A corpus study of English stress rules*

Roger Yu-Hsiang Lo
University of British Columbia

Abstract: The current project extends the phonological search function in Phonological CorpusTools (PCT) (Hall, Allen, Fry, Mackie, and McAuliffe 2016) to the syllable level. This updated search function is then used to quantify different stress patterns, as described in the literature (e.g., Hayes 1982), in English monomorphemic nouns and verbs, based on the phonological transcriptions from the English CELEX2 corpus (Baayen, Piepenbrock, and Gulikers 1995) and the pronunciations from the online Merriam-Webster dictionary. The results indicate that about 90% of nouns and verbs follow the respective dominant patterns when syntactic-categorical information is taken into account. These distributions also suggest a way to characterize a stress system probabilistically, which allows for the incorporation of other information such as syntactic category.

Keywords: lexical stress, stress placement rule, English, corpus, Phonological CorpusTools

1 Introduction

In recent years, language corpora have become valuable sources and an indispensable part of much linguistic research. To handle the vast amount of information usually contained in corpora, computational tools are therefore necessary. The current project aims to address the latter issue by (i) enhancing the functionality of an existing corpus-analyzing software — Phonological CorpusTools (PCT) (Hall et al. 2016), and (ii) demonstrating a possible application of such a tool with an English example. Specifically, I will show how PCT can be used to quantify the stress patterns in a subset of English nouns and verbs.

This paper is organized as follows. The next section gives an overview of PCT and its phonological search feature in an earlier version v1.3.0. Section 1.2 provides a brief description of the English stress rules examined in this study, along with a couple of illustrating examples. The focus of the current study is detailed in Section 2. Section 3 explains the design and use of the enhanced phonological search function. Key information regarding the CELEX2 corpus is also provided in the section. The results are presented in Section 4 and discussed in Section 5. Finally, Section 6 concludes the study.

1.1 Phonological CorpusTools (PCT)

PCT (http://phonologicalcorpustools.github.io/CorpusTools/) is a free open-source tool for performing phonological analysis on transcribed corpora. It is written in the Python programming language and features both a graphical user interface and a command-line interface that allow researchers without programming knowledge to carry out phonological analysis on language corpora. In the context

* I would like to acknowledge Dr. J. Scott Mackie for his assistance in helping me with the design of graphical user interface and thank Prof. Kathleen Currie Hall, Prof. Molly Babel, and Prof. Douglas Pulleyblank for their guidance.

Contact info: roger.lo@ubc.ca

of PCT, a corpus refers to a structured list of words, each of which has a phonological transcription and typically also a token frequency of occurrence within some body of usage.

In PCT, corpora are stored by means of a number of classes[1] of objects that reflect their linguistic hierarchy. The overall organization of the classes of objects in PCT is depicted in Figure A1 in Appendix A. In PCT v1.3.0, there is no class corresponding to syllable, so a word in PCT is structurally represented as a flat sequence of segments, each of which is in turn associated with a feature matrix that specifies its features and feature values. The necessity to add a syllable class is grounded in a number of points. First, for languages that use prosodic information contrastively at the lexical level (e.g., lexical tone in Cantonese and lexical stress in Russian), this prosodic information cannot be reliably imported or stored with the structure in version v1.3.0. An extra class — syllable — is needed if we want to keep all information when importing corpora of these languages into PCT.[2] Second, the phonological search function can benefit from the added syllable class as well. In PCT v1.3.0, phonological search only supports string search at the segmental level; there is no way to specify in what position (e.g., in onset or coda) a desired segment is to be in. For instance, if we configure the function to search for the words that contain the string [pra] in the corpus, the function will return words in which [p] acts as the onset (i.e., [.pra]) and those in which it functions as the coda (i.e., [p.ra]). With an additional syllable class, it is possible to refine the search by explicitly constraining the position within a syllable that certain segments can occur in. Similarly, the properties associated with the syllable, such as stress and tone, can also be added to the search environments.

1.2 English stress patterns

Besides targeting words with higher specificity, another application of an enhanced phonological search is to empirically examine various phonological generalizations regarding lexical stress or tone. In the remaining part of this paper, I will demonstrate one such application, which concerns English lexical stress.

English stress has provoked dense and continuous research (e.g., Chomsky and Halle 1968; Halle and Vergnaud 1987; Hayes 1980; Selkirk 1984). In this study, only a subset of rules that govern the primary stress in morphologically simple (i.e., monomorphemic) nouns and verbs is investigated. Secondary stress is left out of consideration because its distribution shows a more complicated pattern that is outside the scope of the current study. Only nouns and verbs are examined, to the exclusion of adjectives and adverbs, because English nouns and verbs follow distinct rules and thus form a nice contrast with each other. Adjectives and adverbs, as syntactic categories, are less homogeneous in terms of their stress patterns — some follow the stress patterns seen in nouns, while others pattern more closely with verbs (Burzio 1994). Since the goal of the study is to showcase the utility of the

[1] In object-oriented programming, a class is a blueprint for creating objects. A class provides initial values for a state and handles implementations of behavior.

[2] While it is possible to represent some syllabic information by modifying segments, such as using novel symbols or adding extra features to vowels, this strategy cannot entirely replace the functions the syllable class is designed to cover. For instance, while lexical stress or tone may be represented using features on the relevant vowels, it is not straightforward how other syllabic properties, like onset and coda, can be appropriately captured by this scheme. Furthermore, representing lexical stress or tone as contrastive features makes certain kinds of generalizations tricky: for example, PCT can only calculate the functional load of a single binary feature (e.g, [±voice]), which means that if we use something like [±tone3] to represent tones, PCT can only calculate the functional load of [±tone3] but not of the non-binary tone contrasts between, say, Tone2 and Tone3.

phonological search function in PCT after syllable structure is formally incorporated, rather than to provided a detailed description or an account for all attested stress patterns in English, I will restrict myself to only a few patterns that are commonly alluded to in the literature. Furthermore, I only consider monomorphemic words to simplify the discussion, as morphologically complex words, such as compounds and derived words, are subject to further rules (e.g., cyclic stress assignment). The review presented below follows the description in the first chapter of Kager (1989), but leaves out most theoretical details.

Before we dive into the stress rules, some remarks on the distinction between stress *placement* and stress *retraction*, and between *heavy* and *light* syllables are in order. In this paper, stress placement refers to the mechanisms that determine the *rightmost* stress in words, which is also the primary stress in most cases (e.g., *Calif**ó**rnia* /ˌkæ.lɪ.ˈ**fɔ**ː.nɪə/[3]), while stress retraction is used for the mechanisms that assign any remaining stress, for example, as the primary stress in the word ***á**necdòte* /ˈ**æ**.nɪk.ˌdəʊt/.

With regard to the syllable weight distinction in English, part of the distinction is stated with reference to the distinction between short and long vowels — only open syllables with short vowels count as light, and the other types all act as heavy.[4] As is typical in the literature, I assume an SPE-like taxonomy of the vowel system, distinguishing two classes with the feature [±long]. These two sets of vowels in English are illustrated in (1a) and (1b) respectively, following the transcription system adopted in the CELEX2 corpus.

