A heuristic corpus for English word prosody: disyllabic nonce words

Sophie Herment & Gabor Turcsan

Aix-Marseille University, Laboratoire Parole et Langage, Aix-en-Provence, France

sophie.herment@univ-amu.fr, gabor.turcsan@univ-amu.fr

Abstract

It is generally admitted that disyllabic words in English are stressed according to their morphological make-up. While prefixed words show differential behaviour according to major grammatical category, non-derived nouns are allegedly trochaic and underived verbs are either iambic or trochaic following rules of quantity-sensitivity. This paper presents a database which was compiled in order to test native speakers' intuition about the stress of disyllables. 53 nonsense words were created displaying different phonological and morphological structures forced by the spelling. These words were embedded in sentences so that each form appears twice, once as a nominal and once as a verbal form. We recorded 20 speakers reading 106 sentences giving 2120 tokens. The construction of nonce words is the main issue at stake, the paper is therefore concerned with methodological questions regarding the design of a heuristic corpus. The data will be freely available on SLDR for the scientific community.

Index Terms: resources, database, phonology, prosody, disyllables, nonce words.

1. Introduction

It is generally admitted that disyllabic words in English are stressed according to i. grammatical category (noun/verb), ii. syllable weight (light/heavy) and iii. lexical properties (*e.g.* prefixation). While prefixed words show differential behaviour according to major grammatical category (Noun, Verb and Adjective), non-derived nouns are allegedly trochaic and non-derived verbs are either iambic or trochaic following rules of quantity-sensitivity (see [1] for an overview). Table 1 below illustrates the major word classes as far as stress is concerned:

	N	V	A
/10/	paper, fellow	offer, *comfort*	clever, *narrow*
/01/	*parade, debate, July, hotel*	neglect, cajole, *begin*	distinct, extreme

Table 1. *Interaction of syllable weight (heavy H/light L), extrametricality, morphology, analogy and part of speech category: exceptional classes in italics.*

Nouns are generally trochaic regardless of syllable weight (HL *paper*, LH *fellow*), as opposed to verbs, which are either trochaic (LL *offer*) or iambic (LH *cajole*) following syllable weight. Moreover, all verbs ending in a consonant cluster (*neglect*) are late stressed. Disyllabic adjectives behave like verbs. There are systematic exceptions to these patterns, like verbs having a prefix + root structure (*begin*), nouns derived from verbs (*debate*) and verbs derived from nouns (*comfort*), and nouns containing specific endings (*parade*). We can also find lexical exceptions like *hotel* or *July*.

In order to test native speakers' intuition about the stress of disyllables, an experiment was carried out involving reading tasks where nonce words were embedded. Similar experiments have been proposed for Spanish [2] or for Italian [3], languages much alike English in that they also display stress patterns conditioned by either syllable weight (phonology) (see [4]) or language specific lexical properties (morphology) (see [5] for a review).

This paper presents the compilation of the corpus conceived as a heuristic corpus (see [6]), and in particular the making up of the nonsense words. Given that nonce words do not have lexical properties, it is impossible to come up with a comprehensive list reflecting all the above categories. We can test syllable weight and grammatical category relatively easily, but prefixation to some extent only and systematic lexical exceptions not at all.

2. Making up of nonce words

2.1. Syllable weight

We focused first on syllable weight. We made up words combining different syllable weights:

- Light/Light (LL): *befin*;
- Light/Heavy (LH): recane;
- Heavy/Light (HL): *furna*;
- Heavy/Heavy (HH): *hastelk*.

One of the major difficulties of having a balanced nonce word corpus is that English spelling allows various possible pronunciations and therefore different rhyme structures: *manem* could either have an LL structure [məˈnem] or an HL structure [ˈmeɪnəm]. *Calbain* could be pronounced [ˈkælbən] (HL) or [kælˈbeɪn] (HH). *Capult* could be [kəˈpʌlt] (LH) or [ˈkeɪpəlt] (HH) and it is possible to imagine at least 4 different pronunciations for *divey*: [ˈdɪvi] (LL), [ˈdɪveɪ] (LH), [ˈdaɪvi] (HL) and [daɪˈveɪ] (HH). We had to ensure that all types would be sufficiently represented in the database.

We also paid attention to the different possible heavy syllable types: *furnoy* contains a heavy final syllable with a VV type [ˈfɜːnɔɪ], while the final heavy syllable of *ducasp* is of the VC type [djuːˈkæsp].

2.2. Nature of the word final consonant

Following [7]'s claim that certain configurations in final unstressed syllables in English do not seem to exist in verb forms, we also took the nature of the word final consonant into consideration. According to [7], there are no constraints on noun forms, but in disyllabic trochaic verbs, no final cluster and no final schwa plus non coronal sequences can be found. For instance, verbal forms like *meluct* and *lanop* are expected to be stressed on the final syllable because, respectively, of the final consonant cluster and of the non-coronal coda consonant.

