INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The Unit of Speech Encoding: the case of Romanian

Irene Vogel[1], *Laura Spinu*[2]

[1]University of Delaware
[2]University of Western Ontario
ivogel@udel.edu, lspinu@uwo.ca

Abstract

The number of units in an utterance determines how much time speakers require to physically plan and begin their production [1]-[2]. Previous research proposed that the crucial units are prosodic i.e., Phonological Words (PWs), not syntactic or morphological [3]. Experiments on Dutch using a prepared speech paradigm claimed to support this view [4]-[5]; however, compounds did not conform to predictions and required the introduction of a different way of counting units. Since two PWs in compounds patterned with one PW, with or without clitics, rather than a phrase containing two PWs, a recursive PW' was invoked. Similar results emerged using the same methodology with compounds in Italian [6], and it was thus proposed that the relevant unit for speech encoding is not the PW, but rather the Composite Group (CompG), a constituent of the Prosodic Hierarchy between the PW and Phonological Phrase that comprises both compounds and clitic constructions [7]. We further investigate the relevant unit for speech encoding using the same methodology in Romanian. Similar findings support the CompG as the speech planning unit since, again, compounds with two PWs pattern with single words and clitic constructions, not Phonological Phrases which also contain two PWs.
Index Terms: speech encoding, Phonological Word, Composite Group, Romanian

1. Introduction

The articulation of an utterance is planned in groupings or units of speech, the number of such units correlating with the amount of time needed to begin speech production. Unsurprisingly, for example, an utterance with more syllables takes longer to encode than one with fewer syllables (e.g., $u_{\sigma_1}ran_{\sigma_2}gu_{\sigma_3}tan_{\sigma_4}$ vs. tan_{σ_4}) (e.g. [8]). It appears, however, that the number of higher level units is also relevant, so *urangutan* would take less time to encode than a phrase with the same number of syllables (e.g., $the_{\sigma_1}man_{\sigma_2}in_{\sigma_3}green_{\sigma_4}$). The question is thus which units are relevant to the planning time of utterances. It should be noted that we are referring here to the process involved in physically producing, or articulatorily encoding, speech, as opposed to processes associated with the activation or retrieval of lexical items. In this context, Levelt [3] proposed that the crucial type of unit is the Phonological Word, and it was argued that Dutch experimental data supported this view [4]-[5]. In fact, however, the situation was more complicated since, unexpectedly, compounds containing two Phonological Words behaved differently from phrases containing two Phonological Words, patterning instead like single words and clitic constructions. Based on a similar pattern in Italian, it was proposed that the relevant planning unit is instead the Composite Group, a constituent in the prosodic hierarchy between the Phonological Word and Phonological Phrase (e.g., [7]), in the position of the Clitic Group in [9].

Given that there is substantial controversy surrounding this intermediate constituent from the theoretical perspective, behavioral evidence contributes additional insight into the question. We thus investigate a range of word and phrase structures in Romanian to determine whether the earlier findings are more broadly generalizable, and thus further support the role of the Composite Group in the prosodic hierarchy.

1.1. Phonological Units of Encoding in Dutch

In the prepared speech paradigm developed by Wheeldon and Lahiri [4], participants saw a word or phrase on a computer screen and then had to use it in response to a generic question on the next screen. This method ensured that the process being tested was specifically the physical planning or encoding of the response, independently of potential effects of selection of the words themselves. When the question appeared, it was accompanied by three beeps at variable intervals, and the participants were instructed to begin their response once they had heard the third beep. That is, the participants knew in advance what answer they needed to provide, but they could only begin to encode it for production after the last beep. Response times were measured from the last beep to the onset of speech. All responses began with *Ik zoek* 'I seek' and were followed by one of the types of structures in Table 1; type (vi) is an additional structure included in a follow-up study [5].

