The Core of the English Lexicon: Stress and Graphophonology

Véronique Abasq, Quentin Dabouis, Jean-Michel Fournier, Isabelle Girard

The Core of the English Lexicon: Stress and Graphophonology

Véronique Abasq, Quentin Dabouis, Jean-Michel Fournier & Isabelle Girard

published in 2019 in *Anglophonia* [online], 27
URL: http://journals.openedition.org/anglophonia/2317

RESUME

Cet article est une étude des 5000 mots les plus fréquents du lexique de l'anglais. Les régularités accentuelles et graphophonologiques sont évaluées dans le cadre défini par Fournier (2010b). Les résultats confirment l'efficacité de ce modèle et montrent que la plupart des exceptions appartiennent à des classes bien identifiables.

ABSTRACT

This paper is a study of the 5000 most frequent word-forms of the English lexicon. Both stress and graphophonological regularities are evaluated within Fournier's (2010b) framework. The results show the high efficiency of Fournier's model and that a large number of exceptions fall into well-known classes.

Mots-clés : accentuation lexicale, corpus, anglais, phonologie, morphologie

Key words: corpus, English, stress, phonology, morphology,

1. Introduction[1]

This article aims to give the first results of a study on the core of the English lexicon. It focuses on what Schmitt & Schmitt (2014)(Schmitt & Schmitt 2014) call "high-frequency vocabulary", i.e. the 3000[2] most frequent word families (word-forms clearly sharing a base, i.e. inflected forms and transparent derivatives, belong to the same family; e.g. *accepting, acceptance, unacceptable*).

Section 2 introduces the theoretical framework of our study. It presents the basic features of the Guierrian School (§2.1), Fournier's (2010b) systems of stress rules (§2.2) and graphophonological rules (§2.3) as well as our research questions (§2.4). Section 3 focuses on the methodology used for building our corpus and the treatment of its data. Section 4 deals with results yielded by the analysis of the corpus. §4.1 sets these results out according to the morphological, syntactic and syllabic features of the words the corpus comprises. In § 4.2, we discuss cases of isomorphism. And, in §§4.3 and 4.4, we analyse a dataset of 2737 units whose pronunciation is computed directly and discuss results on the stress patterns (§4.3) and graphophonological regularities of vowels with primary stress (§ 4.4) in order to evaluate the efficiency of the rules put forward by Fournier (2010b).

2. Framework

2.1. The Guierrian School

The "Guierrian School" is an approach which was introduced in the seventies by Guierre (1979). Its main characteristics are the use of pronouncing dictionaries to study the phonology of English, the use of morphology (e.g. elements such as suffixation, prefixation or compounding) and orthography (e.g. elements such as orthographic consonant geminates, vocalic digraphs or final mute <e>) when necessary.

This approach focusses on the assignment of lexical stress considering that each lexical unit has its own stress pattern. Lexical units are defined as semantically inseparable units. Therefore, the first step in lexical stress assignment is to make sure that the sequence under consideration is a

[1] We warmly thank all our colleagues for their advice and friendly discussions during ALOES 19ème Colloque d'Anglais Oral at Villetaneuse / Université Paris 13 (April 6th - 7th, 2018)

[2] Of which the 5000 most frequent words give a fair representation.

semantically inseparable unit. Cases as *dark room* or *re-act* are analysed as semantically separable sequences (a noun phrase in the first case, a semantically transparent prefix and a stem in the second) since the meaning of the whole sequence amounts to the combined meanings of its constitutive elements. These sequences have to be separated to complete the stress assignment procedure for each of the lexical units they are comprised of: *dark*, *room*, *re-* and *act*. Cases such as *dark*, *room*, *act* (simplex words), *darkroom* (compound words), *lovingly*, *infernal* (suffixed words, with either free[3] or bound bases), *react* (semantically opaque prefixed constructions) and *re-* (semantically transparent prefixes) are analysed as semantically inseparable units, and the stress assignment procedure is directly applied to each of them. Note that a distinction is also made between autonomous lexical units (words) and non-autonomous lexical units (semantically transparent prefixes), and this difference translates phonologically into different stress levels: primary and secondary, respectively.

Within this approach, only three levels of stress are acknowledged:

- primary stress (annotated with an acute accent, [ˈ] or /1/)
- secondary stress (annotated with a grave accent, [ˌ] or /2/)
- no stress (/0/)

The stress pattern of all lexical units (prefixes, words, suffixed words and compounds) is described as being regulated by the following four general stress principles (Fournier 2007, 2010b):

1. Every lexical unit has one and only one primary stress
2. There can be no sequence of two stresses within a lexical unit
3. No lexical unit can begin with two unstressed syllables
4. Syllables which receive neither stress /1/ nor stress /2/ are unstressed

and the placement of stress is determined by a system of rules that is presented in the next section.

2.2. Fournier's stress rules

Fournier (2010b) put forward a system of stress rules which are based on both morphological and segmental criteria.[4] This system governs the stress placement of monosyllables, disyllables and words of three syllables and more.

Stress placement in monosyllabic lexical units is determined by the only possible rule Monosyllable → /1/, e.g. *dárk*, *róom*, *áct*. Although stress placement in these cases is self-evident, the definition of monosyllables and syllable count deserves consideration. The final consonants found in such sequences as <Cm#> (e.g. *plasm*) and <Cle#> (e.g. *bible*) constitute the nucleus of a phonetic syllable in [ˈplæzm̩] and [ˈbaɪb̩]; they are called 'syllabic consonants' but they should not be analysed as separate syllables. If a strong ending like *-ic* or *-ical* imposes stress on the preceding syllable, it imposes it on the vowel to the left, not on the consonant: *plasm* [ˈplæzm̩]→ *plasmic* [ˈplæzmɪk]; *bible* [ˈbaɪb̩] → *biblical* [ˈbɪblɪk̩].

The final [m̩] in *plasm* and the final [t̩] in *bible* lose their syllabic status. In other words, they are phonetic, not phonological syllables. This means that stress placement rules belong to the phonological level, where *plasm* or *bible* are monosyllabic not disyllabic.

Stress placement in some disyllables and words of three syllables and more is determined by morphology when they are suffixed with endings (e.g. *-ade*, *-ic*, -C_2 + adjectival suffix in -V(C_0(e)), disyllabic suffixes such as *-ity* or -ION[5]) which prevent any reference to a base. The stress pattern of these suffixed words is computed directly, no matter the existence of a base. These endings, called strong endings, are each closely connected to a fixed position of primary stress:

[3] Contrary to prefixes, suffixes never constitute lexical units by themselves.

