Verbless clauses in Italian, Spanish and English: a Treebank annotation

Annamaria Landolfi [1], Carmela Sammarco [2], Miriam Voghera [1]

[1] University of Salerno

[2] University of Venice

Abstract

We present here the first data collected on the AN.ANA.S. Multilingual Treebank (AN.ANA.S._MT), consisting of Italian, English and Spanish task-oriented dialogues and spontaneous speech for a total of about 21300 words. AN.ANA.S. (*Annotazione e Analisi Sintattica*) is a system freely downloadable at the portal www.parlaritaliano.it, developed within a treebank project for the syntactic annotation of both spoken and written texts (Voghera and Cutugno in press). The system is constituent-based and allows the organization of syntactic units within a hierarchical structure, according to XML standards. Constituent relations are coded directly using elements nesting XML properties.

Our research focuses on the treatment of 'non-canonical' syntactic structures, such as verbless sentences, which are usually marginal in the annotation schemes of treebanks. In this paper we will take into account both methodological and quantitative aspects. Firstly, we will compare how verbless sentences are treated in the main annotation systems in the investigated languages. Secondly, we present data on the typology and frequency of such structures in AN.ANA.S._MT.

Keywords: Treebank, syntax, speech, verbless clauses, multilingual corpora.

1. Introduction

The AN.ANA.S. Multilingual Treebank (AN.ANA.S._MT) is a project aimed at the syntactic annotation of spoken and written texts of different Western European languages. In the present work we present data derived from the initial portion of AN.ANA.S._MT consisting of both elicited and spontaneous spoken material in Italian, English and Spanish. The AN.ANA.S. annotation scheme (Voghera and Cutugno in press) has been created to annotate spoken texts, but it is suitable to also annotate written texts. Thus, the scheme, while theoretically conceived of to account for particular features of speech, may be adapted to any kind of text.

The AN.ANA.S. scheme is based on the theoretical assumption that the reference unit for spontaneous speech is the clause instead of the sentence. Thus, it is a tagging system created for the syntactic analysis of the clause, both in written and in spoken texts. AN.ANA.S. makes use of the manual software XGATE, which generates a database of texts in the XML format. The set of annotation rules is given to the software using a DTD (document type definition). The DTD is composed of a series of elements, which represent the syntactic nodes (sentence, clause, the different phrases, etc.). Each element has its own tag (for example: clause) and a list of attributes (for example: type, n. of phrases, etc.).

Unlike other annotation schemes, there is not a separated morphological level of analysis, even though some morphological information on the phrase heads is given in the list of attributes belonging to the different phrases.

The list of attributes also gives functional information (for example, the subject or object role in the NP tagset), constituents' order information, information on the internal properties of constituents, and lexical information.

The data we present here are based on the spoken section of AN.ANA.S._MT. The corpus has been built taking into account two relevant dimensions of variation: contextual and interlinguistic. In fact, it consists of both elicited and spontaneous speech of three different languages. The size of the different sections is reported in Tab. 1.

AN.ANA.S._MT Corpus		*Words*	*Clauses*
Spoken Italian	Task-oriented dialogues	3700	743
	Spontaneous speech	3570	518
Spoken English	Task-oriented dialogues	3900	748
	Spontaneous speech	3635	622
Spoken Spanish	Task-oriented dialogues	3656	871
	Spontaneous speech	3460	615
Total		21921	4117

Table 1: internal composition of the corpus annotated in the AN.ANA.S. Multilingual Treebank.

2. Verbless utterances in treebanks

Corpus-based studies on spoken language (Biber et al.,1999; Cresti and Moneglia, 2005) have shown that more than 30% of utterances are verbless. Verbless sequences belong to very different types and occur in both informal and formal speech (Cresti and Moneglia, 2005) as well as in written registers (Fiorentino, 2004; Ferrari et al., 2008).

However, these kinds of structures are usually considered marginal in linguistic data treatment. Surprisingly enough, despite the considerable number of existing treebanks, scant attention is paid to such structures. This is due to the difficulty of inserting such structures in the current syntactic models, which are all more or less verb-centric; i.e. they consider the presence of a verb a basic requirement to identify a sentence or a clause. The result is that in the spoken language, above all in dialogues or conversation, many linguistic structures that completely fit with whole utterances are automatically considered as non-sentential material.

