ne* in three sentences of context (we included two sentences before the matching one) from the "la Repubblica Corpus", a large corpus of Italian newswire text (Baroni et al., 2004). The data is already marked up with part-of-speech, lemma, and sentence boundary information. We split this set in two and used 150 samples to develop the annotation guidelines (see Section 2. above) and the remaining 150 to validate the scheme and to carry out some preliminary analysis.

3.2. Method

Two annotators independently marked up a total of 150 occurrences of *ne*, extracted in three sentences of context. The annotation was performed using a customised version of GATE (Cunningham et al., 2002). Inter-annotator agreement was assessed for each class considering one of the annotator's data as the gold standard (G) and the other's as test data (T), and then calculating the f-score, the harmonic mean of precision and recall. Note that the f-score is symmetric, since precision(G,T) is equal to recall(T,G). Additionally, we calculated the kappa score for all of the classification tasks, i.e. the *ne* type assignment, all of the antecedent features, and the parallelism for the predicate.

3.3. Results

Table 1 reports the f-scores and the kappa for the inter-annotator agreement. Among the pre-selected instances of *ne*, there was one case of disagreement on whether a given instance was to be annotated or not (idiom). Overall, with a kappa score of .869 the annotation of *ne* types is very reliable, but not entirely straightforward. Quite surprisingly, it appears to be more difficult to annotate reliably the predicate than the antecedent. This might be due to less specific guidelines and to the fact that sometimes one annotator opted for a stricter extent whereas the other included adverbs, for example. Nevertheless, all classes yielded satisfactory agreement figures (> .7). Note that for antecedents and predicates, partial agreements (the "entity" identified as the antecedent/predicate was the same for both annotators but the extent was different) were not considered as correct. There are 12 such cases for the antecedents, and 24 for the predicates. The reported figures are 'strict' (only fully agreed on cases are taken as correct).

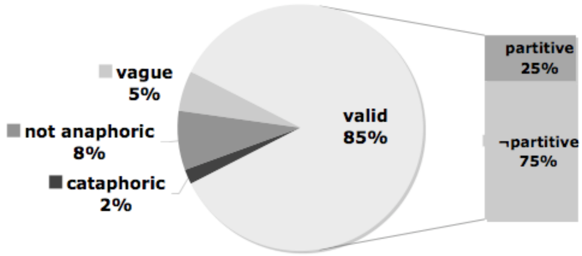

Figure 4: Overview of *ne* annotation in the development data. Under "valid" we group the 124 purely anaphoric *ne* with an expressed antecedent within the extracted window (124).

4. Corpus and Discussion

In order to produce a gold standard corpus, a reconciliation phase followed the independent annotation. Whenever no agreement could be reached on a specific case, this would be classified as "disagreed", and the level of disagreement (*ne*, antecedent, or predicate) was also specified. There were four such cases, all at the antecedent level. These, as well as eleven idiomatic uses, were excluded from any further analysis, leaving us with a total of 135 annotated samples. Eight of these cases were marked as *vague* and three as *cataphoric* (see Section 2.), thus leaving us with a total of 124 purely anaphoric *ne* with an expressed antecedent within the extracted window. The distribution of the annotated cases is summarised in Figure 4.

The data shows that *ne* is mainly used as a non-partitive anaphoric pronoun. Partitive uses, which are the most studied in the syntactic literature, only cover 25% of the valid cases. Partitive uses seem also to be more prone to exhibiting parallelism between the predicate of the *ne* and that of the antecedent (see Example 5).

(5) L'antipatico ritratto di Andrea del Castagno (ex diecimila lire) varrà dieci lire. Il Bernini (cinquantamila) **ne** *varrà cinquanta*.
The unpleasant portrait of Andrea del Castagno (ex ten thousand lire) will be worth ten lire. The Bernini (fifty thousand) *will be worth fifty* **[lire]**.

We can also observe that the overwhelming majority of antecedents are NPs (ca. 93%), and of them, most play an oblique grammatical role. VPs and Ss can be found (see Example 6) but are very rare. Figure 5 summarises this data. Antecedent modification appears to be present in just over 50% of the cases.