(1) a. Underlyingly short vowels

p**i**t	/ɪ/	p**a**t	/æ/	p**o**t	/ɒ/
p**e**t	/ɛ/	p**u**tt	/ʌ/	p**u**t	/ʊ/

b. Underlyingly long vowels

b**ea**n	/iː/	b**ay**	/eɪ/	p**eer**	/ɪə/
b**ar**n	/ɑː/	b**uy**	/aɪ/	p**air**	/ɛə/
b**or**n	/ɔː/	b**oy**	/ɔɪ/	p**oor**	/ʊə/
b**oo**n	/uː/	n**o**	/əʊ/		
b**ur**n	/ɜː/	br**ow**	/aʊ/		

Note that, since the phonological transcription in the English CELEX2 corpus is mainly based on the Received Pronunciation (RP), the vowel in words like *burn* is transcribed as /ɜː/ and treated as [+long]. All the example words in the paper are transcribed using the phonological transcriptions from the CELEX2 corpus.

The rules that will be covered in the next two sections are listed in Figure 1, and each rule defines a set of stress patterns. Figure 1 also shows the derivation order of these rules; the order matters because there are overlaps among the environments specified in different stress rules, and some words could therefore be categorized into more than one stress pattern. I follow the extrinsic orderings proposed in Hayes (1982) when applying the rules, so if the lexical stress of a word matches

[3] Throughout the paper, in addition to marking stress on the orthography, as is the convention in the literature, I also transcribe the words using International Phonetic Alphabet (IPA) to ensure maximal clarity. Primary stress is marked with the acute accent (*á*) on the orthography and [ˈ] in IPA, and secondary stress with the grave accent (*à*) on the orthography and [ˌ] in IPA.

[4] I do not consider the issues related to ambisyllabic consonants in this study, as ambisyllabicity is sometimes used in the literature (e.g., Chomsky and Halle 1968) as a strategy to explain apparent exceptional words.

multiple rules, it will be treated as falling into the stress rule/pattern that is further up on the derivation order.

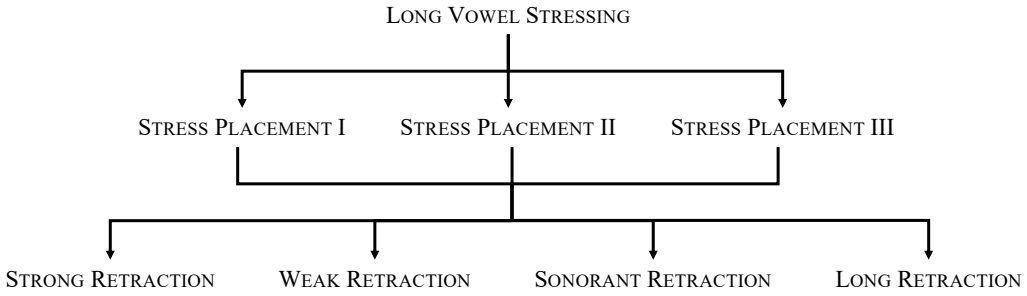


Figure 1: Extrinsic orderings for stress placement rules. The order starts from the very top, so STRESS PLACEMENT I follows LONG VOWEL STRESSING for example.

1.2.1 Stress placement rules

The first generalization (not shown in Figure 1), which applies across all lexical categories (i.e., nouns, verbs, adjectives, and adverbs) and with all segmental compositions, is that lexical monosyllabic words are stressed. Since all monosyllabic words necessarily satisfy this rule, I will not discuss this rule further.

The second generalization (2) is that words whose final syllable contains a long vowel always have a stress on the final syllable. Nouns that demonstrate this rule include *kangaróo* /kæŋ.gə.ˈ**ɹu**ː/ and *magazíne* /mæ.gə.ˈ**zi**ːn/, and verb examples include *maintáin* /meɪn.ˈ**teɪn**/ and *eráse* /ɪ.ˈ**ɹeɪz**/.

(2) LONG VOWEL STRESSING: long vowels in final syllables are stressed.[5]

The next rule concerns a subset of the noun class that consists of polysyllabic words whose final syllable does not contain a long vowel. The stress placement rule in this set of words can be described as in (3). Using *America* /ə.ˈmɛ.ɹɪ.kə/ to illustrate the process, we first note that the penultimate syllable /ɹɪ/ is light, so we place the primary stress on the antepenultimate syllable /mɛ/. On the other hand, in *horízon* /hə.ˈɹaɪ.zn̩/, the penultimate syllable /ɹaɪ/ is heavy, therefore it is the syllable that hosts the primary stress.

(3) STRESS PLACEMENT I (nouns)
- (i) Stress the antepenultimate (if present) provided the penultimate is light,
- (ii) Otherwise stress the penultimate.

However, there are a number of nouns that seem to violate (3): typically, these exceptions are made up of the nouns ending in /ɪ/, as in (4a), or in a syllabic sonorant consonant, as in (4b). Following the rule in (3), the words like *faculty* and *calendar* should have been pronounced as **facúlty*

[5] A more precise description that includes secondary stress would be: the final stress on long vowels in final syllables is always present, either as the primary or a secondary stress (e.g., *árchìve* /ˈɑː.ˌ**kaɪv**/ and *ánecdòte* /ˈæ.nɪk.ˌ**dəʊt**/). In Section 1.2.1, we focus on the instances with the primary stress on the final syllable; in Section 1.2.2, the focus will be on the instances with a secondary stress on the final syllable.

*/fəˈkʌl.tɪ/ and **carbúncle* /kɑːˈbʌŋ.kl̩/, but instead they are pronounced with the primary stress in the first syllable.

(4) a. Exceptional nouns ending in /ɪ/

fáculty	ánarchy	áutopsy	híerarchy	óligarchy
/ˈfæ.kl̩.tɪ/	/ˈæ.nə.kɪ/	/ˈɔː.təp.sɪ/	/ˈhaɪ.ə.rɑː.kɪ/	/ˈɒ.lɪ.gɑː.kɪ/

b. Exceptional nouns ending in a syllabic sonorant consonant

cárbuncle	árchangel	cálendar[6]	sálamander	cýlinder
/ˈkɑː.bʌŋ.kl̩/	/ˈɑːk.eɪn.dʒl̩/	/ˈkæ.lən.dɻ/	/ˈsæ.lə.mæn.dɻ/	/ˈsɪ.lən.dɻ/

As opposed to nouns, verbs appear to follow a different set of rules, which are summarized in (5). To illustrate, consider the verbs *develop* /dɪˈvɛ.ləp/ and *torment* /tɔːˈmɛnt/. We first discount the final consonant in both words, yielding /dɪ.vɛ.lə/ and /tɔː.mɛn/ respectively. Now the final syllable /lə/ in /dɪ.vɛ.lə/ is light, so we stress the penultimate syllable /vɛ/. On the other hand, the final syllable /mɛn/ in /tɔː.mɛn/ is still heavy, so the syllable /mɛn/ (or /mɛnt/) receives the primary stress.

(5) Stress Placement II (verbs)
- (i) Discount the final consonant,
- (ii) Stress the penultimate (if present) provided the final syllable is light,
- (iii) Otherwise stress the final syllable.

