Effective Communication without Verbs? Sure! Identification of Nominal Utterances in Italian Social Media Texts

Gloria Comandini
Università di Trento
Trento, Italy
gloria.comandini@unitn.it

Manuela Speranza, Bernardo Magnini
Fondazione Bruno Kessler
Trento, Italy
{manspera,magnini}@fbk.eu

Abstract

English. Nominal utterances are very frequent, especially in social media texts, and play a crucial role as they are very dense from a semantic point of view. In spite of this, their automatic identification has received little to no attention. We have thus developed a framework for the annotation of nominal utterances and created the manually annotated corpus COSMIANU (Corpus Of Social Media Italian Annotated with Nominal Utterances), which could be used to train automatic systems.

Italiano. *Gli enunciati nominali sono un fenomeno linguistico molto frequente, specialmente nello scritto dei social media, e di cruciale importanza, data la loro alta densità semantica. Tuttavia, ben poca attenzione è stata dedicata al loro riconoscimento automatico. In quest'ottica, questo lavoro illustra le guidelines per l'annotazione manuale degli enunciati nominali da noi sviluppate e presenta il corpus dell'italiano dei social media da noi annotato con gli enunciati nominali (COSMIANU), utilizzabile per addestrare sistemi automatici.*

1 Introduction

Syntactic declarative constructions built around a non-verbal head (as in, for example, "What a nice movie!") are very common linguistic phenomena in many Indo-European, Slavic and Semitic languages (such as Latin, Hebrew, Arabic, Russian, English, Spanish, and Italian), as well as in Finno-Ugric and Bantu languages (Benveniste, 1990; Simone, 2013). Not all of these nominal constructions can be unanimously considered sentences, although they can surely be considered utterances, defined as concrete units of actually produced text, devoid of any pre-determined syntactic or semantic form (Sabatini and Coletti, 1997; Adger, 2003; Graffi, 2012; Ferrari, 2014).

It has been clearly shown that nominal utterances (NUs) occur with relatively high frequency not only in spoken language (Cresti, 1998; Landolfi et al., 2010; Garcia-Marchena, 2016) but also in written texts. Literary and journalistic prose certainly offer some fine examples of NUs (Mortara Garavelli, 1971; Dardano and Trifone, 2001), but nonetheless texts produced with computer mediated communication (CMC) or, more generally, within social media, are also a fertile ground for this phenomenon. In fact, NUs are extremely important from the semantic point of view as they allow speakers or writers to provide a lot of information using only a few words (high semantic density), often without any explicit hierarchical relationship (Sornicola, 1981; Ferrari, 2011a), which is a typical feature of CMC (Ferrari, 2011b).

Yet NUs pose significant challenges when it comes to both their automatic processing, because of the absence of a verbal head, and identification, due to the fact that they can have diverse syntactic structures, containing, for example, dependent clauses with finite verbs.

So far, little or no attention has been paid to the identification and processing of NUs in NLP areas such as information extraction/retrieval, sentiment analysis, and opinion mining. However, in order to address newly emerging challenges, these research fields could greatly benefit from tackling NUs specifically. This is the case, for instance, with aspect-based sentiment analysis, which aims to identify the main (e.g., the most frequently discussed) aspects (e.g., food, service) of given target entities (e.g., restaurants) and the sentiment expressed towards each aspect, instead of detecting the overall polarity of a text span (as sentiment analysis usually does). Similarly, argumen-

tation mining, which takes one step forward with respect to opinion mining by extracting not only information about people's attitudes and opinions, but also about the arguments they give in favor of and against their target entities (e.g., products, institutions, politicians, celebrities, etc.), could dramatically improve by focusing on NUs, which are often used, just like slogans, as the most emphatic part of the argumentation.

As a first step towards enabling automatic systems to process NUs, we have developed a complete framework for their annotation, and have created the Corpus Of Social Media Italian Annotated with Nominal Utterances (COSMIANU), which will be freely distributed with a Creative Commons (CC-BY) licence and can therefore be used to train automatic systems.

In this paper, we first summarize the main criteria adopted for the annotation of NUs (Section 3); in Section 4 we describe the annotated corpus; in Section 5 we present the results of some preliminary experiments on automatic identification of NUs, and finally, in Section 6, we draw some conclusions.

