Phonological evidence for morphological complexity in English proper names

Quentin Dabouis

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

Phonological evidence for morphological complexity in English proper names

Quentin Dabouis, Université Clermont Auvergne – LRL (UR 999)

RESUME

Cet article présente les résultats d'une analyse d'un important corpus de noms propres complexes anglais tels que Cambridge, Manchester *ou* Washington, *et qui est inspirée de travaux existants sur les noms de lieux en néerlandais (Köhnlein 2015). Ces noms fournissent de nouvelles données montrant que les mots qui n'ont pas un sémantisme transparent peuvent se comporter phonologiquement comme des mots complexes. Ceci est démontré à partir de l'analyse des prononciations de plus de 3500 noms propres, au cours de laquelle huit propriétés phonologiques ont été étudiées. Les résultats montrent que ces noms propres ont des caractéristiques phonologiques qui sont plus semblables à celles des mots complexes (composés et mots à suffixes neutres) qu'à celles des mots sans structure interne. L'analyse que je propose suppose que ces mots sont effectivement complexes morphologiquement, et que cette complexité est reflétée dans la structure en domaines phonologiques. Cette structure interne est supposée acquise à travers l'exposition aux formes récurrentes qui forment ces noms propres ainsi qu'à travers l'identification de caractéristiques phonologiques « anormales ». Une possible représentation lexicale de ces mots est proposée. Celle-ci postule un stockage analytique de ces noms (Bermúdez-Otero 2012) et la coindexation entre les différents niveaux de représentation de l'entrée lexicale (Jackendoff 1997, 2002).*

SUMMARY

This paper presents the results and an analysis of a large dataset of complex English proper names such as Cambridge, Manchester *or* Washington, *inspired by previous work on Dutch place-names (Köhnlein 2015). Those names provide new evidence that words that do not have transparent semantics may still behave phonologically as complex words. The evidence comes from a dictionary-based study of the pronunciations of over 3500 proper names in which eight phonological characteristics were found to resemble those observed in complex words (compounds or words with neutral suffixes) rather than those found in simplex words. The analysis that I propose posits that such words are indeed complex morphologically, and that this is reflected in their phonological domain structure. This internal structure is assumed to be learned through the exposure to the recurring constituents in such names and through the identification of 'anomalous' phonological characteristics. A possible lexical representation is proposed that involves analytic listing (Bermúdez-Otero 2012) and coindexation between the different levels of representation in a lexical entry (Jackendoff 1997, 2002).*

Mots-clés : nom propre, interface morphologie-phonologie, toponymes, opacité sémantique, phonologie anglaise

Keywords: proper name, morphology-phonology interface, toponyms, semantic opacity, English phonology

1. Introduction[1]

All theories of phonology integrate an interface with morphology, and they do so through various theoretical devices, either procedural (e.g. cycles or phases) or representational (e.g. alignment between morphological boundaries and prosodic constituents, insertion of phonological material) (see Bermúdez-Otero (2012) and Scheer (2011) for discussion of the dual aspect of the morphology-

phonology interface). A major question for the study of the morphology-phonology interface is what Scheer (2011) calls the 'mapping puzzle', i.e. "which particular morpho-syntactic (or semantic) configuration produces a phonological effect. And in turn, which phonological effects are induced by boundaries". The most common approach seems to be to use morpheme-based approaches to morphology, in which the phonology only refers to units that one could call "prototypical" morphemes, i.e. minimal signs within the words that have both form and an identifiable meaning. However, there are constituents of word structure which do not fit this "prototypical" definition of the morpheme, and so that view of morphology has led to a lack of study of the possible phonological behaviours of morphological structures that have opaque semantics.

This paper is largely inspired by a study on Dutch place-names by Köhnlein (2015), who argues that, even though such words are semantically opaque, their behaviour is best analysed if one assumes that they are morphologically complex. My aim is to see whether the same can be said about English proper names such as *Cambridge, Chesterfield* or *Washington.* I will start by discussing some of the available evidence regarding the phonological effects of opaque morphological structures (§2) and will present the exact aims of the paper and the hypotheses that will be tested (§3). The methodology will then be presented in §4 and the results detailed in §5. I will finish by a proposal on how to account for the data (§6) before concluding the paper (§7).

2. Opaque morphology and phonology

2.1. Opaque prefixed words in English

The best studied type of opaque morphological structure that is known to impact the phonology in English is opaque prefixed words such as *contain, refuse* or *submit*. These words are historically prefix + root constructions, mostly inherited from Romance languages. Their constituents are "clearly recurrent elements at the level of form, but their meanings are essentially unclear" (Plag and Balling 2020). Those words have been treated as complex in early generative phonology (notably in Chomsky and Halle (1968), in which such words are assumed to contain a specific boundary, '=') and in the Guierrian School (see Fournier (2010b), Guierre (1979) and Trevian (2015), among many). The available evidence on the phonological behaviour of such words has been synthetised in Dabouis and Fournier (submitted b). Let us summarise the main points that they reviewed in the literature:

- Primary stress in words that are not nouns is almost never on the prefix (e.g. *consérve, devélop, entertáin*[2]), and that sometimes leads to stress patterns that diverge from those of simplex words (Chomsky and Halle 1968: 94; Dabouis and Fournier 2023; Fournier 2007; Guierre 1979; Halle and Keyser 1971: 37; Liberman and Prince 1977), while prefixed nouns only probabilistically diverge from truly simplex words. Guierre (1979) reports that 93% of non-prefixed disyllabic nouns have initial stress (e.g. *béauty, rádish, vísa*) while this is true of only 81% of opaque prefixed disyllabic nouns (e.g. *ábsence, cóngress, dígest*).
- The diachronic evolution of primary stress in verb-noun pairs (e.g. *convict, concrete, exile*), as Sonderegger and Niyogi (2013) report that words which share a prefix tend to evolve in a similar direction. They note that this "suggest[s] that it is a shared morphological prefix rather than simply shared initial segments which correlates with trajectory similarity".
- Opaque prefixed words have reduced vowels in the initial pretonic position, even in closed syllables (e.g. /ə/*dvánce, c*/ə/*ndénse, s*/ə/*btráct*), contrary to what is found in non-prefixed words (e.g. *l*/a/*mpóon, p*/ɔ/*ntíficate, t*/ɛ/*chníque*). This has been observed in numerous works on English phonology (Chomsky and Halle 1968: 118; Guierre 1979: 253; Halle and Keyser 1971: 37; Halle and Vergnaud 198: 2397; Hammond 2003; Hayes 1982; Liberman and Prince 1977: 284-85; Pater 2000; Selkirk 1980) and has been confirmed empirically in a multifactorial study of vowel reduction using large-scale dictionary data Dabouis and Fournier (submitted a).
- Opaque prefixed words tend not to have reduced vowels in the final syllable of disyllabic words with initial primary stress. Guierre (1979: §4.2.6) observes that reduction is the rule in

non-prefixed words (e.g. *báll*/ə/*st*, *hárv*/ɪ/*st*, *hón*/ɪj/ – 87% of words have reduced vowels) but that non-reduction is the dominant pattern in opaque prefixed words with initial primary stress (e.g. *cóntr*/ɑː/*st*, *díg*/ɛ/*st*, *surv*/ɛj/ – 84% do not have reduced vowels).[3]

- Opaque monosyllabic prefixes favour weak stress preservation. In words such as *accessibility, deliberation* or *reliability*, which are derived from a base with second-syllable primary stress (here, *accéssible, delíberate* and *relíable*), Arndt-Lappe and Dabouis (submitted) report that there is more second-syllable stress (that is, stress identity with the base) in words with such prefixes than in words without (e.g. *anticipation, domestication, municipality*), in both dictionary data and elicitation data.
- The phonotactics of the medial consonant clusters found in such words display "irregularities". Guierre (1990) and Hammond (1999: §3.3) report that certain clusters found in these words are never attested in simplex words (e.g. /bs/ in *absence*, /bv/ in *obvious* or /dh/ in *adhere*), thus indicating the presence of a morphological boundary.

There is therefore ample evidence that these words differ from simplex words and are better treated as complex words, even though their behaviour also diverges from that of prefixed words with transparent semantics (e.g. *co-author, deconstruct, realign*).[4] Therefore, opaque morphological structure can have phonological effects, and it is worth investigating other comparable interactions.

2.2. Place-names in Dutch and Central Catalan

Such similar interactions have been described by Köhnlein (2015) in an analysis of the phonological behaviour of Dutch place-names. Köhnlein starts by noting that "[i]n linguistic theory, […] names have been largely neglected so far", especially regarding their synchronic internal structure. He reviews a number of languages in which place-names display complex structures, especially in Germanic languages, for which "it is well established that many Germanic place-names are etymologically complex and display compound-like structures or show signs of affixation, whether their semantic surface structures are synchronically opaque or not".