Stress Placement II in (5), however, can extend beyond verbs. For instance, the stress in the nouns in (6a) patterns like the verb *develop*, and the nouns in (6b) have the stress pattern similar to the verb *torment*.

(6) a. Nouns that pattern like the verb *develop*

molásses	proféssor	vanílla	confétti	banána
/məʊˈlæ.sɪz/	/pɹəˈfɛ.sə(ɹ)/	/vəˈnɪ.lə/	/kənˈfɛ.tɪ/	/bəˈnɑː.nə/

b. Nouns that pattern like the verb *torment*

cemént	evént	resúlt	dessért	ellípse
/sɪˈmɛnt/	/ɪˈvɛnt/	/ɹɪˈzʌlt/	/dɪˈzɜːt/	/ɪˈlɪps/

We now turn to the last stress placement rule. Stress Placement III in (7) differs from Stress Placement I in that the final syllable is taken into account, and it differs from Stress Placement II in the sense that the final consonant is taken into consideration when determining syllable weight. Words that belong to this pattern include the nouns *guitár* /gɪˈtɑː/ and *chiffón* /ʃɪˈfɒn/, and the verbs *acquiésce* /æ.kwɪˈɛs/ and *caréss* /kəˈɹɛs/.

(7) Stress Placement III (idiosyncratic, mainly nouns)
- (i) Stress the penultimate (if present) provided the final syllable is light.
- (ii) Otherwise stress the final syllable.

[6] The words *calendar*, *salamander*, and *cylinder* are relevant only in the English varieties where the final syllable contains a syllabic /ɻ/.

1.2.2 Stress retraction rules

So far we have focused on the English data that have only one stress (i.e., the primary stress), located on the rightmost stressable syllable. In this section, I will present data that involve secondary stress, in addition to primary stress. Even though in this study I do not focus on secondary stress, the presence of a secondary stress does influence the placement of the primary stress. For instance, in *húrricàne* /ˈhʌ.ɹɪ.ˌkeɪn/, the secondary stress on the syllable /ˌkeɪn/ acts as the reference for the primary stress on the syllable /ˈhʌ/ (See below for more detail).

The first two retraction rules — Strong Retraction and Weak Retraction — are described in (8) and (9) respectively. Weak Retraction differs from Strong Retraction in that Weak Retraction is a quantity-sensitive (i.e., weight-sensitive) style of retraction, while Strong Retraction is insensitive to the weight of the penultimate syllable. Some nouns that exemplify the Strong Retraction (8) are *ánecdòte* /ˈæ.nɪk.ˌdəʊt/ and *húrricàne* /ˈhʌ.ɹɪ.ˌkeɪn/, while verb examples include *récognìze* /ˈɹɛ.kəg.ˌnaɪz/ and *sátisfỳ* /ˈsæ.tɪs.ˌfaɪ/. Weak Retraction pattern (9) can be seen in nouns like *cýanìde* /ˈsaɪ.ə.ˌnaɪd/ and *peróxìde* /pə.ˈɹɒk.ˌsaɪd/.

(8) Strong Retraction (nouns and verbs)
If there is a stress in the final syllable,
(i) Stress the antepenultimate (if present),
(ii) Otherwise stress the penultimate.

(9) Weak Retraction (nouns)
If there is a stress in the final syllable,
(i) Stress the antepenult (if present) provided the penultimate is light,
(ii) Otherwise stress the penultimate.

A third stress retraction pattern that has drawn much attention in the literature is the so-called Sonorant Retraction. This retraction pattern involves a mixture of Strong and Weak Retraction, and can be summarized in (10). Some examples that comply with Sonorant Retraction are *mérchandìse* /ˈmɜː.tʃən.ˌdaɪz/ and *ámpersànd* /ˈæm.pə.ˌsænd/.

(10) Sonorant Retraction (nouns)
If there is a stress in the final syllable,
(i) Stress the antepenultimate (if present) provided the penultimate has a sonorant consonant as coda and is preceded by exactly one syllable,
(ii) Otherwise stress the penultimate.

The final primary stress retraction pattern to be discussed is Long Retraction (11), which separates the primary stress from the final secondary stress by a distance of two stressless syllables. Words following this pattern are rare; some examples include the noun *cátamaràn* /ˈkæ.tə.mə.ˌɹæn/ and the verb *amélioràte* /ə.ˈmiː.ljə.ˌɹeɪt/.

(11) Long Retraction (rare among nouns and verbs)
If there is a stress in the final syllable,
(i) Mark the penultimate as extrametrical,
(ii) Stress the pre-antepenultimate (if present) provided the antepenultimate is light,
(iii) Otherwise stress the antepenultimate.

2 The current study

As mentioned in Section 1, this study has two goals. The first one is to to build into PCT a representation for syllables, which enables the software to extract and store relevant information at the syllabic level from corpora. With syllabic information in place, the phonological search function can be enhanced by allowing the user to input parameters related to syllable structure. The second goal is to demonstrate the utility of the enhanced phonological search function, using the English stress patterns reviewed above as an example.

To push the second goal a bit further, the distribution information obtained allows for characterizing English lexical stress from a probabilistic perspective. Of course, given that only the stress patterns in monomorphemic nouns and verbs are examined, any characterizations drawn from the results cannot be generalized over the entire English lexicon. Nonetheless, the conceptualization presented in Section 5 can be straightforwardly extended when the stress patterns of the whole lexicon are considered.

3 Methods

The first half of this section provides a high-level description of the implementation and design of the syllable class and phonological search function; the second half details the features of the English CELEX2 corpus and the procedure of extracting relevant information using PCT. Due to the length and scope of this paper, a lot of technical details on the implementation of the syllable class and phonological search function are necessarily glossed over. The interested reader is referred to the software GitHub page (https://github.com/PhonologicalCorpusTools/CorpusTools), where all the code is publicly available.

3.1 Representation of syllable in PCT

Similar to the corpus class being an assembly of words, an inventory, and some attributes pertaining to a corpus, the syllable class can be viewed as consisting of a collection of attributes and a set of operations related to a syllable. The attributes include the onset, nucleus, coda, and depending on the language, stress and tone of a syllable. The operations include, for instance, asking a syllable for its onset or stress. In essence, the syllable class is therefore a computational way to capture the general formal properties of a linguistic syllable. The relationship between the syllable class and the other classes in PCT can be found in Figure A1 in Appendix A.

Once the syllable class is in place, we need a way to extract syllabic information from a corpus. So far there is no algorithm in PCT that can automatically delimit syllables from a plain transcription; PCT requires that syllables be delimited prior to the corpus being imported into the software. Once syllables are singled out, PCT then parses the segments in a syllable to corresponding syllabic constituents.

To assign a segment in the syllable to the proper constituent (i.e., onset, nucleus, or coda), PCT uses a simple strategy that hinges on the feature [±syllabic]. When loading a corpus into PCT, the user needs to specify a feature system to go with the corpus (e.g, SPE or Hayes (Hayes 2009)) and a mapping between the transcription used in the corpus and the feature system. Typically, a syllable only has one [+syllabic] segment that corresponds to the nucleus, with all the other segments being [−syllabic]. When a corpus with delimited syllables is imported into PCT, the software takes the

[+syllabic] segment as the nucleus, and all the [−syllabic] segments before the [+syllabic] segment are treated as the onset while all the [−syllabic] segments after the [+syllabic] segment are treated as the coda. As for stress and tone, the user specifies the corresponding symbols when importing the corpus (e.g., the user might tell PCT that the single quotation mark ’ maps to the primary stress, and the double quotation mark ” to a secondary stress).

