Word stress processing integrates phonological abstraction with lexical access – an ERP study

Karolina Broś[ab], Martin Meyer[d], Maria Kliesch[bc] and Volker Dellwo[b]

[a] Neurolinguistics Lab, Institute of Applied Linguistics, University of Warsaw

[b] Phonetics and Speech Sciences Group, Institute of Computational Linguistics, University of Zurich

[c] Zurich Center for Linguistics, University of Zurich

[d] Department of Psychology, University of Zurich

Abstract

It is unclear whether word stress in a language is stored as part of the word or whether it is generated by a rule. We test the generativist hypothesis of lexical storage stating that only unpredictable stress is stored in long-term memory against the contrasting usage-based approach assuming that all phonetic information regardless of its (un)predictability is stored in the mental lexicon together with the word. In a correctness judgment task involving correctly and incorrectly stressed penults and antepenults, we found that incorrectly stressed penults do not evoke an N400 effect, whereas incorrectly stressed antepenults do: there is increased negativity with a peak latency around 350-600 ms from word onset. Only changes to words with exceptional stress cause lexical inhibition, hence exceptional but not default stress markers are stored in the lexicon. Additionally, differences in processing patterns between the N400 and the late positivity component window point to an integration of two stages of word processing: pre-lexical stress recognition and stress-to-meaning matching. The results of the study support the view that stress should be understood as abstract phonological information.

Keywords: Spanish stress, stress perception, lexical inhibition, N400, lexical storage, exemplar theory, generative phonology

Corresponding author:

Karolina Broś, PhD
k.bros@uw.edu.pl
Neurolinguistics Lab
Institute of Applied Linguistics
University of Warsaw
Dobra 55, 00-312 Warszawa

1. Introduction

Word stress processing is a complex mechanism that has been subject to a vigorous debate in both psychological and linguistic circles (e.g. Cutler et al., 1997; Dupoux et al., 2001; van Donselaar et al., 2005; Domahs et al., 2014; Rahmani et al., 2015). In language, stress can be linked to semantic information. Word meanings in many languages are often differentiated merely by stress. It is also used as a cue in segmenting speech into words, which has been demonstrated in studies on both adults and infants (Norris et al., 1997; Jusczyk et al., 1999; Hanulíková et al., 2010). Phonetically, it is associated with several acoustic cues marking the length, tone and intensity of the stressed vowel. In contexts encompassing more than individual words, however, it is difficult to disentangle stress from accent (i.e. phrasal prominence). At the same time, stress seems to be processed somewhat differently in perception than in production. Numerous studies report that different cues seem to be crucial for the realisation of stress than those extracted from the signal in auditory processing (compare e.g. Ortega-Llebaria & Prieto, 2007; Ortega-Llebaria & Prieto, 2009 and Torreira et al., 2014 for Spanish).

In this paper we asked whether stress is abstract phonological information or whether it is stored as part of a lexical item. The processing of stress as an abstract category means that stress in e.g. the word 'category', which falls on the first syllable, is not purely phonetic, and hence changeable from speaker to speaker and from one communication situation to another. Speakers of English have a mental representation of stress extracted from this phonetic information. Thus, if they hear the word 'category' stressed other than on the first syllable, they perceive it as incorrectly stressed because it is against the stress assignment rules pertaining to English. Additionally, in some cases, correct stress assignment will play a role in differentiating meanings or grammatical categories, e.g. *project* (noun with stress on the first syllable, verb with stress on the second). The question of storage concerns the lexical status of stress, i.e. whether the stress is permanently attached to a word and its meaning, and hence memorised, or whether it is derived in word processing because it is predictable to some degree and subject to the grammatical rules of a given language. The answer to this question will depend on the language, but most importantly on whether stress is necessary to recognise a given word and its meaning or constitutes additional, purely prosodic information. This is strictly connected with two leading approaches to lexical storage in the field of linguistics. On the one hand, generative phonology (Chomsky & Halle, 1968) assumes that only unpredictable information is stored in the lexicon (i.e. memorised), which makes material derivable by rules, such as predictable stress, redundant. On the other hand, usage-based models (Bybee, 2001; Pierrehumbert, 2001) assume a rich lexicon in which all phonetic and sociolinguistic detail concerning a given word is memorised together with its meaning. As a consequence, different predictions on the status and semantic importance of stress are made. With this in mind, we decided to test the generativist hypothesis *vis à vis* the usage-based model in the course of a neurophysiological experiment measuring event-related potentials (ERPs), to advance our knowledge concerning the link between phonology and semantics in the brain. The well-established N400 is the component of interest in this study as it is closely related to meaning extraction in speech processing (Kutas & Hillyard, 1984).[1]

As a testing ground, we used Spanish, which is a language with variable yet unevenly distributed stress. It shows a great prevalence of penults (words stressed on the penultimate, i.e. second-to-last syllable) compared to antepenults (stressed on the antepenultimate, i.e. third-to-last syllable) and items with final syllable stress. This restricted variability gives us two crucial features: robust perception and discrimination of stress by native speakers derived

[1] Although not exclusively, confer e.g. Frisch & Schlesewsky (2001), Hinojosa et al. (2005), Bornkessel-Schlesewsky & Schlesewsky (2008). See also Friederici's (2004) review paper.

from direct language experience, and contrastive linguistic behaviours of words with the prevalent penultimate stress compared to the less frequent patterns. As a result, we have a default penultimate stress in the language, and lexical exceptions. In this paper, we used the contrast between words with penultimate and antepenultimate stress to gain access to the phonology-semantics interface, i.e. retrieval of meaning based on auditory input forms, and to observe disruptions caused by shifting stress to an incorrect position. The comparison of listener responses to the two contrastive word types let us decide whether stress is part of the phonological system or is stored in memory as part of the phonetics of a word.

1.1. Theoretical approaches to phonological processing

As signalled above, generative phonology (Chomsky & Halle, 1968; see Kenstowicz & Kisseberth, 1979) involves the generation of attested outputs from abstract mental representations acquired in the process of learning the first language. This knowledge is stored in memory as a set of rules. Various models belonging to this approach share the assumption that predictable information that can be derived by phonological rules is not stored in the mental lexicon. All phonetic detail that is not contrastive (does not differentiate meaning) or that is redundant for the processing of a given word as a stand-alone linguistic unit is therefore excluded from the long-term representation of that word. By contrast, information that is unpredictable and cannot be derived by rules is lexicalised, which means that it has to be stored in long-term memory in the form of phonemes, stress markers and other phonological features. Generative phonology is therefore a theory of abstraction and generalisation. Most importantly for this paper, stress is conceived of as a construct based on a variety of phonetic cues (usually F0, duration, and intensity) which can be either derived by a rule or acquired and stored as part of the phonological representation of the word. In languages such as French, where the word-final position of stress is fully predictable, there is no need to learn stress as a contrastive unit. Consequently, speakers of languages with predictable stress have difficulties differentiating functional stress in contrastive stress languages, a phenomenon often referred to as stress-deafness (Dupoux et al., 1997, 2001; Peperkamp et al., 2010). By contrast, in languages such as German or English, stress can fall on any syllable of the word which makes prediction more difficult. Thus, stress rules have to be acquired and some morphemes have to be lexically marked for stress.

The second approach to phonological processing draws on the theory of exemplars dating back to Semon's (1923/1909) Mnemic Psychology (see also Nosofsky, 1988). Here, the focus is on the actual language use and the effect of frequency and other external factors on sound production and perception. As developed by linguists, exemplar theory (henceforth: ET; Bybee, 2001, 2006) argues against abstract, phonemic representations of words or morphemes. The categoricity of the generative model is abandoned in favour of gradient, lexically diffuse differences in pronunciation which are all stored in the mental lexicon. Based on the assumption that human memory capacity is much richer than linguists might have anticipated, and on the observation that language in use shows much variation and not all observed patterns can be derived by rules, ET postulates that each instantiation of a given word or phrase (exemplar) is stored in memory alongside hundreds of other pronunciations of the same item without any computation or abstraction mechanism.[2] However, the more frequently a given word is heard or produced, the more its representations are strengthened in the lexicon and the easier it is to access. Also, it is more probable that it will undergo some

[2] The proponents of ET refer to generalisation mechanisms but of a different kind: lexemes are stored in associative networks and are categorised in a way that some associations between the phonetics and word meanings are stronger while others weaker. Additionally, some authors refer to *schemas* (lexical connections) based on analogy (Bybee, 2001).

linguistic process. It follows from ET that stress cannot be a derived or abstract category. Instead, it is a bundle of acoustic and auditory features stored with each word.

1.2. Default and exceptional stress in Spanish

As already mentioned, Spanish is a language with free (or variable) stress. In spelling, deviations from the penultimate stress norm (see below), are typically marked by a diacritic. There are numerous minimal pairs or even triplets of words differing only in stress, which suggests that speakers must at least partially learn and store stress information as part of their word memory. Some examples include the word *limite* ('to limit'; penult, 3rd p. present subjunctive), *limité* (final, 1st p. past tense), *límite* (antepenult, sg. noun) or the noun pair *sabana* 'savanna' (penult) – *sábana* 'bed sheet' (antepenult). Nevertheless, the accentual pattern of individual lexical units is predictable to a large extent from either word/syllable or morphological structure. There are also important statistical differences in the occurrence of each attested stress pattern. Traditional accounts (Harris, 1969; Quilis, 1981; Roca, 2006) postulate that Spanish stress is limited to the so-called three-syllable window (final, penultimate, antepenultimate). In the case of nouns and adjectives, a great majority of consonant-final words have final stress, and a great majority of vowel-final words have penultimate stress. Morales-Front (1999, 2014) showed that 64.2% of all the Spanish words are stressed on the penultimate syllable, while antepenults constitute merely 8.09% and should be considered exceptional. Earlier sources provide an even greater discrepancy (78.9% vs. 2.76% according to Quilis, 1981). The pattern is also strongly represented when looking at word length. In disyllabic words, penult prevalence amounts to 70% (Alcina & Blecua, 1975), although the situation is slightly more balanced in first language acquisition. According to Prieto (2006), around half of the words heard by infants are disyllabic, 65% trochaic and the rest iambic. As for trisyllabic words, 70% of them have penultimate stress in Spanish (Sebastián-Gallés et al., 2000).

As a result, we can assume that there is a default penult pattern derivable by rules in the language with lexical exceptions (final and antepenult) that have to be learned (cf. Piñeros, 2016; Martínez-Paricio & Torres-Tamarit, 2018; Baković, 2016). Here we investigated the processing of stress by Spanish speakers and its consequences for the users' grammars. Our aim was to establish whether the default penultimate stress pattern is processed differently than the exceptional antepenult and, consequently, whether the latter but not the former is stored in the mental lexicon to facilitate word retrieval.