Table 1: *Prosodic structure and examples of Dutch Stimuli [4]-[5].*

Type	Example	Gloss
i. Clitic + Word	het water	'the water'
ii. Phrase	vers water	'fresh water'
iii. Word	water	'water'
iv. Full Pronoun	het	'it'
v. Null (= control)	ø	ø
vi. Compound	oog lid	'eyelid'

In Wheeldon and Lahiri [5], both stress-initial and stress-final words were included in type (iii); however stress position turned out to be irrelevant.

Synthesizing the two studies, it was predicted that since types (i), (iii) and (iv) contain one Phonological Word, they would all require longer response times than type (v). Moreover, since types (ii) and (vi) contain two Phonological Words, they were expected to have similarly longer response times than those with one Phonological Word. Instead, it was found that type (vi) (compounds) behaved like the types with one Phono-

http://dx.doi.org/10.21437/Interspeech.2016-1601

logical Word. The explanation given was that while compounds contain two Phonological Words, they must be considered a single (recursive) Phonological Word' in order to maintain the claim that it is the Phonological Word that serves as the unit for speech encoding. This begs the question, however, of how structures with the same number of Phonological Words could exhibit distinct timing patterns, if the relevant unit is in fact the Phonological Word. Moreover, the only account for why the various constructions with and without clitics, on the one hand, and the compounds, on the other hand, exhibit the same timing patterns rests on the fact that the former contain a single Phonological Word regardless of any other content, and the latter are considered to constitute a single Phonological Word due essentially to their morpho-syntactic structure as a single lexical item, despite the presence of two internal Phonological Words.

1.2. Phonological Units of Encoding in Italian

The same prepared speech paradigm was used to test the timing patterns of Italian, a language with quite different prosodic properties [6]. Specifically, the response times were measured for encoding each of the four types of items in Table 2.

Table 2: *Prosodic structure and examples of Italian Stimuli [6].*

Type	Example	Gloss
i. Underived Word	fazzoletto	'handkerchief'
ii. Derived Word	fidanzato	'fiancé'
iii. Compound	ficcanaso	'busybody'
iv. Phrase	faccio nodi	'(I) make knots'

Both underived and derived single words were included to examine whether the internal morphological structure of words affected their encoding times; no difference was found.

Although clitics are common in Italian, they were not examined in this study, as it was more specifically designed to investigate the behavior of compounds, to determine if they pattern with single words or with phrases in terms of encoding times. That is, given that a compound constitutes a Composite Group, as does a single word, if the two structures require the same, shorter, encoding times, this would indicate that the relevant unit is the Composite Group, since the number of Composite Groups is the same (i.e., one). If, however, compounds require longer encoding times, like phrases, this would indicate that the relevant unit is the Phonological Word, since compounds and phrases have the same number of constituents (i.e., two).

The findings showed that the phrase response times were significantly longer than those of the other stimulus types, as in Dutch. This pattern thus provided further support for the proposal that encoding time does not depend on the number of Phonological Words, but rather on slightly larger constituents, specifically Composite Groups.

2. Romanian Study

To further investigate the encoding times of different types of constructions that would compare the suitability of the Phonological Word versus Composite Group as the prosodic unit of speech planning, we tested additional types of structures in Romanian using the same experimental paradigm. Moreover, this experimental study provided an independent means of assessing the theoretical claims advanced regarding the Composite Group in Romanian [10].

2.1. Experiment

2.1.1. Hypotheses

Specifically, we tested the prediction that encoding time would depend on the number of Composite Groups rather than Phonological Words in a string, as formulated in the following hypotheses:

- Hypothesis 1: Constructions consisting of two Composite Groups (= slow group) will take longer to encode than those consisting of one Composite Group (= fast group).
- Hypothesis 2: Different constructions consisting of one Composite Group (= fast group) will show similar encoding times regardless of the number of Phonological Words they contain.

