Appositions versus Double Subject Sentences – what Information the Speech Analysis brings to a Grammar Debate

Horia-Nicolai Teodorescu[1,2] and Diana Trandabăţ[1,3]

[1] Institute for Computes Science, Romanian Academy
[2] Faculty of Electronics and Telecommunications, Technical University "Gh. Asachi" of Iaşi
[3] Faculty of Computer Science, University "Al. I. Cuza" of Iaşi

hteodor@etc.tuiasi.ro, dtrandabat@info.uaic.ro

Abstract. We propose a method based on spoken language analysis to deal with controversial syntactic issues; we apply the method to the problem of the double subject sentences in the Romanian language. The double subject construction is a controversial linguistic phenomenon in Romanian. While some researchers accept it as a language 'curiosity' (specific only to the Asian languages, but not to the European ones), others consider it apposition-type structure, in order to embody its behavior in the already existing theories. This paper brings a fresh gleam of light over the debate, by presenting what we believe to be the first study on the phonetic analysis of double-subject sentences in order to account for its difference vs. the appositional constructions.

Key words: apposition, double-subject construction, formant analysis, speech-based syntactic disambiguation, prosody.

1 Introduction

Grammatical issues are often controversial and leave space to interpretations. We introduce a new method to help decision in such controversial cases, based on the analysis of speech. The main idea is that two different grammatical structures should have different prosodic interpretations, while instances of a single syntactic structure should have similar correspondences in the speech, all other variables kept constant (speaker, environment etc.). All European languages use appositions to emphasize a specific meaning the speaker wishes to convey. Some languages, like the Japanese, Mandarin, Korean and the Thai languages, use for similar purposes specific constructions, named "double-subject constructions" [5], [6]. (For a detailed analysis of the double subject issue in Asian languages, as well as for an extensive list of references on the topic, see [6]). Such constructions are unknown to most modern European languages, like English or French. In the Romanian linguistic community there has been in recent years a debate on some types of sentences which are considered by several researchers [1], [2] and by us a double-subject construction.

The purpose of this paper is to present a detailed analysis on the contrastive prosodic features of the double-subject sentences and apposition constructions in Romanian. The analysis goes beyond the basic prosody, as represented by pitch values and trajectory; it aims to determine the evolution of higher formants and temporal patterns. After presenting the different approaches to double-subject sentences in Section 2, we discuss the methodology behind the double-subject corpus creation and its analysis: annotation, acoustic parameters determination, etc. The results of the prosodic analysis are presented in Section 4, before drawing some conclusions and indicating some further directions.

2 Double-Subject Sentences in Romanian

The semantic arguments of a predicate (the subject, the direct object and the indirect object) can be doubled, in the Romanian language. While the objects are commonly doubled by clitic pronouns (the doubling is sometimes mandatory, like in *L-am văzut pe Ion,* EN: I saw John), the subjects receive, occasionally, and mainly colloquially, a doubling pronoun (not only in Romanian, as Masahiro [6] shows[1]). The doubling of the subject for the Romanian language is a controversial phenomenon: after having long been considered an apposition, Alexandra Cornilescu [2] has reopened the doubling problem, Verginica Barbu [1] has modeled it using HPSG instruments, but until today, there is no unitary consensus. In this context, supplementary information should be gathered on the specificities of the double-subject constructions contrasted both to the single subject sentences and to sentences which include appositions. Specific phonetic constructions for the three cases would be a significant argument for three independent linguistic constructions. What supplementary information the pronouncing brings, from a descriptive perspective, in double-subject phrases, remains an open question. The present paper partially answers this question.

We provide subsequently a few examples of brief sentences with double subject in the Romanian language. To translate these sentences, we use the symbol Ø to mark the place of the missing doubled subject in the sentences translated in correct English. Examples of sentences with double subject are:

(a) Vine ea mama!
*Comes she mom! [Mom Ø is coming!]
(b) „A trecut el așa un răstimp.” (Sadoveanu M.)
*Passed has it thus a time. [A time has Ø thus passed.]