Then, he proceeds to the analysis of Dutch place-names such as *Wageningen* or *Amsterdam*, which many authors have analysed as monomorphemic, and so which could be expected to display phonological behaviours that resemble those of simplex words. Köhnlein analyses complex place-names as being synchronically complex, with an initial constituent that serves "a purely referential purpose" followed by either a toponymic suffix (which is stress-neutral and syllabifies with the stem) or a classifier (a stem which forms a compound and has its own phonological word). Assuming that those words are synchronically complex allows him to account for five phonological behaviours which would otherwise violate the general rules of Dutch phonology:

- Primary stress is normally within a final three-syllable window, but such names regularly have extrafenestral stress (e.g. *Wá.ge.nin.gen*), just as complex words (e.g. *búr.ger.lijk.er* 'pettier', *burger* 'citizen' + *-lijk* (adjectivizing suffix) + *-er* (comparative suffix));
- Schwa is normally always preceded by a stressed syllable (e.g. *pa.li.sá.d*[ə] but **pa.lí.sa.d*[ə]), yet such names regularly have schwas that are not (e.g. *Bún.scho.t*[ə]*n, Nij.me.g*[ə]*n*), just as complex words (e.g. *báby* + *tj*[ə] → *bá.by.tj*[ə] 'baby-DIM');
- Final superheavy syllables (i.e. "syllables which contain either a diphthong or a tense vowel followed by one consonant, or a lax vowel followed by two consonants") attract stress (e.g. *do.cu.mént* 'document', *ba.náan* 'banana'), but place-names in *-drecht* do not have final stress (e.g. *Dór.drecht, Slíe.drecht, Pá.pen.drecht*), just as nominal compounds (e.g. *áu.to.deur* 'car door');
- Superheavy syllables may only occur word-finally, but such names may have internal superheavy syllables (e.g. *Móor.drecht, Zwíjn.drecht, Lóos.drecht*), just as compounds (e.g. *túin.huis* 'garden house');

- Final primary stress is only possible if the final syllable is superheavy, yet certain classifiers (e.g. *-dam, -huizen, -veen*) have final stress (e.g. *Am.ster.dám, Rot.ter.dám*).

Köhnlein then argues that the two constituents are underspecified but still contain some semantic properties: the first constituent contains a "referential pointer" to a unique object in the world (a settlement) and the second constituent bears a feature [+settlement]. This raises the question of how this comes to be, and he argues that recurring sound strings may be identified as classifiers if they recur in a number of different forms. However, Köhnlein does not discuss how cases which do not occur in many forms should be analysed. The analysis that he proposes is based on models which assume fully specified lexical entries, with semantic, morphosyntactic and phonological information (Bermúdez-Otero 2012; Jackendoff 1997, 2002). The internal complexity of words is then assumed to be represented in lexical entries (it is called "analytic listing" by Bermúdez-Otero (2012)). The different units of representation are then related to one another using coindexation, i.e. instructions found in lexical entries to indicate how the units of representation of different modules should be associated. Köhnlein concludes that "the general approach proposed in this paper could be fruitfully applied to place-names in other languages", and it is that observation that has led to the present paper.

Mascaró (2016) makes similar observations for Central Catalan, which displays apparent underapplication of vowel reduction in compounds, even if they are composed of elements which are unattested as free forms. In Central Catalan, "all and only a, ɛ, ɔ, e, o, i, u appear in stressed position" and "[a]ll and only i, u, ə appear in unstressed position". Vowel reduction affects vowels within a stem if they are followed by another stress (e.g. *tr*[ɛ́]*nta* 'thirty' cp. *tr*[ə]*nt*[ɛ́] 'thirtieth'), but this does not affect compounds (e.g. *tr*[ɛ]*nta-c*[í]*nc* 'thirty-five') or phrases (e.g. *tr*[ɛ̀]*nta c*[í]*ncs* 'thirty fives'), in which those vowels are destressed but not reduced. Mascaró shows that this is also true in constructions in which the first element is unattested as an independent word, either in neoclassical constructions (e.g. [nɛuklásik] 'neo-classic') or in constructions in which the first element is unattested elsewhere (e.g. [bɛtəkí] 'here it is').

He argues that such forms are complex even if they are not semantically compositional, and that "there is extensive evidence against strict compositionality", which we reproduce in (1).

(1) a. There are linguistic expressions that have complex structure and noncompositional meaning.

b. *Idioms*. Eng. *kick the bucket*; Cat. *prendre el pèl* 'to fool somebody', lit. 'to take somebody's hair'; Cat. *dinyar-la* 'to die' (**dinyar*, obj. clitic *la* nonreferential)

c. *Inflected forms*. pluralia tantum, Eng. *jean-s*; Cat. *pantalon-s* 'trousers'

d. *Derivatives*. Eng. *prob-able* (cf. *prob-abil-ity*); Cat. *lubr-i*[k] 'lubricious' (cf. *lubr-i*[s]*-itat* 'lubricity')

e. *Compounds*. Eng. *bull's eye*; Cat. *mata-parent* 'Boletus satanas', literally 'kills relative'

Finally, he uses a stratal-cylic analysis which, like Köhnlein's (2015), relies on fully specified lexical representations which include morphological schemas and in which the different parts of lexical entries are associated through coindexation. In transparent constructions, the subparts of semantic, morphosyntactic and phonological representations are all coindexed with one another. His analysis then diverges from that of Köhnlein because, if an element is unattested as a free form or if a construction is semantically opaque, then only the morphosyntactic and phonological levels are coindexed, and the whole form has a semantic representation that is unrelated to those of its

constituents (if any). Köhnlein (2015) assumes that all three levels are systematically coindexed, but that the semantic representations of the constituent of place-names is underspecified, while Mascaró assumes that there need not be coindexation between all levels.

Let us conclude this section by noting that none of these two studies uses large empirical data and, surely, such data would enrich our understanding of the morphophonology of complex names.

3. Aims of the paper

3.1. The phonology of complex English proper names

Many English place-names such as *Canterbury*, *Oxford* or *Bradfield* are known to be historical compounds, and much of the work on English onomastics has been done by the English Place-Name Society, whose work has been compiled by Watts (2004) in his massive dictionary of English place-names. I am not aware of comparable work on complex family names, which may also be complex diachronically (e.g. *Donaldson, Wordsworth*). For both categories, however, semantics are clearly non-transparent or even entirely opaque. By this, I mean that, although some constituents do look like real words, the semantics of such words is purely referential, and does not contain the semantics of its constituents, even associated in an unpredictable way, as is found in compounds. For example, one may see that *Oxford* is historically made up of *ox* and *ford*, but the meaning of *Oxford* does not make any use of the meanings of those constituents. As I will argue in §6.2, the second element simply functions as a classifier and the first element as a referential pointer. This lack of semantic transparency might lead some to assume that such names are not complex morphologically, and would predict that their phonological behaviour should resemble that of simplex words. I have not been able to find a lot of literature on the phonological behaviour of such names, thus confirming Köhnlein's (2015) observation that "in linguistic theory, […] names have been largely neglected so far, despite the fact that they are such obvious language material". The few sources that I have been able to find, and which mention the phonological behaviours of such names will be discussed in the next section.

As we have seen in the previous section, the analysis of comparable Dutch and Central Catalan place-names as simplex words does not hold. My aim is thus to study certain key phonological behaviours of proper names and to establish whether they are consistent with the assumption that such words are simplex or if, on the contrary, they suggest that they should be analysed as complex words. We also saw in the previous section that those previous studies do not base their analyses on large data samples, and so this study of English proper names will be conducted using as large a dataset as possible. This will allow us to get a more fine-grained understanding of the phonological behaviours of such names and how they relate to one another.

3.2. Selected phonological properties

In this paper, I will study eight phonological properties which may differentiate simplex words from complex words. I will first look at two types of words for which the position of primary stress may be informative with regards to morphological structure. Then, four phonotactic characteristics will be investigated, for which the distribution of certain segments or clusters of segments may function as a 'boundary signal'. Then, I will study the realisation of vowels found in the environment for the rule of Trisyllabic Shortening. Finally, I will study vowel reduction in the final constituent. Let us go through those properties and how simplex words and complex words contrast. Note that the only complex structures that will be considered are those for which there exists a strong morphological boundary between the two constituents: words with neutral suffixes and compounds.[5]

The vast majority of the literature on English stress assumes that the main stress should be placed on one of the last three syllables of the words and that the position of that stress, in nouns at least, depends on the weight of the penultimate syllable (see, among many, Burzio 1994; Chomsky and Halle 1968; Halle and Vergnaud 1987; Hayes 1980; Liberman and Prince 1977). This is

sometimes called the 'Latin Stress Rule', and can be stated as in (2), which we reproduce from Moore-Cantwell (2020).