1.3. Goals and hypotheses

Because of the evidence for (quasi)default stress in Spanish whose behaviour is expected to differ from the exceptional stress, the language constitutes a good testing ground for the two dominant grammar models (section 1.1). We therefore designed a study focused on auditory processing and subsequent classification of native Spanish words as either correctly or incorrectly pronounced. An equal number of penults and antepenults was used with both a standard and a deviant pronunciation in a paradigm involving a correctness judgment task. Electrophysiological recordings focused on the N400 effect were made during the study. Incorrect stress was assumed to invoke a more robust negativity in the range of approximately 400 ms from the onset of the stimulus compared to the correctly stressed word (following e.g. Knaus et al., 2007; Domahs et al., 2012b; see also Section 1.4). Given the distributional (frequency) and grammatical differences between penults and antepenults in the language, it was further assumed that a significant difference would ensue in the electrophysiological data between the two stress patterns: the processing of incorrect stress in antepenults should be

more costly (greater negativity). N400 is a component that occurs, among others, in response to a semantic violation (Kutas & Hillyard, 1984; see Friederici, 2004 for a review). Thus, if information concerning stress is derived in online processing and not stored in the mental lexicon, the change of the stressed syllable should not cause major problems (no substantial lexical inhibition). If, however, stress information is stored (or lexicalised), then a mismatch between the memorised and the perceived word will be detected, and more processing steps will be needed to identify the word in question. The experiment described and discussed in Sections 2-4 is therefore aimed at testing the generativist hypothesis that abstract phonological categories are stored in the mental lexicon and only unpredictable phonetic information is lexicalised. If the assumptions of generative phonology are correct, changes to the exceptional pattern should evoke stronger responses because they are interpreted as a lexical violation. Since stress is stored only in the case of exceptional words, changes to the default are not a lexical violation; hence a less pronounced N400 response is expected. Conversely, if the assumptions of the exemplar-based phonology model are correct, then stress information is stored together with segmental and semantic information pertaining to a given word regardless of the stress pattern. Accordingly, there should be no difference in responses to stress shift between the exceptional antepenult and the default.

Control for frequency effects influencing exemplar storage are necessary (the more frequently a given word is experienced, the stronger its representation in memory). Thus, words of matching frequencies were chosen from both patterns (see Section 2.2). Additionally, the frequency of word types should be taken into account. According to ET, more frequent patterns, structures or contexts have an effect on the representation of exemplars in memory. This is often referred to as 'entrenchment' (Bybee, 2006). Given its prevalence, the penultimate pattern must be more present in memory and hence the change of stress from the penult to a different syllable should be more costly for the speaker. This should lead to an opposite effect to the one hypothesised for the generative phonology model. Thus, we postulate that a stronger N400 effect in the case of changing antepenultimate stress supports the generative phonology framework, whereas no difference in the effects or a stronger effect of the change in penultimate stress should be considered evidence for a better applicability of the exemplar model.[3]

1.4. Further assumptions and comparison with previous studies

Our primary assumption is that Spanish listeners perceive stress as an abstract category and identify words based on stress differences. Because Spanish listeners need to distinguish between three stress patterns, it is assumed that they are sensitive to acoustic stress cues and judge stress placement correctly, which is supported by a body of literature (e.g. Peperkamp & Dupoux, 2002; Dupoux & Gallés, 2001; Torreira et al., 2014, Schwab & Dellwo, 2017). In a series of perception studies, it has been shown that Spanish listenerts, as opposed to the French for instance, can correctly identify stress in words. This is also true for infants, who are able to distinguish different stress patterns from nine months of age (Pons & Bosch, 2007). Furthermore, Spanish speakers not only associate stress with a given syllable, but also find it difficult to recognise words if the stress is altered without losing the ability to identify

[3] It should be noted that ET models within linguistics focus on learning, which does not include errors in input data, and on variation and language change, hence natural language processes occurring due to effort minimisation, undershoot of articulatory gestures and context-based assimilation and lenition processes. Word frequency dynamics are very useful in this context in terms of both production and perception, as well as the perception-production loop which can promote or inhibit change (typically, more frequent words undergo phonetic changes first, followed by lower frequency items but at the same time more entrenched, very frequent lexemes often tend to resist change, see Bybee, 2006). In our experiment, illegal stress is used. There are no attested pronunciations in the language that would correspond to our deviants hence an error detection mechanism rather than learning or a regular language processing mechanism must be invoked in speakers' brains. Deviants do not correspond to any existing words. It is unclear how ET interprets this kind of word processing *vis à vis* its claims about the role of both token and type frequency.

stress as a category in itself. In a pilot study on stress and vowel perception, Broś (2015) demonstrated that stress was identified correctly in nonce words, but native words with stress shifted to a different syllable caused confusion. In several cases, the words were not identified correctly, but the stress pattern was recognised. It is therefore of our primary interest to determine to what extent a shift in stress causes lexical inhibition (i.e. difficulty with identifying the word) and whether this inhibition differs as a function of the default/exceptional status of the stress pattern.

To achieve this goal, we must gain access not only to the phonetic and phonological (pre-lexical) processing of the word and the stress realised on it, but also to the semantic activation responsible for linking phonology with meaning. In neurophysiological approaches to the lexical processing of language, paradigms evoking the N400 negativity effect (Kutas & Hillyard, 1984) are typically used for this purpose. Several such studies concerning stress have been conducted to date on a variety of languages. For instance, in their ERP study of explicit and implicit processing of stress errors in German, Knaus et al. (2007) observed a negativity effect (interpreted as N400) in individuals listening to incorrectly stressed trisyllabic stimuli, which can be interpreted as an increased cost in lexical retrieval. Similarly, in their study of metrical violations in Russian, a language with lexical stress, Mołczanow et al. (2013) found negativity effects which are argued to belong to the N400 type and reflect increased costs in lexical processing. Comparable studies were also conducted on languages with fixed stress: Polish and Turkish (Domahs et al. 2012a, 2012b). All those experiments also showed late positivity effects, usually attributed to differentially demanding cognitive tasks or to a general decision-making mechanism concerning stress congruity. In view of these results, we assume that there is a strong case for analysing word stress in conjunction with semantic processing. Cross-linguistic evidence supports the thesis that shifting the stress from its original position has an influence on word retrieval from the lexicon. Even more importantly, there are differences in the processing of words belonging to different stress categories in the same language (Turkish), with a strong indication of the default as the one that is not ingrained enough in the mental representation of the word to cause miscomprehension. This is also in line with the reported 'stress deafness' effects in languages which do not have contrastive stress: stress cues may be superficially processed but do not seem to be acquired or stored in these languages.[4]

There is a possible limitation, however, as to the interpretation of these results with respect to the hypothesised access to word meaning based on correctly or incorrectly pronounced phonetic forms, which is of consequence for the methodology adopted in the present study. For this reason, we will briefly describe the procedure and rationale chosen in previous literature and comment on where our approach differs.

In most of the studies conducted on the subject (Knaus et al., 2007; Magne et al., 2007; Domahs et al., 2008; Domahs et al., 2012a; Domahs et al., 2012b), the focus was on prosodic processing, and more specifically on the perception of stress violations *per se*, not stress-based lexical access. For this reason, the researchers involved chose a correctness judgment task with different types of violations (different directions of stress shift and metrical effects). For instance, Domahs et al. (2008) focused on whether there is a difference in electrophysiological response depending on the foot structure, showing that when incorrect stress causes a different foot parsing of the word in German, this is treated as a strong metrical violation as opposed to foot structure-abiding stress shifts. In their study of Turkish word stress, Domahs et al. (2012a) tested default vs. lexicalised stress patterns showing no effect of foot structure but lexical effects instead. In their study, trisyllabic Turkish words were taken and the stress was shifted to the penult or the antepenult in the case of words with final stress,

[4] But see a recent study by Schwab et al. (2020) concerning L2 stress learnability upon training and its relation to working memory rather than acoustic sensitivity.

and to the penult or the final syllable in the case of penult stress. Words with correct antepenult stress were used as fillers. It is worth noting that final stress is the default in this language, while penults and antepenults are considered lexical exceptions, hence the situation is somewhat similar to the one encountered in Spanish but with the default elsewhere. Since the focus was on the perception of stress deviations, the authors of the Turkish study (similarly to the German and Polish cases) decided to present the target words visually prior to auditory presentation. The aim was to create an expectation as to the syllable that should be stressed and investigate the participants' subconscious (ERPs) and overt (judgement task) responses. The investigators were looking at the differential presence or absence of a positivity component (presumably P300) whose latency might depend on the exact moment in which the created expectation is not matched by input data. As a result, they did find a P3b-like response in reactions to stress violations except when the stress was shifted to a default position, in which case they found an N400 effect instead. This was interpreted as difficulty in lexical processing caused by changing lexically stored stress. In the opposite case, the correct default stress is not lexically stored and hence changing it does not cause problems with semantic processing but does cause a response to the metrical violation.

In the other studies mentioned above, similar results were obtained, with some of the conditions showing a biphasic ERP distribution: negativity and positivity. We attribute these results to the paradigm used and to the specific goals of the respective papers. Given the visual presentation of the stimuli an expectation is created in the participant (which is suitable for the N400 paradigm) but the word is accessed from the lexicon at the very outset of the trial. In this way we can appreciate ERPs related to the degree of violation or the relative predictability of the stimulus, but we cannot test lexical access. In our experiment we decided to use unpredictable stimuli only, similarly to one of the experiments conducted by Mołczanow et al. (2013). In the latter study, the authors wanted to investigate, among others, the lexical status of stress assigned to each of the tested stem types in Russian. To do this, they decided to avoid the lexical expectation created at the outset of the trial and hence did not use visual presentation. As a result, they observed an N400 effect which they interpret as an increased cost of lexical processing rather than the detection of a metrical violation (no priming and no context), followed by late positivity, which they attribute to a task-driven decision-making process.

In view of the above and given the principal goal of our paper (i.e. looking at how abstract representations are reached in lexical processing), we decided to use unpredictable words embedded in a neutral carrier sentence without prior visual presentation. Our aim was to trigger lexical search that happens online, starting from the onset of the word, in accordance with the principles of spoken word processing which is both automatic and incremental (information is updated based on each piece of incoming data, see e.g. van Petten et al., 1999; Deacon et al., 2000; O'Rourke & Holcomb, 2002; Holle et al., 2010).

The assumption of incremental and combinatorial processing reflects very well how word stress is detected. Most importantly, in natural speech there is no particular moment at which stress appears or is absent from a given syllable. Rather, stress is relational, which means that stress differences or anomalies can be detected only based on several phonetic cues that converge in a sequence of syllables rather than based on a specific parameter measured on one particular syllable in separation from its neighbours. Arguments and ERP evidence for this relational property of stress were presented by Domahs et al. (2008). Thus, if no specific expectation as to which syllable is stressed is created at the outset of the trial, the participant starts processing the word at the very beginning and all incoming information is used and combined before lexical access is complete and both its meaning and any deviation from the correct pronunciation are identified. Our assumption is that changes in stress inhibit the process of lexical search, causing an increased negativity response. Since the participants

of the study are asked to decide whether a given word was pronounced correctly or not, we also expect a positivity component related to the task.

We also wanted to pay special attention to a few other issues that may have affected the results of studies by Domahs and colleagues. First, we assume that the performance of the participants in terms of detecting stress deviations will be very good overall, which is based on previous behavioural studies on stress perception in Spanish. In the case of Polish and Turkish, however, the participants' accuracy was quite poor and did not go in line with the electrophysical data. For instance, Turkish speakers were unable to reliably detect violations with final stress during the experiment (52% of correct answers) and fared even worse in a subsequent offline stress identification task (29%). Given that major electrophysiological differences were identified between shifts to the default final position and the opposite, at least some of these discrepancies should be attributed to poor behavioural performance (especially the lack of a positivity effect). As will be shown in the following sections, there was no such confound in our study as Spanish speakers were very good at detecting all shifts in stress.

Another issue consists in the differences between the phonetic parameters measured as cues to stress between correctly and incorrectly stressed syllables. In the Turkish study, there were significant differences in F0 in word-final stress, and in duration and intensity in the antepenult case. Some of these differences were attributed to the difference in syllable structure between words with differing correct stress patterns. This problem is avoided in the present study. We use the same syllable structure for all stimuli, and we made sure that the key phonetic parameters are not significantly different between correct and incorrect conditions. In this way we avoid a situation in which a participant responds to some salient phonetic cue in an incorrectly stressed syllable instead of responding to stress shift. As will be shown in our experiment design, we wanted to make sure that the observed electrophysiological response would not be modulated by any uncontrolled extraneous variables.