2.3. Grammatical category

In order to test the influence of the grammatical category of the word, the same nonce words were embedded in carrier sentences once in a nominal and once in a verbal position:

My Mum likes these ....
She often ... when she's tired.

So as to mask the task, the words were used in a plural or 3rd person singular form in the sentences.

We also embedded the words in a sentence where they were understood as a proper noun (a place name) so as to see if the speakers pronounced the common noun and the proper noun differently, as it is often the case in the lexicon (the proper noun was written with a final -s, like the common noun):

My Mum likes these… vs. *My Mum lives in …*

2.4. Morphological structure

Our word list also contains items that may be associated to a prefix + root construction. Although we are fully aware that testing morphological structure without meaning is a delicate issue, we nevertheless thought it was worth a try. A few words were therefore made up with the common prefixes a-, ab-, ad-, be-, de-, di-, dis-, ex- and re-: *anem, abmone, adnop, bepult, debilk, dilact, disper, exbain, adnop, recane.*

3. Recordings

All in all, following the criteria described above, 53 words were created. We submitted the list to native speakers not participating in the experiment to exclude items that may call for analogical responses with existing items in the lexicon of English.

Each form appears twice, once as a nominal and once as a verbal form, randomly distributed in the test, so that the two word forms are never too close to each other:

My Mum lives in Ducasp.
She often galeafts when she's tired.
My Mum likes these furnoys.
She often calbens when she's tired.

We recorded 20 native speakers of English reading 106 sentences giving 2120 tokens embedded in two sentences.

The 20 speakers were between 20 and 30 years old. They all worked as language assistants at Aix-Marseille University at the time of the recording and they all had a university degree (B.A. or higher). Most of them found the task easy and did not figure out the aim of the experiment. They do not speak the same variety of English but to our best knowledge, while there may be slight differences in vowel reduction patterns, variation of stress placement in disyllables is non-existent.

The recordings took place at Aix-Marseille University, in a recording studio equipped with a Shure SM 58 microphone, a TASCAM M512 mixing desk, related to an iMac with a digidesign Mbox 2 sound card. The software Protools LE 7 was used.

Questionnaires collecting data about the speakers were filled in by each speaker, along with a consent form. The data have been anonymized, each speaker being assigned a code.

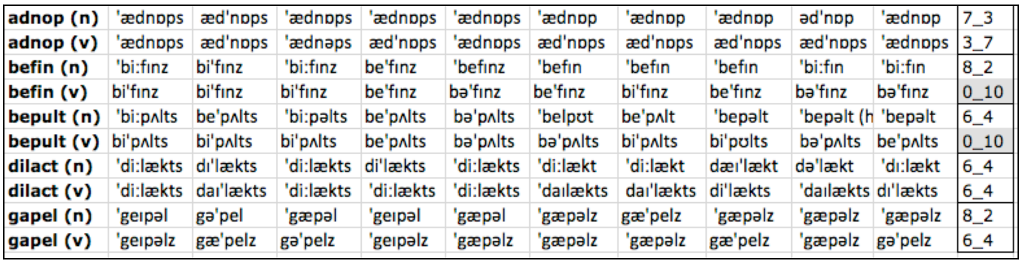
4. Annotation

The results were analysed in an auditory way by two specialists: in the overwhelming majority of cases stress placement was a straightforward issue, accompanied by vowel reduction. The remaining dubious cases, mostly heavy – heavy structures were submitted for judgment to two other trained phoneticians. Figure 1 at the bottom of the page shows an extract of the file containing the results for 10 speakers, with two types of information: the phonetic transcription and the stress pattern. The final column shows the proportion of speakers choosing a trochaic versus iambic pattern.

5. Results and perspectives

The results of our experiment are beyond the scope of this paper and they are detailed in [8]. They confirm our claim that nonce word corpora contribute to our understanding of how languages work. While some of our results confirm generalisations based on random language samples [7] or on dictionary data [9], others refine our knowledge of grammar. Let us just give an example for each type.

- The robustness of the noun/verb dichotomy as far as stress placement is concerned (see table 1 above) is a pleasant surprise given that nonce words do not have meaning. Most of the nouns in our corpus are trochaic (76%) while both iambs and trochees are equally found for verbs (48% are trochees).
- Contrary to what we can see in dictionary data, in our corpus final consonant clusters do not necessarily attract stress for verbs: /01/ *bepult, capult, debilk, galeaft, meluct, nabbast, nabellk*: all LH structures; /10/ *dilact, finlact, foslaint, hastelk*: all HH structures (except *dilact* pronounced /dɪ/). This shows that the observation that a final consonant cluster attracts stress is not an active constraint: verbs displaying this type of final just happen to have a light initial syllable in the English lexicon. Thus a nonce word corpus allows us to separate active dynamic constraints from static lexical patterns.