2.1.2. Participants

The speech of twelve native speakers of Romanian was analyzed in this study. The recordings of some other participants were excluded from the analysis since they included large numbers of inappropriate responses, as described below. All of the participants were university educated speakers of the standard variety of Romanian, between the ages of 21 and 59 years (mean = 38).

2.1.3. Stimuli

The stimuli were grouped into five categories as shown in Table 3: (i) individual words, used as the baseline for single Phonological Word/Composite Group timing, (ii) compounds, (iii) phrases, (iv) verb-clitic constructions, and (v) verb-clitic-clitic constructions. The prosodic structure for each stimulus type is also provided.

Table 3: *Prosodic structure and examples of Stimuli. Phonological Word = ω; Composite Group = κ.*

Type	Example/Pros. Str.	Gloss
i. Word	[[piersicuţă]$_\omega$]$_\kappa$	'little peach'
ii. Compound		
N-A	[[burtă]$_\omega$ [verde]$_\omega$]$_\kappa$	'careless'
V-N	[[pierde]$_\omega$ [vară]$_\omega$]$_\kappa$	'lazy'
N-P-N	[[cal]$_\omega$ de [mare]$_\omega$]$_\kappa$	'seahorse'
iii. Phrase		
N-A	[[burtă]$_\omega$]$_\kappa$ [[plină]$_\omega$]$_\kappa$	'full belly'
V-N	[[pierde]$_\omega$]$_\kappa$ [[timpul]$_\omega$]$_\kappa$	'(he) wastes time'
N-P-N	[[cal]$_\omega$]$_\kappa$ de [[piatră]$_\omega$]$_\kappa$	'stone horse'
iv. One clitic		
V-CL	[[pierzându]$_\omega$ ne]$_\kappa$	'losing us'
v. Two clitics		
V-CL-CL	[[prinde]$_\omega$ ţi le]$_\kappa$	'attach-you-them'

There were 12 items in each of the main category types. Since Romanian has multiple types of compounds, rather than limiting the experiment to one type, three common types were used. We thus also included the same three types of phrases to ensure that the responses to the compounds and phrases were based on similar types of strings. In fact, examination of the results of the subtypes did not reveal significant differences for either the compounds or phrases; see Section 2.3.

All of the items contained four syllables, and were matched to the extent possible for segmental and syllabic structures. Stress was on the penultimate syllable in all cases except for the clitic constructions, where stress fell to the left of the clitics, on the verb.

2.1.4. Procedure

Each subject was tested individually by a native speaker of Romanian in Bucharest. The experiment was conducted using e-prime software [11], and all of the speakers received the stimuli in a different random order. All tests were conducted in a quiet room; however, it was not always possible to eliminate external distractions, a point we return to below.

As in the Dutch and Italian experiments, in both the training session (with different items) and the actual experiment, the construction (i.e. word, compound, phrase, verb-clitic or verb-clitic-clitic) to be read aloud appeared on the computer screen for 1 second. The next screen presented the prompt, that is, a + sign indicating that the subject should prepare to produce the string s/he had just seen. The stimuli were then provided and the participant produced them as they appeared. Differently from the Dutch and Italian experiments, the responses were not uttered in a carrier sentence.

To avoid anticipatory responses, when the + appeared on the screen, a series of 3 beeps at variable intervals was also initiated. The speaker was instructed to wait until the third beep was heard before beginning to produce the response. The time between the third beep and onset of speech was recorded. If no speech was detected within the first 2 seconds after the third beep, the program automatically moved on to the next item.

2.1.5. Questionnaires

It is possible that the frequency and/or naturalness of the stimuli might affect their response rates, so to the extent possible, the stimuli included in the experiment contained common, well-known words. We could not verify this directly, however, since there are no reliable sources for word frequency in Romanian. Moreover, determining frequency in a language with a rich inflectional system poses substantial problems, as discussed in Vogel and Wheeldon [6] in relation to Italian.