2. Material and methods

A correctness judgment task was designed during which the participants were asked to listen to a series of standard and deviant stimuli and decide whether they were correctly pronounced.

2.1. Participants

32 native speakers of Spanish (19 females) aged 19-32 participated in the study after giving their informed consent. None of them reported neurological, language or hearing disorders. All of them were right-handed (attested by the Edinburgh Handedness Scale, Oldfield, 1971). They were paid for participation. Speakers of Spanish specifically from Spain and no other Spanish-speaking country were recruited to provide a more restrictive and representative sample. They come from 14 out of the 17 autonomous regions of Spain and had been residing in Switzerland for no more than 2 years at the time of the experiment.[5] After recording the data, two participants were excluded given an insufficient number of correct answers, the data from further two were excluded due to an excessive number of artefacts, and 1 dataset had to

[5] No differences in the perception of stress between the different regions of Spain have been reported in the literature. We assume that given the Spanish educational system and previous studies, which involved speakers from various regions, the perception of stress in quite uniform within Spain. Our data do not show any consistent differences between speakers.

be removed because of technical problems that arose during the recording. The 27 remaining participants were included in the analyses.

2.2. Stimuli

80 target words were selected for the experiment, 40 per stress pattern (penults, antepenults). All of them were nouns consisting of three open syllables (CVCVCV). They were selected based on frequency data provided by the EsPal database containing 300 million Spanish words annotated semantically, orthographically and phonologically (Duchon et al., in press). The principal frequency metric used was the log count (log10(cnt+1), current minimum value: 0.301030; current maximum value: 7.340494). This was then compared to the frequency per million counts provided by the CORPES (RAE) and CdE (El Corpus del español, Davies, 2002) databases. Apart from the syllabic structure, stress and word length, phonological neighbourhood was also considered to avoid listeners' bias toward certain phoneme combinations. Following thorough corpus research, the target words were selected in accordance with the following criteria:

(a) Proper names were excluded.
(b) Words were chosen so as not to become real Spanish words (lexical competitors) after the stress shift (i.e. after changing the stressed syllable).
(c) Words that have 10 or more phonological neighbours were excluded.
(d) Words which have a phonological neighbour of a higher frequency were excluded.
(e) Words which have a phonological neighbour with the other stress pattern under investigation were excluded.

Most of the words we used have a few phonological neighbours – usually the plural form of the same noun or the feminine/masculine counterpart of the same word, etc. For this reason, we did not exclude such cases. The same applied to words whose phonological neighbours were less frequent and hence less predictable for the hearer. We assume that phonetically similar words with a higher frequency are direct competitors of the words used as stimuli and it is most probable that they will be the first 'in line' in terms of processing in lexical search. Thus, we made sure that only words with less frequent phonological neighbours would be included in the stimuli list so as not to add difficulty to the task at hand (lexical access to unpredictable words). This is in line with a lexicality judgement study conducted on Spanish speakers by Vitevitch and Rodriguez (2005). The same mechanism might be prompted by words which had too many phonological neighbours (too many competitors in lexical search, see van Heuven et al., 1998), hence we excluded all items exceeding 10 similar words.

After preselecting around 50 words per stress pattern, we conducted a small survey among native speakers of Spanish with the aim of excluding those words which were infrequent to the point of not being comprehensible or easily recognised. Based on the survey, the final list of words was prepared. Additionally, given the differences in the frequency of words with antepenultimate stress as opposed to those with penultimate stress, we wanted to ensure that words of matching frequencies were selected. As a result, both frequent and infrequent words were chosen for each stress pattern with very close frequency log counts and no close phonological neighbours (see Appendix).

A reviewer points out that other factors, such as word familiarity and concreteness, might have influenced the results of our study. While it is true that word familiarity can play a role in auditory word recognition and there has been work contrasting it with frequency effects (e.g. Connine et al. 1990), given the focus of our study, namely prosodic and phonological effects, we believe that controlling for word frequency and phonological

neighbourhood should be enough to avoid any unexpected lexical effects. We follow Luce (1986) in assuming that phonological effects are of primary importance, followed by word frequency which might bias lexical selection in lexical tasks. Also, a comparative study provided by Connine et al. (1990) shows that word familiarity, while important, is task-dependent and may be a reflection of post-lexical rather than lexical processing. It also explains around 8.8% of variance in that study, while 40% of variance is explained by the first phoneme of a word in an auditory lexical decision task. That said, we believe that word familiarity should not be dismissed altogether as it seems to show different mental processes than word frequency, although its role has been shown most prominently in research on second language acquisition (e.g. Flege et al. 1996). Also, we assume that the survey conducted on a group of native speakers of Spanish before the final list of stimuli was selected informed us on the familiarity of the proposed words to some extent (participants were asked to judge the relative familiarity of the words and indicate words unknown to them, as well as words that are rarely used, strange or obsolete). As for such variables as concreteness, given the auditory modality of the task and the expectation that the phonological representation is reached in the lexicon, we do not expect any significant effects of such attributes on the way they are processed. We assume that a given word can be accessed from the mental lexicon even if the participant has only a vague idea about its definition. The employed semantic questions did not show any indication of a bias in this respect (see Section 3.1).

To ensure a better signal-to-noise ratio, half of the words from each pattern was to be repeated in the experiment (in a different block). For this purpose, 10 words with the highest frequency and 10 words with the lowest frequency were selected from each group. After the final list of words was prepared, the stimuli were recorded as naturally pronounced by a female native speaker of Spanish in two versions: correctly and incorrectly stressed. The words with antepenultimate stress (coded as APU) had a deviant version with penultimate stress and vice versa, i.e. the words with penultimate stress (PU) had deviant versions with antepenultimate stress. To avoid prosodic phenomena that might affect stress, the words were embedded in carrier sentences where they were not primarily stressed. They were then cut out from these sentences and, after acoustic preparation in Praat (Boersma & Weenink, 2016), spliced into carrier sentences used in the experiment. The prerecorded carrier sentence was *[proper name] pronunció la palabra [target word] otra vez* '[proper name] pronounced the word [target word] again'. Seven proper names were used as the subject of the sentence, all with a similar length: *Pedro, Pablo, Dani, Marta, Laura, Sonia, Lupe*. Similar to target words, the names were spliced into one carrier sentence chosen for the experiment in a way that the time between the beginning of the sentence and the onset of the target word was always 1.58 seconds. The duration of the whole sentence was around 3.5 seconds (the time differed depending on the target word, but these differences were in the range of a few milliseconds at best). The assignment of target words to a particular version of the carrier sentence was random. Some examples of target sentences are presented below (capital letters mark the stressed syllable):

Pedro dijo la palabra seMAna otra vez. (PUs – standard)

Juana dijo la palabra PAjaro otra vez. (APUs – standard)

Lupe dijo la palabra SEmana otra vez. (PUd – deviant)

Pablo dijo la palabra paJAro otra vez. (APUd – deviant)

Apart from sentences with embedded target words (both standards and deviants), 120 additional sentences were created with unrelated distractors (nouns of different types, with different syllable structures and/or word stress, all pronounced correctly). The stimuli were then divided into two blocks with an equal number of sentences each. Each target word was presented once in its correct (standard) form, and once with the shifted stress (deviant). Repeated words occurred once per block, interspersed with other words.

As for the acoustic properties of naturally produced standards and deviants, no differences in intensity (as measured in dB in Praat, Boersma & Weenink, 2016), duration (in ms) or pitch (mean F0) were observed in most stimuli. Around 10 words were either corrected by means of Praat or replaced by another version of the stimulus produced by the native speaker during the recording session. As a result, there were no statistical differences between penultimate standards and deviants or antepenultimate standards and deviants. Similar results were obtained for the corresponding unstressed syllables. Table 1 presents the descriptive summary of the three parameters.

Table 1. Descriptive statistics of the phonetic parameters. Here, standards and deviants of each stressed and unstressed syllable were compared to make sure that there are no statistical differences between them and that deviants can be used as analogues of the standards in the experiment.

	stressed antepenult			stressed penult		
standard	*F0*	222.9 Hz	(21.9)	*F0*	200.9 Hz	(13.7)
	Int.	71.8 dB	(3.4)	*Int.*	69.9 dB	(2.0)
	Dur.	187 ms	(59)	*Dur.*	182 ms	(29)
deviant	*F0*	224.0 Hz	(23.6)	*F0*	203.1 Hz	(6.9)
	Int.	73.0 dB	(2.6)	*Int.*	69.5 dB	(3.0)
	Dur.	196 ms	(46)	*Dur.*	193 ms	(23)
comparison	*F0*	F(1,78)=0.05	p=0.81	*F0*	F(1,78)=0.82	p=0.368
	Int.	F(1,78)=3.1	p=0.08	*Int.*	F(1,78)=0.44	p=0.5
	Dur.	F(1,78)=0.67	p=0.41	*Dur.*	F(1,78)=3.7	p=0.06
	unstressed antepenult			unstressed penult		
standard	*F0*	180.9 Hz	(15.7)	*F0*	267.5 Hz	(11.44)
	Int.	72.8 dB	(2.4)	*Int.*	69.8 dB	(2.5)
	Dur.	190 ms	(34)	*Dur.*	153 ms	(23)
deviant	*F0*	181.9 Hz	(13.4)	*F0*	264.2 Hz	(11.7)
	Int.	72.0 dB	(1.9)	*Int.*	70.8 dB	(2.3)
	Dur.	200 ms	(35)	*Dur.*	151 ms	(23)
comparison	*F0*	F(1,78)=1.62	p=0.2	*F0*	F(1,78)=0.09	p=0.76
	Int.	F(1,78)= 3.41	p=0.07	*Int.*	F(1,78)=2.44	p=0.12
	Dur.	F(1,78)=0.2	p=0.65	*Dur.*	F(1,78)=2.07	p=0.15

2.3. Procedure

The experiment was set up as a direct correctness judgment task. 360 sentences, 240 of which contained target words in either standard or deviant form, were divided into two blocks and presented in a pseudo-randomised order, making sure that the deviant and standard of the same word were not presented one after the other. Each block lasted around 20 minutes. A short preparation session was run first to familiarise the participants with the task and give them instructions. The paradigm was prepared and presented to the participants using Presentation software (Neurobehavioural Systems, www.neurobs.com). Triggers were set at the beginning of each syllable, including the onset of the target word.

The participants were instructed to sit at a desk in front of the computer, listen to the stimuli and decide whether the pronunciation of the target word is correct or incorrect. To make sure that the gaze is steady with no visual stimuli, we set a fixation cross on the screen during the presentation of the audio stimulus. To avoid contamination of the baseline, the fixation cross appeared at the beginning of each trial, 500 ms before the onset of the target sentence. The cross disappeared shortly after the end of each sentence: at 3500 ms from trial onset, at which point a question asking about the correctness of the target word appeared on the screen. The participant was asked to use the left and right arrow keys to answer (left = incorrect, right = correct). Additionally, to make sure that the target words, especially the infrequent ones, were understood by the participants and judged based on both prosodic and semantic information, a semantic question appeared in 10% of the cases. Each question contained a definition of the word requiring a yes/no answer. After the participant's response, an ITI of 2000 ms was set before the onset of the next trial. The participants were allowed to blink throughout the question and post-question periods, whenever the fixation cross was not on the screen. The speakers were set at the same volume for the whole experiment. Each participant was asked at the beginning of the session whether he/she hears the sentences correctly. No adjustments were necessary in any of the cases.