[4] Fournier's work on Guierre's rules essentially focused on the ordering of these rules, which led to a number of adjustments/modifications, mainly: independent computation of the stress position of prefixed non-substantives rather than a specific rule, and merging of the rules of the final stressed vowel and that of the stressed vowel followed by another vowel under C^0.

[5] -ION is the abbreviated notation used to represent the whole family of endings whose common structure is the following: -{i,e,u}+ V(C_0(e)). This includes endings such as <-ion, -ear, -ual, etc>.

- A first class of endings determines the placement of primary stress on the final syllable of the word, i.e. /-1/. It includes orthographic endings such as -$\overline{\text{V'V'}}$(C_0(e)) (e.g. *tabóo, ballóon, papóose* …), those with French or Germanic origins (e.g. *-ade, -ette, -eur, -que, -teen …*) and disyllabic verbs in -Vte (e.g. *creáte, ignite, salúte*).
- A second class of endings imposes primary stress on the penultimate syllable, i.e. /(-)10/. It includes three series of endings: *-ic(s),* -C_2 + adjectival suffix in -V(C_0(e)) (e.g. *inténsive, abýssal, indúlgent …)*, disyllabic suffixes mostly found in learned vocabulary (e.g. *-osis, -itis* …).
- A third class of endings imposes primary stress on the antepenultimate syllable, i.e. /(-)100/; this class of endings is to be found with words of three syllables and more. It includes another group of disyllabic suffixes, viz. nominal *-ity / -ety*, verbal *-ify / -efy*, and adjectival suffixes such as *-ical, -inous* or *-ular*. This third class also includes two other endings:
    - the whole family of endings with two successive vowels <{i,e,u} + V> which share the following common structure -{i, e, u} + V(C_0(e)), e.g. *génuine, invérsion, núclear.*
    - words in -Vte (e.g *ábsolute, décimate, réquisite*), adjectives and nouns in -ence / -ent (e.g. *cónfidence/-ent, dífference/-ent, résidence/-ent*).

All suffixed words which are not affected by any of the strong endings and whose base is free are subject to the Neutral Derivation Law. Once the suffix has been removed, the placement of primary stress is computed by reference to the base. In other words, these suffixed items owe the placement of their primary stress (and actually their whole pronunciation) to their base (e.g. *cárelessness < cáre, cháracterize <chárácter, lóvingly < lóve*).

The stress pattern of opaque prefixed words other than nouns (e.g. *below, decide, develop, understand*) is computed directly on the remaining portion of the word after the prefix has been ignored. The remaining portion of the word either reproduces the pronunciation of the base when it is free,[6] or follows the usual stress assignment rules which are based on segmental features. Fournier calls this the "Germanic Law". These rules also apply to all other disyllables and longer words (i.e. suffixed words with bound roots (e.g. *potent, distant, horrible*), opaque prefixed nouns (e.g. *refuge, revenue*), and words with no identifiable internal structure (e.g. *honest, elephant*). Three general rules are based on the number of syllables displayed by stressable sequences, be they words or roots in semantically opaque prefixed words:

- sequences subject to the rule Monosyllable → /1/ are stressed on their unique syllable (e.g. *cát, dóg, mílk +-cíde, -táin, -spéct*)
- those subject to the rule Disyllable → /10/ are stressed on the penultimate (e.g. *dístant, hónest, pótent, réfuge +-vélop, -término*)
- and others, subject to the general rule called the Normal Stress Rule (NSR), are stressed on the antepenultimate; NSR → /(-)100/ (e.g. *áccident, élephant, révenue + (ad)mínister*)

Although the NSR assigns primary stress to a vast majority of long words of three syllables and more, the stress pattern of some of them is governed by two other rules which dominate the NSR. These rules are based on two segmental configurations which are associated with primary stress on the penultimate:

- words with a consonant cluster in the prefinal position, prefinal C_2 → /(-)10/ (e.g. *appréntice, coriánder, umbrélla*).
- words ending in <t, d, n, s, z> + <a,e,i,o,u>[7], 'Italian' words → /(-)10/ (e.g. *banána, karáte, piáno*).[8]

Fournier's system of stress rules is presented in Figure 1.

[6] This reference part, similar to the Neutral Derivation Law, is not displayed in Figure 1 below, for readability reasons.
[7] The pattern applies to final <e> only in words where it is not silent (e.g. *finale, furore, ukulele*).
[8] The denomination 'Italian words' was coined by L.Guierre, on account of the numerous words in this class that were indeed borrowed from Italian, but it actually includes words from other sources as well. All, and only, words with the spelled out final sequence obey the rule.