The first author proposes that the double-subject sentences convey different meanings, depending on the prosody, for example:

- a neutral pronunciation indicates a non-determination of the time interval.
- a pronunciation accentuating the pronoun “el” (EN: he) indicates that the speaker has an idea about the time interval duration, and that the focus is on the passing of that time, and not on the duration.

[1] There is no definite explanation why not all languages accept the double-subject structure. For these languages, in most of the cases, the doubling of the subject is realized as an apposition. The Romanian language has both double subject and apposition structures.

- if the sentence is further developed, it can bring a further specification of the interval. For example, in the development „A trecut el aşa un răstimp de lung, încât..." (EN: A so long time has thus passed, that…), the duration of the interval is specified in a certain way.

(c) O şti el careva cum să rezolve asta.
*would know he someone how to solve this. [He would know Ø how to solve this.]

Different pronunciations may mark either the fact that the speaker does not know who is the person mentioned („el"), either that he knows, but has no intention on telling to the audience (when the accent is on „careva", EN: someone), or clearly specifies, by an apposition, who is envisaged, if the sentence is developed as „O şti el careva, Ion, cum să rezolve asta" (EN: He, John, would know how to solve this). Notice that such a sentence, including both apposition and double subject, is a strong argument in favor of the existence of the double subject constructions as a distinct linguistic structure.

For the examples b) and c), the interpretation is that the information must be partially known by the auditorium (knowledge at the generic level, but not at the level of instantiation with a concrete individuality).

(d) Mama vine şi ea mai târziu.
*Mom is coming also she later. [Also mom is coming later.]

(e) Mama ştie ea ce face.
*Mom knows she what is doing. [Mom knows what she is doing.]

Examples d) and e) are considered by some linguists [1] as constructions with doubled subject, while other authors [2] consider them particular structures of the Romanian language. We intend to compare them to examples a) – c) to see if there are differences in their prosodic realizations.

In this context, we recorded a set of sentences bearing doubled subject for a comparative analysis of the prosody in sentences with doubled/simple subject and appositions, in order to observe the modifications involved by the doubling of the subject. This paper aims to bring clarifications on the change of prosody in double-subject sentences in comparison with simple sentences and appositions.

3 Methodology

A principle we propose and use here is that consistent distinctions at the phonetic level between two specific syntactic constructions reflect and represent an argument to distinguish at the syntactical level between the two constructions. In order to realize a correlation between the semantic charge carried by a sentence and the representation of its subject, the five sentences presented in Section 2 have been recorded by 15 speakers. The database is freely accessible on the web site of the Romanian Sounds Archive [8]. The Romanian Sounds Archive contains over 1000

distinct recordings, available in various accuracy and encoding formats (more methodological aspects are given in [10], in this volume).

Apart the archive itself, the site hosts also documentations regarding the description of the technical modalities and conditions (protocols) involved by the realization of the archive. Namely, the database contains two types of protocols:

- The documentation protocol, which contains the speaker profile (linguistic, ethnic, medical, educational, professional information about the speaker), and a questionnaire regarding the speaker's health, especially concerning the pathologies of the phonating tract.
- The recording protocol, containing information about the noise acceptable values, the microphone, the soundboard, and the corresponding drivers.

3.1 Double-Subject Spoken Database

After subjects have been informed about the objectives of the project, they signed an informed consent according to the Protection of Human Subjects Protocol of the U.S. Food and Drug Administration and to the Ethical Principles of the Acoustical Society of America for Research Involving Human Subjects. The speakers' selection was tributary to the Archive's constraints (the documentation protocol).

The recordings (sound files) corresponding to the simple subject, double-subject, and apposition sentences have been recorded according to the methodology explained in the recording protocol of the Romanian Sound Archive [8]. The recordings were performed using the GoldWave™ application [3], with a sampling frequency of 22050Hz [10].

The speakers[2] have recorded several variants of the five sentences mentioned in Section 2; the sentences have been uttered with neutral tone, accentuation of the doubling pronouns, focuses on the words next the pronouns, and respectively the extension of the sentences.