(2) *Latin Stress Rule for English*
If a word's penultimate syllable is heavy, then it takes penultimate main stress. If the penultimate syllable is light, then the word takes antepenultimate main stress.

A syllable is assumed to be heavy if it contains a long vowel or if it closed by a consonant. Complex words are known to regularly violate (2), as words which contain neutral suffixes show an underapplication of this rule and do not have penultimate stress even if the penult is heavy (3a) (see, among many, Raffelsiefen 2005). The same is true of many nominal compounds which bear primary stress on their first constituent (3b).

(3) a. *cáptaincy, cónstantly, devélopment, nóvelty, párenthood, thóusandfold, wízardry…*
b. *búlletproof, mótorway, nóteworthy, páperwork, wínning-post, wár-weary …*

Such words are also known to show extrafenestral stress, i.e. primary stress that is leftwards of the antepenultimate syllable (4), if the second constituent is monosyllabic and the first constituent has antepenultimate stress (4a), or if the second constituent is disyllabic and the first constituent has penultimate stress (4b).

(4) a. *advénturousness, àspirátionally, chárácterless, símilarly, váriableness …*
b. *ánybody, créditworthy, cópyholder, ínfantryman, wátermelon, yéllowhammer …*

Therefore, finding non-penultimate stress in proper names with a heavy penult or extrafenestral stress will possibly be an indication that these words are better analysed as complex words rather than as simplex words. However, it will not be categorical evidence for complexity, as there are known lexical exceptions to those stress rules (e.g. *chárácter, líberty, mínister, álligator, cáricature, nécessary*). The only paper that I have been able to find on how proper names are stressed is very ancient and does not deal with those issues (Hempl 1896). Hempl only notes that many of what he calls 'conglomerates' such as *Altenburg*, *Newport* or *Pittsburg* "came to have the stress of real compounds", but does not say how that may differ from simplex words.

The second type of property that I will be investigating is the phonotactics of proper names. The basic assumption which I will adopt is that illegal clusters or anomalous distributions for individual segments may be interpreted as 'boundary signals', in the spirit of Trubetzkoy's (1936) 'Grenzsignale', i.e. phonological signals of morphosyntactic structure. This type of property has been evoked in §2.1 for opaque prefixed words, some of which have word-internal consonant clusters which are not normally found in simplex words. For example, Guierre (1990) claims that the only clusters of two obstruents which are attested in simplex words are clusters of two voiceless obstruents. In her study of the phonological properties of compounds, Allen (1980) notes that there may be consonants that do not assimilate when placed at the boundary between the two constituents of a compound (e.g. *goo*[s]*e-barn, pa*[n]*-cleaner*), although she notes that such 'phonological distortions' may weaken over time and even disappear. Kaye (1995) also notes that suffixation may generate clusters that are otherwise unattested, the most extreme case being a word such as *sixths*, with the cluster /ksθs/. He mainly discusses final clusters, but the same can be said of medial clusters (e.g. /df/ does not occur in simplex words, but can be found in suffixed words such as *dreadful* or *handful*). Therefore, the first phonotactic property that I will study is that of medial clusters.

The second phonotactic property concerns the anomalous distributions of two individual segments: alternations between /ɪj/ and /ɪ/ (or /ə/), and occurrences of the velar nasal before a vowel or a non-velar consonant. The symbol /i/ is used by Wells (2008) in his dictionary to represent the

possible neutralisation between /ɪj/ and /ɪ/ in unstressed positions. In several varieties of English word-final /ɪ/ has undergone 'happY-tensing' and now has a more tense realisation that is now mostly perceived as the FLEECE vowel (e.g. *happ*/ɪj/, *cit*/ɪj/, *diversit*/ɪj/). I will assume here that Wells's /i/ represents a reduced (some would say unstressed) realisation of /ɪj/. That vowel is also the realisation of unstressed <i> or <e> before vowels (e.g. *r*/ɪj/*ality*, *nucl*/ɪj/*ar*, *glor*/ɪj/*ous*). Several have noted that words that end in /ɪj/ (5a) maintain that vowel preconsonantally when they are the first constituent of a compound (5b), but that there is variability in suffixed words, some which maintain /ɪj/ (5c), while others alternate between /ɪ/ and /ə/ (5d) (Allen 1980; Cruttenden 2014: 114; Halle and Mohanan 1985; Herment 2010).[6]

(5)	a.	b.	c.	d.
	bod/ɪj/	*bod*/ɪj/*guard*	*bod*/ɪj/es	*bod*/ɪ ~ ə/*ly*
	fanc/ɪj/	*fanc*/ɪj/*work*	*fanc*/ɪj/es	*fanc*/ɪ ~ ə/*ly*
	bab/ɪj/	*bab*/ɪj/*-sit*	*bab*/ɪj/*hood*	-
	beaut/ɪj/	*beaut*/ɪj/ *contest*	*beaut*/ɪj/*es*	*beaut*/ɪ ~ ə/*ful*

Therefore, the occurrence of word-internal preconsonantal unstressed /ɪj/ may be taken as a boundary signal, indicating that it is at the end of a morphological constituent.

Then, the velar nasal /ŋ/ is known to be found before velar stops (e.g. *finger, single, tinker, monkey*) or word-finally (e.g. *sing, hang, wrong*) in simplex words. It may occur prevocalically in words containing vowel-initial neutral suffixes (6a)[7] or in compounds (6b).

(6) a. /ˈlɔŋɪʃ/ *longish*, /ˈsɪŋɪŋ/ *singing*, /ˈswɪŋə/ *swinger*, /ˈrɔŋə/ *wronger*
b. /ˈhaŋawt/ *hangout*, /ˈhaŋˌəwvə/ *hangover*, /ˌhɪːɪɪŋɪmˈpɛːd/ *hearing-impaired*

Therefore, the distribution of the velar nasal can be another cue to the presence of a morphological boundary, among illegal medial consonant clusters (i.e. clusters involving /ŋ/ not followed by a velar stop; e.g. /ˈkɪ**ŋl**ɪj/ *kingly*, /ˈlɔ**ŋh**and/ *longhand*, /ˈlə**ŋf**əl/ *lungful*), but also when it is found in prevocalic positions, as in (6).

The last phonotactic property that I will study, in line with what Köhnlein (2015) has done for Dutch, is the presence of word internal superheavy syllables. A superheavy syllable is a syllable which has either a long nucleus and a coda consonant or which has at least two coda consonants. In simplex words, the occurrence of word-internal superheavy syllables is strongly restricted. As described by Harris (1994: 69) and Harris and Gussmann (1998), among others, such syllables are only found with a long nucleus (i.e. a diphthong or a pure long vowel) and only if:

(7) a. the following consonant is a fricative or a sonorant (e.g. /ˈkawnsəl/ *council*, /ˈdɛjndʒə/ *danger*, /ˈojstə/ *oyster*, *pastry* /ˈpɛjstɪj/), and
b. if it is a sonorant, the consonant must be homorganic with the following onset, and the place of articulation is almost invariably coronal (e.g. /ˈkawnsəl/ *council*, /ˈloːndə/ *launder*, /ˈʃəwldə/ *shoulder*, /ˈbəwldə/ *boulder*).

Such restrictions are regularly violated in complex words, either compounds (7a) or suffixed words (7b).

(8) a. /ˈ**gɛjm**plɛj/ *gameplay*, /ˈ**hɑ**ːtfɛlt/ *heartfelt*, /ˈ**so**ːltbɔks/ *saltbox*
b. /ˈ**ɑ**ːtfəl/ *artful*, /ˈ**mo**ːltstə/ *maltster*, /ˈ**sɛjf**tɪj/ *safety*

In a footnote, Harris (1994: 69, fn. 60) says that *Cambridge*, which also violates these restrictions, "can be set aside as a compound name with a historical word-level internal boundary".

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

Giegerich (1999: 275-276) also notes that internal superheavy syllables are very common in place-names.

> A frequent historical characteristic of names, seldom shared by other lexical items, is that they may be obscured compounds. While there is no reason to suppose that such items are morphologically complex in synchronic terms, their segmental make-up and possibly 'irregular' syllabification are a residue from earlier complexity. This phenomenon is also found, for example, in plant names *(honewort, loosestrife, thalecress, groundsel)*. For the same diachronic reason, personal names may display irregular syllabification: *Edward* should have an ambisyllabic /d/ (compare *dwell)* but syllabifies as *Ed.ward*. All these forms in fact contain (one or more) cranberry morphs. The phonological behaviour of such items becomes regular if they are assumed to have retained an internal ']' bracket (Allen 1980), which automatically licenses a preceding consonant -- a diacritic solution, of course, but one that has strong diachronic motivation.
>
> Giegerich (1999: 276)

One might interpret Giegerich's analysis in the light of that proposed by Mascaró (2016), and which we have discussed in §2.2. Retaining a ']' bracket while not being "morphologically complex in synchronic terms" could be taken to mean that such names are semantically non-transparent but complex at the levels of form: morphological (which need not correlate with semantics) and phonological. These observations suggest that many names may contain internal superheavy syllables, and that their presence signals morphological complexity, especially if those syllables violate the restrictions in (7).