Most treebanks do not have a unitary treatment of verbless utterances and mostly concentrate on predicative verbless clauses, such as those composed of an NP subject and an adjectival predicate:

(1) What a wonderful world!

We analyzed the Penn Treebank (Taylor et al., 2003), the Italian Syntactic-Semantic Treebank (ISST) (Montemagni et al., 2003) and the Prague Dependency Treebank (PDT) (Böhmová et al., 2003).

The Penn Treebank, a constituency treebank which deals with both written and spoken texts, has a level of functional annotation in which any predicate that is not a VP bears the functional tag – PRD. Thus, if the set of functional tags specifies functions at phrase level, there is no indication on the verbless (and, consequently, on the predicative vs. non-predicative) status of a given sentence at sentence level.

The ISST and the PDT both annotate written texts only and encode dependency relations [1].

The ISST functional dependency annotation takes predicative verbless clauses into account. Consider the annotation of the string (2) [2]

(2) Queste le principali raccomandazioni
lit. "these the main recommendations"
sogg (, questo)
pred (, raccomandazione)
mod (raccomandazione, principale)

Here, the pred functional relation exists between the nominal predicate le raccomandazioni and an absent copula.

If the construction is considered elliptical and the absent constituent can be retrieved, it is placed before the comma and written in capital letters. Consider the annotation of (3):

(3) Mario mangia la mela, Gianni l'arancia
lit. "Mario eats the apple, Gianni the orange"
sogg (mangiare, Mario)
pred (mangiare, mela)
sogg (MANGIARE, Gianni)
pred (MANGIARE, arancia)

In the Prague Dependency Treebank, predicative verbless clauses are in fact considered to be verbal clauses and are called constructions with an empty verb. Their governing node bears the wording #EmpVerb. Some examples are:

(4) Cigarette? (noun phrases acting as questions in the sense of offers and invitations);
(5) What a beautiful day ! (exclamatory constructions with *what a*);
(6) In Praha, at five o' clock (adverbials) [3].

As far as non-predicative verbless clauses are concerned, a noteworthy example is that of the PDT, which provides labels for vocatives, interjectional clauses (interjections, yes-no particles and well introducing a clause), and subject-case clauses (which are basically composed of a NP and can be either independent or embedded). The governing node of these kinds of clauses will bear the label VOCAT, PARTL and DENOM respectively [4]. The other two analyzed treebanks do not seem to have any particular label for such constructions.

Thus, the treebanks analyzed above either mark the predicative status of certain constituents without assigning a specific tag to the verbless sequence, as in the case of the Penn and the ISST, or they assimilate predicative verbless clauses to verbal clauses, as in the case of the PDT.

AN.ANA.S. proposes a unique treatment of all verbless utterances. Apart from what we could clearly define as pure nominal sentence (Benveniste, 1950; Hengeveld, 1992) i.e. a verbless predicative clause, it is possible to recognize many different types of verbless sequences, which do not have a pragmatic value only, but share with verbal sentences a) syntactic autonomy and b) prosodic features (Giordano and Voghera, in press). We consider all these sequences clauses,

[1] The PDT is a dependency treebank, while the ISST has got a level of constituency annotation and a level of functional annotation which encodes dependency relations.

[2] In *Documento di Specifiche Tecniche di SI-TAL Manuale Operativo*, http://www.ilc.cnr.it/tressi_prg/papers/Treebank1_1.pdf.

[3] In *Annotation of English on the tectogrammatical level*, http://ufal.mff.cuni.cz/~toman/pedt_manual/ch04s03s01.html#pic232slov6.

[4] Even though if an independent subject-case clause is a parenthesis, it will bear the PAR label.

regardless of their internal constituency. Consequently, the AN.ANA.S. annotation scheme has a unique ‘verbless clause’ tag to annotate both syntactic units which can be interpreted as an elliptical form of sentence, and syntactically autonomous sequences which do not have verbal projection or that are simply base-generated constituents (Progovac, 2006), such as discourse markers, hesitations, greetings, etc.