2 Related work

The first corpus-based study of NUs was part of the C-ORAL-ROM project, a multilingual (Italian, French, Portuguese and Spanish) corpus composed by 1,200,000 words of spontaneous speech, created in order to describe the prosodic and syntactic structures of romance languages (Cresti et al., 2004).

Relatively similar is the study conducted on the AN.ANA.S Multilingual Treebank, consisting of 21,300 words of spontaneous speech and task-oriented dialogues in Italian, English and Spanish, manually annotated in order to identify verbless clauses (Landolfi et al., 2010).

In more recent work, Garcia-Marchena (2016) uses the Spanish open-source corpus CORLEC[1] to manually identify and classify over 7,000 verbless utterances in a detailed taxonomy.

While the above-mentioned studies all address verbless sentences and clauses, the phenomenon in which we are interested is wider and includes more complex syntactic structures, partly because we address nominal utterances, which is a wider set than verbless utterances (in our perspective, in fact, the main clause of a NU can govern dependent clauses with finite verbs). For this reason we devised a complete annotation framework. Moreover, to the best of our knowledge, our work is the first attempt towards a corpus-based study of NUs on written texts (Cresti (2004), Landolfi et al. (2010), and Garcia-Marchena (2016) address spoken language).

3 Annotation Framework

In the following, we provide a brief summary of the annotation framework we devised for the manual annotation of NUs, which is based on the literature on NUs in Italian (Mortara Garavelli, 1971; Ferrari, 2011a; Ferrari, 2011b). For a thorough description (and plenty of annotated examples), see the document "Linee guida per l'annotazione degli enunciati nominali" (in Italian) [2].

3.1 NU Identification

According to the annotation schema we propose, every utterance whose main clause is non-verbal, i.e. it does not contain a finite verb (see (1)), is marked as a Nominal Utterance (NU); note, however, that a non-verbal main clause can contain non-finite verbs, such as infinitive and/or participial forms and gerunds (see (2), (3), and (4)).

(1) <NU>Felicissima per il suo ritorno!</NU>
[Very happy about his return!]

(2) <NU>Ma impegnarsi di più?</NU>
[Why not put more effort into it?]

(3) <NU>Spariti i negozi, l'edicola, il posteggio.</NU>
[Shops, news stand, and car park, all gone.]

(4) <NU>Facendo due conti.</NU>
[Doing the math.]

3.2 Coordination of main clauses

When the main clause of an utterance bears a coordination relation to another clause, the NU is annotated as follows:

- If both are non-verbal, the extent of the NU includes them both (see (5));

[1] CORLEC, Corpus Oral de Referencia de la Lengua Española Contemporánea, available from: http://www.lllf.uam.es/ING/Corlec.html

[2] This document is available for consultation from http://tiny.cc/auhvvy

- If one is verbal and the other one is non-verbal, the extent of the NU includes only the non-verbal one (see (6)).

(5) <NU>Acqua a dirotto e tutti a casa!</NU>
[Too much rain and everyone home!]

(6) <NU>I lavori prima,</NU> e poi si cena.
[Chores first, and then we'll eat dinner.]

Due to their peculiar syntactic structure, NUs with coordination are further marked with the attribute "verbal-coordinate" (coordination of verbal and non-verbal clauses) or "non-verbal-coordinate" (coordination of non-verbal clauses).

3.3 NUs with subordinate clauses

Non-verbal subordinate clauses are included in the extent of an NU, as in (7), whereas verbal subordinate clauses are not, as in (8) and (9).

(7) <NU>Che bello partire tutti quanti!</NU>
[Great to leave all together!]

(8) <NU>Felice</NU> che ti sia piaciuta.
[Glad you liked it.]

(9) Siccome piove, <NU>tutti a casa.</NU>
[As it is raining, everyone home.]

NUs with verbal subordinate clauses are marked with a specific attribute, i.e., "verbal-subordinate".

3.4 Ellipses

As explained above, NUs are utterances whose main clause is non-verbal, i.e. it does not contain a finite verb. Unlike in other NUs, in ellipses it is always possible to infer the omitted verb (Mortara Garavelli, 1971; Ferrari, 2010), since the omitted verb is exactly the same as the one in the preceeding utterance.