Since we were not able to control for the familiarity or naturalness of our stimuli, we collected additional information from the participants at the end of the experiment, in case it was relevant to the findings. Specifically, we administered a questionnaire in which we asked the participants to indicate the age at which they believed they learned each target word (i.e., the lexical items that appeared alone and in phrases, as well as the compounds), choosing from the following age categories: 1-3, 4-6, 7-10, 11-14, 15-18, over 18. This information was taken as an indication of the familiarity of the items since age of acquisition, like frequency, has been documented to have effects on adult lexical processing, including speech recognition [12], speech production [13], and latency performance on naming and lexical decisions tasks [14]. The clitic constructions could not be assessed in terms of acquisition, so for these items, we asked the participants to rate them on the basis of a naturalness scale of 1 (least natural) to 5 (very natural).

2.2. Analysis

The response times (RTs) recorded by the e-prime software were analyzed for all speakers. As in the previous investigations, we excluded individual measurements that were initiated before the third beep or too quickly after it (i.e., RT less than 100 ms.) as well as those that were timed out (i.e., took longer than the maximum of 2 seconds). Based on the questionnaires, all of the lexical items were well known by the participants, and we found no effect due to the average age at which participants indicated that they had learned them. By contrast, there were three items involving clitics (1 verb-clitic, 2 verb-clitic-clitic) that we excluded based on a combination of atypically long RTs and low naturalness scores on the questionnaires.

A more challenging situation that arose with analyzing the data was a consequence of the less than ideal experimental facilities. As mentioned, even though testing was done in a quiet environment, it was not always in the same location, and it was not always possible to completely exclude distractions. Moreover, it was clear at the time of testing (carried out by one of the authors) that in some cases the participants were distracted and were not fully attending to the task. Since response times crucially depend on the participants' consistent concentration on the task, we did not wish to include unreliable results in the analysis. Rather than subjectively determine which participants were felt to have been distracted, we developed a response-based criterion to identify the participants who were most likely not to have been fully attending to the task. Several participants were found not to show any significant differences among the response times to the various types of constructions, but the specific characteristic we considered grounds for excluding a participant's responses was a failure to show a statistical distinction between the single word and the phrasal categories. This response pattern was considered inappropriate, since no model of speech production predicts that single words and multiple word phrases will be treated identically with regard to speech encoding times.

The results presented here are thus based on our analysis of the data of the twelve participants who were deemed to be reliable as far as focus or attention to the task was concerned. In our analysis, we used a repeated measures ANOVA to explore whether there was a significant main effect of Construction Type on Reaction Time. To find out specifically which of the means (i.e., for the construction types word, compound, phrase, verb-clitic, and verb-clitic-clitic) were significantly different from each other, we used a Bonferroni post-hoc analysis.

2.3. Results

Table 4 shows the mean reaction times in milliseconds, measured as the time between the third beep and onset of speech, and standard deviations for the five construction types. Figure 1 represents the means in a graph format.

Table 4: *Means and standard deviations for each construction type (in milliseconds).*

Category	Mean	St. Dev.
Word	560	208
Compound	558	211
Phrase	631	226
Verb-clitic	578	210
Verb-clitic-clitic	585	223

As can be seen, the (single) word and compound categories have the shortest reaction time (mean 560 ms and 558 ms, respectively) and the phrase category displays the longest (631 ms). The remaining two categories, verb-clitic and verb-clitic-clitic fall in between these two 'poles', with mean reaction times of 578 ms and 585 ms; however, they are much closer in their timing patterns to the words and compounds than to the phrases.