2.4. EEG recording and pre-processing

The BioSemi ActiveTwo EEG system (ActiveTwo, BioSemi B.V., Amsterdam, Netherlands) was used to record continuous EEG data from 128 channels with Ag/AgCl electrodes plugged into an elastic cap (Electro-Cap International Inc. Eaton, OH, USA). Two ocular electrodes were placed below the left and right canthi to record vertical eye movements. The sampling rate was 512 Hz. One active and one passive electrode served as ground and reference electrodes at central scalp positions. Impedances were kept equal to or below 20kΩ for all electrode sites, and an online band pass filter of 0.1-100Hz was applied. EEGLAB Toolbox (Delorme & Makeig, 2004) combined with the ERPLAB software (Lopez-Calderón & Luck, 2014) were used to extract and pre-process the ERPs. The data was filtered offline with a low cut-off filter of 0.1Hz (12dB) and a high cut-off of 30Hz (48dB), then baseline-corrected and re-referenced to the average of two electrodes close to the mastoid region labelled B10 (~P6) and D32 (~P7). The 50Hz frequency was filtered out using the Cleanline plugin from the EEGLAB software developed by T. Mullen. Before epoch extraction, an independent component analysis (ICA) was performed and used to correct blinks and saccades (Jung et al., 2000). Bad electrodes were interpolated spherically based on neighbouring electrode data.

1200 ms epochs were extracted with a baseline period of 200 ms based on predefined events marking the onset of the word for each of the four conditions: standard penults (PUs), deviant penults (PUd), standard antepenults (APUs) and deviant antepenults (APUd). Only epochs corresponding to correct responses were included in the pool. This was followed by epoch-based artefact rejection via visual inspection, with a rejection threshold set at 25% in line with previous literature. A total of 86% of all epochs was included in the statistical

analysis, i.e. between 49 and 53 trials out of the original 60 depending on the condition (87% of PUs, 85% of PUd, 89% of APUs, and 83% of APUd trials).

2.5. Statistical analysis

Given the selected paradigm and the expected EEG activity, two general time windows were preselected (350-550 ms and 600-900 ms after the onset of the antepenultimate or penultimate syllable, respectively). It was expected that latency differences might occur depending on the stress pattern (PU vs. APU), which was not confirmed during the visual inspection of the grand averages. Based on original ERPs and on difference waves, we observed that to identify the stress of the target word, participants had to wait until the second syllable was heard in each of the cases (see the Discussion section). We also corrected the windows of interest to 350-600 ms for the expected negativity effect, and 600-950 ms for the subsequent positivity.

Given the previous literature and the typical scalp distributions of the components under investigation, we selected the principal regions of interest (ROI): topographically central and parietal-distributed scalp channels. We also included frontal electrodes for comparison as similar activity was reported there in the studies on stress shift cited herein. Fz, Cz and Pz channels were taken into account, together with 6 sites surrounding each of them. Thus, the total number of channels included in statistical analyses was 21 (see Fig. 1). To further improve the SNR of the obtained recordings, we averaged the signals from all the selected channels from each scalp region and then extracted mean amplitude values for the two time windows (3×2). We then carried out a series of repeated measures ANOVAs with the use of R software (R Development Core Team, 2008) with factors *stress* (PU, APU), *condition* (standard, deviant), and *region* (frontal, central, parietal). Detailed information on each of the performed analyses is provided in Section 3.

The behavioural results were analysed based on reaction times (RTs) recorded from the beginning of the answer period (repeated measures ANOVAs with factors *stress* and *condition*, *lme* function in R).

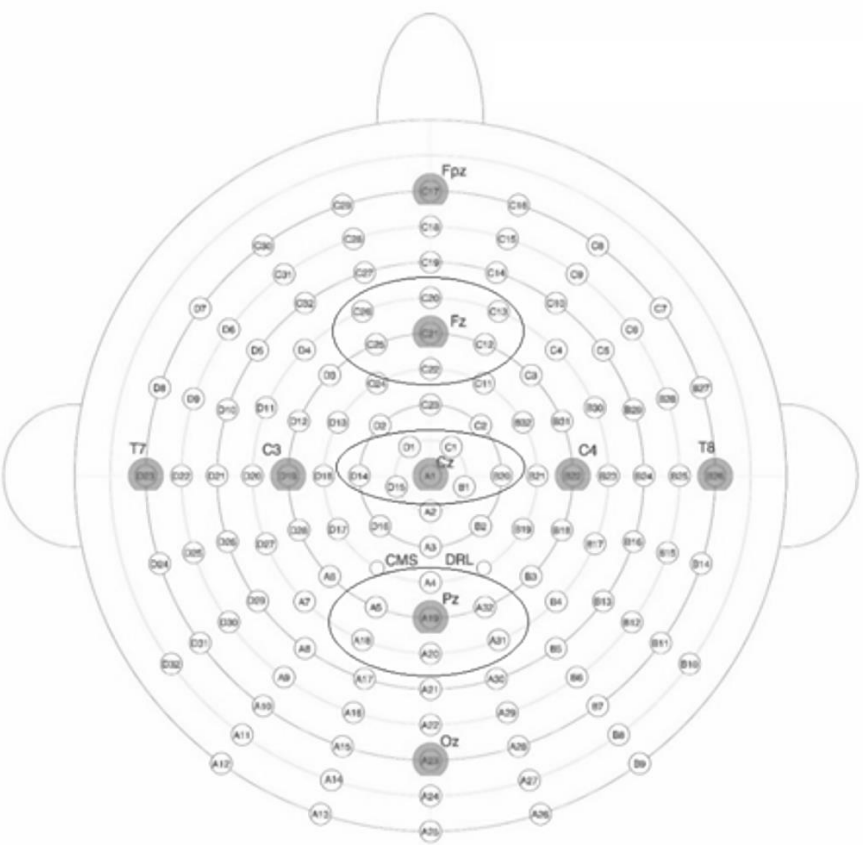

Fig. 1. BioSemi system 128-channel scalp distribution with three electrode pools of interest marked in black.

3. Results

3.1. Behavioural results

Both accuracy and reaction times were recorded during the trials based on button pushing. Accuracy scores were used to decide whether the participants were successful at the task and to reject trials with incorrect answers from subsequent analysis. Given the nature and goal of the study, we set the threshold of response accuracy at 75%. First, we had to make sure that the participants recognise the words and the stress associated with them correctly, i.e. well above chance levels (50%). Second, to ensure a good SNR we could not admit participant data containing less than 75% of the original material.

General accuracy was very high, as expected. 30 participants had an average of 9 misses in the experiment (230.8 correct answers out of 240). The misses ranged from 2 to 28, depending on the person. As for the stimuli, correct PU and APU trials gave between 100% and 88.3% accuracy. As for the deviant stimuli, APUd trials seemed to cause the most problems (from 100% to 68.3% in one case, 95.3% on average) while PUd trials where the second most difficult group (from 100% to 81.6% accuracy, 97.9% on average). Two participants had very low accuracy scores and hence their results were excluded from further analysis.[6]

The statistical analysis of accuracy scores (*aov* function in R) shows a significant effect of *condition* (standard vs. deviant, p = 0.0235) but not *stress pattern*, and a significant

[6] Interestingly, participant A2 seemed to be unable to detect incorrect stress in almost all of the cases (0.6 accuracy in the case of PUd and APUd trials). The overall accuracy score for this person, including correctly stressed, words was 51.6%. This may be due to the reported early exposure to Latin American rather than Peninsular Spanish before moving to Spain at the age of 4. The other participant fared slightly better but had more than 50% of incorrect answers in the case of the APUd condition, which made it impossible to include the data in the statistical analysis (after artefact rejection the data were too noisy).

interaction between the two variables (p = 0.0108). In the model combining standard/deviant and stress pattern results in a single 4-level condition factor with a random effect of participant (*lme* function in R), the Bonferroni-corrected *p* values show a significant difference between the APUd and both APUs and PUs conditions (p = 0.002055 and p = 0.000894, respectively), and no other effect. This means that, as noted descriptively above, the APUd condition is especially difficult and caused most errors in stress correctness detection.

As for the accuracy of responses to the semantic questions, it was also very high (93.5% on average, from 1 to 9 incorrect answers out of 39), which means that the participants understood or knew the words that were used as stimuli. As mentioned earlier, the use of low frequency words might raise doubts about whether the speakers would actually access the words from the lexicon and make informed judgments about the stress rather than random or intuitive ones. The semantic accuracy scores and post-experiment conversations with the experimenter suggest that some of the words were 'strange', 'rarely used' or 'difficult to decipher in terms of meaning', but nevertheless mostly recognisable as known to the user. Only rarely did the participants report that some word did not exist or was unknown to them. Thus, the general conclusion concerning accuracy scores is that the participants of the study were familiar with the words used in the experiment and their meanings and were able to access them from the lexicon and judge whether they were stressed correctly or incorrectly. Hence, the data are suitable for both RT and ERP analysis.

The reaction times were measured from the onset of the question, which appeared on the screen 3500 ms from the beginning of each trial, irrespective of word length. Each participant had to press one of the buttons as soon as they saw the question and had enough time to respond (trial duration was set at 10000 ms). It is worth mentioning that differences in word length were negligible in general and cancelled out by a few extra milliseconds of silence before the appearance of the question on the screen. Since each target word was presented exactly 2080 ms from trial beginning, measuring reaction times from word onset gives the same statistical effects.

Mean reaction times per condition are as follows: 504 ms for APUs, 636 ms for APUd, 514 ms for PUs and 559 ms for PUd. Two observations can be made here. First, the penultimate stress pattern requires a bit more time to process in the standard condition, possibly due to the fact that the stressed syllable appears later than in the antepenultimate condition, although the duration of the first syllable (around 170-200 ms) is much more than this difference. Second, the difference in the reaction times (and hence the lag between the standard and deviant conditions) is much greater in the case of the exceptional APU (132 ms) than in the case of the default PU (45 ms). Statistical tests confirm this, showing a significant effect of *condition* ($F(3,78) = 4.415$, $p = 0.0064$). Pairwise comparisons (Tukey) with Bonferroni-corrected *p* values show that there is a significant effect in the case of APUd when compared to APUs (p = 0.0066) and PUs (p = 0.0155). No other comparison reached significance. We interpret this result as evidence for a significant difference in responses to deviants depending on the stress pattern. The difference in reaction times in response to deviants as opposed to standards resulted significant in the APU (i.e. exceptional) condition only. Furthermore, the time lag in the response is similar regardless of the standard we compare the deviant to, which means that APUd stimuli are particularly difficult to process and stress shift from the antepenultimate to the penultimate syllable inhibits word comprehension. The opposite change causes only a slight lag in the response. These results match those of accuracy scores.

3.2. EEG results

The aim of the ERP analysis was to determine whether there is an effect of stress shift on word processing, and whether this effect is different depending on the stress pattern. We also wanted to know whether the task produced a late positivity effect given increased processing effort needed to judge the correctness of deviant stimuli, and if so, whether it differed between the two stress patterns. Figures 2 and 3 show grand averages of event-related potentials obtained for PU and APU-type words, respectively. The time windows of interest are marked accordingly.