Ellipses are marked, using the specific attribute "ellipsis", both when the preceeding utterance is written by a different user, as in (10) and when it is written by the same user, as in (11).

(10) Cosa vorresti per cena? [What would you like for dinner?]
<NU>Una pizza!</NU> [A pizza!]

(11) Cosa voglio??? [What do I want???]
<NU>Del rispetto!</NU> [Some respect!]

	#sentences	#words	#tokens
Blogs	1,178	16,054	18,874
Forums	1,331	15,168	18,105
Newsgroups	1,395	15,045	19,109
Soc. networks	1,057	7,770	9,923
Total	4,961	54,039	66,011

Table 1: Data about COSMIANU.

4 Annotations in COSMIANU

COSMIANU contains texts taken from the Web2Corpus_IT (Chiari and Canzonetti, 2014), a balanced Italian corpus of 1,050,000 words consisting of social media texts of five types, i.e., blogs, forums, newsgroups, chats, and social networks. In particular, we focused on semi-synchronous forms of CMC, i.e. blogs, forums, newsgroups, and social networks (Pistolesi, 2004), and randomly chose 24 files (six from each of the four selected categories), for a total of 54,039 words.

These texts consist of discussions between users across a large number of themes (from politics to popular singers). Thus in most cases, users interact with each other creating a dialogic enviroment rich in verbal crossfires and quotes. This kind of interactions are a particularly fertile ground for ellipses and NUs in the form of greetings, which are usually very frequent in spoken language.

Automatic pre-proccessing of the corpus, for which we used the TextPro suite of NLP tools (Pianta et al., 2008), consisted of tokenization and sentence-splitting and resulted in 4,961 sentences and 66,011 tokens (see Table 1 for more detailed data).

The manual annotation was then performed by an expert annotator using the Content Annotation Tool (CAT) (Bartalesi Lenzi et al., 2012). The annotation effort, for an expert annotator, consisted of two weeks of work.

In order to evaluate the inter-annotator agreement, a subpart of the corpus consisting of 5,193 tokens was annotated by a second annotator. The resulting Dice coefficient is 87.40. Both annotators identified 127 NUs, 111 of which are common (evaluation based on exact match).

Table 2 reports, for both the whole corpus and for each subcategory, the total number of NUs and the number of NUs marked with each specific attribute, i.e. "verbal-coordinate", "non-verbal-

	NUs	Verbal coord.	Non-verb. coord.	Verbal subord.	Ellipsis	Simple NUs
Blogs	261	30	15	32	37	194
Forums	263	36	13	23	34	190
Newsgroups	196	33	21	17	35	122
Social networks	304	41	9	19	31	231
Total	1,024	140	58	91	137	737

Table 2: Distribution of NUs in the four social media categories.

	Verbal coord.	Non-verb. coord.	Verbal subord.	Ellipsis
Verbal coord.	-	7	13	38
Non-verb. coord.	7	-	11	10
Verbal subord.	13	11	-	26
Ellipsis	38	10	26	-
no other attribute	82	30	41	63
Total	140	58	91	137

Table 3: Attribute co-occurrence.

coordinate", "verbal-subordinate", and "ellipsis" (NUs that are not marked with any attribute, such as (1), (2), (3), and (4), are referred to as "simple NUs").[3]

In the whole corpus we annotated 1,024 NUs, which means that 20,6% of the sentences contain an NU. This percentage is lower than those reported by Cresti (2004) (38,1%) and Landolfi et al. (2010) (28%). This can be explained by the fact that the above-mentioned studies focus on spoken language, where interrupted strings, brachyologies and turn-taking cues are more frequent with respect to written language. Still, this percentage shows that the nominal style is well represented in written informal Italian, most likely due to its linguistic economy and to its high semantic density, which are particularly useful for expressing emphasis (see (12)).

(12) <NU>Dichiarazione da Mr. Hyde!</NU>
[A statement worthy of Mr. Hyde!]

In addition, the large number of NUs marked as coordinate, either "verbal" (140 NUs) or "non-verbal" (58 NUs) shows that parataxis is constant throughout these texts. In fact, NUs appear to be extremely suitable to the parataxis typical of CMC; furthermore, they are often isolated, i.e., free from hierarchical syntactic bonds. This also explains why NUs can be composed of a series of denotative elements simply listed without any explicit hierarchical bond, as in (13), in a way that reminds one of a list of keywords.