The current study uses the CELEX2 English corpus, which comes with phonological transcriptions where syllables are delimited with the hypen - and primary and secondary stresses are marked with the single quotation mark ’ and the double quotation mark ” respectively. Using the procedure described above, all the syllables in the corpus are correctly parsed.

3.2 Phonological search function in PCT

In this section, I enumerate the important features of the updated phonological search graphical user interface after incorporating the newly-added syllable class. The main dialog window that pops up when the user clicks on *Phonological search...* from the option menu is shown in Figure 2.

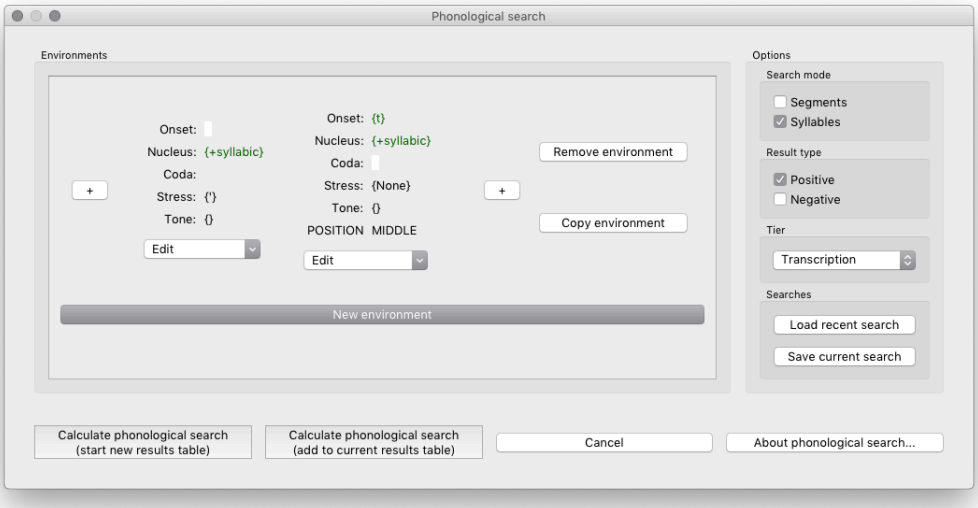


Figure 2: Phonological search graphical user interface as of version v1.4.0 in PCT. The specification here shows the phonological environment for /t/-flapping in General American Accent.

On the right-hand side, the user can set the search type to be *Positive* or *Negative*. A positive search means that the returned words will have elements that match the specified environments, while a negative search will return words that do *not* have any elements matching the environments — the complement of the positive results in the corpus. This negative search function is particularly useful for the current study, because it gives exceptional tokens to a specified stress pattern. Furthermore, the user can specify whether the search should be conducted at the segmental or syllabic level. If the segment mode is chosen, the user can input a sequence of segments and boundary symbols, and the entities in the loaded corpus that match the specified sequence will show up in the result window. In this mode, which is equivalent to the original phonological search function up until version 1.3.0,

all the information pertaining to syllable is discarded, so this type of search is suitable when the user only wants words that contain a certain sequence of segments, without constraining the position of the sequence in a syllable. On the other hand, by switching to the syllable mode, the user can further modify the searching environments by adding an unspecified syllable (i.e., a generic syllable that matches any of the different syllable types present in the corpus) or constructing a customized syllable in the popped *Construct syllables* window, the snapshot of which is shown in Figure 3.

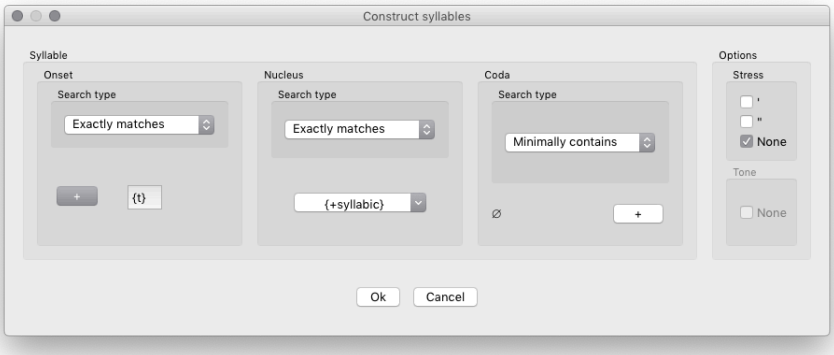


Figure 3: Dialog window for constructing a syllable in PCT. This window specifies a non-stressed syllable with /t/ in the onset and an unspecified coda.

In the window, the user is able to define the types of onset, nucleus, and coda with a syllable they intend to search for. For each syllable constituent, the user can additionally choose between four search types — *Exactly matches*, *Minimally contains*, *Starts with*, and *Ends with*. *Exactly matches* returns the results that include the specified pattern in the slot precisely. *Minimally contains*, in contrast, will return the words that have at least the specified sequence in the particular slot. For instance, if the user specifies the onset of the syllable to be /sp/ under *Minimally contains*, then the returned results will not only have the words with the /sp/ onset (i.e., *speed* or *span*), but will also have the words with onset such as /spr/ or /spl/ (i.e., *spring* or *split*). On the other hand, if the user chooses *Exactly matches*, then only the words having the /sp/ onset will be returned. *Starts with* returns the tokens that contain the matching strings at the initial position in the slot, while *Ends with* returns those that contain the matching strings at the final position. Two special specifications are worth mentioning here. When the user does not specify anything in the onset or coda, and *Exactly matches* is chosen, the results will only give the words that have zero onset (e.g., *egg* or *ink*) or coda (e.g., *ray* or *fly*). On the other hand, when no onset or coda is specified but *Minimally contains* is chosen, the results will contain words with all types of onset or coda, including zero onset or zero coda.

One additional feature when specifying segments is the *Set negative* option, shown in Figure 4. This feature is similar to the *Negative* search type: when this feature is selected, the results will exclude all words with the specified segments in this place.

Figure 4: Segmental specification options

3.3 English CELEX2 corpus

Now that the interface of phonological search is introduced, this section summarizes the key features of the English CELEX2 corpus and the inclusion criteria for the tokens used in the subsequent analyses.

The CELEX2 corpus is chosen over other corpora that are claimed to have more accurate word-frequency information, such as SUBTLEXus (Brysbaert and New 2009), because it contains syllabification, stress, and morphological constituent information the current study relies on and word-frequency information is not used in any case. The sources used to compile the English CELEX2 corpus include the *Oxford Advanced Learner's Dictionary of Current English* (Hornby, Cowie, and Lewis 1974) with about 41,000 lemmas and the *Longman Dictionary of Contemporary English* (Procter 1978) with around 53,000 lemmas. The corpus contains orthographic, phonological, morphological, syntactic, and frequency information for each token. The orthographic and phonological representations are both based on British varieties. The frequency information in the corpus is based on the 17.9 million token COBUILD/Birmingham corpus (https://collins.co.uk/pages/elt-cobuild-reference).