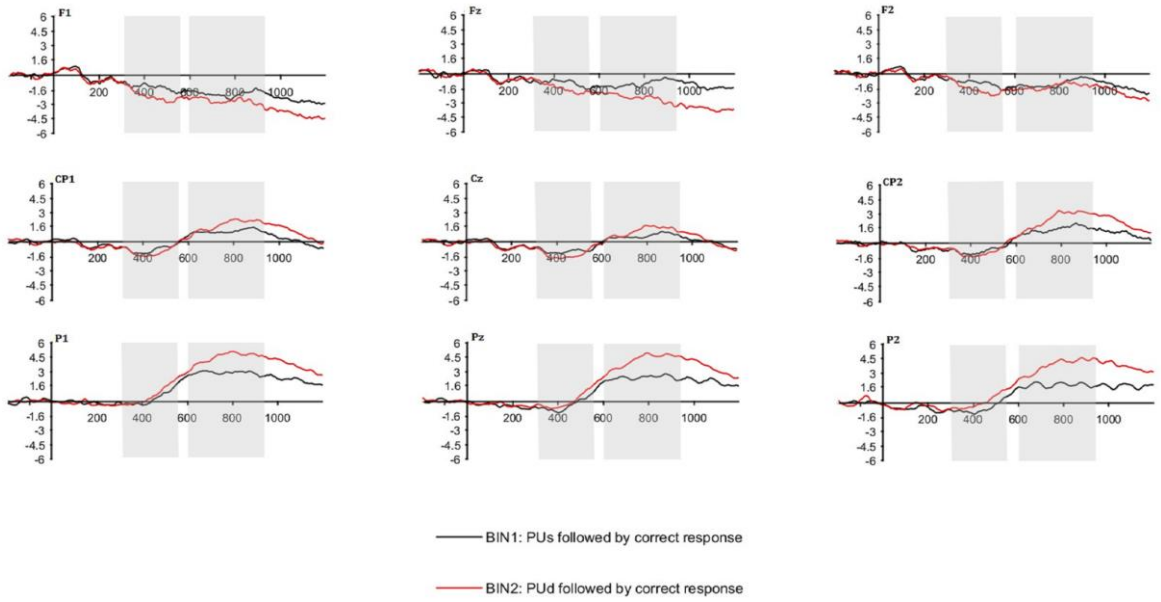

Fig. 2. Grand average ERPs of the penultimate trials from 3 electrode sites per ROI. Positive values are plotted up.

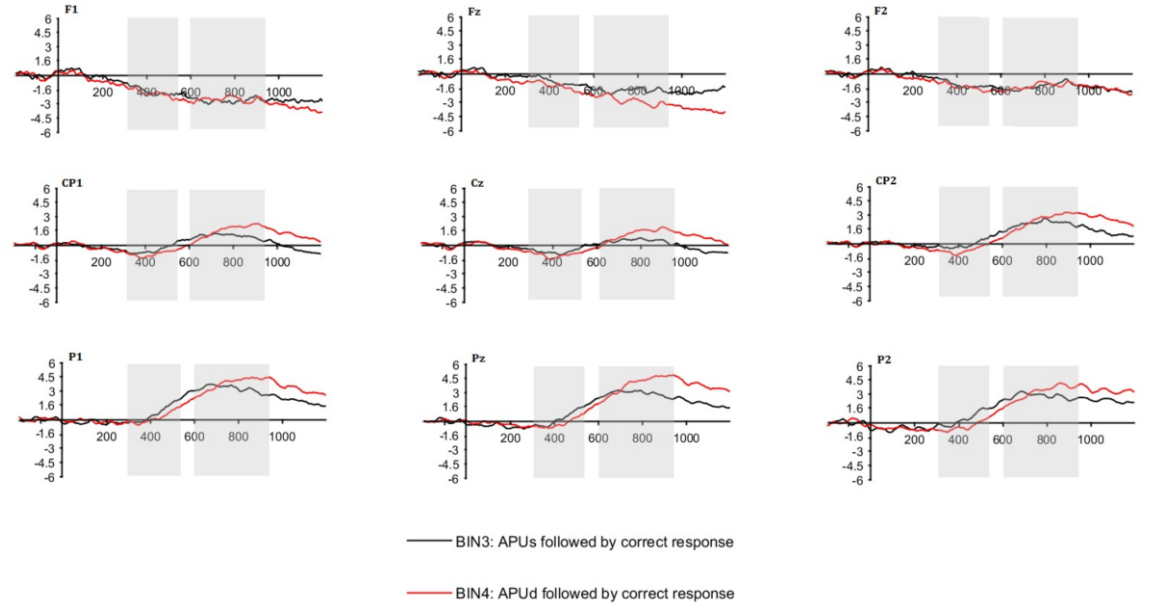

Fig. 3. Grand average ERPs of the antepenultimate trials from 3 electrode sites per ROI. Positive values are plotted up.

As can be appreciated from the inspection of the above figures, there was no activity corresponding to the processing of target words over the frontal electrodes. Furthermore, whereas the negativity in the first TW seems to be more pronounced over the central electrode sites, the subsequent positivity effect is greater over the posterior regions of the scalp. To test the significance of these effects, we first carried out a general 2 × 2 × 2 × 3 repeated measures ANOVA with the factors *stress* (PU, APU), *condition* (standard, deviant), *time window* (TW, 350-600 ms, 600-950 ms) and *region* (frontal, central, parietal). A positive main effect of both TW ($F(1,26) = 46.36$, $p < 0.001$) and ROI ($F(2,52) = 30.85$, $p < 0.001$) was found, hence we continued with the analysis of the TW and the ROI separately.

N400 (350-600 ms): No significant effect of either stress or condition was found over the frontal electrodes ($F(1,26) = 0.844$, $p = 0.367$ and $F(1,26) = 0.025$, $p = 0.874$, respectively). Over the central electrode sites, there was a significant main effect of stress ($F(1,26) = 9.124$, $p = 0.006$) and condition ($F(1,26) = 5.206$, $p = 0.03$) with no interaction. Over the posterior region, there is a significant main effect of stress ($F(1,26) = 6.718$, $p = 0.015$) but not of condition ($F(1,26) = 0.132$, $p = 0.719$), and a significant interaction between the two ($F(1,26) = 24.77$, $p < 0.001$). Given these results, we excluded frontal electrodes from further analysis.

To disentangle stress from condition and to see whether there was a negativity effect on each of the stress patterns, we ran separate analyses for the PU and the APU trials. A 2 × 2 repeated measures ANOVA with the factors *condition* (standard, deviant) and *region* (central, parietal) showed a highly significant main effect of condition ($F(1,26) = 20.38$, $p < 0.001$) and region ($F(1,26) = 30.36$, $p < 0.001$) but no interaction ($F(1,26) = 0.68$, $p = 0.417$) for the APU stress pattern, which means that although there are differences in mean amplitude values between the two regions of interest, there was no difference in the N400 effect (see the interaction plot in Fig. 4). As for the PU trials, condition did not reach statistical significance ($F(1,26) = 1.562$, $p = 0.222$), which means that there was no N400 effect in this case. There was, however, a significant main effect of region ($F(1,26) = 23.63$, $p < 0.001$) and a condition × region interaction ($F(1,26) = 23.56$, $p < 0.001$) which shows that whereas the mean amplitude over the central sites was slightly smaller in the deviant condition, it was greater than in the case of the standard over the posterior region, contrary to all expectation (see Fig. 4). The analysis of the posterior pool data alone gave $F(1,26) = 9.523$, $p < 0.01$, which likely represents a P3 effect. Narrowing the time window for the analysis of PU trials to 300-500 ms did not cancel the effect. This was additionally confirmed via inspection of the difference waves which show a positive instead of a negative inflection in the posterior sites (see Fig. 5). In view of these facts, the reversal of the standard-deviant dynamic might suggest a different reaction to PU as opposed to APU deviants. Apparently, PU words pronounced with the stress on the first syllable are not in violation of any expectation about the prominence of the second syllable, which would further corroborate our hypothesis about the N400 as a proxy for the lexical processing of stress. In the PU case, we seem to be dealing with a positivity effect with an earlier onset compared to APU that is most likely part of the later positivity discussed below.

A repeated measures ANOVA carried out for the mean amplitudes of the PU and APU difference waves further corroborated our findings: there is a significant main effect of stress pattern ($F(1,26) = 12.89$, $p = 0.001$). Thus, we can conclude that our main hypothesis concerning the N400 effect in APU vs. PU words was confirmed. Not only is there a

difference between the two stress patterns, but there is also no significant N400 effect in the case of the default as opposed to the purportedly lexicalised type of words.

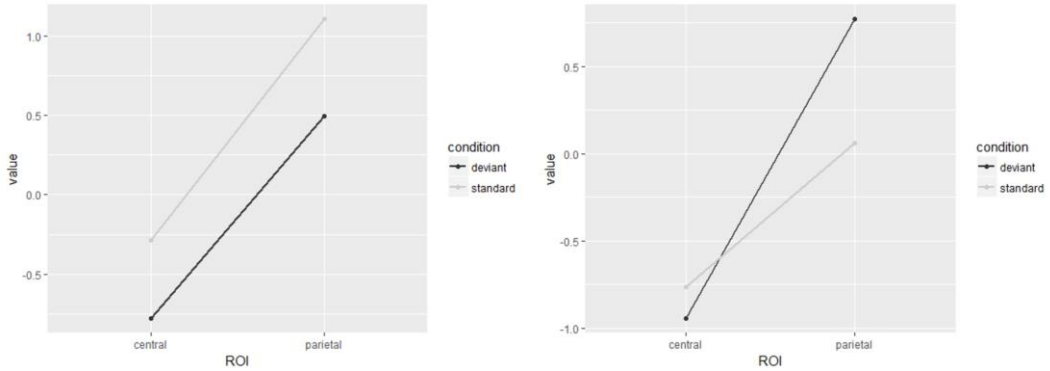

Fig. 4. Interactions of condition and region in APU (left) and PU trials (right). The *y* axis represents ERP values.

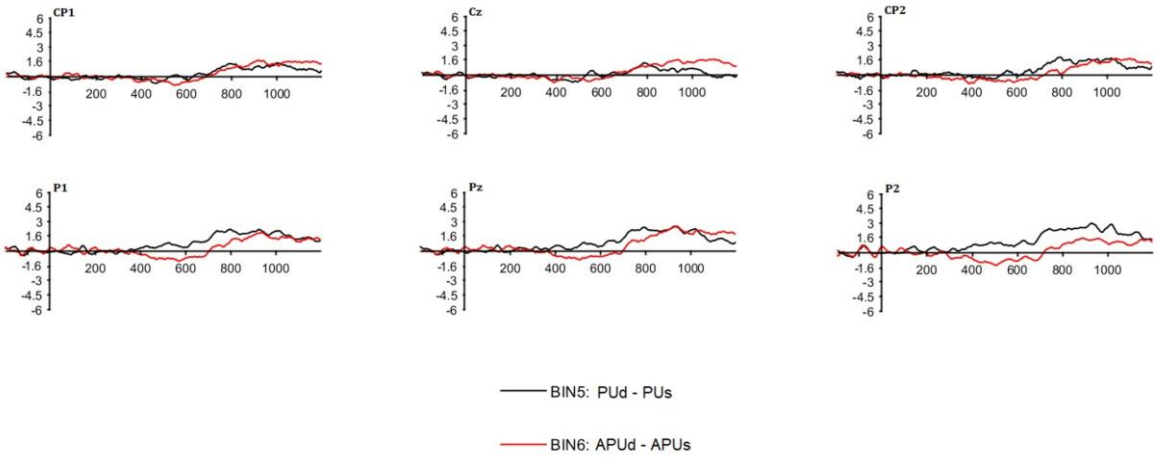

Fig. 5. Difference waves for the two stress patterns showing deviant minus standard ERPs.

Apart from investigating the N400 effect separately for each of the stress patterns and looking at their interactions, we also performed a general 2 × 2 × 2 ANOVA on the centroparietal data taken altogether. This gave us a main effect of stress ($F(1,26) = 13.9$, $p < 0.001$) and an interaction between stress and condition ($F(1,26) = 12.88$, $p = 0.001$) but no main effect of condition ($F(1,26) = 1.192$, $p = 0.285$). Based on these results and on the interaction plots, we could conclude that whereas deviants did not differ much from one stress pattern to another, the standards showed much lower values in the case of PU compared to APU. In other words, stress mattered in the standard condition only, which can be interpreted as a kind of 'levelling' of the negativity effect between the two stresses.