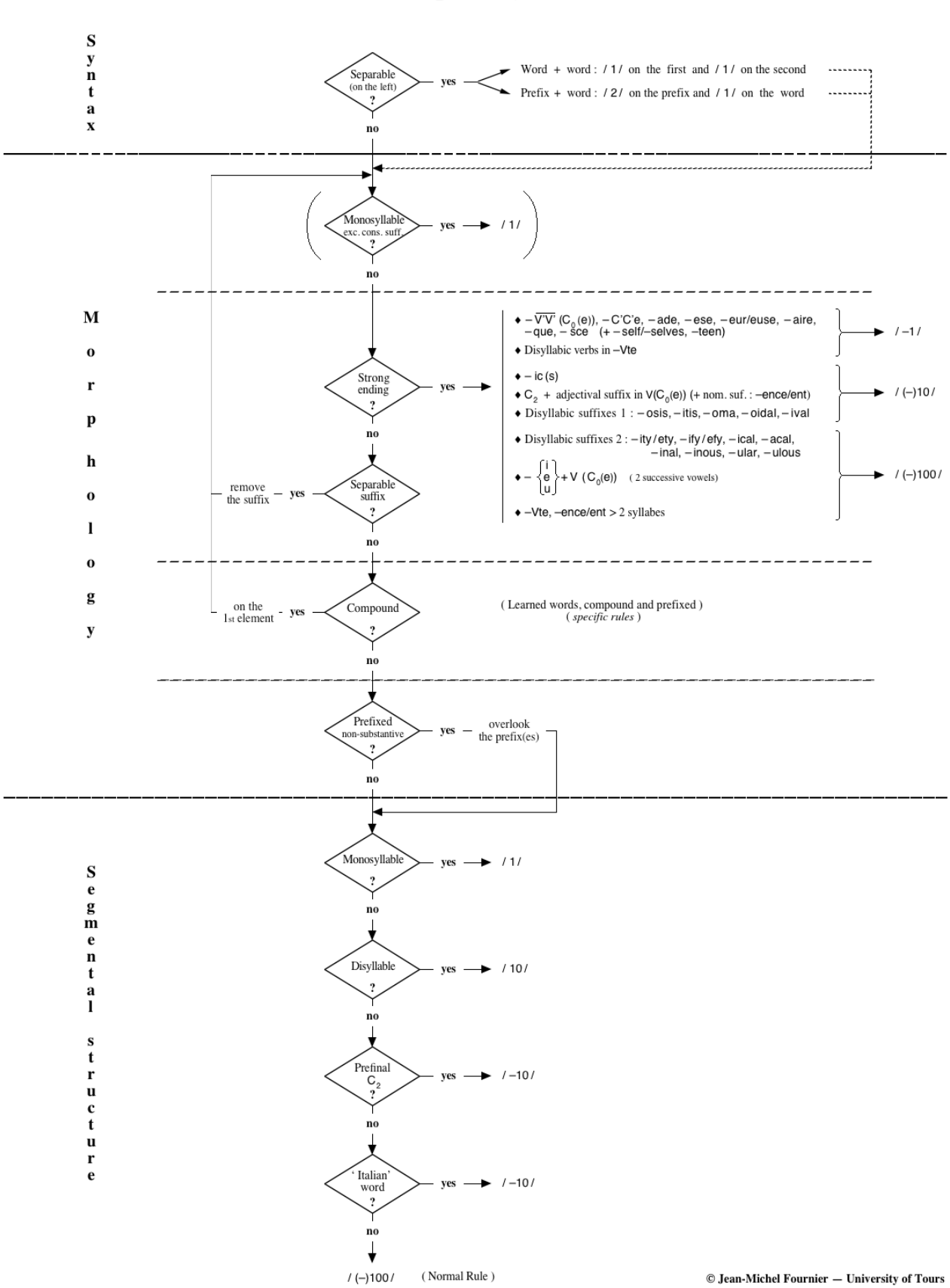

Figure 1. Fournier's system of stress rules

2.3. Fournier's contextual and graphophonological rules

In the analyses below, we will use the terminology displayed in Table 1.

$\overset{r}{V}$	$\breve{V}$	**Monographs**	$\bar{V}$	$\bar{V}^r$	**Digraphs**
r vowel	checked vowel	**<V>[9]**	free vowel	r-coloured free vowel	**<$\overline{VV}$>**
[ɑː]	[æ]	<a>	[eɪ]	[eə]	<ai, ay / ei, ey>
[ɜː]	[e]	<e>	[iː]	[ɪə]	<ea, ee / ie**>
[ɜː]	[ɪ]	<i>	[aɪ]	[aɪə]	<ie*, ye>
[ɔː]	[ɒ]	<o>	[əʊ]	[ɔː]	<oa**, oe*>
[ɜː]	[ʌ (ʊ)]	<u>	[(j)uː]	[(j)ʊə]	<e(a)u, ew/ ue*>
			[ɔː]	[ɔː]	<au, aw>
			[uː]	[ɔː]	<oo>
			[ɔɪ]		<oi, oy>
			[aʊ]	[ɔː (aʊə)]	<ou, ow>

*: final **: non-final

Table 1. Correspondences between orthography and pronunciation for stressed vowels in RP English (after Fournier, 2010b: 98)

The table shows that each orthographic vowel can have:

- four[10] different *values* when it is a monograph (e.g. r, checked, free and r-coloured free);
- two different *values* when it is a digraph (e.g. free and r-coloured free).

While values are displayed in columns, the possible realisations of a given orthographic vowel (e.g. [ɑː], [æ], [eɪ], [eə] for <a>), i.e. the lines, are said to share a *quality*.

Crucially, the Guierrian School sees vowel values as being predictable from 3 main parameters: the nature of the stressed vowel (e.g. a digraph, <u>), the context to the right of the vowel, and its position from the end of the word. Therefore, vowel values are phenomena that the theory seeks to predict along with the position of stresses. Similarly to stress rules, Fournier (2010b) presents a whole set of reading rules which determine the way English vowels in stressed syllables are pronounced. In his system, contextual rules explain the values of stressed digraphic or monographic vowels and graphophonological rules account for their various phonetic realizations.

The contextual rules are organized in two groups which correspond to two successive levels of analysis:

- the rules of the first group are characterized by a level of analysis which is limited to characteristics of the stressed syllable itself.
- the rules of the second group are characterized by a level of analysis which extends over the stressed syllable itself and centres on characteristics of its surrounding context. At this level, the stressed vowel necessarily appears in a ˈVCV context.

At each level, the value of the stressed vowel is determined by rules that are either spelling sensitive, context sensitive or rank sensitive according to the three parameters mentioned above. Apart from the spelling sensitive rule of the first group which governs a digraphic vowel $\overline{VV} \rightarrow \bar{V}$, all the other rules govern monographic vowels.

[9] Angle brackets are used for orthography.

[10] That does not include foreign free vowels (e.g. *ban*[ɑː]*na, alb*[iː]*no, blas*[eɪ]).

Determining parameter	**Rule**	**Description**	**Examples**
Spelling	$\overline{VV} \rightarrow \bar{V}$	the stressed vowel is a digraph and has a free value in all positions	*b**oo**merang, hydr**au**lic, s**ea** …*
Context	$C^0 \rightarrow \bar{V}$	the stressed V is either final or followed by another V; it has a free value	*ch**a**os, l**i**on, m**e**…*
	$C_2 \rightarrow \breve{V}$	the stressed V is followed by a C cluster (at least 2 Cs) other than <rC>; it has a checked value[11]	*m**y**stery, n**e**st, van**i**lla*
	$rC \rightarrow \overset{r}{V}$	the stressed V is followed by <rC> (with C ≠ r); it has an r value	*c**u**rtain, f**o**rtunate, sh**i**rt*
	alC/olC → [ɔː]/[əʊ]	the stressed V <a> or <o> is followed by a C cluster < l + l#, t, d, k >	***a**ll, h**a**lt, b**a**ld, w**a**lk tr**o**ll, v**o**lt, g**o**ld, f**o**lk*
Rank	$C\# \rightarrow \breve{V}$	the stressed V is final and followed by a single C other than <r>; it has a checked value	*c**a**t, p**e**rmit, p**e**t*
	$r\# \rightarrow \overset{r}{V}$	the stressed V is final and followed by <r>; it has an r-value	*c**a**r, n**o**r, s**i**r*

Table 2. Graphophonological rules of the first group

The context sensitive rules of the first group are dependent on the number of consonant(s) placed after the stressed vowel. This number ranges from “no C at all” to “at least 2 Cs”. The vowel rank sensitive rules are based on the final character of a stressed vowel followed by a single consonant other than ‘r’ or <r> itself.[12]

[11] Consonant clusters include <x>, orthographic geminate consonants <C’C’> as well as any group of consonants except <Ch>, <Cr> and <C + syllabic l/r>.