The analysis in the current study is based on the English lemma lexicon[7], as opposed to wordform. Lemma is an abstract way of representing a whole inflectional paradigm. For the noun class, it is the singular form that is used to represent the lemmas, except for pluralia tantum (e.g., *scissors*) for which the plural form is used. As for the verb class, it is the first person singular present-tense form that is used. Therefore, while words like *character* or *develop* are included in the analysis, their inflectional forms like *characters* or *developed* are excluded from the current analysis and discussion.

As mentioned above, only lemmas belonging to the noun or verb class are selected, and among the nouns and verbs, only monomorphemic lemmas[8] are subject to analysis. The selection of monomorphemic lemmas is made on the basis of the morphological status codes in the CELEX2

[7] The criteria the corpus uses to distinguish homophonic lemmas are (i) orthography of the wordforms (e.g., the nouns *peek* and *peak* are two different lemma), (ii) syntactic class (e.g., the noun *water* and the verb *water* are two different lemmas), (iii) inflectional paradigm (e.g., the noun *antenna*, meaning radio aerial, and the noun *antenna*, which is an anatomical feature of some insects, are different lemmas because the former has the plural *antennas*, while the latter has the plural *antennae*), (iv) morphological structure (e.g., the noun *rubber*, from *rub* + *-er*) and the noun *rubber*, which is the elastic substance, are two distinct lemmas, and (v) pronunciation of the wordforms (e.g., the verb *récount*, meaning to count again, and the verb *recóunt*, meaning to tell a tale, would be different lemmas).

[8] Pluralia tantum are inconsistently (mis)classified as monomorphemic in the corpus. As manually checking all of them requires going through the entire corpus, the ones that are labeled as monomorphemic are still included in the analysis.

corpus, the hierarchy of which is shown in Figure B1 in Appendix B — only the lemmas that are coded as either *R* (i.e., root) or *M* (i.e., monomorphemic) are kept, though upon careful inspection on a small chunk of tokens, a small fraction of polymorphemic lemmas are mistakenly classified as monomorphemic in the original CELEX2 encoding.[9] Furthermore, in the cases where a single lemma can have multiple pronunciation variants (e.g., /ˈvɪtəmɪn/ and /ˈvaɪtəmɪn/ for *vitamin*), only the first primary pronunciation variant is selected, which corresponds to a citation form based on RP. In total, the final number of nouns matching these criteria is 4,052 and the number of verbs is 998.

3.4 Procedure

The goal of the procedure is to classify the nouns or verbs selected from the CELEX2 corpus into various stress patterns. Specifically, I inspect each of the patterns derived by the stress rules and count how many words fit those patterns. The order of derivation follows that shown in Figure 1.

To use the phonological search function in PCT as a means to extract different stress patterns, I employ an "additive" approach. Broadly speaking, this approach includes adding stress placement rules into the phonological search one at a time and then recording the number of matching words with each additional rule, so we can deduce the proportion of the corpus associated with each rule. This procedure is used for the four rules that only involve the primary stress — Long Vowel Stressing (2), Stress Placement I for nouns (3), Stress Placement II for verbs (5), and Stress Placement III (7). For the rules that also make reference to secondary stress (i.e., all the retraction rules), the words have to be manually checked to see if they match any rules, as secondary stress is not consistently marked in the CELEX2 corpus. Instead, the secondary stress information is obtained from the online Merriam-Webster dictionary.[10] For example, the noun *anecdote* is transcribed as /ˈæ.nɪk.dəʊt/ in the CELEX2 corpus without any secondary stress, but the Merriam-Webster dictionary lists the pronunciation of the word as /ˈæ.nɪk.ˌdəʊt/, with a secondary stress on the final syllable.

In the rest of this section, I illustrate these steps with the nouns from the corpus. Following the ordering in Figure 1, the first step is to determine the number of nouns that fall within the category of Long Vowel Stressing (2). The chart in (12) shows the specification of the searching environment for this rule. Here the searching environment defines a primarily stressed syllable containing a long vowel at the end of the word. Notice that the combination of *Minimally contains* with no segments (i.e., ∅) in the onset and coda will match zero or more segments in these positions. Together, this environment singles out the words that have a final syllable whose nucleus is a long vowel.

---

[9] I do not know the exact number of misclassified lemmas, as this requires checking each token manually, which is not practical given the size of the lexicon. To estimate the proportion of misclassified lemmas, I randomly selected 100 lemmas and recorded the number of misclassified tokens — out of these 100 lemmas, two are misclassified. These misclassified lemmas are still included in the analysis. However, it should be noted that these misclassified lemmas do not necessarily violate stress rules. For instance, a misclassified item *introspect* /ɪn.tɹəʊ.ˈspɛkt/ consists of two morphemes, but it still follows Stress Placement II for verbs.

[10] I use the pronunciations from the Merriam-Webster dictionary, which are based on GAA, because both the online Oxford English Dictionary and Cambridge Dictionary, whose pronunciations are based on RP, do not transcribe secondary stress in their pronunciations. The classification results therefore inevitably reflect the stress patterns of a mixture of American and British varieties. This is admittedly not ideal; future research should ensure the consistency in phonological transcription data.

(12) Environment specification for the LONG VOWEL STRESSING rule

Onset	Nucleus	Coda	Stress	
minimal	exact	minimal	exact	#
∅	{+syllabic, +long}	∅	prim.	

The next rule to be considered is the STRESS PLACEMENT I rule for nouns. This rule has to be decomposed into several environment specifications. The specification in (13a) corresponds to (3-II), where the antepenultimate receives the primary stress because the penultimate is light. Notice that a light syllable is expressed through restricting the nucleus to be a short vowel and the coda to be empty. The environments (13b) through (13d) map to (3-III), which assigns the primary stress to penultimate when the penultimate is heavy (i.e., (13b) and (13c)) or when the penultimate is light in a two-syllable word (i.e., (13d)).

(13) Environment specification for the STRESS PLACEMENT I rule

a.

Onset	Nucleus	Coda	Stress	Onset	Nucleus	Coda	Stress		
minimal	exact	minimal	exact	minimal	exact	exact	exact	σ	#
∅	{+syllabic}	∅	prim.	∅	{+syllabic, −long}	∅	none		

b.

Onset	Nucleus	Coda	Stress		
minimal	exact	minimal	exact	σ	#
∅	{+syllabic, +long}	∅	prim.		

c.

Onset	Nucleus	Coda	Stress		
minimal	exact	minimal	exact	σ	#
∅	{+syllabic, −long}	{−syllabic}	prim.		

d.

	Onset	Nucleus	Coda	Stress		
#	minimal	exact	exact	exact	σ	#
	∅	{+syllabic, −long}	∅	prim.		

Next, we separate the nouns that match STRESS PLACEMENT II (5). The specifications in (14a) and (14b) correspond to (5-II), which target final syllables that are light after the final consonant is ignored. The specification in (14c) picks out the nouns in which the final syllable is heavy. In this case, the syllable has to have at least two coda consonants to be heavy, as the final consonant is overlooked.