(6) Ma frattanto su Palazzo Vecchio continua a sventolare la bandiera del pentapartito. Andiamo a sentire cosa *ne* pensa il sindaco.
But in the meantime, the pentapartito flag keeps on flying on Palazzo Vecchio. Let's go see what the mayor *thinks* **of this**.

The data described so far is intended to inform the analysis of *ne* from a linguistic point of view as well as be used for the development of resolution algorithms. Since any system developed on this data would need to be tested on unseen data, at a later stage we annotated an additional set of 100 samples to be used for evaluation. The snippets were extracted in the same fashion described in Section 3.1. above and from the same corpus ("la Repubblica"). Given the satisfactory agreement on the annotation of the development data, the test set was marked up by one annotator only. The figures with respect to *ne* types are reported in Table 2. The final test set contains a total of 79 valid samples.

Table 2: Overview of test set

ne type		#occurrences
partitive		28
	ante type	
	SN	27
	SV	0
	S	1
non-partitive		51
	ante type	
	SN	42
	SV	1
	S	8
total valid		79
cataphoric		4
vague		6
not anaphoric		11
total extracted		100

5. Resolution Tasks

Resolving *ne* anaphora involves at least two issues: finding the antecedent and determining whether there is a partitive or non-partitive interpretation in order to get the full semantics. Whether these two subtasks should form a sequential procedure (exploiting the results of one subtask in the other) is a non-trivial issue. Additionally, if this were to be the case, the order which resolution should follow is not straightforward. On the one hand, knowing the antecedent would help determine whether there is a partitive or non-partitive interpretation (a plural antecedent is more likely to point to a partitive interpretation, for instance) ; on the other hand, knowing whether there is a partitive reading or not would contribute to finding the antecedent (with a partitive *ne* a plural noun phrase might be a more likely antecedent, all other things being equal, for example).

5.1. Antecedent selection

For assessing the difficulty of the antecedent selection task, we observed some basic phenomena in the data and developed a simple baseline.

In the training data, the antecedent is found in most cases in the very same sentence where the *ne* appears (77), less frequently in the previous one (42) and rarely in two before (5). In 11 cases the antecedent was even further back and could not be annotated since outside the extracted snippet. Among the remaining 124 valid samples, in the overwhelming majority of cases, as observed, the antecedent is an NP (116: 94%), with the remaining cases being either VPs or full sentences, in equal proportion. For nouns, the average distance between the correct antecedent and the anaphor is

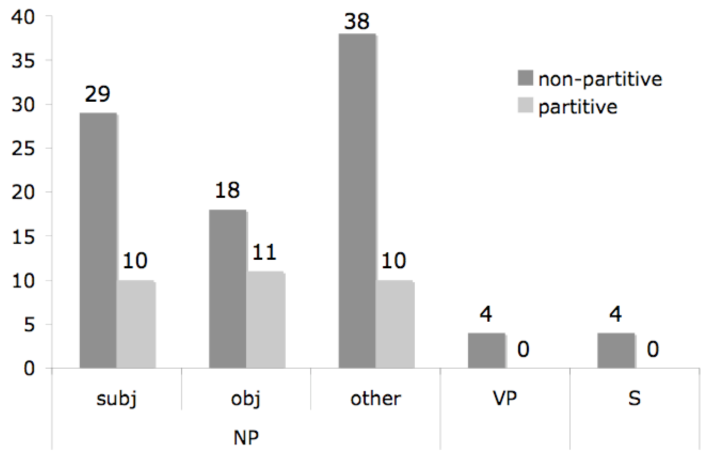

Figure 5: Characteristics of the antecedents of *ne* in the development data

11.2 tokens but only 1.6 nouns.[3]

Building on these observations, as well as on experience on other kinds of non-standard anaphora, we developed a recency-based baseline which for each *ne* selects as antecedent the preceding closest noun. Obviously, this simple approach would never get the correct antecedent whenever this is a VP or an S, by definition. Therefore, the upper-bound on the development data is 94%, which corresponds to the proportion of NP antecedents. On the development data, this baseline finds 37 correct antecedents, yielding an accuracy of 29.8%.