It is possible that the curious result is due to what we shall call the 'two-syllable window' of stress processing. More specifically, given the nature of the stimuli and the task, we assume that the hearer needs exactly two syllables of each word to decide which stress pattern is applied. Both in standard and deviant stimuli, the stressed syllable is not necessarily longer and of higher pitch than the unstressed one. Note that in APU words the mean F0 of the stressed antepenultimate syllable is 222.9 Hz in standards and 224.0 Hz in deviants. The

unstressed syllables of these words have 267.5 Hz and 264.2 Hz values, respectively, which means that the pitch is quite high at the beginning and steadily rising. The intensity values are very similar (see Table 1). As for the duration, it is between 187 and 196 ms in the stressed syllable and falls to 151-153 ms in the unstressed one. In the case of PU words, the second (stressed) syllable is equally long as the first (unstressed) one, and often even shorter (182-193 ms vs. 190-200 ms). At the same time, the pitch is rising from a lower value in the unstressed first syllable (~180 Hz) to the stressed penult (~200 Hz) but is never as high as in the APU word type. Again, the intensity values are similar and do not seem to play an important role in disjunction from the other two parameters. Thus, PU words have lower pitch values than APU words but both word types show a rise from the first to the second syllable. This rise, however, is much greater in APU words (~40 Hz). At the same time, whereas APU words show the expected difference in the length of the stressed with respect to the unstressed syllable, PU words do not confirm this relation.

Based on the above comparison we can expect that upon hearing the first syllable Spanish speakers cannot determine with any degree of certainty whether this syllable is stressed in the word. Judging by the duration, they cannot make a decision since it is more or less the same for PU and APU words.[7] Focusing on the pitch, they might make an initial prediction given the higher starting value in APU stimuli, but they still have to wait to hear the second syllable to determine how much of a rise there is. Furthermore, it is worth mentioning that according to Nooteboom (1997), the human perceptual system cannot reliably distinguish pitch differences below three semitones, and one semitone corresponds to approximately 12 Hz in stimuli with a mean frequency of 220 Hz. Consequently, speakers probably need combined information concerning pitch and duration fluctuations between the two consecutive syllables (and possibly intensity as well) in order to detect and respond to a given stress pattern. This is in line with our data, which show no latency difference in the electrophysiological response to PU and APU words.

Late positivity (600-950): The general 2 × 2 × 2 repeated measures ANOVA with mean amplitude from the 600-950 ms time window as the dependent variable showed a significant effect of condition ($F(1,26) = 23.05$, $p < 0.001$) but not stress ($F(1,26) = 0.125$, $p = 0.726$), with an interaction between the two ($F(1,26) = 4.721$, $p = 0.039$). Note that this is in exact opposition to what we observed at the earlier time window (see Fig. 6). There was also a significant main effect of region ($F(1,26) = 30.65$, $p < 0.001$) and an interaction between the condition and region ($F(1,26) = 15.57$, $p < 0.001$).

Analyses by region are similar to the ones run for the 350-600 ms TW: there was a significant effect of condition in the central sites ($F(1,26) = 8.201$, $p = 0.008$) and no main effect of stress nor interaction. In the parietal region, there was a significant effect of condition ($F(1,26) = 34.54$, $p < 0.001$) and an interaction ($F(1,26) = 8.882$, $p = 0.006$). The persistent effect of condition was further confirmed by the analysis of difference waves, which we interpret as an indication that correctness judgment occurs at this stage.[8] Whereas

[7] It might be argued that vowel duration should be compared here instead of syllable duration. Indeed, in most languages, including Spanish, vowel length is considered to be a primary cue to stress. However, the situation is somewhat complicated as different cues are used in perception depending on the vowel. Ortega-Llebaria et al. (2008) argue that the vowel [i] is perceived as stressed or unstressed mainly based on intensity while stress in the low vowel [a] is distinguished based on duration. In our data, the length of the consonant does affect syllable duration (stressed penult vowels have the length of 94 ms, while unstressed first syllable vowels – 87 ms, which means that stressed vowels **are** longer). This, however, does not cancel the fact that the speaker has to wait until the second syllable to compare the durations, especially that different vowels occur depending on the word. 17 of the APU words, and 13 of the PU words have short vowels in the first syllable (either [i] or [u]). As for the consonants, the situation is quite balanced between the two word types: around 25 words from each paradigm have a short sound at the beginning of the first and second syllable in each paradigm. Yet it must be noted that any differences between PU and APU in this respect cancel out because each type of words was presented as both a standard and a deviant.

[8] There appears to be a significant effect of stress in the difference waves, which is best visible in the posterior electrode pool, but when the two ROI are taken separately, there is no main effect of stress in either region and only the same interaction effect in the parietal sites ($F(1,26) = 8.882$, $p = 0.006$).

the prevalence of stress effects in the first TW points to the processing of prosody (stress pattern) separately from the rest of the information necessary to perform the task, later on the hearer has to decide whether what (s)he heard was correct or incorrect. At this later stage, phonological-semantic integration must have taken place.

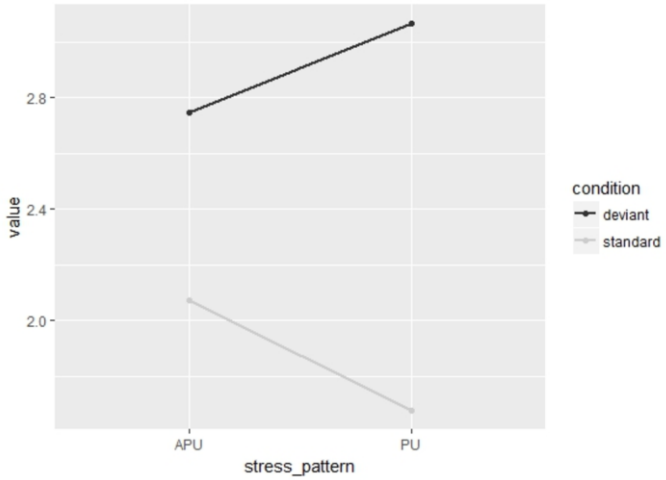

Fig. 6. Interaction plot of the stress and condition effects in the late positivity time window. Here, the difference in the amplitude between the standards and deviants is closer in the APU condition. It is more robust in PU words.

4. Discussion

As evidenced by the analysis of the event-related potentials, Spanish speakers respond to stress shift differently depending on the stress pattern. Moving the stress from the penultimate to the antepenultimate syllable does not result in an N400 effect. The response to such deviants is not significantly different from the response to the corresponding standards. At the same time, shifting the stress in the opposite direction evokes a negative response whose amplitude is significantly lower than the response evoked by the corresponding standards. We tend to interpret this result as an N400 effect observed in the range of 350-600 ms from the onset of the target word, which roughly corresponds to the end of the second syllable. In line with our initial assumptions and hypotheses, we consider this evidence in favour of the default vs. exception dichotomy. Penults behave as true defaults whose underlying abstract representations are not indexed with stress information. Instead, the stress is inferred (or computed) from grammatical rules concerning default stress assignment. Antepenults, on the other hand, must be stored together with the information concerning the syllable that is stressed. Deviation from this lexicalised stress is costly for the hearer: accessing the semantic information about the word is more difficult, hence the enhanced negative response.

Thus, the data may support the generative phonology framework which assumes that only unpredictable information is stored in the mental lexicon. There is no evidence that stress is processed indiscriminately regardless of the stress pattern, following the principles of exemplar theory. While frequency effects play a role in speech processing, especially speech perception (Savin, 1963; Pierrehumbert, 2001; Pisoni et al., 1985; Goldinger et al., 1992), when they are controlled for, grammar is the decisive factor. Grammatical operations, which translate acoustic detail and auditory cues into abstract features and phonological constituents, are therefore an indispensable element of online language analysis and cannot be limited to mere statistical inference. This conclusion finds support in previous studies on the perception of stress in Russian and Turkish (Mołczanow et al., 2013; Domahs et al., 2012a), among

others. As mentioned in the introduction, these studies show N400-like negativity effects interpreted as pointing to problems with the lexical processing of exceptional word types. We add evidence from yet another language with variable stress, using a more restricted study design aimed at evoking a direct response to lexical vs. non-lexical deviations. The differential N400 response as per stress pattern and similar positivity effects in both cases confirm differential stress processing depending on the grammatical status of a given stress pattern and hindered lexical access in exceptional cases only. They also show incremental perception of stress as a relational property of language, corroborating the conclusions drawn by Domahs et al. (2008).

Furthermore, our study confirms that stress should be conceived as an abstract category and disentangled from both segmental phonetic information and semantics. Although necessary to mark and differentiate meanings, at least in languages with variable stress, it should be treated as a separate entity belonging to the realm of phonology as hearers respond to it separately from the meaning of the word.[9] In our data, we have seen a shift in focus between stress and correctness recognition. The earlier time window of 350-600 ms presents stress-driven effects, while at later latencies we can see a more correctness-driven response. This is in line with a bottom-up speech perception approach (Norris et al. 2000), according to which the hearer analyses incoming speech cues and makes predictions concerning the underlying category (be it sound, syllable or stress, depending on the time window). With each cue, sound and syllable, the hearer gets more and more information concerning the stress pattern used in the audio stimuli. The main effect of stress but not condition in the N400 window tells us that recognising the stress pattern itself is crucial at this step. At the same time, the difference in responses between the PU and the APU words already at this stage suggests that inferences are also made about the word's meaning. Semantic information must therefore be accessed to some extent as well. We deem this intermediate step crucial for there being a semantic expectancy violation of some kind, leading to the N400 effect.

At a later stage, the hearer needs to decide whether the identified stress matches the word accessed from the lexicon. The late positivity effect observed in the data reflects a top-down wrap-up process that integrates prosodic and semantic information and allows the hearer to decide whether the experienced word was correct or incorrect. The posterior distribution of the positivity effect suggests that it is closely related to the task. A significant rise in the amplitude of the ERPs was observed after hearing deviant stimuli compared to the standards, which means that increased processing must have been involved in analysing the stimuli. This is confirmed by the behavioural data: hearers needed more time to judge the correctness of the words whenever the stress was shifted.[10] Consequently, increased processing costs and the need to put together pre-lexical and lexical data provided by the whole word are probably the main contributors to the LPC wave. This finds support in studies suggesting that positivity in the range of 600 ms from stimulus onset reflects cognitive control processes and error monitoring (e.g. van Herten et al., 2005; Kolk & Chwilla, 2007).

However, it cannot be completely ruled out that the negativity that peaks at Cz between 350-600 ms may be partly affected by the subsequent positivity. The difference waves for correctly and incorrectly pronounced words are most strongly enhanced at the Pz electrode, while the negativity is clearly smaller compared to Cz. One could argue that the

[9] As shown by studies focusing on pseudowords and differences between real word and pseudoword processing, e.g. Honbolygó & Csépe (2012), Ylinen et al. (2009), as well as by studies involving stress processing by infants with different language backgrounds, e.g. Friederici et al. (2007).

[10] Interestingly, the increased difficulty that APUd trials caused as opposed to PUd that was revealed by the behavioural data is not confirmed by the LPC results. In fact, there is a greater difference in the amplitude of the signal between the standards and deviants in the PU stress pattern than in the 'more difficult' APU. It must be noted, however, that these effects may be hidden by what happens in the earlier time window. Since there is no N400 effect in PU words, there is no significant negative inflection in the 300-600 ms time window. In the case of APU, given the N400 negativity, the APUd response signal is positive-going later on but does not reach values comparable to those of PUd responses.

clear disambiguation of the lexical information is only possible at this later stage. In this sense, the late positivity would be the more reliable neuromarker for the process of differentiation between the conditions. In this case, however, this process would not be lexical but triggered by general cognitive mechanisms. Further research would be necessary to clarify this question.