(13) <NU>Buon senso, etica, vincere tanto per vincere.</NU>
[Common sense, ethics, winning for winning's sake.]

Looking at the distribution of NUs in the four subcategories, we see that social networks have the highest number of NUs (304), despite having a significantly lower number of tokens than blogs, forums and newsgroups. This probably depends on the high perceived communicative economy typical of social networks (Cosenza, 2014), which leads writers to produce short, almost telegraphic, texts.

In Table 3 we report the co-occurence of NU attributes by pairs[4] in order to show how diverse syntactic structures NUs can have. Particularly interesting is the presence of 38 NUs containing ellipses coordinated with a verbal clause; in fact, the ellipsis usually follows the verbal clause, whose verb is implied in a contrastive context. Additionally, ellipses can support a verbal subordinate clause (in our corpus we have 26 cases), which usually adds further information in favor of the contrastive utterance (see (14)).

[3] Notice that a single NU can be marked with more than one attribute.

[4] Although we have case where NUs have been marked with up to four attributes, we only focus on co-occurrence by attribute pairs.

(14) Non è un edificio specifico, <NU> ma una tipologia architettonica </NU> che caratterizza l'URSS.
[It is not a specific building, but an architectural typology that characterizes the USSR.]

5 Automatic Identification of NUs

We used COSMIANU to train an open source SVM classifier, YamCha[5], and performed some preliminary experiments on NU identification. As training data, we selected 44,170 tokens (i.e. about 2/3 of the corpus) while maintaining the same proportion of blogs, forums, newsgroups, and social networks over the whole corpus. We used the remaining part of the corpus (21,841 tokens) as a test set. In these preliminary experiments we also included the NUs that appear in the text as metadata, which are annotated and marked with the specific tag "metadata" in COSMIANU, as shown in Example (15) [6]. The training set and the test set thus contain respectively 1,775 and 1,058 NUs.

(15) <NU> Data: 27/09/2010. </NU>
[Date: 09/27/2010.]

We pre-processed the data using the TextPro suite (Pianta et al., 2008) and performed a number of experiments combining the following basic features: two-word window context (W2), three-word window context (W3), token (Tok), lemma (Lem), and Part-of-Speech (Pos).

Configuration	Prec.	Rec.	F1
Baseline	33.80	27.13	30.10
W2+Tok+Lem+Pos	79.80	67.96	73.40

Table 4: Results on NU identification.

Table 4 reports, in terms of Precision, Recall, and F1, the results we obtained with the baseline configuration (the system identifies only the NUs in the test set that also appear in the training set) and those we obtained with the best configuration, i.e. using all the features and a two-word window context. With the latter, the classifier identified 901 NUs, of which 719 are correct (exact match), thus reaching an F1 of 73.40% and outperforming the baseline by over 43 points.

[5]Yet Another Multipurpose CHunk Annotator. Website: http://chasen.org/ taku/software/yamcha/

[6]Metadata usually refer to when and where a certain message has been written; although "metadata" NUs are very frequent in the corpus (more than 60% of the total), they are not particularly interesting from a linguistic point of view and we did not include them in the counts of Section 4.

6 Conclusion and Future Work

This work shows how common NUs are in written informal language, as well as how important they are in conveying semantically dense concepts in emphatic informative peaks, which could be useful for many NLP fields (e.g., argumentation mining and aspect-based sentiment analysis).

By creating COSMIANU, an Italian corpus annotated with NUs, and making it freely available to the research community, we made a first step towards the development of automatic tools for the identification and classification of NUs. In our preliminary experiments on NU identification (performed using an SWM classifier), with our best configuration, we obtained a performance of 73.40% in terms of F1 on all NUs (i.e. including metadata).

In the future, we intend to further expand COSMIANU, both in terms of its size and in terms of the annotations it includes, hoping that this will encourage more research on this extremely common, and yet almost neglected, linguistic phenomenon. We also plan to work on the analysis and automatic recognition of NUs, especially when they are used to convey hate speech, in the form of racist, sexist, homo/transphobic or classist slogans and insults.