(a) RO: Vine ea mama! EN: Mom Ø is coming!
(b) RO: A trecut el așa un răstimp EN: A time has Ø thus passed.
(c) RO: O știe el careva cum să rezolve asta.
EN: He would know Ø how to solve this.
(d) RO: Mama vine și ea mai târziu. EN: Also mom is coming later.
(e) RO: Mama știe ea ce face. EN: Mom knows what she is doing.

Corresponding variants of the five mentioned sentences with simple subject and appositions have also been recorded. Every speaker pronounced each sentence three times, following the archive recording protocol (see for details [8]).

[2] Fifteen speakers have been recorded for the double subject analysis. The results discussed in Section 4 consider only seven subjects, subject #1, subject #2, subject #12, subject #13 (female) and subject #5, subject #6, subject #15 (male), selected because they all work in academic/university environment, and should therefore be more familiar with the linguistic structures of the Romanian language.

3.2 Analysis Methodology

We performed the analysis of the double subject in two steps. The first step requires finding and correlating the double sentences parameters with the corresponding simple sentences parameters. The second phase envisages the contrastive analysis between double subject and appositions.

The sentences have been annotated using the Praat™ software [7] at phoneme level. Then, the syllable, word, sentence, subject position, and articulation type level were easily created. After the annotation, the pitch and the formants (F0–F4) are determined for the sentence vowels and semi-vowels. For a determination as precise as possible, a segment of the vowel fulfilling the following conditions is selected:

- The selected segment should be a central area, where there are no transitions of the formants to those of the joined phonemes;
- The formant's frequency should not present big fluctuations. The fluctuations of the formants and their correlation to the double subject will be analyzed in a subsequent stage;
- The formant's contour should not contain interruptions.

Unfortunately, various analysis tools provide different results. This is due to the fact that there is no single definition for these parameters for non stationary signals (as the speech signal is), various tools using different ad hoc definitions. Therefore, we have used several programs, namely Praat™ [7], Klatt analyzer™ [4], GoldWave™ [3] and WASP™ [11] to determine the acoustic parameters. The obtained results, discussed in the next section, use a mean of the values obtained with the four analysis programs.

4. Double-Subject Sentences Analysis

The hypothesis that motivated this analysis is to provide prosodic evidence of the fact that double subject sentences and appositional constructions are two linguistically different phenomena.

We analyzed therefore the values of the formants and duration of the vowels for seven subjects from our database for the constructions:

- Vine mama (EN: Mom is coming) – simple subject
- Vine ea mama (EN: Mom Ø is coming) – doubled subject
- Vine ea, mama (EN: She, mom, is coming) – apposition.

We realize that an analysis over seven subjects can have no claims on generality, but it represents a good start for the pioneering contrastive analysis on the specificities of the Romanian double subject and apposition constructions.

Fig. 1 presents the relative deviation for the sentences "Vine ea mama" vs. "Vine mama" for the seven speakers. The relative deviation σ_r was computed as:

$$\sigma_r = \frac{\Delta F0_k^v}{F0^v} \text{ and } F0_k^v - F0^v = \Delta F0_k^v$$

where v represents each vowel in the sentence, and k each of the seven speakers.

For each subject (1, 2, 5, 6, 12, 13 and 15) we computed the pitch values for double subject sentences (DS) and the corresponding simple subject ones (SS). Thus, the first

bar in graph represents DS_1, the double subject sentence for subject 1, the second gar represents the pitch values for simple subject construction (SS) for subject 1, etc.