[12] Note that, in more classical terms, the r# and rC rules can be grouped together in a rule referring to a coda r and that the C# and C_2 rules can be grouped together in a rule referring to a coda different from r.

Determining parameter	**Rule**	**Description**	**Examples**
Spelling	u → $\bar{V}$	The stressed vowel is spelt <u>; it has a free value.	*ac**u**te, constit**u**tion, cr**u**cify*
Context	-V# → $\bar{V}$	The stressed syllable is followed by a final vowel; it has a free value.	*ar**o**ma, b**a**by, b**i**ke*
	-ic(s)# → $\breve{V}$	The stressed syllable is followed by the suffix -ic(s).	*ang**e**lic, oce**a**nic, t**o**nic*
	-{i,e}V(C_0(e)) # → $\bar{V}$	The stressed syllable is followed by the pattern -{i,e}V(C_0(e))#. The stressed vowel must not be <i, y>.	*appr**e**ciate, spont**a**neous, z**o**diac*
Rank	Luick[13] → $\breve{V}$	The stressed syllable is antepenultimate; the stressed vowel has a checked value.	*aust**e**rity, c**y**linder, r**i**tual*
	Prefinal → $\bar{V}$	The stressed syllable is penultimate and the word has more than two syllables; the stressed vowel has a free value.	*cath**e**dral, hor**i**zon, neur**o**sis*
	-V{s, x}# → $\bar{V}$	In disyllables, the stressed syllable is followed by the pattern -V{s, x}#; the stressed vowel has a free value.	*cr**i**sis, m**o**tus, m**a**trix*
	-iC# → $\breve{V}$	In disyllables, the stressed syllable is followed by the pattern -iC#; the stressed vowel has a checked value.	*cr**e**dit, f**i**nish, s**o**lid*

Table 3. Graphophonological rules of the second group. The stressed monographic vowel necessarily appears in a ˈVCV context.

The spelling sensitive rule of the second group (u → free vowel) is valid whatever the length of the word and the rank of the stressed syllable. However, the scope of all the other rules of that group is limited to disyllabic and other polysyllabic words. Stressed vowels governed by context sensitive rules are conditioned by 3 determining contexts: -V#, -ic(s)# and -{i,e}V(C_0(e))#, two of which are also strong endings. Those governed by syllable rank sensitive rules are conditioned by the "syllabic position" parameter. The stressed vowels of words of three syllables and more are governed by Luick's rule and the rule of the prefinal. Those of disyllabic words in /10/ are governed by two contextual subrules characterized by two specific patterns, -V{s, x}# and -iC#.

Fournier's system of graphophonological rules is presented in Figure 2.

[13] This rule is named after Luick (1898), who first described it.

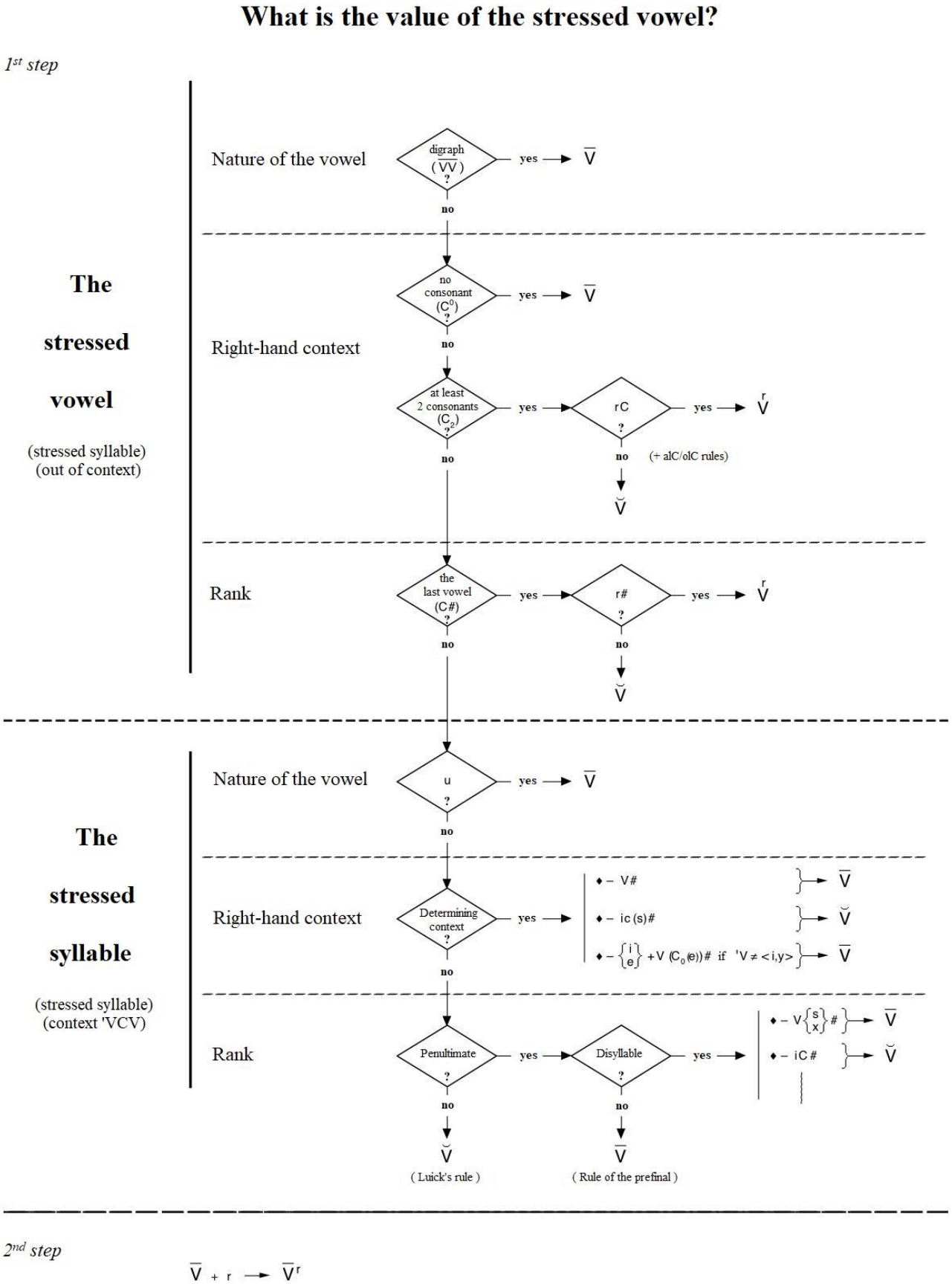

Figure 2. Fournier's system of graphophonological rules

2.4. Research questions

Our goal is to study the morphological, phonological and graphophonological properties of the core of the English lexicon in order to determine what structures and generalisations are predominant in that part of the vocabulary. The basis of our study consists in evaluating the efficiency of the stress rules and graphophonological rules put forward by Fournier (2007, 2010b) as we wonder whether this evaluation could provide valuable insights for morphological and phonological research and also for second language teaching. We expect these analyses might contribute to establish which generalisations should be taught in priority to L2 learners of English.