Acknowledgments

We would like to thank Isabella Chiari for providing us the Web2Corpus_IT, from which we selected the raw texts to build COSMIANU. We also thank our colleagues Roberto Zanoli and Rachele Sprugnoli for their valuable advice and contributions in performing the experiments and defining the annotation guidelines.

References

David Adger. 2003. *Core Syntax: A Minimalist Approach*. Oxford University Press.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 333–338, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Émile Benveniste. 1990. *Problemi di linguistica generale*. Mondadori, Milano, Italia.

Isabella Chiari and Alessio Canzonetti. 2014. Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. In E. Garavelli and E. Suomela-Härmä, editors, *Dal manoscritto al web: canali e modalità di trasmissione dell'italiano*, pages 595–606. Franco Cesati Editore, Firenze, Italia.

Giovanna Cosenza. 2014. *Introduzione alla semiotica dei nuovi media*. Laterza, Bari, Italia.

Emanuela Cresti, Fernanda Bacelar do Nascimento, Antonio Moreno-Sandoval, Jean Véronis, Philippe Martin, and Khalid Choukri. 2004. The C-ORAL-ROM CORPUS. A Multilingual Resource of Spontaneous Speech for Romance Languages. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Prococeedings of the 4th LREC Conference*, pages 575–578, Paris, France. European Language Resources Association (ELRA).

Emanuela Cresti. 1998. Gli enunciati nominali. In M. T. Navarro, editor, *Atti del IV convegno internazionale SILFI (Madrid 27-29 giugno 1996)*, pages 171–191, Pisa. Franco Cesati Editore.

Maurizio Dardano and Pietro Trifone. 2001. *La nuova grammatica della lingua italiana*. Zanichelli, Milano, Italia.

Angela Ferrari. 2010. Enunciati ellittici. *Enciclopedia dell'Italiano*. `http://www.treccani.it/enciclopedia/enunciati-ellittici_(Enciclopedia-dell'Italiano)/`.

Angela Ferrari. 2011a. Enunciati nominali. *Enciclopedia dell'Italiano*. `http://www.treccani.it/enciclopedia/enunciati-nominali_(Enciclopedia_dell'Italiano)/`.

Angela Ferrari. 2011b. Stile nominale. *Enciclopedia dell'Italiano*. `http://www.treccani.it/enciclopedia/stile-nominale_(Enciclopedia-dell'Italiano)/`.

Angela Ferrari. 2014. *Linguistica del testo. Principi, fenomeni, strutture*. Carocci, Roma, Italia.

Oscar Garcia-Marchena. 2016. Spanish Verbless Clauses and Fragments. A corpus analysis. In Antonio Moreno Ortiz and Chantal Pérez-Hernández, editors, *CILC 2016. 8th International Conference on Corpus Linguistics*, volume 1 of *EPiC Series in Language and Linguistics*, pages 130–143. EasyChair.

Giorgio Graffi. 2012. *La frase: l'analisi logica*. Carocci, Roma, Italia.

Annamaria Landolfi, Carmela Sammarco, and Miriam Voghera. 2010. Verbless clauses in Italian, Spanish and English: a Treebank annotation. In S. Bolasco, I. Chiari, and L. Giuliano, editors, *Statistical Analysis of Textual Data. Proceedings of the 10th International Conference on Statistical Analysis of Textual Data (JADT 2010)*, pages 450–459. Roma, Italia, June 9-11.

Bice Mortara Garavelli. 1971. Fra norma e invenzione: lo stile nominale. In Accademia della Crusca, editor, *Studi di grammatica italiana*, volume 1, pages 271–315. G. C. Sansoni Editore, Firenze, Italia.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro tool suite. In *Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference*, Marrakech, Morocco, May 28-30.

Elena Pistolesi. 2004. *Il parlar spedito. L'italiano di chat, e-mail e sms*. Esedra, Padova, Italia.

Francesco Sabatini and Vittorio Coletti. 1997. *Dizionario Italiano Sabatini-Coletti*. Giunti, Firenze, Italia.

Raffaele Simone. 2013. *Nuovi fondamenti di linguistica*. McGraw-Hill, Milano, Italia.

Rosanna Sornicola. 1981. *Sul parlato*. Il Mulino, Bologna, Italia.