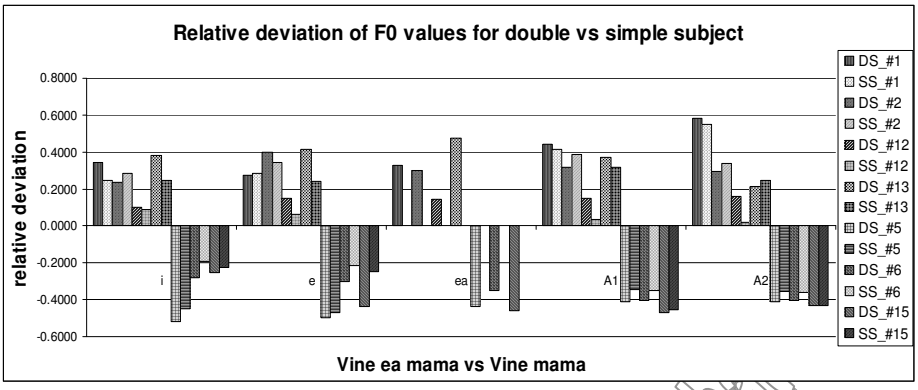

Fig. 1. Pitch values for the double subject and simple subject constructions

When comparing simple subject with double subject constructions, an increasing tendency of the F0 values in the simple subject sentences vs. double-subject sentences was observed. The major differences in the pitch values are visible for the vowels in unaccented syllables [see for details 9].

As for the other formants, it looks that they are fluctuating and carries no double subject information. However, they carry information about the speaker [9]. Also, no significant differences were found in the duration comparison.

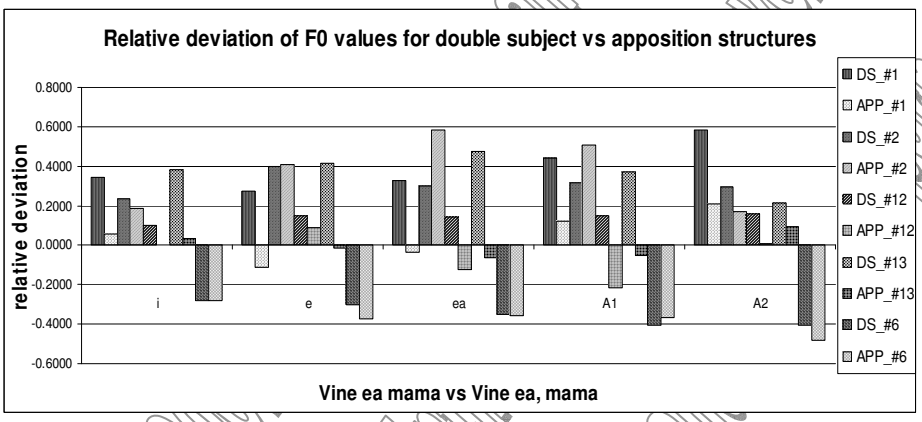

Fig. 2. Pitch values for the double subject and apposition constructions

When comparing apposition structures with double subject structures, we must emphasize that the formant values bring no definite difference. Fig. 2 shows the pitch values for the double subject and the apposition sentence for five of the seven considered subjects. The pitch tendency has no obvious pattern. The data recordings we have annotated and analyzed are not sufficient to draw pertinent statistic conclusions. However, our hypothesis on different patterns for different syntactic constructions is confirmed by the duration of the vowel. Fig. 3 shows the significant

difference between the double subject construction and the apposition. If, in the double subject case, vowel duration is around 0.100s (with some minor exception to the end of the sentence), the sentence containing an apposition bear a strong accentuation of the word the apposition refers to ("ea" in our case). Thus, the duration of the "ea" diphthong is around 0.400s, four times bigger than for double subject. An important observation is that the apposition structures had a very big pause (about 0.400s) before the apposition, corresponding to the comma, break. The comma break was annotated as an individual entity, not as included in the "ea" pronoun or in the "mama" apposition.