The next phonological characteristic that I will consider is whether or not names obey the rule called Trisyllabic Shortening (also called Trisyllabic Laxing or Luick's rule). This rule states that vowels that are at least the third syllable from the end of the word should be short (e.g. *c*/a/*lendar*, /ɛ/*lephant*, *v*/ɪ/*negar*). That rule applies in certain complex words if they contain suffixes that are known to affect stress, as in (9a). However, Wennerstrom (1993) shows that this is not true of compounds (9b). This rule is also known not to apply to words with neutral suffixes (9c).

(9) a. *div*/ɑj/*ne – div*/ɪ/*nity*, *prof*/ɛj/*ne – prof*/a/*nity*, *ser*/ɪj/*ne – ser*/ɛ/*nity*
b. *s*/ɪj/ *patrol* (**s*/ɛ/ *patrol*), *t*/ɛj/*ble loom* (**t*/a/*ble loom*), *pr*/əw/ *mobile* 'conveyance for pros' (**pr*/ɔ/ *mobile*)
c. ˈ/ɛj/*gelessness* (*ˈ/a/*gelessness*), ˈ*p*/ɛː/*rentless* (*ˈ*p*/a/*rentless*), ˈ*f*/ɑj/*nally* (*ˈ*f*/ɪ/*nally*)

However, like stress rules, that rule is known to have lexical exceptions, in both simplex (e.g. /ɛj/*pricot*, *h*/ɑj/*bernate*, /əw/*beron*) and complex words (e.g. *h*/ɑj/*phenate*, *n*/əw/*tify*, *ob*/ɪj/*sity*). Therefore, finding words in which that rule does not apply will only be relevant probabilistically, if the proportion of words with long vowels is significantly higher than in the rest of the lexicon. Deschamps (1994) evaluates the proportion of words that obey the rule at 92%, so the proportion of 'irregular' items will have to be well above 8% for it to be considered significant.

Finally, innumerable reference works on English phonology have noted that the general tendency for compounds with primary stress on their first constituent is for their second element not to contain reduced vowels (10a) (except in rare cases such as *breakfast* /ˈbrɛkfəst/ or *cupboard* /ˈkəbəd/), while many nominal suffixes do have reduced vowels (10b), though not all (10c).

(10) a. *airport* /ˈɛːpoːt/ (*/ˈɛːpət/), *blackbird* /ˈblakbəːd/ (*/ˈblakbəd/, *website* /ˈwɛbsɑjt/ (*/wɛbsət/)
b. *blindness* /ˈblɑjndnəs/, *endowment* /ɪnˈdawmənt/, *graceful* /ˈgrɛjsfəl/
c. *childhood* /ˈtʃɑjldhəd/, *bucketful* /ˈbəkɪtfəl/, *falsehood* /ˈfoːlshəd/

It is that property of compounds which has led many phonologists to assume that the second constituent forms a phonological domain, while suffixes integrate the phonological domain of the

base (though there is some controversy in Prosodic Phonology regarding how; see §6.2). Allen (1980) notes that compounds show some variability and that there are cases like (11a), which contrast with 'regular' cases such as (11b).

(11) a. *mainland* /-lənd/, *Iceland* /-lənd/, *fireman* /-mən/, *strawberry* /-b(ə)rɪj/, *Dartmouth* /-məθ/
b. *bear-land* /-land/, *farm-land* /-land/, *tax-man* /-man/, *bush-berry* /-bɛɪɪj/, *river-mouth* /-mawθ/

She notes that compounds such as those in (11a) are "semantically non-transparent, to a greater or lesser extent; many are names for specific items, persons or places, rather than general category names", and this indeed includes the kind of proper names that are the focus of this paper.[8] She also notes that those words "appear to be concatenations of a lexical item plus an element which is clearly related to a lexical item, but which has lost some of the semantic characteristics of the free lexical item". Eventually, she proposes that "the word-like second element of compounds like *chairmən* have been reanalyzed as suffixal elements of some type", and compares this to the historical development of the suffixes *-dom* and *-hood*, which were free words in Old English.

While the general tendency for compounds to have full vowels in their final element contrasts with simplex words, it is important to note that simplex words do not systematically have reduced vowels in their final syllable. Deschamps *et al.* (2004: 219) identify the four contexts in (12), for which reduced vowels are generally prohibited in the final syllable.

(12) a. digraphs: *shadow* /ˈʃadəw/, *guffaw* /ˈgəfoː/, *cashew* /ˈkaʃɯw/ (excepted certain suffixes, e.g. *-ous*, *-our*, *-ey*)

b. final <o> or <u>: *lotto* /ˈlɔtəw/, *photo* /ˈfəwtəw/, *menu* /ˈmɛnjɯw/, *zebu* /ˈzɪjbɯw/…

c. <__Ce#>: *analyse* /ˈanəlɑjz/, *accelerate* /əkˈsɛləɹɛjt/, *turpentine* /ˈtəːpəntɑjn/, *fragile* /ˈfɹadʒɑjl/, *microbe* /ˈmɑjkɹəwb/, *finite* /ˈfɑjnɑjt/… (excepted certain suffixes, e.g. *-age*, *-ace*, *-ate*$_{N/ADJ}$)

d. <__x#>: *climax* /ˈklɑjmaks/, *equinox* /ˈɛkwɪnɔks/, *ibex* /ˈɑjbɛks/…

That last phonological property shows less clear-cut differences between complex and simplex words as there is variation for both types of words. However, the presence of full vowels in the second constituent of a proper name, outside of the environments of (12), may be taken to indicate compound-like phonological behaviour. A reduced vowel in that constituent will be more difficult to interpret, as it could be taken to mean that the words are analysed as simplexes or as constructions involving a neutral suffix. The other characteristics that we have described may be used to choose between those two analyses.

The phonological characteristics that will be investigated in the rest of the paper and which we have detailed in this section are summarised (and slightly simplified) in Table 1.

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

		Simplex	Neutral suffix	Compound
Primary stress	> 2 syllables, closed penult	Penult	Possibly before penult	
	> 3 syllables	Within final three-syllable window	Possibly extrafenestral	
Phonotactics	Medial clusters	Restricted	Unrestricted	
	Preconsonantal stressless /ɪj/	No	Possible	Yes
	/ŋ/	__ {C[+velar], #}	Unrestricted	
	Internal superheavy syllable	Strongly restricted	Possible	
Trisyllabic Shortening		Yes	No (if the base has a long vowel)	
Vowel reduction in final constituent		Yes		No

Table 1. Summary of the phonological characteristics studied in this paper

4. Methodology

4.1. Data collection and annotation

As noted in §2.2, one limitation of existing studies on proper names is the lack of large-scale data. Therefore, in the tradition of the Guierrian School[9] (Dabouis *et al.* 2023), I have sought to collect a sample of complex proper names that is as extensive as possible. However, this endeavour has turned out to be far from easy, as there exist lists of place-names, but they do not contain pronunciation information, and pronunciation dictionaries do not indicate that words are proper names other than through the use of capitalisation. Moreover, it was necessary to make decisions on which words would qualify for this study, as etymological information is not always available and so historical morphological complexity could not be used as a criterion to define which words to include.

Therefore, I chose to manually collect the British pronunciations given in Wells (2008) for words starting with a capital letter and which contain a recurrent initial or final orthographic sequence, that is to say any sequence that occurs in at least two forms listed in Wells (2008), either initially or finally. Only words which do not contain an internal dash or space were collected. The issue was then to decide which sequences to search for. I started by collecting items with common final sequences (e.g. *-ton*, *-ham*, *-ford*), which in turn led to the identification of certain initial sequences (e.g. having *Southam* led to all forms in *South-*, then to all final elements associated with *South-*: *Southborough*, *Southcott*, *Southdown*, etc.). This initial search was then complemented by a search through an automatic extraction of all nouns listed in Wells (2008), in which common recurrent final sequences that had not yet been identified were collected. The list of final constituents was then completed using those listed on Keith Briggs's website[10], who bases his work on that of the English Place-Name Society. Finally, during the first analyses of the phonotactic properties of the words, I searched for a number of medial consonant sequences, which allowed for the identification of words belonging to more minor classes that had avoided detection thus far.