In the repeated-measures ANOVA, there was a significant main effect of construction type on response latencies, $F(4, 44) = 5.62$, $p=0.01$. Moreover, the post-hoc comparisons showed

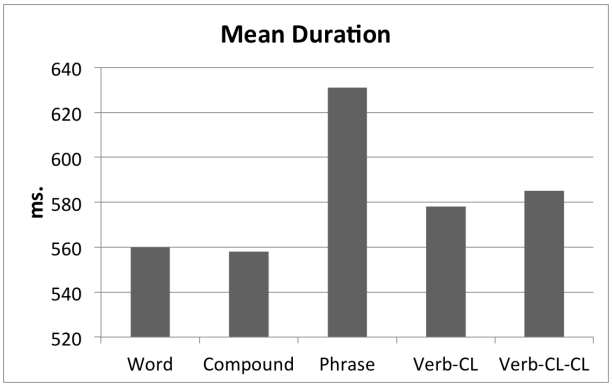

Figure 1: *Reaction times for for each construction type (n=12).*

that the phrase category differed significantly from all other categories. No other significant differences were found.

As was seen in Table 1, where we provide examples of the stimuli, there are different subtypes within the groups of compounds and phrases. To be sure that the internal composition of the stimuli did not result in different timing patterns, we also compared the mean reaction times by prosodic structure (N-A, V-N, N-P-N) within the Compound and Phrase categories. No significant differences were found among these sub-groups, indicating that the compounds and phrases were treated as unified categories independently of the characteristics of their components.

Taken together, the results fall into the same general categories that were observed in the previous investigations. That is, a clear distinction emerged between short and long RT groups; there were no responses with a null element to serve as the "shortest" time measure, as in the Dutch structure type (v) in Table 1. Thus, in the short RT group, we find those structures that consist of one Composite Group: individual words, compounds (composed of two Phonological Words) and constructions with a verb and one or two clitics. Within this short RT group, moreover, the different constructions consisting of one Composite Group showed similar encoding times regardless of the number of Phonological Words or other material they comprised. By contrast, the long RT group contained only the phrases, those structure consisting not only of two Phonological Words, but of two Composite Groups.

3. Discussion

Our results support both of our hypotheses, and as such, they are also consistent with the encoding patterns observed previously for Italian and Dutch. The significant difference in encoding time for single words (i.e., the clear case of one Phonological Word and one Composite Group), as compared to two-word phrases (i.e., the clear case of two Phonological Words and two Composite Groups), established, first that the experiment revealed a direct connection between prosodic structure and speech encoding time. In addition, the category that crucially tests the role of the Phonological Word vs. the Composite Group as the appropriate prosodic constituent for speech encoding, compounds, confirmed that the latter is the relevant unit. That is, since compounds contain two Phonological Words, but only one Composite Group, they demonstrated that the encoding time was predicted by the latter, not the former. Specifically, it was seen that the response times for the various compound types were essentially the same as those for the single words, even though they differ in the number of internal Phonological Words. If the number of Phonological Words is what determines the amount of time needed for encoding, the compounds would have instead behaved like the phrases, with two Phonological Words. In addition, the two types of clitic constructions showed similar reaction times to the single words and compounds; the slightly longer RTs with the clitic constructions were not statistically different from the others in the short RT group. While it is predicted by a Phonological Word count that the clitic constructions would show similar behaviors to single words, without the Composite Group, it is not predicted that these constructions would also show similar behaviors to compounds.

At first glance, it appears that there is a possible alternative explanation for the similar behavior of compounds and individual words. That is, it could be argued that since both constitute a single lexical item, the similar encoding times are dependent on the lexical item count. While such an account could describe the patterns observed in our findings, as well as those of the Dutch and Italian experiments, it runs counter to all the previous research that has argued, following Levelt [3], for prosodic constituents being the units relevant to speech encoding, not morpho-syntactic units. Moreover, while the number of lexical items could also account for the clitic construction RTs, it does not provide a positive explanation for why clitics constructions behave in the same way as single words and compounds. That is, simply ignoring clitics or other function words in calculating response times on the grounds that they are not lexical items leads to the observed timing patterns; however, the definition of the Composite Group as containing a lexical item or compound, as well as any stray material such as clitics and function words, specifically predicts that all of these types of constructions will show similar timing patterns.