Therefore, any verbless sequences with clause features depends on an unique verbless clause node.

3. Data analysis

In the AN.ANA.S._MT corpus, we basically distinguish five types of verbless clauses (Vless_C):

1. Classical predicative Vless_Cs (dirhematic, with or without circumstantials):

(7) It. *Anche lì, volontario, non volontario il fallo di mano di Iuliano* (lit. “even there, voluntary, not voluntary the handball of Iuliano”) ;
(8) Engl. *What about his face?*;
(9) Spa. *Yo cualquier equipo español menos el Barça* (lit. “me$_{\text{(NOM.)}}$, all Spanish football teams but not Barça”).

2. Argumental and non argumental-clauses, both embedded or part of a predication which is distributed over several turns, belonging to either one or more speakers:

(10) It. E e credo *che di lí* (lit. “and (I) believe that from there”);
(11) Engl. *The the questions* that that you…;
(12) Spa. A. *De qué color es el cabello ?* (lit. “What color is the hair? ”)
B. *De quién ?* (lit. “whose? ”)
A. *Del señor* (lit. “of the man’s”)
B. *Negro* (lit. “Black”).

3. Elliptical structures [5]:

(13) It. I contrari sono stati cinquantadue, *mentre ben ottantacinque gli astenuti* (lit. “the opponents have been fifty-two, while well eighty-five the abstainers”);
(14) Engl. A. Ok and the two windows are they on the right side or the left side?
B. *On the right on the right side*;
(15) Spa. A: Cuántos dedos tiene ? (lit. “How many fingers does he have ? ”)
B: *Cuatro dedos* (lit. “four fingers”).

4. Isolated phrases or words, both syntactically and semantically [6]:

(16) It. *Sindrome spalla mano* (lit. “syndrome shoulder hand”);
(17) Engl. *Two pointed ears*;
(18) Spa. *La caravana* (lit. “The caravan”).

5. All cases of DMs, simple yes/no answers, interrupted strings, exclamations [7]:

(19) It. *Ecco!* (lit. “here it is; that’s why”);
(20) Engl. *Oh well*;
(21) Spa. *Vale* (lit. “well”).

[5] Elliptical structures could also include argumental constituents that are pieces of predication distribuited over more utterances.

[6] In this type we have also included cases of repetitions of entire phrases by another speaker, which are very frequent in the task-oriented dialogues. These repetitions do not add anything to the construction of a predication either in syntactic or in semantic terms.

[7] A similar classification can be found in Giordano and Voghera (in press).

Obviously, the distinction among the five types is not always straightforward. If it is easy to identify a classical Vless_C as predicative, it is not the same for the extremely variegate class of verbless sequences.

Firstly, our data shows that, especially in elicited speech, the phenomenon of "building a predication" over different turns is a very frequent phenomenon:

(22) It. A: *Poi sotto alle ultime due a destra* (lit. "then under the last two on the right")
B: *Sì altre due* "yes another two".

As far as the different functions of short yes/no answers are concerned, we tagged this element as 'predicative phrase' when the element 'yes' or 'no' could be considered as 'pro'-sentence:

(23) It. *Credo di sì* (lit. "(I) believe of yes") "I think so", where "sì" is a predicative phrase.

Instead, if the same element represents a replica, or a discourse marker (above all at the beginning of the turn with a phatic function) we have tagged these elements with DM or ISO tags, clarifying its grammar category and underlining that this element is syntactically isolated.

(24) Spa. *Sí hablo general Cristina* (lit. "yeah I speak in general Cristina").

Moreover, the classical dirhematic Vless_C itself is not forcedly composed of a NP subject and a PredP. In our corpus we also have examples of dirhematic clauses composed of NP+ a predicative PP (prepositional phrase) or NP+ a predicative adverbial:

(25) Engl. *Gravy over the table*;
(26) It. *Queste canzoni qui* (lit. "These songs here"), where "qui" has no translation but has the function of intensifying the demonstrative adjective.