Words that are marked as being borrowed from another language (e.g. *Brandenburg*, *Goncourt*, *Hanover*) were not included as I am here focusing on English proper names. Wells does this by indicating the source language and giving a phonemic transcription of the word in the source language (e.g. the entry for *Bundesbank* reads as follows: "ˈbʊnd əz |bæŋk ˈbʌnd —German [ˈbʊn dəs |baŋk]"). Trademarks (e.g. *Merrydown*) were not included either. Names referring to places that are not in countries in which English is the main official language (e.g. *Carlsberg*, *Heligoland*, *Leghorn*, *Freetown*) were excluded during data cleaning.

This led to the collection of 3579 words, whose formal characteristics will be described in the next section, before we turn to their phonological characteristics in §5. The full dataset can be accessed on OSF at the following link: https://osf.io/t3pwy/.

The identification of internal constituents was done on an orthographic basis, through the identification of the final constituent first. The first constituent was identified as what was left once the final constituent was removed, although minor adjustments were made whenever it appeared that orthographic simplifications have occured, for example to avoid adjacent identical consonants. For example, items that end in *-ham* are sometimes attached to a constituent ending in <th>, as shown by the pronunciation, but they are not spelled with <thh> (e.g. *Southam* /ˈsawðəm/ was analysed as *south* + *-ham*). Many first constituents end in <-s> and, in the cases in which there exists other words in the dataset with an identical first constituent except for the <s> (cp. e.g. *coul – couls, din – dins, king – kings, peter – peters*), the <s> was not counted as part of the first constituent nor of the second. In the case of words ending in <-ston>, it was assumed that the segmentation is X + *-ston* rather than X*s* + *-ton* on the basis that the first constituents identified in that way may:

- occur in other words without the <s> (e.g. *Cranston* cp. *Cranbrook, Cranfield, Cranwell*);
- occur in a near homograph without the <s>, analysed as X + *-ton* (e.g. *Alton – Alston; Coulton – Coulston; Dalton – Dalston; Wigton – Wigston*);
- be 'free' in the sense described in the following section if <s> is excluded (e.g. *Beeston, Grimston, Johnston*), exception made of *Charleston*, which was analysed as *Charles* + *-ton*.

Final constituents were coded as 'free' or 'bound' depending on whether or not there exists a homographic freestanding word listed in Wells (2008) (e.g. *apple* in *Appleby* is treated as 'free' because there is a freestanding word *apple*).[11] Similarly to the segmentation procedure described in the previous paragraph, if there exists a near homograph that consists of the first constituent minus <s>, the constituent was treated as free. In some cases, that leads to rather unintuitive classifications, as in the case of *-ton*, which is not related etymologically to the word *ton*. However, it is unclear what synchronic criteria should be used to say that, for example, *Oxford* contains the free word *ford*, but *Washington* does not contain the word *ton*. That classification of constituents as 'free' or 'bound' is therefore to take with caution, and I mainly take it to mean that it is possible that certain constituents may be perceived to be related to existing words while others may not, and that it could make a difference in their phonological behaviour. Therefore, looking only at 'bound' items is a guarantee that the item is quite unlikely to be influenced by an existing word, while it is much less clear how we should interpret the 'free' category, beyond a potential influence from existing words to which certain items may be associated.

For ease of presentation, the way that the annotation was made regarding the different phonological behaviours investigated in this paper will be presented before the results corresponding to each of those behaviours in §5.

4.2. Description of the dataset

Let us now describe some of the formal characteristics of the dataset. A first observation is that there are more different first constituents than there are second constituents, and that the latter generally occur in more different words than the former, as shown in Table 2.

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

	First constituents	Second constituents
Number of different forms	2338	170
Maximum number of different words in which they appear	18	406
Average number of different words in which they appear	1.5	21.1
Number of constituents that occur in only one word	1758	29
Proportion of free constituents	1245/2338 – 53%	151/170 – 89%
30 most frequent constituents (by decreasing type frequency)	*new, ash, black, south, stan, green, west, har, al, whit, water, war, wood, long, king, fair, brad, bur, bar, win, white, north, col, mar, stock, red, kirk, hol, hor, broad*	*ton, ley, son, ham, ford, man, field, by, land, den, bury, ston, well, don, wood, dale, worth, wick, bridge, more, way, shire, ville, sham, town, ling, hurst, kin, head, gate*

Table 2. Main distributional properties of first and second constituents in the dataset

If we now consider the internal characteristics of constituents, some of them may occur in the dataset without the other constituent. There are 46 words which have a first constituent that is attested in the dataset without the second constituent (13a), and each such already complex constituent only occurs once, except *Welling* which appears in two words. Seventeen of those are county names in *-shire*. The same is true of the second constituent of 31 words, with 15 different constituents (13b)

(13) a. ***Alden****ham*, ***Auden****shaw*, ***Barset****shire*, ***Bedford****shire*, ***Belling****ham*, ***Berwick****shire*, ***Bexley****heath*, ***Bingham****ton*, ***Boreham****wood*, ***Braden****ton*, ***Buckingham****shire*, ***Burton****wood*, ***Cambridge****shire*, ***Carling****ford*, ***Chorley****wood*, ***Clifton****ville*, ***Edwin****stowe*, ***Gloucester****shire*, ***Gorman****ston*, ***Hailey****bury*, ***Hereford****shire*, ***Hertford****shire*, ***Hodgkin****son*, ***Hopkin****son*, ***Huntingdon****shire*, ***Jackson****ville*, ***Jenkin****son*, ***Leicester****shire*, ***Monmouth****shire*, ***Moreton****hampstead*, ***Newton****more*, ***Northampton****shire*, ***Nottingham****shire*, ***Oxford****shire*, ***Pembroke****shire*, ***Penton****ville*, ***Ponson****by*, ***Rowland****son*, ***Simpkin****son*, ***Stafford****shire*, ***Tompkin****son*, ***Watkin****son*, ***Welling****borough*, ***Welling****ton*, ***Wensley****dale*, ***Worcester****shire*

b. *Ash**burton**, Cleck**heaton**, Crick**howell**, Fen**stanton**, Finc**hampstead**, Glen**gormley**, Higgin**botham**, Hun**stanton**, Inver**gordon**, Kim**bolton**, Leck**hampton**, Little**hampton**, Moreton**hampstead**, North**hampton**, Nun**eaton**, Oak**hampton**, Oke**hampton**, Rams**botham**, Rath**farnham**, Rock**hampton**, Roe**hampton**, Row**botham**, Side**botham**, Somer**leyton**, South**hampton**, Throg**morton**, War**burton**, West**houghton**, Wheat**hampstead**, Winter**botham**, Wolver**hampton***

One interesting combinatorial property of those constituents is that 75 of them may occur either as a first or a second constituent (14).

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

(14) *bank, beck, bent, berg, berry, bottom, bourn, bourne, brook, brooke, brough, burgh, burn, burton, by, caster, castle, cheap, chester, church, cock, den, don, down, fleet, ford, gate, glen, grave, ham, head, heath, hill, horn, horse, hough, kirk, lake, leigh, ley, ling, lock, long, low, man, moor, more, over, port, porth, sea, shot, side, smith, staple, stead, sted, stock, stone, thorn, thorpe, ton, town, wade, wark, water, way, wear, well, white, wick, win, wood, wych, wyn*

Finally, the division of the data according to the free or bound nature of the two constituents is shown in Table 3. As can be seen, words that are entirely made up of bound constituents are quite rare (90 words - less than 3% of the data).

		Second constituent		
		Free	Bound	Total
First constituent	Free	2021	90	1412
	Bound	1412	154	2175
	Total	3343	244	3587

Table 3. Distribution of words depending on the free or bound nature of their two constituents

5. Results

5.1. Primary stress

5.1.1. Words longer than two syllables with a closed penult

I will focus on words in which the penult is heavy because it is closed[12], as heavy penults are generally formed by the concatenation of the two constituents. The existing literature points that certain consonants may have variable weight (Burzio 1994: 61-62; Giegerich 1999: 264; Halle and Vergnaud 1987: 257; Moore-Cantwell 2020; Selkirk 1984: 127). Syllables that contain consonants which may be syllabic have been claimed to make syllables heavy if stressed but not if unstressed, as the consonant then occupies the nucleus, and the syllable has no coda. Syllables for which the vowel is followed by an /sC/ cluster are also problematic because such clusters may syllabify tautosyllabically or heterosyllabically. Finally, in a non-rhotic variety such as the one we are analysing, it is an open question whether or not etymological (and orthographic) /ɹ/ should be assumed to be still present underlyingly. Therefore, the 644 words in the dataset which may be analysed as closed were sorted based on the nature of the consonant which may close the penult. In total, 597 words (93%) have antepenultimate stress, while only 28 (4%) have penultimate stress, and the nature of the consonant closing the penult makes no difference. Examples of words with antepenultimate stress (15a), penultimate stress (15b), final stress (15c) and preantepenultimate stress (15d) are shown in (15).