3. Methodology

The first 5000 most frequent word-forms were extracted from the SUBTLEX-UK corpus (Van Heuven *et al.* 2014).[14] This was achieved by ordering the data in descending order using the "DomPoSFreq" column of the database, which gives the frequency of the dominant part of speech for each item. We chose to study word-forms and not lemmas to be able to measure the overall proportion of morphologically complex items in the core of the English lexicon, and we wanted this to include inflected forms. Note that the SUBTLEX-UK corpus lists syntactic categories. However, this information was occasionally manually corrected for a few words which were tagged as names when they are in fact not proper names (e.g. *empire, minister, united*).

The data was then coded for morphological structure. So as to include certain opaque morphological structures which have been shown to impact the phonology (Dabouis 2017), the online *Oxford English Dictionary* was used to establish the presence of historical affixes, as no established method for identifying opaque morphological constituents in synchrony exists yet. We coded stress patterns using a numerical transcription: /1/ for primary stress, /2/ for secondary stress and /0/ for unstressed syllables (e.g. *acádemy* would be coded /0100/). Finally, all the items were coded for the stress rules and graphophonological rules put forward by Fournier (2010b): what law or rule they are supposed to follow and the different types of exceptions. Stress patterns and vowels were taken from Wells (2008). Only the main pronunciation for British English was included.

108 entries were left out:

- 56 "non-lexical" entries (e.g. *s, ya, f, oh, ha, ah, wow…*);
- 22 syntactic constructions (e.g. *n't, m* (← {be}), *gonna, wanna, gotta, innit…*);
- 24 acronyms (e.g. *UK, BBC, TV, NHS, UN, NATO…*)
- 6 entries absent from Wells (2008): *cha, lau, nok, tok, tombliboo, yay*

The corpus analysed in the following sections therefore contains 4892 entries.

4. Results

In the following sections, we detail the results yielded by the analysis of the corpus. We start by reviewing a few general facts about the structure of the corpus in terms of morphological structure, syntactic categories and word length (§4.1). In §4.2, we discuss words which can be analysed as owing their pronunciation to that of another word, because they have morphological structures in which the pronunciation of the base is unaffected (e.g. neutral suffixes, separable prefixes, compounds). We then turn to the stress patterns (§4.3) and graphophonological regularities of vowels with primary stress (§4.4) observed in the remainder of the corpus.

4.1. Structure of the corpus

4.1.1. Morphology

Let us first consider the morphological structures of the words found in the corpus. The different structures found in the corpus can be found in Table 4.

[14] This corpus is based on television subtitles from nine British channels broadcasted between January 2010 and December 2012.

Category	Description	Examples	Count	%
Suffixed	Contains a suffix	*active, enemies, swimming*	1995	41%
Simplex	No identifiable structure	*alien, force, Kate, round*	1874	38%
Suffixed and prefixed	Contains both a prefix (transparent or opaque) and a suffix	*arrival, included, unlikely*	442	9%
Prefixed-Opaque	Prefixed construction with an opaque meaning	*accept, intend, protect, refer*	287	6%
Compound	Made up of two free bases[15]	*anybody, gentleman, network*	182	4%
Truncation	Truncated form	*Chris* (← *Christopher*) *Jenny* (← *Jennifer*)	61	1%
Neoclassical compound	Made up of Latinate or Greek bound roots	*apology, democrat, telephone*	22	0%
Adverbial particle	First element is an adverbial (generally locative) particle	*downstairs, income, overnight*	19	0%
Prefixed-Transparent	Prefixed construction with a compositional meaning (the base might itself be prefixed)	*disagree, incorrect, unable*	16	0%

Table 4. The different morphological categories identified in the corpus

A few observations can be made. First, the two biggest classes are suffixed words and words with no identifiable structure. This is rather unsurprising considering that inflection is marked through suffixation in English and, as will be seen in §4.2, more than half of the suffixed words in the data contain an inflectional suffix. It can also be expected that the most frequent words should have simple morphological structures and that the most complex words should have lower frequencies. This is in fact true even in this small corpus: as can be seen in Figure 3, as a bigger proportion of the data is considered, the proportion of simplex words decreases, from 56% for the first thousand words to 38% in the whole corpus.

[15] A few words included in this category are actually historical compounds whose constituents might be bound in contemporary English (e.g. ***Cam****bridge*, ***Eng****land*, ***Liver****pool*)

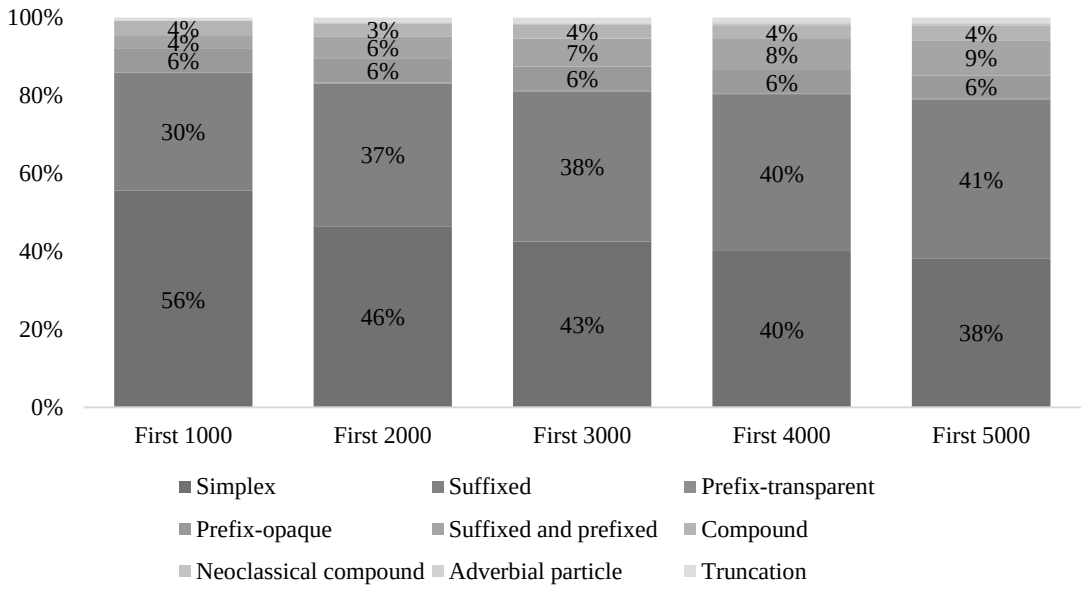

Figure 3. Morphological categories in different samples of the data (percentages lower than 3% are not shown)

It could be surprising to see that the most underrepresented morphological structure is that of semantically transparent prefixed constructions, with only 16 items in the corpus. This may not actually be that surprising considering that it is very productive morphology and that most of these structures are not lexicalised.