We decided to include cases of repetitions of entire phrases and cases of further explanations by another speaker in the sub-class "Isolated phrases or words, both syntactically and semantically". However, some of these cases are not always easily distinguishable by mere cases of ellipsis, for example:

(27) Engl.: A: Oh we've found a difference?
B: *The second difference.*

We calculated the percentage of Vless_Cs on the total amount of all clauses and the different frequency of the various types. We have also tried to compare the quantitative data across the texts of the different languages and spoken situations. Tab. 2 and 3 show data on the total amount of Vless_Cs in our corpus:

VERBLESS CLAUSES	*ITALIAN* *Sp.It.*	 *El.It.*	*ENGLISH* *Sp.En*	 *El.En.*	*SPANISH* *Sp.Spa.*	 *El.Spa.*	*Total*
Vless_Cs	65	292	102	224	122	367	1172
Vless_Cs per language	357 (28%)		326 (24%)		489 (33%)		

Table 2: Amount of Vless_Cs per language

VERBLESS CLAUSES	*Spontaneous Speech* *Italian*	 *English*	 *Spanish*	*Elicited Speech* *Italian*	 *English*	 *Spanish*
Vless_Cs	65	102	122	292	224	367
Total	289 (25%)			883 (75%)		

Table 3: Amount of Vless_Cs in spontaneous vs. elicited speech

Tab. 4 shows the percentages of five verbless types we recognized in our corpus:

VERBLESS	*ITALIAN*		*ENGLISH*		*SPANISH*		*Total*
CLAUSES	*Sp.It.*	*El.It.*	*Sp.En*	*El.En.*	*Sp.Spa.*	*El.Spa.*	
Classical predicative	4 (6%)	4 (1%)	6 (5%)	25 (11%)	25 (21%)	51 (14%)	115 (10,%)
Arg/non-arg-clauses	11 (17%)	49 (17%)	3 (3%)	17 (7%)	8 (6,5%)	59 (16%)	147 (12,5%)
Elliptical structures	13 (20%)	45 (15%)	14 (14%)	39 (17%)	7 (6%)	26 (7%)	144 (12,3%)
Isolated phrases	29 (45%)	53 (18%)	18 (18%)	42 (19%)	37 (30%)	92 (25%)	271 (23%)
DM/ yes/no	8 (12%)	141 (48%)	61 (60%)	101 (45%)	45 (36,5%)	139 (38%)	495 (42,2%)
Total	65	292	102	224	122	367	1172
Total in each language	357		326		489		

Table 4: Percentage of different types of Vless_Cs from the total amount of clauses.

Although these data derived from the initial portion of AN.ANA.S._MT, it is possible to recognize some general trends.

1. The percentage of Vless_Cs in our data confirms previous studies (Biber et al., 1999; Cresti and Moneglia, 2005) and shows that Vless_Cs cannot be considered a disfluencies phenomenon, but cover nearly one third of speech.
2. As stated in other corpus-based studies (Cresti, 2005), classical predicative clauses represent only 10% of the total amount of Vless_Cs. Thus, the treebanks which only deal with this kind of structure actually take into account a very small part of the phenomenon.
3. The frequency of Vless_Cs is strongly influenced by the context of enunciation (elicited vs. spontaneous): Vless_Cs occur more frequently in elicited dialogues in all three languages. This is definitely due to the specific task required of the speakers, who were forced to engage in frequent question-and-answer exchanges to “spot the differences”. The typical task-oriented dialogues features also affect the higher frequency of occurrence of elliptical structures.
4. The total percentage of Vless_Cs does not present relevant differences across the three languages and types 4 and 5 are the most frequent of all the considered languages. However, differences can be found in the distribution of the other types that deserve further and deeper investigations.

4. Conclusions

The data on the frequency and typology of Vless_Cs in AN.ANA.S._MT suggest some quantitative and qualitative considerations. Quantitative data confirm that in speech Vless_Cs constitute one third of total clauses in all the considered languages: Italian, English and Spanish. In fact, verbless utterances are particularly well-suited to the semiotic and cognitive conditions in which speech naturally takes place, i.e. rapid and on-line linguistic production/reception. This shows that speakers, according to different communicative situations and exigencies, produce a variety of structures that cannot be easily subsumed under canonical sentencehood representation (Voghera, 2008; Progovac et al., 2006). From a qualitative point of view, the partially different distribution of the vary types of Vless_Cs in the three languages suggest the need of further investigations to assess whether it may be attributed to structural linguistic differences or to the difficulty to apply rigorous criteria to distinguish such categories. This is a necessary step to produce the theoretical basis for a richer treebank annotation tagset, which could consider Vless_Cs as proper syntactic objects and not only as reduced forms of canonical sentences.