(15) a. *Ábingdon, Ápplegarth, Báskerville, Bédfordshire, Cálverley, Cásterbridge, Dóncaster, Dúggleby, Éckersley, Édmondson, Félixstowe, Géraldton, Hámington, Hémingway, Hinchingbrook, Hópkinson, Íllingworth, Ísherwood, Lázonby, Létterman, Máidenhead, Múggleton, Névinson, Níckleby, Ópenshaw, Óswestry, Próvincetown, Puddletown, Rémington, Rútherford, Sánderstead, Sílverstone, Símington, Tínkerbell, Tódmorden, Únderwood, Úpminster, Wáterstone, Wátkinson…*

b. *Carshálton, Cleckhéaton, Dumbárton, Fenstánton, Glengórmley, Kimbólton, Lohgílphead, Sarándon, Winstánley…*

c. *Àberpórth, Bìrkenhéad, Bòrehamwóod, Bùrtonwóod, Hòrsmondén, Ìnvernéss, Jòrdanhíll, Kèlvinsíde, Lòchearnhéad, Pèrranpórth, Pèterhéad*

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

d. *Ándersonstown, Búckinghamshire, Frédericton, Hétherington, Húntingdonshire, Kídderminster, Nóttinghamshire, Óllerenshaw*

We can compare the rate of antepenultimate in this subset to that observed in words which do not have a closed penult. There are 505 such words, and 426 (84%) of them have antepenultimate stress while only 32 (6%) have penultimate stress. If anything, the effect of the closed penult goes in the opposite direction as that predicted by the Latin Stress Rule. Potential effects of the free nature of the two constituents were tested but no significant effects were found.

5.1.2. Words longer than three syllables

The dataset contains 70 words which may be pronounced with at least four syllables, and 45 of them (62%) may be stressed on the preantepenultimate syllable. Among those words, 25 have *-bury, -burgh, -brough* or *-borough* as their second constituent (16a), and that element is usually pronounced as a monosyllable, even though it may be pronounced disyllabically: /bəɹɪj ~ bɹɪj/ *-bury*, /bəɹə ~ bɹə ~ ˌbəɹə/ *-burgh* and *-borough*, /bəɹə ~ bɹə ~ bɹəf/. Three other words may undergo syncope elsewhere in the word (16b), and one word may undergo compression (16c).[13] The remaining 16 may not undergo such reductions (16d).

(16) a. *Ábbotsbury, Áddlebrough, Álconbury, Ámondbury, Áttenborough, Áttleborough, Bráckenbury, Cánonbury, Cánterbury, Chánctonbury, Cónisborough, Édinburgh, Fráserburgh, Glástonbury, Gúnnersbury, Háileybury, Hélensburgh, Híldenborough, Íngleborough, Músselburgh, Péndlebury, Ráttenbury, Wáterbury, Wédnesbury, Wéllingborough*

b. *Frédericton, Hétherington, Márgerison* (variant *Margérison*), *Óllerenshaw* (variant *Òllerénshaw*)

c. *Hárrietsham*

d. *Ándersonstown, Búckinghamshire, Círencester, Dérwentwater, Gódmanchèster* (variant *Gòdmanchéster*) *Hígginbotham, Hígginbottom, Húntingdonshire, Kidderminster, Líttlehampton* (main pronunciation *Lìttlehámpton*), *Míckleover, Nóttinghamshire, Sómerleyton, Wínterbotham, Wólverhampton* (main pronunciation *Wòlverhámpton*)

If we assume that the reductions that are possible for the words in (16a-c) are phonetic processes, then that means that these words have four phonological syllables, and so have phonological preantepenultimate stress. This is based on the assumption that there may be mismatches in the number of syllables in the lexical representation (which often correlates to the number of orthographic syllables) and the number of phonetic syllables. Various processes of reduction or epenthesis may affect how many syllables a word contains phonetically.

The remaining 25 words have either antepenultimate stress (17a), penultimate stress (17b) or final stress (17c).

(17) a. *Basútoland, Caernárfonshire, Christópherson, Clackmánnanshire, Damáraland, Fazáckerley, Ìnvermóriston, Mashónaland, Namáqualand, Northámptonshire, Northúmberland, Nyásaland, Phizáckerley, Wéstminster*[14]

b. *Àbercónway, Àberdéenshire, Bàllycástle, Càrrantúohill, Ìnvergórdon, Kènnebúnkport, Mòretonhámpstead, Nàrragánset, Rùmpelstíltskin, Tìllicóultry*

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

c. *Shòeburynéss*

As was seen in §3.2, preantepenultimate stress may arise in complex words if the final constituent is disyllabic and the first constituent has penultimate stress or if the final constituent is monosyllabic and the first constituent has antepenultimate stress. First, note that there is one word in (17), *Invermoriston*, for which the construction is *Inver-* + *-moriston* (cp. *Invergargill*, *Invergarry*, *Invergordon*), and so preantepenultimate stress is not necessarily predicted as the second constituent is trisyllabic. Then, can it be assumed that some of the patterns observed here are due to phonological identity with an embedded word (or at least one perceived to be embedded)? There are 37 words for which the first constituent is free and, in 36 cases, stress is on the same syllable as in the related word (e.g. *cánon* ↔ *Cánonbury*, *Húntingdon* ↔ *Húntingdonshire*, *Damára* ↔ *Damáraland*, *Northámpton* ↔ *Northámptonshire*). The only exception to this is *Christópherson* (cp. *Chrístopher*). In the remaining 35 words, preantepenultimate stress is possible in 20 words, so phonological identity with existing words cannot explain all the patterns observed here.

5.2. Segmental phonotactics

5.2.1. Word-medial clusters

In order to determine whether the words in the dataset contain internal consonant clusters which are unattested word-medially in simplex words, I used the sound search tool provided in the CD-ROM version of Wells (2008). If a cluster was found to occur only in clearly complex words (i.e. words that contain productive affixes or compounds) or in the type of proper names that is the focus of this paper, then the word containing that cluster was marked as containing an illegal cluster. For example, (18) shows the results returned by the search of five different clusters found in the data.[15]

(18) /bʃ/**:** *job-sharing, Trubshaw*

/θb/: *clothbound, Cuthbert, Cuthbertson, deathbed, deathblow, earthborn, Lethbridge, Lothbury, mothball, northbound, Northbrook, Rathbone, Rothbury, Southborough, southbound, toothbrush*

/mzb/: *Amesbury, Bloomsbury, gemsbok, Grimsby, Malmesbury, Ormesby, Ormsby, Ramsbotham, Ramsbottom, Samlesbury, Williamsburg*

/mzd/: *Domesday, Doomsday, Helmsdale, Holmesdale, Lumsden, Ramsden*

/ftsb/: *Shaftesbury* (for that word, if the cluster simplifies to /fsb/, there is only one other word that contains the same cluster: *wolfsbane*)

More than a third of the dataset (1299/3579) contains such illegal clusters. Let us detail the clusters found. For two-consonant clusters, there are 132 different clusters found in 797 different words, as shown in Table 4. We can add 24 clusters found in 37 words in which the two-consonant cluster is preceded by an orthographic <r>. Let us point out that many of those are highly marked, such as clusters with obstruents which disagree in voicing (e.g. /kd/ in *Buckden*, /gh/ in *Waghorn*) or geminates (e.g. /tt/ in *Levittown*, /kk/ in *Cockcroft*).

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

	p	t	k	b	d	g	f	θ	s	ʃ	h	v	ð	z	ʒ	m	n	ŋ	j	w	r	l
p			2	1	1		2			3	1					5						
t		2	5	5		1	10				14	1				6						3
k	3		1	8	4		11			6	14	1				9	4					
b			1		3				8	1						1				1		
d	1		6	27	1	5	11		5	7	12					19				14		40
g		2		3	4		2		2	2	2					8	2					
f					1				1							2	1					
θ	1		2	3	1	1	4		2		4					2				7		3
s				6	1						3											
ʃ		2		5			3				2	2				9	1			6		
v	1			1	3					1												4
ð				1																		
z		1			8	2										2				5		27
ʒ																						
tʃ			1	1	2		6									2				3		5
dʒ	1		2				1		2	1	2					3				8		
m			1					1		2										7		
n	4		8	42		1	22				27					15				27		
ŋ		88		5	11		13		1		5					1	1			7	1	13
l								2			7									1		

Table 4. Illegal word-medial two-consonant clusters found in the dataset (more frequent combinations are shown in darker grey). The first column represents the first consonant of the cluster, and the first line represents the second

In the dataset, we can find 416 words which contain 163 different three-consonant clusters. We could add to those 12 clusters involving <r> followed by three consonants, found in 12 words. The types of clusters found are detailed in Table 5. There too, there are many highly marked sequences, such as the 31 clusters which involve voicing disagreements between two adjacent obstruents (e.g. /ldh/ in *Guildhall*, /ŋzf/ in *Kingsford*).

