AN ACOUSTIC AND DISTRIBUTIONAL APPROACH TO DISFLUENT REPETITIONS IN ROMANIAN SPONTANEOUS SPEECH

Maria Candea[1] & Oana Niculescu[2]
[1]Sorbonne Nouvelle Université, Laboratory Clesthia EA 7345, France
[2]Romanian Academy Institute of Linguistics "Iorgu Iordan - Al. Rosetti", Romania
maria.candea@sorbonne-nouvelle.fr; oeniculescu@yahoo.com,

Studies dealing with disfluencies in spontaneous speech predominantly focus on English data, while among Romance languages most of the research is drawn from well-resourced languages such as French or Spanish [1]. Consequently, in this presentation we propose a preliminary acoustic and distributional analysis of *identical repetitions* (i.e. immediate and identical repeat of spoken material, e.g. 'din din' *[from from]*; IR henceforth) based on Romanian speech data. Phonetic analyses of Romanian spontaneous speech phenomena are still rare, in part due to a lack of available speech aligned corpora. As a result, our pilot study is carried out on 4hs of addressed monologues pertaining to 4 adult native speakers, 2 female (110 minutes), 2 male (130 minutes), 30-40 years of age, without any speech impairments, extracted from a larger Romanian speech corpus recorded and hand-annotated by [2]. To our knowledge, this would be the first applied linguistic research on disfluent repetitions carried out on the only speech aligned corpus available for spontaneous Romanian spoken data.

A total of 629 repetitions as immediate repeats were manually extracted from the corpus, with 72% produced by male speakers and 28% derived from female monologues. Our results show that in 79% of IR, the linguistic format of the repetition was a single word. The remaining tokens consist of multiple word repeats distributed as following: two-word (16%, N = 101), three-word (3.5%, N = 22), four-word (0.6%, N = 4), five-word (0.2%, N = 1), six-word (0.5%, N = 3), and seven-word sequences (0.2%, N = 1). Similar to previous findings on the topic in Romance languages ([3], [4], [5], among others), in over 98% of the extracted data the *reparandum* (RM) was repeated only once, while two *repairs* (RR) surfaced in 1.3% of the cases. There was only one instance of a two-word seven repeat utterance pertaining to a female speaker.

When taking into account the analytic vs synthetic typological parameter, the results for this Romanian corpus differ from prior linguistic data pertaining to other Romance languages as modern French and Spanish. In this context, Romanian occupies an intermediate position, especially within the nominal case morphology [6]. For this reason, *prepositions* represent the most frequent category of repeated function words (31%, prep. 'de' *[of]* being the most common, having multiple functions within the discourse [7] – compared to only 19% in French [4]), followed by *adverbs* (18%, neg. 'nu'*[not]*) and *conjunctions* (14%, 'să' *[to]*). In our data, *determiners* have a lesser frequency within identical repetitions (3% compared to 17% in spoken French [4]).

We observed that, in up to 51% of IR, there is no pause between RM and RR. In our preliminary analysis, we distinguished between short silent pauses (under 200ms) and long silent pauses (above 200ms). When a pause occurs in the *interregnum* (IM) [8], it is more often a long pause (68% a long pause vs 29% a short one), while in the remaining cases we have encountered either two long pauses as well as a combination between short and long pauses (1% each). From a temporal perspective, the median duration of an IR extends to 873ms, with a range of 4305ms (213ms minimum duration and 4518ms maximum duration; see *Table 1* for data related to duration as a function of repetition form and number of repeats). When a pause is present in the IM, the mean duration of the repeat is 1348ms (± 744ms), while the absence of a pause correlates with a decrease in the overall duration of the IR (749ms ± 359ms).

In our study we also focus on the interaction between IR and other disfluencies such as prolongations (found in 75% of the data) and pause fillers (occurring in only 23% of the cases). We document the frequency, distribution and acoustic correlates of these DFs in connection to IR present within the Romanian speech corpus under investigation.

While some findings appear to be language independent (related to speech planning phenomena), others are language specific (due to typological differences in connection to IR output), individualising Romanian in the context of Romance languages.

	1 word	2 words	3 words	4 words	5 words	6 words	7 words
Frequency	497	101	22	4	1	3	1
Mean ± Std.	925 ± 524	1,286 ± 767	2,079 ± 903	1,670 ± 240	1,479 ± NaN	2,527 ± 1,092	3,812 ± NaN

Table 1a. *Sequence duration (ms) as a function of repetition form*

	1 repeat	2 repeats	7 repeats
Frequency	620	8	1
Mean ± Std.	1,032± 646	1,420± 391	3,543 ± NaN

Table 1b. *Sequence duration (ms) as a function of number of repeats*

References

[1] Trandabăț, D., E. Irimia, V. B. Mititelu, D. Cristea, and D. Tufiş (2012). *The Romanian language in the digital age*, In G. Rehm & H. Uszkoreit (eds.), META-NET White Paper Studies, Berlin: Springer, 1-87.

[2] Niculescu, O. (2021). "Developing linguistic resources for Romanian written and spoken language", In Rebreja, P., Onofrei, M., Cristea, D., Tufiş, D. (eds.) *Proceedings of the 16th International Conference „Linguistic Resources and Tools for Natural Language Processing"*. Iaşi: Editura Universității „Alexandru Ioan Cuza", 21-36.

[3] Candea, M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané. Étude sur un corpus de récit en classe de français*. [Unpublished PhD Thesis]. Université Sorbonne Nouvelle – Paris III.

[4] Dister, A. (2014). «parler sans accent pour moi c'est sans sans sans bafouiller» Quelles répétitions de formes en français parlé?, In *SHS Web of Conferences*, EDP Sciences, VIII, 2017-2031.

[5] Moreno-Sandoval, A., L. Campillos Llanos, D. T. Toledano (2012). "A quantitative study of disfluencies in formal, informal and media spontaneous speech in Spanish", In *Proceedings of IberSPEECH 2012*. VII *Jornadas en Tecnología del Habla*. Madrid, 21st-23rd November 2012, 164-173.

[6] Pană Dindelegan, G. (2013). "Flexiunea cazuală - între analitic şi sintetic", In *Limba Română*, II, 159-173.

[7] Mardale, A. (2010). *Les prépositions fonctionnelles du roumain : Etudes comparatives sur le marquage casuel*, Paris : L'Harmattan, ISBN: 9782296241268

[8] Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*, Doctoral dissertation, University of California, Berkeley.