(14) Environment specification for the STRESS PLACEMENT II rule

a.

Onset	Nucleus	Coda	Stress	Onset	Nucleus	Coda	Stress	
minimal	exact	minimal	exact	minimal	exact	exact	exact	#
∅	{+syllabic}	∅	prim.	∅	{+syllabic, −long}	∅	none	

b.

Onset	Nucleus	Coda	Stress	Onset	Nucleus	Coda	Stress	
minimal	exact	minimal	exact	minimal	exact	exact	exact	#
∅	{+syllabic}	∅	prim.	∅	{+syllabic, −long}	{−syllabic}	none	

c.

Onset	Nucleus	Coda	Stress	
minimal	exact	minimal	exact	#
∅	{+syllabic, −long}	{−syllabic} {−syllabic}	prim.	

Subsequently, we extract from remaining words that satisfy Stress Placement III (7) with the environment specifications listed in (15). Note that in (15b), we restrict the primary-stressed final heavy syllable to those containing a [−long] vowel and closed by one or more consonants. This is due to the fact that the words containing a final syllable that is heavy because of a [+long] vowel should have been identified by the Long Vowel Stressing rule specified in (12) (i.e., adding another corresponding rule is redundant here).

(15) Environment specification for Stress Placement III rule

a.

Onset	Nucleus	Coda	Stress	Onset	Nucleus	Coda	Stress	
minimal	exact	minimal	exact	minimal	exact	exact	exact	#
∅	{+syllabic}	∅	prim.	∅	{+syllabic, −long}	∅	none	

b.

Onset	Nucleus	Coda	Stress	
minimal	exact	minimal	exact	#
∅	{+syllabic, −long}	{−syllabic}	prim.	

The rest of the nouns that do not satisfy any of the four stress placement rules above are then checked manually to see if they fall into any of the retraction patterns described in (8) through (11), using secondary-stress information from the online Merriam-Webster dictionary.

The same procedure is also carried out for the verbs in the corpus. The results obtained from the aforementioned steps are discussed in the next section.

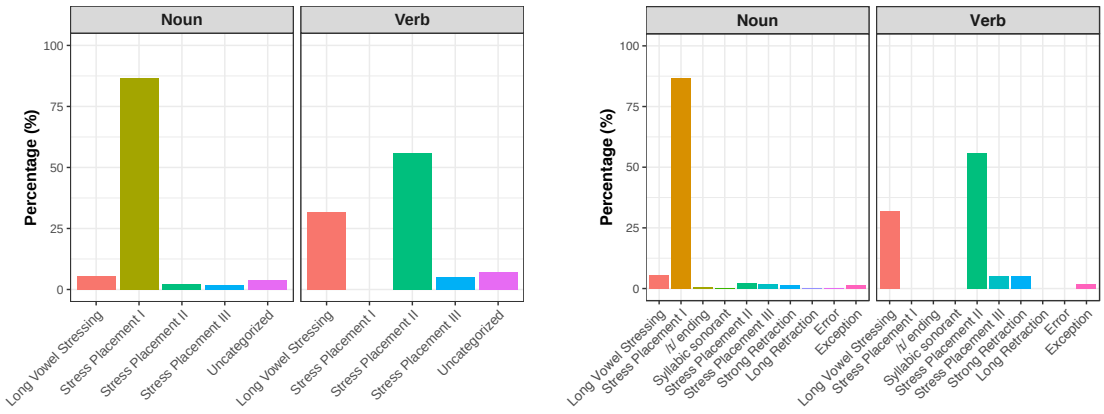
4 Results

The raw results output by PCT are summarized in Table 1 and visualized in Figure 5a. The results show that, for example, around 5.4% of all the nouns have the stress pattern that follows Long Vowel Stressing, that is, these nouns have a long vowel in the final syllable, *and* the primary stress is on the final syllable. Next, around 86.6% of the nouns have the stress patterns as described in Stress Placement I, and *after* these stress patterns are accounted for, about 2.3% of the nouns follow the stress patterns specified in Stress Placement II. Note that the nouns and verbs belonging to the "uncategorized" group are simply the ones that do not match any of the four stress patterns listed right above. Note also that since Stress Placement I is claimed to apply to nouns only and is therefore not included in the searching environments for the verbs, naturally this category is not applicable to verbs here (but I will have more to say about this point in Section 5).

The words in the "uncategorized" group are then manually examined and classified, based on secondary-stress information from the online Merriam-Webster dictionary. The results after the manual classification are tabulated in Table 2 and plotted in Figure 5b. As discussed in Section 1, nouns ending in /ɪ/ (e.g., *acrimon**y***, *alimon**y***, and *epilepsy*) or in a syllabic sonorant consonant (e.g., *badmin**ton*** and *partici**ple***) tend to be exceptional to Stress Placement I. These nouns are singled out and listed under Stress Placement I, as if we disregard the final syllable when applying the

Table 1: Distributional results output by PCT. The raw counts are in parentheses.

Rule	Noun	Verb
Long Vowel Stressing	5.4% (218)	31.9% (318)
Stress Placement I	86.6% (3,509)	—
Stress Placement II	2.3% (93)	56.0% (559)
Stress Placement III	1.9% (78)	5.1% (51)
Uncategorized	3.8% (154)	7.0% (70)
Total	100% (4,052)	100% (998)



(a) Output by PCT

(b) After manual classification

Figure 5: Distributions of stress patterns in monomorphemic nouns and verbs. The height of a bar represents the percentage of tokens falling in the stress pattern specified by the labeled stress rule. The heights of bars in a panel therefore sum to 100%. Note that there are no verbs in the categories defined by Stress Placement I, /ɪ/ ending, and Syllabic sonorant, as these are claimed to only apply to nouns.

stress rules, then they will conform to this stress pattern (Hayes 1980). The rest of the nouns and verbs are further subclassed into different retraction types. As the specifications for different retraction types can overlap to different extents (e.g., between STRONG and WEAK RETRACTIONS, and between STRONG and SONORANT RETRACTIONS), in cases where more than one possibility exists, STRONG RETRACTION will take precedent, given that the environments defined in the STRONG RETRACTION pattern are most general. However, even after these retraction patterns are considered, there is still a small percentage of nouns (around 1.3%) and verbs (around 1.9%) that seem to be exceptional with regard to the rules discussed in the paper.

Table 2: Distributional results output after manual classification. The raw counts are in parentheses.