4.1.2. Syntactic categories

The syntactic categories of the words in the corpus, as given by SUBTLEX-UK, are shown in Figure 4. The most striking fact about the distribution of the data is that close to half of the words of the corpus are nouns.

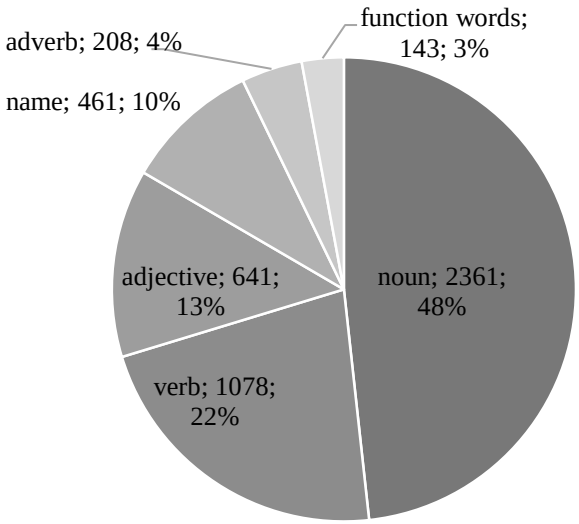

Figure 4. Syntactic categories found in the corpus

4.1.3. Word length

We also looked at the word length of the words of the corpus. Quite unsurprisingly, the most frequent words in English are short, with less than 8% that are longer than three syllables.

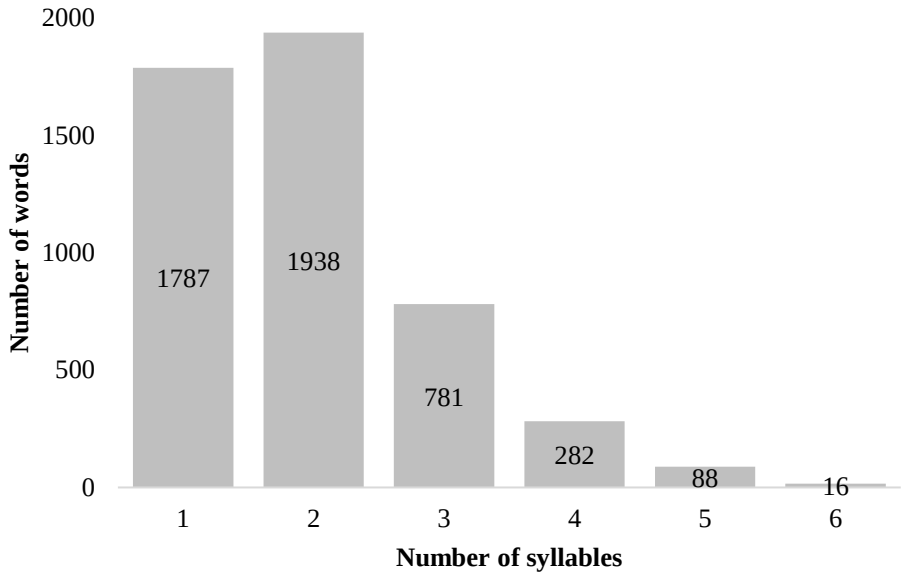

Figure 5. Distribution of the data by word length

4.2. Isomorphism

In Fournier's model, certain words preserve the pronunciation of their base in different configurations. This is referred to as "computation by reference", meaning that the pronunciation of a given word is computed by referring to the pronunciation of its base. This is to be opposed to "direct computation", in which the pronunciation of the word is computed directly from its structural properties. In this section, we discuss only words whose pronunciation is isomorphic to that of their base, i.e. is computed by reference. We identified five classes of words which should follow this mode of computation:

- Prefixed constructions
  - with compositional meaning (21 words): the addition of a prefix does not affect the pronunciation of the base. Constructions with an initial adverbial particle behave in the same way and were included in this category (e.g. *in**crédible**, out**stánding**, un**cómfortable***).
  - with non-compositional meaning (184 words): words other than nouns generally obey the Germanic Law, i.e. stress is placed on their roots, never on their prefixes. If the root is bound, it was included in the analysis in the following sections, as if it were an independent word (123 words; e.g. *ap**péar**, de**vélop**, re**dúce***), but not if it is free (61 words; e.g. *be**cóme**, en**córage**, per**fórm***). Among constructions with bound bases, five do not follow the Germanic Law as they have primary stress on their prefix (*difficult, énter, récognise, rélevant, súffer*). These are treated as a whole in the following sections.
- Suffixed constructions with neutral suffixes (1861 words): the pronunciation of the base is normally left unaltered, following the Neutral Derivation Law (e.g. ***annóunce**ment, **descríbe**d, **guést**s, **ímportant**ly, **pártner**ship, **smók**ing*). 1823 words regularly follow this law and only 38 exceptions are found in the data: 13 irregularly inflected forms (e.g. *gone, dealt, fifth)* and 25 isolated exceptions to the Neutral

Derivation Law (e.g. *máintenance* < *maintáin*, *n*[æ]*tional* < *n*[eɪ]*tion*, *rélative* < *reláte*). These exceptions are included in the rest of the analyses as if they were independent words.

- Compounds (198 words): primary stress generally falls on the first element, and this element is pronounced as if it were pronounced on its own (116 words; e.g. ***évery****day*, ***láy****out*, ***wóod****land*). The data also contains 26 right-stressed compounds (e.g. *old-****fáshioned***, *wher****éver***). 46 historical compounds were also treated as compounds in the analysis and, if the stressed element is bound, it was treated as if it were an independent unit and is included in the following sections (e.g. ***Éng****land*, ***Hólly****wood*, ***Nótting****ham*), among which two are right-stressed (*Bel****fást***, *Sout****hámpton***). Non-suffixed neoclassical compounds, which are also normally stressed on their first element, were also included in this category and their first element is included in the analyses of the following sections (10 words; e.g. ***héli****copter*, ***phóto****graph*, ***téle****phone*).
- Truncated forms (54 words): they are not analysed by Fournier. They were treated as isomorphic units considering that they usually preserve the stressed syllable of their base (52 words; e.g. *Dánny* ← *Dániel*, *Ráy* ← *Ráymond*, *Térry* ← *Térrence*). We found two non-isomorphic truncations (*Lisa* [ˈ**liːsə**] ← *Elizabeth* [iˈ**lɪz**əbəθ], *Mo* [ˈ**məʊ**] ← *Morris* [ˈ**mɒ**rɪs]), which were therefore analysed as independent units.