4.1. Theoretical implications

The combination of the N400 and the LPC effects reported above is compatible with the theory of speech perception set forth by Poeppel and colleagues (2008). According to this model, language processing requires a series of bottom-up and top-down operations, as well as a set of abstraction steps. Acoustic and articulatory information is extracted from the signal and translated into distinctive features that are the building blocks of phonological constituents, serving as 'representational primitives' linked to both the physical aspects of speech and to meanings. Thus, the model is compatible with the generativist view of generalising speech patterns and storage of abstract representations. Crucially, the extraction mechanism is of a dual nature: segment- and syllable-level processes take place in parallel according to two different time resolutions, which is consistent with neuroscientific literature (Poeppel, 2003; Boemio et al., 2005; Schonwiesner et al., 2005). The speech-to-concept mapping takes place based on internally synthesised representations (*analysis by synthesis*). Possibly, intermediate representations are needed at the interface between the feature-based underlying representation linked to meaning and the auditory stimulus consisting of the acoustic waveform. The model also assumes top-down feedback to previous steps. As a result, we can imagine that in our case the spectro-temporal cues corresponding to syllable prominence embedded in the acoustic signal of the stimulus are transposed onto an abstract representation of 'stress' as a discrete category differentiating words or meanings. Then, combined with the segmental representation of the word built up in parallel to stress extraction, the 'stress' representation is linked to candidates from a set of mental representations that best match the input. Any mismatch in stress will be processed online based on feedback mechanisms predicted by the theory and, upon the final recognition of the word, correctness judgment will follow.

It should be mentioned that although Poeppel et al. (2008) strongly advocate the abstractionist approach to speech processing, they do not deny the role of frequency or statistical modelling of speech or speaker-specific information, suggesting that core, categorical representations might be accompanied by more gradient periphery information, both contributing to lexical processing. We follow Poeppel et al. (2008) in admitting that some assumptions of the exemplar model can be adopted given cross-linguistic evidence for frequency effects involved in language processing (see Bybee, 2006 and citations therein). Also, the frequency effects observed specifically for Spanish in relation with our experiment should be considered evidence for some level of statistical inference and storage of language-external information.

In this context, it should be mentioned that apart from exemplar-based models of phonology, some generativist attempts at placing statistical and other grammar-external factors inside phonological computation have been made, usually in the context of language variation and change. For instance, Coetzee & Kawahara (2013) and Coetzee (2016) argue that given omnipresent variation in natural languages that is due to a range of social and pragmatic factors (age, education, gender, register, situational context), as well as other phonology-external variables (e.g. syntax, word type and frequency effects), phonological processes have to be modulated by extraneous variables. These variables contribute to inter-

and intra-speaker variation rather than change the way phonological features and processes are represented in the mental lexicon. This view converges with the one proposed by Poeppel and colleagues in that there is a core categorical grammar and all gradient, statistical effects based on language experience constitute peripheral information that helps in lexical processing. Thus, a hybrid model involving both generative representations and usage data reflecting the distribution of forms and patterns in the language might be postulated. Such a model would involve placing an additional building block in the lexical processing architecture. We can imagine that the speech signal consisting of acoustic cues is analysed in speech perception and key parameters are extracted from it incrementally, in a multi-time resolution fashion (at the level of segments and at the level of syllables). This information is translated into phonological features that determine phonological contrasts (consonants vs. vowels, labial vs. coronal sounds, stressed vs. unstressed syllables, etc.). While this is being done, candidate lists are generated and compete in a cohort manner. Thus, hypotheses are made about the words that are heard and either rejected or supported by the incoming data. The process is continuous and feedback-based until an unambiguous match is reached in lexical search. Crucially, candidate lists are based on peripheral information concerning word probability in a given context, word type and frequency, speaker identity and other factors that influence perception and either accelerate or impede lexical search. The usage-based portion of this hybrid model consists of several contributing factors. In spoken word perception, phonotactic information concerning possible sound sequences in the language and phonological neighbourhood effects giving rise to stronger competitors in lexical search are assumed to play a vital role. In this way abstract feature extraction from acoustic data is fine-tuned by usage-based considerations. Lexical access is facilitated by the probability distributions within a hearer's lexicon.

5. Conclusions

The aim of this paper was to investigate lexical access in stress processing vis à vis two major theoretical approaches: generative phonology and exemplar theory. Spanish was chosen as a testing ground for this purpose as it shows both default and exceptional stress, and its speakers are known to be sensitive to word stress. The rationale of the electrophysiological experiment conducted on Spanish speakers was inspired by a series of stress perception studies led by Ulrike Domahs and colleagues, albeit with some important changes in focus and design. The results of our study confirm previous findings concerning stress processing in speech perception, as well as the assumption that Spanish speakers are sensitive to stress differences. We corroborated our hypotheses concerning the difference in processing between default and exceptional stress, observing an N400 effect in the latter case only, as well as a positivity effect in both cases. Thus, we found direct evidence for the generative phonology framework.

In broader terms, our results are in support of an integrative view of speech processing, i.e. the assumption that auditory cues are extracted from the acoustic speech signal prelexically and then integrated with the semantic information computed or accessed from memory based on consecutive pieces of incoming data (phones, syllables, morphs, finally whole words). At the same time, several of these information extraction mechanisms work side by side. We found support for such processing steps in the case of word stress correlated with meaning based on the reaction times, and on the electrophysiological data from an earlier and a later time window. This integration mechanism requires abstraction, i.e. working out of intermediate mental representations of both segmental and prosodic information. Given the discrepancy between the two tested stress patterns, we conclude that a purely usage-based approach to stress processing has to be rejected in favour of the

generativist model assuming the phonological status of stress as an entity that is separate from segmental and semantic information of any lexical item.

Conflict of interest
No conflict of interests.

Acknowledgements
We would like to thank our collaborators and colleagues from the Psychology Department and from the Phonetics Lab at the University of Zurich for helpful discussion, as well as technical help. Our thanks are owed especially to Nathalie Giroud, Sandra Schwab, Lei He, Thayabaran Kathiresan, Ira Kurthen, and Carlota de Benito Moreno.

Funding
The project was financed by the Swiss Federal Government (Swiss Federal Government Excellence Scholarship for 2017/2018, ref. no. ege_2017.0335), and by the Phonetics Laboratory of the University of Zurich.

Appendix

Wordlist: APU = ANTEPENULTIMATE, PU = PENULTIMATE
All of the words are nouns, with the exception of *sólido*, *química*, *lógica* and *pícaro* which can also function as adjectives. The frequency counts are given for all occurrences of a given word. There is a significant difference between the 20 most frequent and the 20 least frequent words from each list (frequency per million: t = 5.0484, df = 39.053, p = 1.072e-05; log count: t = 13.018, df = 67.342, p < 2.2e-16). At the same time, there are no statistical differences between stress types (t = -0.072327, df = 77.977, p = 0.9425).

APU	**frq**	**log count**	**PU**	**frq**	**log count**
música	217.287086	4.825270	cabeza	220.039119	4.830736
cámara	134.498676	4.616958	carrera	159.660114	4.691435
década	79.022643	4.385999	caballo	82.492738	4.404663
método	71.965483	4.345374	llegada	72.878495	4.350849
sábado	71.026478	4.339670	cadena	68.664344	4.324982
código	68.176971	4.321888	mirada	68.150978	4.321723
lógica	40.016565	4.090505	delito	39.204276	4.081599
máquina	37.930609	4.067257	minuto	35.714686	4.041116
química	28.010945	3.935608	tabaco	27.393606	3.925931
mérito	23.936508	3.867350	locura	24.046979	3.869349
pájaro	17.698135	3.736237	pelota	17.512933	3.731669
sólido	15.482213	3.678154	botella	15.618677	3.681964
cúpula	14.816136	3.659060	rodilla	12.164828	3.573452
cólera	14.267029	3.642662	gallina	13.916121	3.631849
párrafo	13.678933	3.624385	camisa	13.724421	3.625827
célula	13.363765	3.614264	pasaje	13.480735	3.618048
pánico	10.787187	3.521269	pureza	10.767692	3.520484

bóveda	10.426531	3.506505	veneno	10.241329	3.498724
víbora	5.552802	3.232996	gusano	5.650277	3.240549
sótano	5.192146	3.203848	dilema	5.361102	3.217747
médula	4.880227	3.176959	ballena	4.854234	3.174641
sátira	4.548814	3.146438	vereda	4.656036	3.156549
túnica	4.532568	3.144885	natura	4.623544	3.153510
látigo	4.243393	3.116276	follaje	4.292131	3.121231
sílaba	3.668293	3.053078	cuchara	3.723529	3.059563
pícaro	3.388866	3.018700	cerezo	3.369371	3.016197
ráfaga	3.356375	3.014521	gitano	3.362873	3.015360
séquito	3.288143	3.005609	cigarro	3.249153	3.000434
sínodo	2.654558	2.912753	coyote	2.583076	2.900913
tópico	2.670804	2.915400	chiquillo	2.612319	2.905796
zócalo	2.251663	2.841359	carroza	2.329642	2.856124
títere	2.082707	2.807535	mucosa	2.066461	2.804139
lóbulo	1.504358	2.666518	chorizo	1.514105	2.669317
sésamo	1.280166	2.596597	boquilla	1.312658	2.607455
búfalo	1.231429	2.579784	pijama	1.172944	2.558709
dígito	1.224931	2.577492	cuneta	1.163197	2.555094
pétalo	0.939005	2.462398	penique	0.958500	2.471292
sílice	0.922759	2.454845	filete	0.903264	2.445604
lúpulo	0.672575	2.318063	papiro	2.105451	2.812245
cháchara	0.435386	2.130334	remesa	0.428888	2.123852

References

Alcina, J. & Blecua, J. M. (1975). *Gramática española*. Barcelona: Editorial Ariel.

Baković, E. (2016). Exceptionality in Spanish stress. *Catalan Journal of Linguistics, 15,* 9-25.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat. Neurosci., 8,* 389-395.

Boersma, P. & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program].

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on semantic P600 effects in language comprehension. *Brain Research Reviews 59(1),* 55–73.

Broś, K. (2015). Percepción de acento y acortamiento vocálico en español. *Itinerarios, 22,* 13-34.

Bybee, J. (2001). *Phonology and language use.* Cambridge: Cambridge University Press.

Bybee, J. (2006). *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Coetzee, A. (2016). A comprehensive model of phonological variation: grammatical and non-grammatical factors in variable nasal place assimilation. *Phonology 33,* 211-246.

Coetzee, A. & Kawahara, S. (2013). Frequency biases in phonological variation. *Natural Language and Linguisic Theory 31,* 47-89.

Connine, C.M., Mullennix, J, Shernoff, E & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology. Learning, memory, cognition 16(6),* 1084-96.

Cutler, A., Dahan, D., van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech 40(2),* 141-201.

Davies, M. (2002). *Corpus del Español*. Available at http://www.corpusdelespanol.org.

Delorme, A. & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. (pdf, 0.7 MB) *Journal of Neuroscience Methods, 134*, 9-21.

Domahs, U., Wiese, R., Bornkessel-Schlesewsky, I. D., & Schlesewsky, M. (2008). The processing of German word stress: Evidence for the prosodic hierarchy. *Phonology*, 25, 1-36.

Domahs, U., Genc, S., Knaus, J., Wiese, R., and Kabak, B. (2012a). Processing (un-)predictable word stress: ERP evidence from Turkish. *Language and Cognitive Processes, 28 (3)*, 335-354.