If the antecedent type were to be known, and the baseline could assign, where appropriate, the closest verb or the previous sentence (based on the pre-existing sentence boundary markup in the corpus) rather than closest noun, the overall accuracy on the development data would be 0.347, with the breakdown per type reported in Table 3.

Table 3: Accuracy of the recency-based baseline on the development data if the antecedent type was known

antecedent type	#cases	# correct	accuracy
NP	116	37	0.319
VP	4	4	1.000
S	4	2	0.500
all	124	43	0.347

When run on the test data, which includes 79 valid samples, the recency baseline correctly identified 20 antecedents, achieving an overall accuracy of 29%.

5.2. Determination of *ne* type

Although there seem to be some indicators of (non)partitiveness, the determination of the *ne* type is not that simple. In spite of the amount of attention that partitive uses have received in the literature, our data shows that non-partitive occurrences are far more frequent. Indeed, a most-frequent-use baseline, which would assign a non-partitive interpretation to all occurrences of *ne*, would achieve an accuracy of 75% on the development data and 65% on the test data.

A preliminary analysis of the development data has shown that two important indicators of a partitive use are (i) the parallelism between the *ne*'s predicate and the antecedent's predicate and (ii) a plural antecedent. However, for such features to be exploited, we would need not only to know the antecedent already (see discussion above), but also to parse the data so as to find the relevant predicates. Dependency parsing for Italian has recently witnessed new interest thanks to the organisation of a shared evaluation task (Bosco et al., 2007). The best parser yielded an accuracy of around 87% (Lesmo, 2007), suggesting that we might be able to use reasonably precise syntactic information, even when this is acquired automatically. Future work will explore these avenues.

6. Outlook

We aim at extending this work in two directions. On the one hand, we would like to increase the size of our corpus so as to have larger figures (even for rarer phenomena) that would allow extensive experimenting with statistical modelling. We also plan to extend the annotation to clitic uses of *ne*, which cover approximately one fourth of the total occurrences in the "la Repubblica Corpus" and have been left out in the present study, but might show an interesting, and possibly different, behaviour. On the other hand, once assessed the difficulty of the resolution tasks by means of the two simple baselines that we have described in this paper, we are moving forward to developing more linguistically motivated resolution algorithms, both in a statistical and symbolic fashion.

7. References

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

[3] We could not count full NPs since our data is not chunked, but we counted as intervening nouns between antecedent and anaphor only NOUN-tagged tokens that were not included in the antecedent NP.

Corpus of Newspaper Italian. In Proceedings of LREC 2004.

A. Belletti and L. Rizzi. 1981. The syntax of ne: some implications. The Linguistic Review, 1:117–154.

Cristina Bosco, Alessandro Mazzei, and Vincenzo Lombardo. 2007. Evalita Parsing Task: an analysis of the first parsing system contest for Italian. Intelligenza Artificiale, IV(2):30–33.

L. Burzio. 1986. Italian Syntax: A Government-Binding Approach. Reidel, Dordrecht.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics.

Daniel Hardt. 1997. An Empirical Approach to VP Ellipsis. Computational Linguistics, 23(4):525–541.

Leonardo Lesmo. 2007. The rule-based parser of the nlp group of the university of torino. Intelligenza Artificiale, IV(2):46–47.

Natalia N. Modjeska. 2002. Lexical and grammatical role constraints in resolving other-anaphora. In Proc. of DAARC 2002, pages 129–134, Lisbon, Portugal.

Tadashi Nomoto and Yoshihiko Nitta. 1993. Resolving zero anaphora in japanese. In Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics, pages 315–321, Morristown, NJ, USA. Association for Computational Linguistics.

L. Renzi, G. Salvi, and A. Cardinaletti, editors. 2001. Grande Grammatica di Consultazione dell'Italiano – Voll. I-III. Il Mulino.

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

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. Computational Linguistics, 26(4), December.