We also found 60 words whose pronunciation is expected to be computed directly and yet follow computation by reference: 60 prefixed multicategorial words which adopt a stress pattern that is not typical of their category (e.g. *amóunt*, *concérn*, *deféat*, *prómise*). These were treated as isomorphic with the related word of the other category, e.g. the noun *surpríse* owes its final stress to the verb *surpríse*.

Overall, we found 2155 words (44% of the data) whose pronunciation is isomorphic with that of their base. As their pronunciation is determined by referring to that of their base (or that of another of their categories), they will not be dealt with in the following sections, which only deal with words or bound roots whose pronunciation is determined by direct computation (2737 units). Note that the biggest class of isomorphic words is suffixed words, with 1823 words (85% of isomorphic words) which obey the Neutral Derivation Law. Among those, 1382 (76%) have an inflectional suffix (*-ed*, *-s*, *-ing*, *-er* or *-est*).

4.3. Stress

In this section, we detail the efficiency of the stress system put forward by Fournier in the dataset of the 2737 units whose pronunciation is computed directly. The results are organized by word-length.

The dataset contains 1448 monosyllabic units, including 143 bound roots from prefixed words or compounds (e.g. *(be)líeve*, *(in)vólve*, *Món(day)*, *Zéa(land)*). These have primary stress on their only syllable and need not be discussed any further.

There are 755 disyllabic units in the dataset, including 31 bound roots (e.g. *(con)síder*, *(de)vélop*, *Lánca(shire)*, *Nótting(ham)*). Three rules can account for the stress data: first, two groups of strong endings, associated with either final or penultimate stress; second, the Normal Stress Rule for disyllables, according to which disyllabic units should have first-syllable stress. The distribution of the data is shown in Table 5.[16] The overall exception rate is 4% with 34 exceptions in total.

[16] In this table and all the ones that follow, exception rates are shown in percentages only when the total amount of units exceeds 100.

Rule	Count	Exceptions	Examples
Strong endings /-1/	20	3	*creáte, caréer, uníque, techníque, ballóon, paráde…* Exc: *cóffee, décade, chárlotte*
Strong endings /10/	24	0	*músic, públic, méntal, áctive, sénsible, réntal…*
Dissyllables /10/	711	31 (4%)	*báby, márket, mínute, séven, kítchen, éarly…* Exc: *políce, exámple, campáign, evént, machíne…*

Table 5. Stress placement in disyllabic units

The dataset contains 534 words of three syllables or more. There are no bound roots of this length. Six rules account for stress placement in these words. First, there are the three groups of strong endings associated to either final, penultimate or antepenultimate stress. Second, there is the rule of prefinal C_2, according to which words with a prefinal consonant cluster should have penultimate stress. Third, there is the rule of so-called "Italian" words, according to which words that ends with the structure <t, d, n, s, z> + <a, e (non-silent), i, o, u> should have penultimate stress. Finally, there is the Normal Stress Rule, which states that all other words of three syllables or more should have antepenultimate stress.

Pattern	Rule	Count	Exceptions	Examples
/(-)100/	Strong endings	310	8 (3%)	*míllion, spécial, polítical, expérience, commúnity, évidence, partícular…* Exc: *Èuropéan, télevision, idéal, muséum, pássionate, María, oppónent*
	Normal Stress Rule	137	22 (16%)	*évery, fámily, líbrary, díscipline, América, Cámeron, díamond, pósitive…* Exc: *idéa, sécretary, párliament, Obáma, nécessary, Pàkistán…*
/-10/	Strong endings	39	2	*fantástic, èconómic, expénsive, ìndepéndent, enórmous, efféctive, intérnal…* Exc: *pólitics, cátholic*
	Prefinal C_2	32	14	*impórtant, rèferéndum, advántage, disáster, Septémber, advénture* Exc: *mínister, índustry, cháracter, pénalty, ínternet, cháncellor…*
	"Italian" words	7	0	*potáto, aníta, piáno, tomáto, banána, Fióna, Àrgentína*
/-1/	Strong endings	9	2	*rèferée, Jàpanése, àuctionéer, Àberdéen, cìgarétte…* Exc: *commítee, ámateur*

Table 6. Stress placement in words of three syllables or more

The overall exception rate is 9%, a third of which are exceptions to the prefinal C_2 rule. This is not surprising considering Fournier's (2010a) claim that this rule actually only applies efficiently to "foreign" sub-parts of the vocabulary.[17] This type of vocabulary is more specialized than non-

[17] Namely Modern Latin, and relatively late borrowings from Southern Romance languages: Italian, Spanish, Portuguese…

Latinate vocabulary and it is to be expected that it should be underrepresented in this sample of the most frequent words in English. Out of 22 exceptions to the Normal Stress Rule, 12 have phonological pre-antepenultimate stress but are usually realized phonetically with antepenultimate stress because of the (often common) elision of a medial unstressed vowel or syneresis (e.g. *órdin(a)ry, párl(i)ament, témp(e)rature*). Most words are found with the structures associated to antepenultimate stress (447/534; 84%). In this whole subset, the proportion of antepenultimate stress is 82%, that is including regular words with antepenultimate stress or exceptions to penultimate or final stress. The biggest source of antepenultimate stress are strong endings. The single class of -{i, e, u}+V(C_0(e)) accounts for 196 words with antepenultimate stress (+ 5 exceptions). Among those, the *-ion* suffix accounts for 122 words (62% of total). The other common strong endings found in the data are: *-ent/-ence* (28 words + 1 exception), *-ety/-ity* (25 words), *-Vte* (19 words + 1 exception) and *-ical* (13 words).

4.4. Graphophonology

The system of graphophonological rules proposed by Fournier was evaluated in the same dataset as that used for stress rules. The detailed results are shown in Table 7 and are discussed below.