Domahs, U., Knaus, J., Orzechowska, P. & Wiese, R. (2012b). Stress "deafness" in a language with fixed word stress: an ERP study on Polish. *Frontiers in Psychology, 3*, 439.

Domahs, U., Knaus, J., Shanawany, H., Wiese, R. (2014). The role of predictability and structure in word stress processing: an ERP study on Cairene Arabic and a cross-linguistic comparison. *Frontiers in Psychology 5*, 1151.

van Donselaar, W., Koster, M., Cutler, A. (2005). Exploring the role of lexical stress in lexical recognition. *Quarterly Journal of Experimental Psychology 58*, 251-273.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., Carreiras, M. (in press). EsPal: One-stop Shopping for Spanish Word Properties. *Behavior Research Methods*.

Dupoux, E., Pallier, C., Sebastián-Gallés, N. & Mehler, J. (1997). A destressing 'deafness' in French? *Journal of Memory and Language, 36* 406–421.

Dupoux, E., Peperkamp, S. & Sebastian-Gallés, N. (2001). A robust method to study stress "deafness". *Journal of the Acoustical Society of America, 110(3)*, 1606-1618.

Flege, J.E., Takagi, N. & Mann, V. (1996). Lexical familiarity and English-language experience affect Japanese adults' perception of /r/ and /l/. *Journal of the Acoustical Society of America 99(2),* 1161-1173.

Friederici, A.D. (2004). Event-related brain potential studies in language. *Current Neurology and Neuroscience Reports 4(6),* 466-70.

Friederici, A.D., Friedrich, M. & Christophe, A. (2007). Brain responses in 4-month-old infants are already language specific. *Current Biology 17,* 1208–1211.

Frisch, S. & Schlesewsky, M (2001). The N400 reflects problems of thematic hierarchizing. *Neuroreport 12,* 3391–3394.

Goldinger, S.D., Luce, P.A., Pisoni, D.B. & Marcario, J.K. (1992). Form-based priming in spoken word recognition: the roles of competition and bias. *J. Exp. Psychol. Learn. Mem. Cogn., 18,* 1210-1238.

Hanulíková, A., McQueen, J.M., Mitterer, H. (2010). Possible words and fixed stress in the segmentation of Slovak speech. *Quarterly Journal of Experimental Psychology 63(3),* 555-579.

Harris, J. (1969). Spanish phonology. Cambridge, MA: MIT Press.

Harris, J. (1983). *Syllable Structure and Stress in Spanish*. Cambridge, MA: MIT Press.

Hinojosa, J.A., Moreno, E.M., Casado, P., Muñoz, F. & Pozo, M.A. (2005). Syntactic expectancy: an event-related potentials study. *Neuroscience Letters 378,* 34–39.

Hochberg, J. (1988a). First steps in the acquisition of Spanish stress. *Journal of Child Language, 15,* 273-292.

Hochberg, J. (1988b). Learning Spanish stress: developmental and theoretical perspectives. *Language, 64(4),* 683-707.

Honbolygó, F. & Csépe, V. (2013). Saliency or template? ERP evidence for long-term representation of word stress. *International Journal of Psychophysiology, 87(2),* 165-72.

Inkelas, S. & Orgun, C.O. (2003). Turkish stress: A review. *Phonology, 20(1),* 139-161.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M.J., Iragui, V. & Sejnowski, T.J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology, 37,* 163-178.

Jusczyk, P.W., Houston, D.M., Newsome, M. (1999). The Beginnings of Word Segmentation in English-Learning Infants. *Cognitive Psychology 39(3),* 159-207.

Kenstowicz, M. & Kisseberth, Ch. (1979). *Generative Phonology*. Cambridge, Ms: Academic Press.

Knaus, J., Wiese, R., & Janßen, U. (2007). The processing of word stress: EEG studies on task-related components. *Proceedings of the International Congress of Phonetic Sciences 2007* (pp. 709–712). Saarbrücken.

Kolk, H & Chwilla, D. (2007). Late positivities in unusual situations. *Brain and Language, 100,* 257–261.

Kutas, M. & Hillyard, S.A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307,* 161–163.

Lleó, C., & Arias, J. (2006). Foot, word, and phase constraints in first language acquisition of Spanish stress. In F. Martínez-Gil & S. Colina (Eds.). *Optimality Theoretic studies in Spanish Phonology* (pp. 470-496). Amsterdam/Philadelphia: John Benjamins.

Lopez-Calderón, J., & Luck, S.J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in human neuroscience, 8*, 213.

Luce, P.A. (1986). *Neighborhoods of words in the mental lexicon*. Unpublished doctoral dissertation, Indiana University, Bloomington, IN.

Magne, C., Astésano, C., Aramaki, M., Ystad, S., Kronland-Martinet, R., & Besson, M. (2007). Influence of syllabic lengthening on semantic processing in Spoken French: Behavioral and electrophysiological evidence. *Cerebral Cortex, 17*, 2659-2668.

Martínez-Paricio, V. & Torres-Tamarit, F. (2018). Trisyllabic hypocoristics in Spanish and layered feet. *Natural Language & Linguistic Theory* (online first).

Michelas, A., Frauenfelder, U.H., Schön, D. & Dufour, S. (2016). How deaf are French speakers to stress? *Journal of the Acoustical Society of America, 139*, 1333.

Mołczanow, J., Wiese, R., Domahs, U. & Knaus, J. (2013). The lexical representation of word stress in Russian: Evidence from event-related potentials. *The Mental Lexicon 8(2),* 164-194.

Morales-Front, A. (1999). El acento. In R. Nuñez-Cedeño & A. Morales-Front (Eds.). *Fonología Generativa Contemporánea de la Lengua Española,* 1st ed. (pp. 203-230). Washington, DC: Georgetown University Press.

Morales-Front, A. (2014). El acento. In R. Nuñez-Cedeño, S. Colina & T. Bradley (Eds.). *Fonología Generativa Contemporánea de la Lengua Española,* 2nd ed. (pp. 235-265). Washington, DC: Georgetown University Press.

Nooteboom, S. (1997). Prosody of speech: Melody and rhythm. In W.J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 640–673). Oxford: Blackwell.

Norris, D., McQueen, J. M. & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behav. Brain Sei. 23,* 299-325.

Norris, D., McQueen, J.M., Cutler, A., Butterfield, S. (1997). The Possible-Word Constraint in the Segmentation of Continuous Speech. Cogniive Psychology 34, 191–243.

Nosofsky, R.M. (1988). Similarity, frequency and category representation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14,* 54-65.

Oldfield, R.C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9(1),* 97-113.

Ortega-Llebaria, M. & Prieto, P. (2007). Disentangling stress from accent in Spanish: Production patterns of the stress contrast in de-accented syllables. In P. Prieto, J. Mascaró & M-J. Solé (Eds.), *Segmental and prosodic issues in Romance Phonology* (pp. 155-176). Amsterdam: John Benjamins.

Ortega-Llebaria, M., Prieto, P., & Vanrell, M.M. (2008). Perceptual evidence for direct acoustic correlates of stress in Spanish. In J. Trouvain & W.J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1121–1124). Saarbrücken, Germany, 6–10 August 2007.

Ortega-Llebaria, M. & Prieto, P. (2009). The perception of stress in Castilian Spanish. The effects of sentence intonation and vowel type. In M. Vigário, S. Frota & M.J. Freitas (Eds.), *Interactions in Phonetics and Phonology* (pp. 35-69). Amsterdam/Philadelphia: John Benjamins.

Peperkamp, S., and Dupoux, E. (2002). A typological study of stress "deafness". In C. Gussenhoven & N. Warner (Eds.) *Laboratory Phonology* (pp. 203–240). Berlin: de Gruyter.

Peperkamp, S., Vendelin, I., & Dupoux, E. (2010). Perception of predictable stress: a crosslinguistic investigation. *Journal of Phonetics, 38*, 422–430.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137-157). Amsterdam: John Benjamins.

Piñeros, C-E. (2016). The phonological weight of Spanish syllables. In R.A. Núñez Cedeño (Ed.) *The syllable and stress: Studies in honor of James W. Harris* (pp. 271–314). Berlin: De Gruyter.

Pisoni, D.B., Nusbaum, H.C., Luce, P.A, & Słowiaczek, L.M. (1985). Speech Perception, Word Recognition and the Structure of the Lexicon. *Speech Commun., 4(1-3),* 75–95.

Poeppel, D. (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Commun., 41*, 245-255.

Poeppel, D., Idsardi, W. & Wassenhove, W. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society B, 363,* 1071-1086.

Pons, F. & Bosch, L. (2007). The perception of lexical stress patterns by Spanish and Catalan infants. In P. Prieto, J. Mascaró & M-J. Solé (Eds.) *Segmental and prosodic issues in Romance Phonology* (pp. 199-218). Amsterdam: John Benjamins.

Prieto, P. (2006). The relevance of metrical information in early prosodic word acquisition: A comparison of Catalan and Spanish. *Language and Speech, 49(2),* 233-261.

Quilis, A. (1981). *Fonética acústica de la lengua española*. Madrid: Gredos.

Rahmani, H., Rietveld, T., Gussenhoven, C. (2015). Stress "Deafness" Reveals Absence of Lexical Marking of Stress or Tone in the Adult Grammar. Plos One 10(12), e0143968.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

REAL ACADEMIA ESPAÑOLA: Banco de datos (CORPES XXI) [online]. *Corpus del Español del Siglo XXI (CORPES)*. http://www.rae.es.

Roca, I. (2005). Saturation of parameter settings in Spanish stress. *Phonology, 22(3),* 345–394.

Roca, I. (2006). The Spanish stress window. In F. Martínez-Gil & S. Colina (Eds.) *Optimality Theoretic studies in Spanish Phonology* (pp. 470-496). Amsterdam/Philadelphia: John Benjamins.

Savin, H. (1963). Word-Frequency Effect and Errors in the Perception of Speech. *The Journal of the Acoustical Society of America, 35,* 200.

Sebastián-Gallés, N., Martí, M.A., Carreiras, M. & Cuetos, F. (2000). LEXESP: Léxico informatizado del español (online database). Barcelona: Ediciones Universitat de Barcelona.

Semon, R. (1923/1909). *Mnemic Psychology.* London: George Allen & Unwin Ltd.

Schonwiesner, M., Rubsamen, R. & von Cramon, D.Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *Eur. J. Neurosci., 22,* 1521-1528.

Schwab, S., & Dellwo, V. (2017). Intonation and talker variability in the discrimination of Spanish lexical stress contrasts by Spanish, German and French listeners. *The Journal of the Acoustical Society of America*, *142*(4), 2419–2429.

Schwab, S., Giroud, N., Meyer, M. & Dellwo, V. (2020). Working memory and not acoustic sensitivity is related to stress processing ability in a foreign language: An ERP study, *Journal of Neurolinguistics, 55*, 100897. https://doi.org/10.1016/j.jneuroling.2020.100897.

Torreira, F., Simonet, M., & Hualde, J.I. (2014). Quasi-neutralization of stress contrasts in Spanish. In N. Campbell, D. Gibbon & D. Hirst (Eds.), *Proceedings of Speech Prosody 2014*, 197-201.

van Herten, M., Kolk, H.H.J. & Chwilla, D.J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research, 22,* 241–255.

van Heuven, W. J., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of memory and language 39(3),* 458-483.

Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders 3(1),* 64-73.

Ylinen, S., Strelnikov, K., Huotilainen, M. & Näätänen, R. (2009). Effects of prosodic familiarity on the automatic processing of words in the human brain. *International Journal of Psychophysiology, 733,* 362–368.