Rule	Noun	Verb
LONG VOWEL STRESSING	5.4% (218)	31.9% (318)
STRESS PLACEMENT I	86.6% (3,509)	—
/ɪ/ ending	0.7% (27)	—
Syllabic sonorant	0.3% (12)	—
STRESS PLACEMENT II	2.3% (93)	56.0% (559)
STRESS PLACEMENT III	1.9% (78)	5.1% (51)
STRONG RETRACTION	1.3% (54)	5.3% (53)
LONG RETRACTION	0.1% (5)	0% (0)
Error	0.1% (5)	0% (0)
Exception[11]	1.3% (51)	1.7% (17)
Total	100% (4,052)	100% (998)

Based on the numbers from Table 2 and Figure 5b, a few generalizations can be made with respect to the stress patterns of nouns and verbs. With regard to nouns, the majority of them (i.e., 86.6%) indeed follows the STRESS PLACEMENT I rule that has been described specifically for English underived nouns in the literature. As for the verbs, again a predominant portion of verbs (i.e., 31.9% + 56.0% = 87.9% since the LONG VOWEL STRESSING environments also fit those for Stress placement rule II) adheres to the verbal stress pattern described in previous studies. The remaining 10 percent of nouns and verbs are unevenly distributed over a number of sub-patterns or are exceptional in terms of its stress pattern. If we treat the respective main stress patterns for nouns and verbs as the default, then these exceptional nouns and verbs are "exceptional" in the sense that they do not follow the respective main stress patterns. To put it in another way, these exceptional tokens require some kind of *lexical marking* in their underlying form to indicate that they belong to sub-patterns in terms of the position of their primary stress.

5 Discussion

One of the focuses of the current project is on the development of a tool that enables linguists to extract stress information in PCT. This tool allows for searching automatically for various syllable types and stress patterns. Using this tool, the second goal of this project can be achieved — quan-

[11] The exceptional verbs are listed in Appendix C.

tifying the distribution of stress patterns in English monomorphemic nouns and verbs. The results indicate that around 90% of the tokens in both syntactic categories follow the respective dominant patterns identified in the literature, while the remaining 10% are scattered unevenly across different subpatterns.

On a more theoretical note, one could ask how we are able to characterize lexical stress, given the distributional data (keeping in mind that these data are only from monomorphemic nouns and verbs). There are, in fact, two underlying questions associated with this issue. The first question involves how to quantify the stress predictability given the frequency distribution over different stress patterns. The second question concerns how to quantify the contribution of different linguistic factors on the distribution of stress patterns. To answer the first question, let us consider the two distributions (of nouns and verbs respectively) in Figure 5a. How do we quantify the stress predictability associated with these two distributions? One way to do this is through a measure called homogeneity, which is defined as follows:

$$\text{homogeneity} = \sum_{c \in \text{syntactic cat.}} \frac{n_c}{n_{\text{all}}} Q(N_c) = \sum_{c \in \text{syntactic cat.}} \frac{n_c}{n_{\text{all}}} \left[-2 \sum_{k \in \text{patterns in } c} \mathbb{P}(k) \log_2(\mathbb{P}(k)) \right]$$

In the definition above, $\mathbb{P}(k)$ stands for the probability of occurrence of the k-th pattern in the syntactic category c. Note that the value of homogeneity actually decreases as the system becomes more homogenous. The homogeneity associated with the distributions in Figure 5a is therefore 1.90 (bits):

$$\begin{aligned}
\text{homogeneity} = {} & (\frac{4052}{4052+998})(-2)[(0.054)(\log_2 0.054) + (0.866)(\log_2 0.866) \\
& + (0.023)(\log_2 0.023) + (0.019)(\log_2 0.019) + (0.038)(\log_2 0.038)] \\
& + (\frac{998}{4052+998})(-2)[(0.319)(\log_2 0.319) + (0.560)(\log_2 0.560) \\
& + (0.051)(\log_2 0.051) + (0.070)(\log_2 0.070)] \\
\approx {} & 1.90.
\end{aligned}$$

Compared with a hypothetical situation where the patterns in a syntactic category are evenly distributed (i.e., 20% for each pattern in nouns and 25% for each pattern in verbs, which results in a homogeneity of 4.51 bits), the distribution in Figure 5a has a much lower homogeneity value, meaning that the distribution is much more predictable.

Now to answer the second question, we need to compare two sets of distributions, one when the linguistic factor interested is not taken into consideration and the other when the factor is considered. In the present case, the linguistic factor we are interested in is syntactic category (i.e., nouns vs. verbs). The two sets of distributions in question and their corresponding homogeneity are summarized in Table 3. The contribution of syntactic category on the predictability of stress patterns is therefore $2.21 - 1.90 = 0.31$ (bits), which means that including syntactic category information increases the predictability/homogeneity of stress patterns by 0.31 (bits). Following the same procedure above, we can similarly derive the impact of different factors (e.g., non-stress-neutral affixes) on stress patterns.

To sum up the discussion, with the assistance of language corpora and computational tools, we can not only obtain a more accurate landscape of the distribution of stress patterns at the descriptive

Table 3: Distribution of stress patterns, with and without syntactic category information

Rule	Noun + Verb	Noun	Verb
Long Vowel Stressing	10.6% (536)	5.4% (218)	31.9% (318)
Stress Placement I	78.8% (3979)	86.6% (3,509)	—
Stress Placement II	5.9% (298)	2.3% (93)	56.0% (559)
Stress Placement III	1.5% (78)	1.9% (78)	5.1% (51)
Uncategorized	3.2% (159)	3.8% (154)	7.0% (70)
Total	100% (5050)	100% (4,052)	100% (998)
Homogeneity (bit)	2.21	1.90	

level, but it also allows us to characterize lexical stress using information-theoretically informed concepts.

6 Conclusion

The current work builds a syllable representation in PCT and extends the phonological search function to syllable level. To show how the updated phonological search function could aid linguistic research, I use the function to extract words from a subset of the English CELEX2 corpus that conform to various stress patterns described in the literature. The results indicate that the stress patterns in monomorphemic nouns and verbs are both dominated by the pattern claimed to be canonical for the respective category. I further argue that stress systems could be analyzed in a probabilistic fashion and could be quantified using distributional information.

References

Baayen, Harald R., Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX lexical database. CD-ROM.

Brysbaert, Marc, and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41:977–990.

Burzio, Luigi. 1994. *Principles of English stress*. Cambridge: Cambridge University Press.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York City, NY: Harper & Row, Publishers.

Hall, Kathleen Currie, Black Allen, Michael Fry, Scott Mackie, and Michael McAuliffe. 2016. Phonological corpustools, version 1.2. Computer program.

Halle, Morris, and Jean-Roger Vergnaud. 1987. *An essay on stress*. Cambridge, MA: MIT Press.

Hayes, Bruce. 1980. A metrical theory of stress rules. Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Hayes, Bruce. 1982. Extametricality and English stress. *Linguistic Inquiry* 13:227–276.

Hayes, Bruce. 2009. *Introductory phonology*. West Sussex: Wiley-Blackwell.

Hornby, Albert Sydney, Anthony Paul Cowie, and Jack Windsor Lewis. 1974. *Oxford advanced learner's dictionary of current English*. Oxford: Oxford University Press.

Kager, René. 1989. *A metrical theory of stress and destressing in English and Dutch*. Dordrecht: Foris Publications.

Procter, Paul. 1978. *Longman dictionary of contemporary English*. Harlow: Longman.