First consonant	Second consonant	Third consonant	Number of cluster types	Examples
Obstruent	Liquid	Glide	5	dlw, psw, dzw, …
		Obstruent	12	dlʃ, plt, slt, …
	Obstruent	Glide	6	ksw, tsw, sθw, …
		Liquid	20	θkl, dbɹ, tfɹ, …
		Nasal	2	ksm, stm
		Obstruent	32	ksb, ptf, gst, …
Nasal	Obstruent	Glide	8	ndw, ŋzw, nθw…
		Liquid	8	ntʃl, nzl, ŋbɹ, …
		Nasal	4	ndm, dzm, ndʒm, ntʃm
		Obstruent	37	ntʃb, ŋzb, ndf, ...
Liquid	Nasal	Glide	1	lmw
		Obstruent	1	lmh
	Obstruent	Glide	3	ldw, lzw, lθw
		Liquid	3	lkl, lzl, ldɹ
		Nasal	3	ldm, ltm, lbɹ
		Obstruent	18	ldb, lzb, ldf…

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

Table 5. Illegal word-medial three-consonant clusters found in the dataset, organised by manner of articulation

Finally, the dataset contains 39 words with one of 30 different four-consonant clusters (e.g. /ldst/ in *Wealdstone*, /kskɪ/ in *Foxcroft*).

Quite expectedly, the presence of such clusters correlates with the free nature of the two constituents, as can be seen in Figure 1, in which it can be seen that the highest proportion of words containing an illegal cluster is found in words for which both constituents are free (41%) while the lowest is found in words in which both constituents are bound (27%).[16]

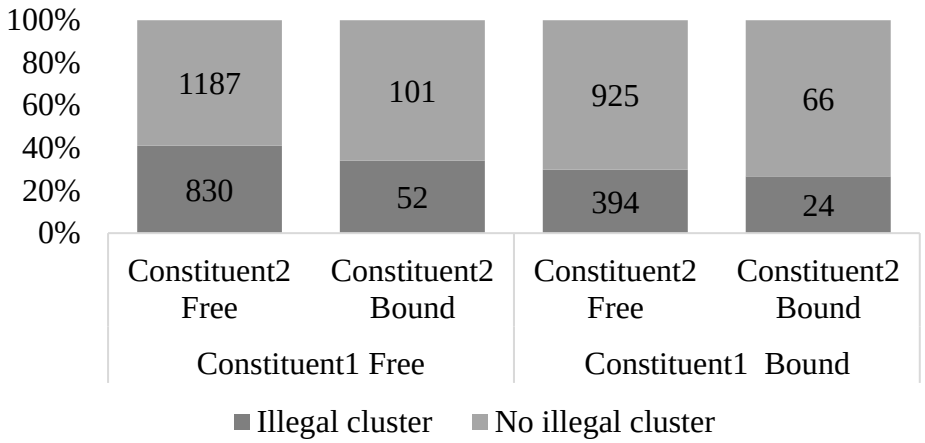


Figure 1. Presence of an illegal medial consonant cluster depending on the free or bound status of the two constituents

5.2.2. Preconsonantal unstressed /ɪj/

To find words for which there might be a word-internal preconsonantal unstressed /ɪj/, I searched for the 6 possible spellings of this vowel, <e>, <ey>, <ay>, <i>, <ie> or <y>, at the end of the first constituent. There are 128 such words in the dataset, and 38 of them have /i/ in that position in Wells, which are interpreted here as stressless realizations of /ɪj/. One additional word shows variation: *Eddystone* /ˈɛdɪstən/ or /ˈɛdɪjstəwn/. The remaining 89 words have /ɪ/ or /ə/ (henceforth, I will be referring to those as having /ɪ/).

(19) /ɪj/: *Aviemore, Ballycastle, Barrymore, Berryman, Bexleyheath, Bollywood, Chorleywood, Corriedale, Dixieland, Ferneyhough, Fernyhough, Gettysburg, Gulliford, Haileybury, Hartlepool, Heaviside, Hollywood, Holyhead, Holyport, Holywell, Honeycomb, Hunniford, Lillywhite, Lossiemouth, Maryport, Merriman, Merseyside, Mudeford, Murrayfield, Palfreyman, Sauchiehall, Shoeburyness, Spennymoor, Sunnyside, Tarrytown, Thorneycroft, Wensleydale.*

/ɪ/: *Aitchison, Alfreton, Armidale, Armistead, Baddeley, Baltimore, Bideford, Bigelow, Caldecote, Cheriton, Colyton, Corydon, Crediton, Davison, Denselow, Drakelow, Dungeness, Edison, Fenimore, Fentiman, Finlayson, Grandison, Halifax, Harrison, Haviland, Henryson, Honiton, Horniman, Hutcheson, Jameson, Jamieson, Lattimore, Malleson, Maryland, Morrison, Murchison, Pattison, Penistone, Renishaw, Runciman, Sotheby, Tennison …* (89 words)

In order to determine if there are any structural properties that can explain these differences in the data, I tested whether significant relationships with those properties could be found. Two factors appear to be related to the presence of /ɪj/: the free nature of the first constituent[17], and whether or not the second constituent is reduced. The free nature of the second constituent was also tested, but no

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

significant effect was found. As will be seen in §5.4, the final element may be full, reduced, varying between those two options or contain an uninterpretable vowel (<i> realised as /ɪ/). One word falls in the last category, *Mandeville* (with /ɪ/), and so it was left out of the following analyses. *Eddystone* was also left out to simplify the analysis.

I ran statistical tests in R (R Core Team 2022) using the function `ctree` from the package `party` and found indeed that those two factors were significant predictors of the presence of /ɪj/. The conditional inference trees generated by that function is represented in Figure 2. In this statistical tool, each node represents a statistically significant split in the data.

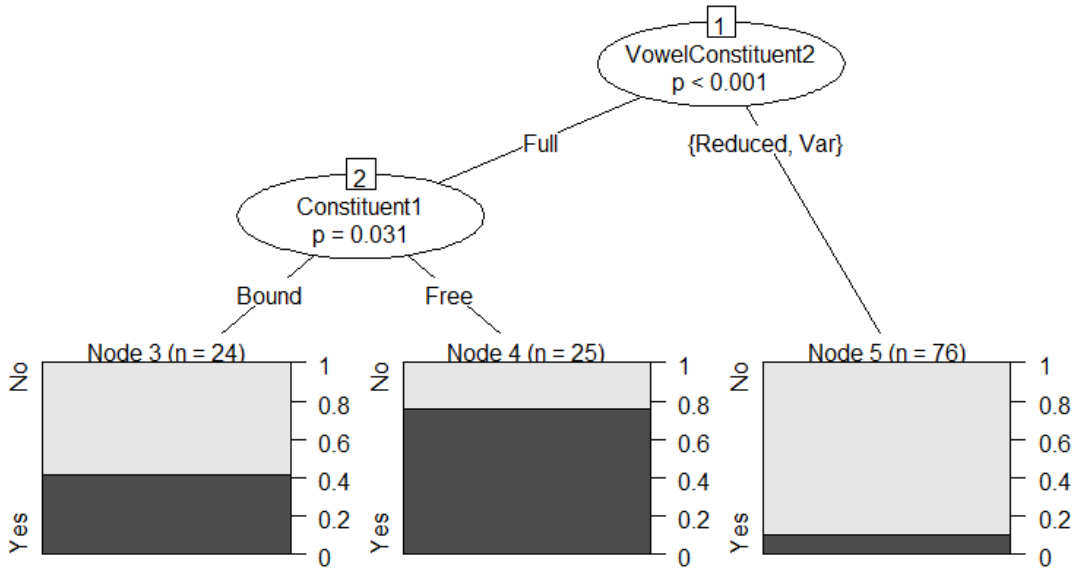


Figure 2. Conditional inference tree for the presence of /ɪj/ (in dark grey) depending on the free nature of the first constituent and the nature of the vowel (reduced or full) of the second constituent

Figure 2 shows that /ɪj/ is more likely to be present if the second constituent is reduced than if it is not or undergoes variation and, among those for which it is reduced, /ɪj/ is more likely to be present if the first constituent is free.[18] That result can be taken to mean that words with a free first constituent are indeed related to the corresponding freestanding words, even in the absence of a clear semantic relationship.

5.2.3. The distribution of /ŋ/

Some of the occurrences of /ŋ/ outside of its regular environments have been dealt with in §5.2.1. There are 15 types of illegal medial clusters found in 141 words which involve /ŋ/ followed by a consonant which is not a velar stop. Examples are shown in (20).