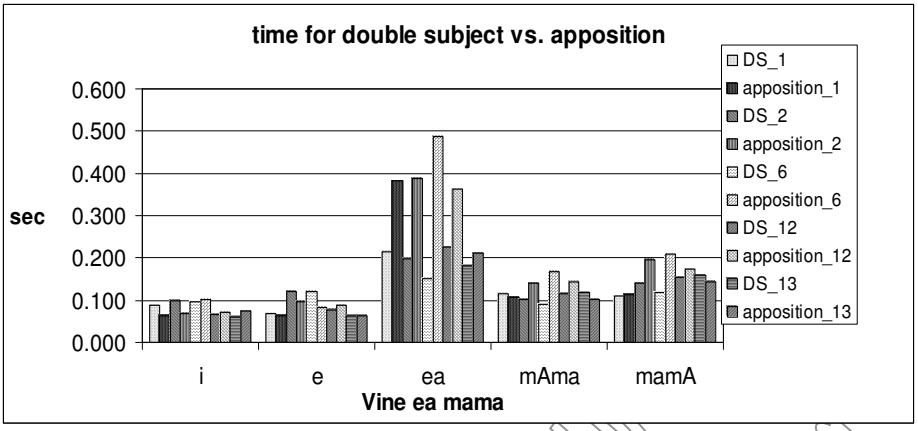

Fig. 3. Duration for the vowels in apposition and double subject construction

After analyzing double / simple subject and apposition constructions, we believe that the hypothesis we have started with is proven. There is a clear difference between the double subject and the apposition constructions. When beginning to pronounce a structure, the speaker has already a prosodic pattern: the pitch contour (higher pitch for simple subject structures, lower values for double subject) or the duration of the vowels (normal for double sentences, more than double for appositions).

5. Conclusions and further work

We have proposed a method to validate hypotheses on the difference between syntactical constructions based on marked differences in the prosody of spoken sentences incorporating such constructions. Specifically, we proposed to use prosodic differences as an argument in deciding when two constructions are different. We have analyzed the influence of the double-subject construction on the prosody in the Romanian language. The analysis involved short sentences which are parallel in the sense that they are identical up to the use of double-subject or apposition constructions.

The main conclusion which can be derived from this preliminary research is that the two syntactic constructions differ in a consistent way from a prosodic point of view. Namely, the word that the apposition explains has duration four times bigger

than normal simple subject sentences or double subject constructions. A second conclusion is that the frequency of the pitch and the central frequency of first formant are different in the two constructions, but both the way of changing and the change amplitude depend significantly on the speaker. These differences represent an argument supporting the existence of double subject construction in the Romanian language – the only Latin, moreover the only non-Asian language exhibiting such a construction.

In the future, we will analyze more recordings in order to confirm these findings and to detect an inter-speaker patterning.

Acknowledgment. This research is part of the Romanian Academy "priority research" topic "Cognitive Systems".

References

1. Barbu, V.: Double subject constructions in the Romanian language. A HPSG perspective (in Romanian), in Aspects of the Romanian language dynamics(in Romanian), vol. II, Ed. Universității Bucharest (2003) 73–79
2. Cornilescu, A.: The Double Subject Construction in Romanian. Notes on the Syntax of the Subject, Revue Roumaine de Linguistique, no. 3–4 (1997) 1–45
3. GoldWave – Audio software, http://www.goldwave.com
4. KPE80 – A Klatt Synthesiser and Parameter Editor, http://www.speech.cs. cmu.edu/comp. speech/Section5/Synth/klatt.kpe80.html
5. Kumashiro, Toshiyuki, Langacker, Ronald W., Double-subject and complex-predicate constructions. Cognitive Linguistics 14–1 (2003) 1–45
6. Masahiro, O.: Analyzing Japanese double-subject construction having an adjective predicate. Proceedings of the 16th conference on Computational linguistics – Volume 2, Copenhagen, Denmark (1996) 865 – 870
7. Praat: doing phonetics by computer, http://www.praat.org
8. Romanian Sounds Archive http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/ index.htm
9. Teodorescu, H-N, Feraru, M, Trandabăț, D.: Studies on the Prosody of the Romanian Language: The Emotional Prosody and the Prosody of Double-Subject Sentences. In Burileanu, Teodorescu (Eds.), Advances in Spoken Language Technology, Romanian Academy Press (2007) 171–182
10. Teodorescu, H-N, Feraru, M.: A study on Speech with Manifest Emotions, Proc. TSD 2007, Lecture Notes in Computer Science, Springer-Verlag (2007) in this volume
11. UCL Phonetics & Linguistics, WASP – Waveform Annotations Spectrograms and Pitch, http://www.phon.ucl.ac.uk/resource/sfs/wasp.htm