References

Annotation of English on the tectogrammatical level. http://ufal.mff.cuni.cz/~toman/pedt_manual/ch04s03s01.html#pic232slov6.

Benveniste E. (1950). La phrase nominale. In *Bulletin de la Société Linguistique de Paris*, 41. Also in *Problèmes de linguistique générale*. Paris: Gallimard, pp.151-167.

Biber D., Conrad S. and Leech G. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Böhmová A., Hajič J., Hajičová E. and Hladká B. (2003). The PDT: a 3-level annotation scenario. In Abeille, A., editor, *Treebanks: Building and Using Parsed Corpora.* Dordrecht-Boston-London: Kluwer Academic Publishers, pp. 103-127.

Corpus AN.ANA.S. 3. In *Parlaritaliano.* http://www.parlaritaliano.it/parlare/.

Cresti E. (2005). Enunciato e frase: teoria e verifiche empiriche. In Biffi, M., Calabrese, O. and Salibra, L., editors, *Italia Linguistica: discorsi di scritto e di parlato. Scritti in onore di Giovanni Nencioni.* Siena: Prolagon, pp. 249-260.

Cresti E. and Moneglia M. (editors) (2005). *C-Oral-Rom. Integrated Reference Corpora for Spoken Languages*. Amsterdam-Philadelphia: Benjamins.

Documento di Specifiche Tecniche di SI-TAL Manuale Operativo. http://www.ilc.cnr.it/tressi_prg/papers/Treebank1_1.pdf.

Ferrari A., Cagnetti L., De Cesare A.-M., Lala L., Mandelli M., Ricci C and Roggia E. (2008) *L'interfaccia lingua-testo: natura e funzioni dell'articolazione informativa dell'enunciato*. Alessandria: Edizioni dell'Orso.

Fiorentino G. (2004). Frasi nominali nel parlato dialogico: problemi empirici e teorici. In Albano Leoni, F., Cutugno, F., Pettorino, M. and Savy, R., editors, *Il parlato italiano.* Napoli: D'Auria Editore, CD-ROM [B05].

Giordano R. and Voghera M. (in press). Frasi senza verbo: il contributo della prosodia. In *Atti del Convegno internazionale della Società Internazionale di linguistica e filologia italiane Sintassi storica e sincronica dell'italiano*, Basilea 2008.

Graffi G. (2001). *200 years of syntax : a critical survey*. Amsterdam: Benjamins.

Hengeveld K. (1992). *Non-verbal Predication*. Berlin-NewYork: Mouton de Gruyter.

Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Lenci A., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M.T., Saracino D., Zanzotto F., Mana N., Pianesi F. and Delmonte R. (2003). Building the Italian Syntactic-Semantic Treebank. In Abeille, A., editor, *Treebanks: Building and using Parsed Corpora.* Dordrecht-Boston-London: Kluwer Academic Publishers, pp. 189-210.

Progovac L., Paesani K, Casielles E. and Barton E. (2006). *The syntax of nonsententials: multidisciplinary perspectives*. Amsterdam Philadelphia: Benjamins.

Taylor A., Marcus M. and Santorini B. (2003). The Penn Treebank: an overview. In Abeille, A., editor, *Treebanks: Building and Using Parsed Corpora.* Dordrecht-Boston-London: Kluwer Academic Publishers, pp. 5-22.

Voghera M. (2008). Progettare la grammatica del parlato. In Pettorino, M., Giannini, A., Vallone, M. and Savy, R., editors, *La comunicazione parlata*. Napoli: Liguori, pp. 1696-1714.

Voghera M. and Cutugno F. (in press). AN.ANA.S.: aligning text to temporal syntagmatic progression in Treebanks. In *Proceedings of the fifth Corpus Linguistics Conference*, Liverpool 20-23 July 2009.