4. Conclusions

In sum, we considered the suitability of the Composite Group and the Phonological Word as the relevant prosodic unit of speech planning based on an experiment with a variety of constructions in Romanian. To permit comparison with previous studies, we employed the same type of experimental paradigm using prepared speech that was previously employed to assess the timing units in Dutch and Italian [4]-[5]; [6]. Crucially, this paradigm specifically addresses the mechanism involved in the physical production or articulatory encoding of speech, as opposed to the process associated with the activation or retrieval of lexical items.

Our hypotheses to the effect that the Composite Group is in fact the relevant unit for speech encoding were confirmed since we found that structures consisting of two Composite Groups took significantly longer to encode than constructions with one Composite Group, regardless of the internal composition of the latter – even when this involved two Phonological Words in the case of compounds. Thus, despite other prosodic differences among Romanian, Dutch and Italian, the fact that all three languages revealed analogous encoding patterns for words, compounds, phrases and clitic constructions, provides additional evidence to the growing body of research, both theoretical and experimental, that demonstrates the need for, and validity of, a constituent (i.e., Composite Group) in the prosodic hierarchy between the Phonological Word and the Phonological Phrase.

5. References

[1] Sternberg, S., Monsell, S., Knoll, R. L. and Wright, C. E. 1978. "The latency and duration of rapid movement sequences: comparisons of speech and typewriting". *Information processing in motor control and learning*, ed. George E. Stelmach, pp. 117–152. New York: Academic Press.

[2] Sternberg, S., Wright, C. E., Knoll, R. L. and Monsell, S. 1980. "Motor programs in rapid speech: Additional evidence". *The perception and production of fluent speech*, ed. Ronald A. Cole. 507–534. Hillsdale, NJ: Erlbaum.

[3] Levelt, W. J. M. 1989. "Speaking: from intention to articulation". Cambridge, MA: MIT Press.

[4] Wheeldon, L. and Lahiri, A. 1997. "Prosodic Units in Speech Production". *Journal of Memory and Language* 37, pp. 356–381.

[5] Wheeldon, L. and Lahiri, A. 2002. "The minimal unit of phonological encoding: prosodic or lexical word". *Cognition* 85. pp. B31– B41.

[6] Vogel, I. and L. Wheeldon. 2010. "Units of speech production in Italian". In S. Colina, A. Olarrea and A.M. Carvalho (eds.) *Romance Linguistics 2009*. Philadelphia: John Benjamins. pp. 95–110.

[7] Vogel, I. (2009). "The Status of the Clitic Group". In J. Grijzenhout and B. Kabak (eds.) *Phonological Domains: Universals and Deviations*. Berlin: Mouton de Gruyter, pp. 15–46.

[8] Meyer, A. S., Roelofs, A. and Levelt, W. J. M. (2003). Word length effects in object naming: The role of a response criterion. *Journal of Memory and Language*, 48(1), 131–147.

[9] Nespor, M. and I. Vogel (1986/2007). "Prosodic Phonology". Dordrecht: Foris.

[10] Vogel, I. and L. Spinu (2009). "The domain of palatalization in Romanian". In P. J. Masullo, E. O'Rourke and C.H. Huang (eds.) *Selected Papers of the 37th Conference on Linguistic Studies of Romance Languages*. Philadelphia: John Benjamins, pp. 307–320.

[11] Schneider, W., Eschman, A., and Zuccolotto, A. 2001. "E-Prime Reference Guide". Pittsburgh: Psychology Software Tools, Inc.

[12] Garlock, V. M., Walley, A. C., and Metsala, J. L. (2001). "Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults". *Journal of Memory and language*, 45(3), pp. 468–492.

[13] Ellis, A. W., and Lambon Ralph, M. A. (2000). "Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5), pp. 1103–1123.

[14] Cortese, M. J., and Khanna, M. M. (2007). "Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words". *The Quarterly Journal of Experimental Psychology*, 60(8), pp. 1072–1082.