adnop (n)	'ædnɒps	æd'nɒps	'ædnɒps	'ædnɒps	'ædnɒps	'ædnɒp	'ædnɒp	'ædnɒp	əd'nɒp	'ædnɒp	7_3
adnop (v)	'ædnɒps	æd'nɒps	'ædnəps	æd'nɒps	'ædnɒps	æd'nɒps	æd'nɒps	æd'nɒps	æd'nɒps	'ædnɒps	3_7
befin (n)	'bi:fɪnz	bi'fɪnz	'bi:fɪnz	be'fɪnz	'befɪnz	'befɪn	'befɪn	'befɪn	'bi:fɪn	'bi:fɪn	8_2
befin (v)	bi'fɪnz	bi'fɪnz	bi'fɪnz	be'fɪnz	bə'fɪnz	be'fɪnz	bi'fɪnz	be'fɪnz	bə'fɪnz	bə'fɪnz	0_10
bepult (n)	'bi:pʌlts	be'pʌlts	'bi:pəlts	be'pʌlts	bə'pʌlts	'belpʊt	be'pʌlt	'bepəlt	'bepəlt (h	'bepəlt	6_4
bepult (v)	bi'pʌlts	bi'pʌlts	bi'pʌlts	be'pʌlts	bə'pʌlts	bə'pʌlts	bi'pʌlts	bi'pʊlts	bə'pʌlts	be'pʌlts	0_10
dilact (n)	'di:lækts	dɪ'lækts	'di:lækts	di'lækts	'di:lækts	'di:lækt	'di:lækt	dæɪ'lækt	də'lækt	'dɪ:lækt	6_4
dilact (v)	'di:lækts	daɪ'lækts	'di:lækts	'di:lækts	'di:lækts	'daɪlækts	daɪ'lækts	di'lækts	'daɪlækts	dɪ'lækts	6_4
gapel (n)	'geɪpəl	gə'pel	'gæpəl	'geɪpəl	'gæpəl	'gæpəlz	gæ'pelz	'gæpəlz	'gæpəlz	'gæpəlz	8_2
gapel (v)	'geɪpəlz	gæ'pelz	gə'pelz	'geɪpəlz	'gæpəlz	'gæpəlz	'gæpəlz	gæ'pelz	'gæpəlz	gə'pelz	6_4

Figure 1: *Results for a few words of the experiment (10 speakers)*

In this paper we also want to insist on the idea that a heuristic corpus should not necessarily serve one purpose and fall into oblivion. As mentioned in the introduction, the list of nonce words was created to test native speakers' intuition about the stress of disyllables in English but it can be used for other prosodic purposes. Rhythm can be an interesting issue: one of the speakers almost always stresses the second syllable of the words and reduces the first syllable of the verbs, not of the nouns. This behaviour is mysterious and definitely worth investigating. Although we could not find any significant difference in stress placement according to the origin of the speakers, vowel reduction patterns might have something to do with varieties: some speakers make more reductions on unstressed syllables than others. Moreover, some syllable types display more vowel reductions than others and it would be interesting to try and understand why. Generally speaking, anyone interested in phonological strength relations and more specifically in the interaction of word prosody and the licensing of segmental features (see [10] and the references therein) will find valuable material in the annotated corpus.

These are only possible lines of research and we think that a nonce word corpus can be helpful for the scientific community. This is the reason why it will soon be made freely available on the Speech Language Data Repository (http://www.sldr.org).

6. References

[1] Halle, M., "The Stress of English Words 1968-1998", Linguistic Inquiry 29, 539-568, 1997

[2] Bárkányi, Zs., "A fresh look at quantity sensitivity in Spanish", Linguistics 40, 375-394, 2002.

[3] Krämer, M., "Main stress in Italian nonce nouns", In D. Torck, and W. L. Wetzels [Eds], Romance Languages and Linguistic Theory 2006, Amsterdam and Philadelphia: John Benjamins, 127-141, 2009.

[4] Hyman, L., "A Theory of Phonological Weight", Stanford: CSLI publications, 2003.

[5] Hulst, H.G., van der, "Word accent", in H. van der Hulst [Ed], Word prosodic systems in the languages of Europe, Berlin & New York: Mouton de Gruyter, 3-116, 1999.

[6] Scheer, T. "Le corpus heuristique : un outil qui montre mais ne démontre pas", Corpus [On line], 3 | 2004, http://corpus.revues.org/210.

[7] Hammond, M., "English Phonology", Oxford: Oxford University Press, 1999.

[8] Turcsan, G. & Herment, S., "Making sense of nonce word stress in English", Proceedings on line of the 3rd international conference on English Pronunciation: Issues and Practices (EPIP3), Murcia, Spain, May 8-10, 2013. https://sites.google.com/site/epip32013/home

[9] Fournier, J-M., "Manuel d'anglais oral", Paris: Ophrys, 2010.

[10] Nasukawa, K. & Backley Ph. [Eds]. "Strength Relations in Phonology", Berlin and New York: Mouton de Gruyter, 2009.