Selkirk, Elisabeth O. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Appendix A Organization of classes of objects in PCT

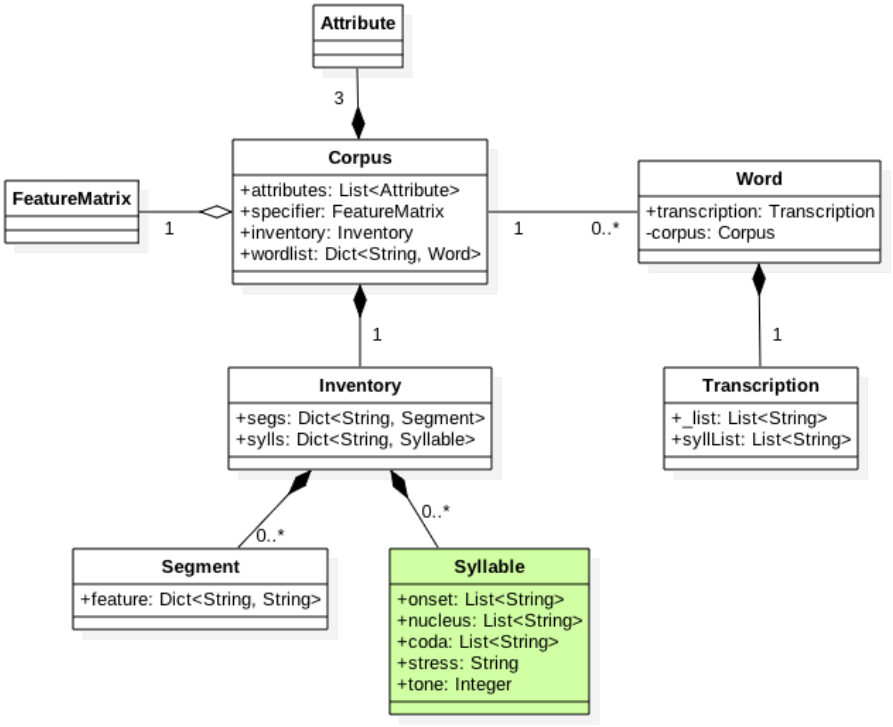


Figure A1: The Unified Modeling Language (UML) diagram of the relation between the existing object classes (the white boxes) and the new syllable class (the green box) in PCT as of version v1.4.0. In UML, a line between two classes mean *association* so that the two classes in a model can communicate with each other. A line with an empty diamond represents *aggregation*, where the child class (e.g., FeatureMatrix) can exist independently of the parent class (e.g., Corpus). A line with a solid diamond means *composition*, which typically specifies a part-whole dependency relationship between the two connected classes. For instance, Word is the parent class of Transcription. Note that the end with a diamond marks the parent class. The number associated with each line is the relationship's *multiplicity*, with 1 meaning "exactly one instance" and 0...* "zero or more instances". In aggregation and composition a child class can only have one parent class, but a parent class can have more than one child class. In this diagram, Corpus can only have one child Inventory instance but can be associated with multiple Word instances; however, each Word can only be associated with one Corpus.

Appendix B Morphological analysis in the English CELEX2 corpus

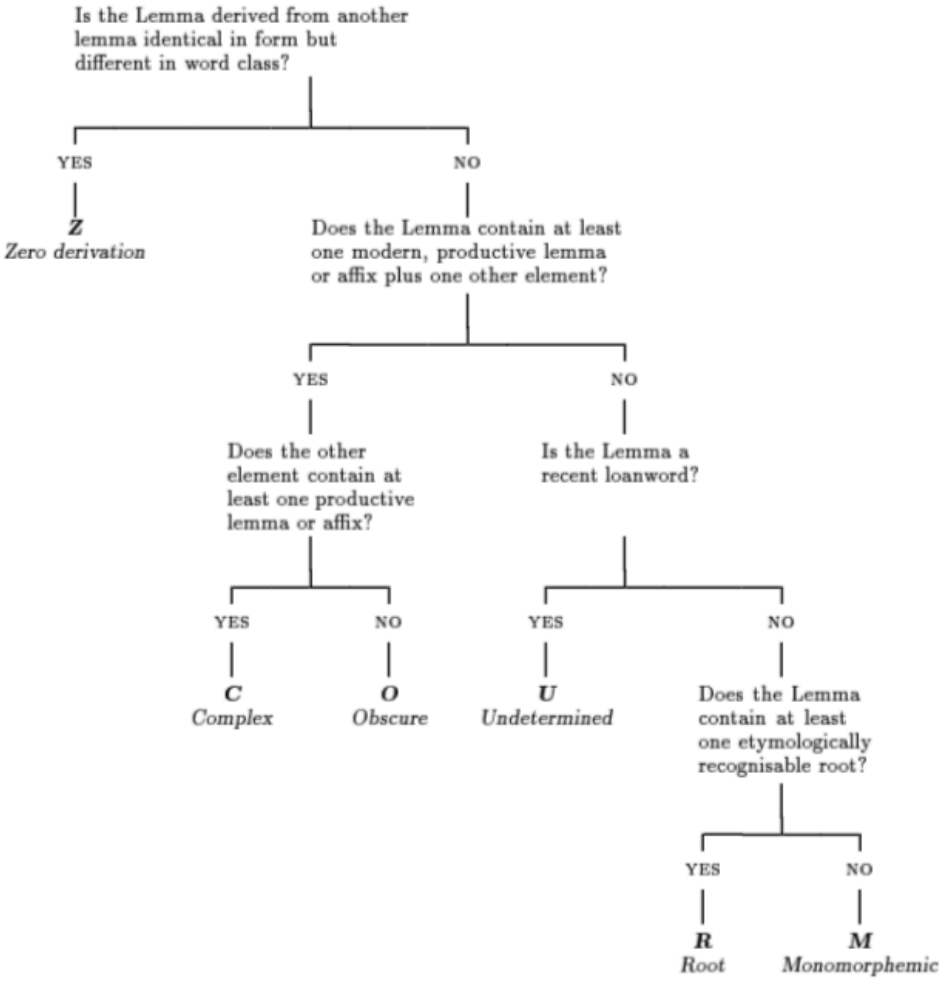


Figure B1: Morphological analysis in the English CELEX2 corpus

Appendix C Verb tokens with exceptional stress pattern

Table C1: Verbs with exceptional stress pattern

Lemma	Transcription in CELEX2	Lemma	Transcription in CELEX2
affranchise	/əˈfɹæn.tʃaɪz/	*flummox*	/ˈflʌ.məks/
argue	/ˈɑː.gjuː/	*gerrymander*	/ˈdʒɛ.ɹɪ.mæn.də/
balance	/ˈbæ.ləns/	*importune*	/ɪmˈpɔː.tjuːn/
benefice	/ˈbɛ.nɪ.fɪs/	*jettison*	/ˈdʒɛ.tɪ.sn̩/
challenge	/ˈtʃæ.lɪndʒ/	*license*	/ˈlaɪ.səns/
continue	/kənˈtɪ.njuː/	*metamorphose*	/ˌmɛ.təˈmɔː.fəʊz/
contribute	/kənˈtɪɹ.bjuːt/	*scavenge*	/ˈskæ.vɪndʒ/
discipline	/ˈdɪ.sɪ.plɪn/	*warrant*	/ˈwɒ.ɹənt/
experience	/ɪkˈspɪə.ɹɪəns/[12]		

[12] CELEX2 treats /ɪə/ here as belonging to the same nucleus.