Rule	Count	Exceptions	Examples
$\overline{V}\overline{V} \rightarrow \overline{V}$	492	65 (13%)	*óut, nów, wáy, néed, dáy, hóuse, méan, fóod…* Exc: *góod, agáin, tóok, héart, déath, dóuble…*
$C^0 \rightarrow \overline{V}$	78	1	*bé, mé, só, twó, whý, scíence, muséum, Rýan…* Exc: *dóes*
$C_2 \rightarrow \breve{V}$	838	74 (9%)	*thínk, nèxt, stíll, sórry, hélp, hístory, Rússia…* Exc: *chánge, ásk, níght, fínd, sígn, táste, móst…*
$rC \rightarrow \overset{r}{V}$	162	1	*wórld, párt, mórning, pérson, gírl, túrn, towárds…* Exc: *Wórcester*
alC/olC → [ɔː]/[əʊ]	30	1	*áll, óld, hálf, wálk, fólk, cálm, póll…* Exc: *sháll*
$C\# \rightarrow \breve{V}$	252	10 (4%)	*lót, mán, jób, stóp, untíl, Japán, canál…* Exc: *báth, páth, hígh, bóth, trúth, Iráq, Irán…*
$r\# \rightarrow \overset{r}{V}$	17	0	*fár, prefér, guitár, Nór(man), stír, fór…*
$u \rightarrow \overline{V}$	56	3	*úse, Júlia, dúring, húman, músical, Júdith,…* Exc: *súgar, cúshion, stúdy*
$-V\# \rightarrow \overline{V}$	322	40 (12%)	*hére, táke, quíte, níce, pláce, náme, impróve…* Exc: *háve, óne, véry, móney, cíty, líve, sémi…*
$-ic(s)\# \rightarrow \breve{V}$	17	1	*mágic, specífic, históric, èconómic…* Exc: *básic*
$-\{i,e\}V(C_0(e))\# \rightarrow \overline{V}$	86	7	*média, Victória, période, negótiate, comédian…* Exc: *Dániel, Itálian, fáshion, ónion, spécial…*
Luick → $\breve{V}$	176	4 (2%)	*fámily, évidence, América, pólicy, délicate…* Exc: *évening, fávourite, cólonel, líbrary*
Prefinal → $\overline{V}$	4	1	*Octóber, mèdiéval, oppónent* Exc: *imágine*
$-V\{s, x\}\#$	9	3	*bónus, crísis, Dávis, fócus, Jésus, mínus* Exc: *Bóris, Páris, Thómas*
-iC#	21	3	*Cólin, fínish, límit, Phílip, sólid, Róbin…* Exc: *Ápril, Dávid, évil*
No known rule	177		*móther, néver, wáter, lócal, lábour, dózen…*

Table 7. Graphophonological regularities in the corpus

The overall exception rate is below 8% (214 exceptions). It is to be noted that the highest proportions of exceptions are found in the largest inventories while more restricted inventories tend

to display high efficiency rates. Many of these exceptions belong to well-identified classes of words. First, for digraphs, there are only 3 exceptions out of 65 which are not one of the following three digraphs: <ea>, <ou/ow>, <oo>. These are notoriously irregular and, as will be seen below, are also often exceptional in quality. Second, many of the exceptions to the C# and C_2 rules belong to subsets known to resist the general behaviour of the whole class. Out of the 84 exceptions, there are 36 words which obey the tendency of "ask" words (e.g. *áfter, clάss, dάnce, mάster*). Note that these words are regular in many English dialects. Moreover, 24 words belong to one of the 5 following sub-classes: <igh(t)#>, <ange#>, <ind#>, <Vgn#> and <Vste#>. These are known to resist the C# and C_2 rules and could be called sub-rules rather than exceptional sub-classes, in which case these 24 words would be seen as regular.

The exceptions we have discussed so far are exceptions in value, i.e. exceptions to the rules. There are also exceptions in quality, i.e. irregular spelling-to-sound correspondences. In the data, we find 153 (6%) such exceptions. Like exceptions in value, most of them form coherent sub-groups. Once again, the three digraphs <ea>, <ou/ow> and <oo> account for a significant share of the exceptions, with 86 exceptions (e.g. *dead, journey, slow, foot*). There are also 44 words in which <o> is realized as if it were a <u> (e.g. *love, come, front, work*), which leaves only 23 cases of isolated exceptions (e.g. *any, busy, pretty*).

To conclude this section, let us now briefly discuss the 177 words for which no known rule applies. Among those, there are five monosyllabic words that end in <-es>: *clothes, James, Jones, Thames, Wales*. All but *Thames* have free vowels. The remaining 172 units are trochaic disyllables with a monograph followed by a single consonant. Among those, 99 (58%) have checked vowels (e.g. *désert, ólive, pétrol, Trévor*), 69 (40%) have free vowels (e.g. *éven, fínal, Péter, récent*) and three have other vowels, two of which can be assimilated to "ask" words (*w*[ɔ:]*ter, r*[ɑ:]*ther, f*[ɑ:]*ther*). Therefore, even in this restricted part of the vocabulary, no clear tendency can be observed in these words.

5. Conclusion

In this study, we have found that two thirds of the 5000 most frequent words of English contain more than one morphological constituent and that close to half have a pronunciation that is isomorphic to that of their base. We have tested Fournier's stress and graphophonological rules on non-isomorphic vocabulary and have found surprisingly low exception rates (between 6 and 9%), as one may expect high-frequency vocabulary to be more prone to contain exceptions.

We hope that this study will be useful to teachers of English as a foreign language who may wish to teach only the main stress and graphophonological generalisations and who will now have access to the relative importance of each generalisation along with exception rates and common examples. We also hope to be able to turn the corpus used in this study into a training tool to practice the application of these rules.

References

Dabouis, Quentin. 2017. Semantically Opaque Prefixes and English Phonology. *14th Old World Conference in Phonology* (20-22th February, Düsseldorf).

Fournier, Jean-Michel. 2007. From a Latin syllable-driven stress system to a Romance versus Germanic morphology-driven dynamics: in honour of Lionel Guierre. *Language Sciences* 29. 218–236.

Fournier, Jean-Michel. 2010a. Accentuation lexicale et poids syllabique en anglais : l'analyse erronée de Chomsky et Halle. *8ème Rencontres Du Réseau Français de Phonologie* (Juil).

Fournier, Jean-Michel. 2010b. *Manuel d'anglais oral*. Paris: Ophrys.

Guierre, Lionel. 1979. *Essai sur l'accentuation en anglais contemporain : Eléments pour une synthèse*. Université Paris-VII dissertation.

Luick, Karl. 1898. Beiträge zur englischen Grammatik III. Die Quantitäts Veränderungen im Laufe der englischen Sprachentwicklung. *Anglia* 20. 335–362.

Schmitt, Norbert & Diane Schmitt. 2014. A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching* 47(4). 484–503.
Van Heuven, Walter V.J., Pawel Mandera, Emmanuel Keuleers & Marc Brysbaert. 2014. Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* (67). 1176–1190.
Wells, J.C. 2008. *Longman Pronunciation Dictionary* 3rd ed. London: Longman.