(20) /ŋb/ *Bassingbourn, Pangbourne, Springburn, Wellingborough…*
/ŋd/ *Abingdon, Haslingden, Langdale, Kingdon…*
/ŋf/ *Chingford, Langford, Lingfield, Stringfellow…*
/ŋl/ *Bletchingley, Langley, Madingley, Sherlock …*
/ŋt/ *Blessington, Eddington, Lexington, Wallington …*

As for prevocalic occurrences of /ŋ/, all 33 of them are found with the second constituent *-ham*. The full list of words is given in (21).

(21) *Allingham, Bellingham, Billingham, Bingham, Birmingham, Buckingham, Collingham, Coningham, Cuningham, Cunningham, Effingham, Erpingham, Etchingham, Fotheringham, Framingham, Framlingham, Gillingham, Hoveringham, Immingham, Ingham, Kingham, Langham, Manningham, Nottingham, Rockingham, Sandringham, Sheringham, Uppingham, Walsingham, Wokingham, Woldingham, Wolsingham, Worlingham*

Note that such unusual occurrences of /ŋ/ are found more often in words whose first constituent is bound (58%), although words with a bound first constituent are a minority (39%).

5.2.4. Internal superheavy syllables

A superheavy syllable can be defined as a syllable which contains at least either a long vowel and a coda or a short vowel and a branching coda. This is based on the widely accepted idea that, in English, short vowels contribute one mora (= one weight-bearing unit), long vowels contribute two moras and a coda consonant contributes one mora (Hayes 1989; Hyman 1985; McCarthy and Prince 1996).[19] An issue is then that of the syllabification of medial consonant clusters. Following common assumptions, I followed the principle of Maximal Onset, which states that as many consonants as possible should be syllabified in the onset, and the Law of Initials (Vennemann 1988), which states that word-medial onsets should resemble word-initial onsets. Therefore, complex medial clusters were split according to these principles, even if the remaining rime violates what Venneman (1988) calls the Law of Finals, which states that word-medial rimes should resemble word-final rimes. If those two laws were to be applied strictly, many words would have to be analysed as having unsyllabifiable medial sequences. In itself, that difficulty informs us on the peculiar phonotactics of the words in the dataset.

With that coding adopted, 753 words have internal superheavy syllables, and that represents over a fifth of the dataset. The different types of rimes found in the data are shown in Table 6.

Rime structure	Number of words	Examples
VCC	209	*Astley* /ˈ**ast**.lɪj/, *Lansdown* /ˈ**lanz**.dawn/, *Melksham* /ˈ**mɛlk**.ʃəm/
VCCC	11	*Chelmsford* /ˈ**tʃɛlmz**.fəd/, *Helmsdale* /ˈ**hɛlmz**.dɛjl/, *Huntsville* /ˈ**hənts**.vɪl/
VVC	427	*Nailsea* /ˈ**nɛjl**.sɪj/, *Townsend* /ˈ**tawn**.zɛnd/, *Wheatstone* /ˈ**wɪjt**.stən/
VVCC	102	*Colnbrook* /ˈ**kəwl**n.bɹʊk/, *Knightsbridge* /ˈ**nɑjts**.bɹɪdʒ/, *Mountford* /ˈ**mawnt**.fəd/
VVCCC	3	*Barnoldswick* /bɑːˈ**nəwldz**.wɪk/, Saintsbury /ˈ**sɛjnts**.bəɹɪj/, *Shaftesbury* /ˈ**ʃɑ**:**fts**.bəɹɪj/
VCC/VVC	1	*Alnmouth* /ˈ**aln**.mawθ/ ~ /ˈ**ɛjl**.mawθ/

Table 6. The rimes found in words with superheavy syllables (the assumed syllable boundaries are indicated with a dot)

In many cases, what contributes to the weight of the syllable is that the first constituent ends in an <s> (sometimes an etymological genitive). Such cases represent 26% of all the words with superheavy syllables (203/753), a figure that goes up to 51% among words which have at least two coda consonants (166/326) and 63% among words which have a rime with four moras (73/116).

We saw in §3.2 that certain types of superheavy syllables are attested in simplex words and therefore cannot be taken as evidence for complexity (see (7)). These are only of the VVC type, and I will assume all other types to be illegal word-internal superheavy syllables. Among VVC syllables, 68 have a coda consonant that is not a fricative or a sonorant (e.g. *Boardman, Coupland, Yorkshire*), 19 have a sonorant coda that is non-coronal (e.g. *Bloomberg, Farnley, Wormwood*) and 119 have a sonorant coda that is not followed by a homorganic consonant (e.g. *Bournemouth, Queenborough, Soulbury*), and sometimes the last two characteristics occur at the same time (e.g. *Armley, Bloomfield, Urmston*). That leaves 232 words for which the structure is consistent with (7), and so for which we cannot consider that the presence of a superheavy syllable is evidence for complexity. If we add up the 195 illegal superheavy syllables in VVC to the other types of superheavy syllables, we get a total of 521 words which contain an illegal superheavy syllable.

As for previously examined characteristics, the presence of internal superheavy syllables correlates with the free nature of the two constituents. As can be seen in Figure 3, the highest proportion of words with an internal superheavy syllable is found in words in which the two constituents are free (23%) while the lowest proportion is found in words in which the two constituents are bound (13%).[20]

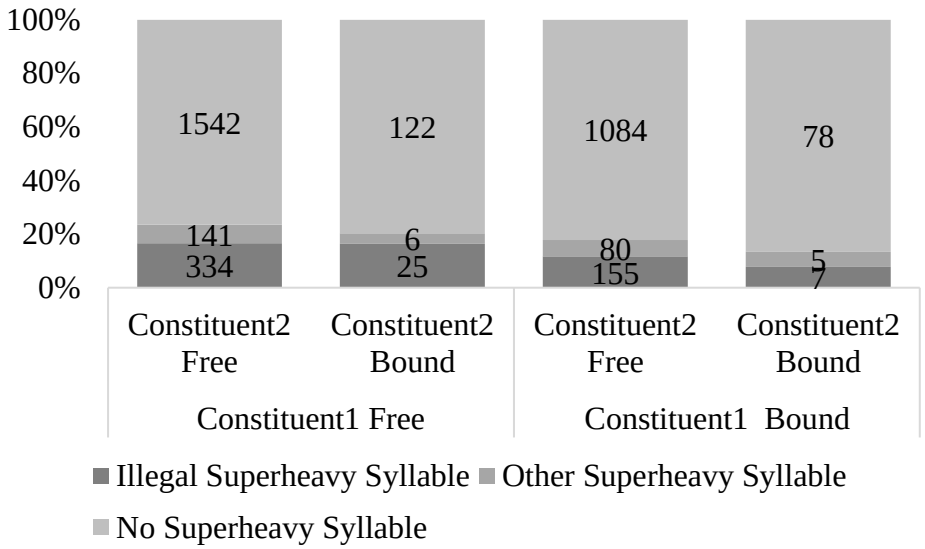


Figure 3. Presence of an internal superheavy syllable depending on the free or bound status of the two constituents

5.3. Trisyllabic Shortening

To identify the words which are susceptible to obey Trisyllabic Shortening, I adopted a restrictive view of that rule by eliminating all potential influences from orthography, as vowels spelled with digraphs generally have long vowels, and monographs followed by two identical orthographic consonants generally have short vowels (Dabouis 2023; Deschamps 1994; Fournier 2010b). Words with orthographic <u> are also known not to obey this rule and were not considered. Therefore, I looked for stressed vowel monographs other than <u> which are followed by a single consonant and are at least three syllables from the end of the word.

The dataset contains 252 words which fit these conditions, and 176 have short vowels (22a), 65 do not (22b) and 11 vary (22c).

(22) a. /a/*bingdon, B*/ɛ/*veridge, D*/ɛ/*lacourt, L*/ɪ/*verpool, Ph*/ɪ/*lipson, R*/ə/*therfield…*
b. *Bl*/ɛj/*kenham, D*/ɛj/*vidson,* /ɪj/*denbridge,* /əw/*diham, P*/ɪj/*tersham …*
c. *B*/ɛ ~ ɪj/*contree, Dr*/ɑː ~ a/*kensberg, Mash*/ɔ ~ əw/*naland, S*/ɑj ~ ɪ/*mington …*

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

Therefore, about a third of relevant words may have a long vowel. This is clearly above the 8% reported by Deschamps (1994). However, as for other characteristics examined thus far, it is possible that the presence of a long vowel is attributable to the fact that the first element is perceived to be related to a freestanding word which itself has a long vowel. Therefore, I tested whether the free or bound nature of the first constituent correlates with a significant difference in the vowels found in this environment. If we regroup words with long vowels with those which show variation and test the difference with those with short vowels only, the difference between free and bound first constituents is statistically significant ($\chi^2 = 13.734$, df = 1, $p < .001$). The distribution of the data is shown in Figure 3.[21]

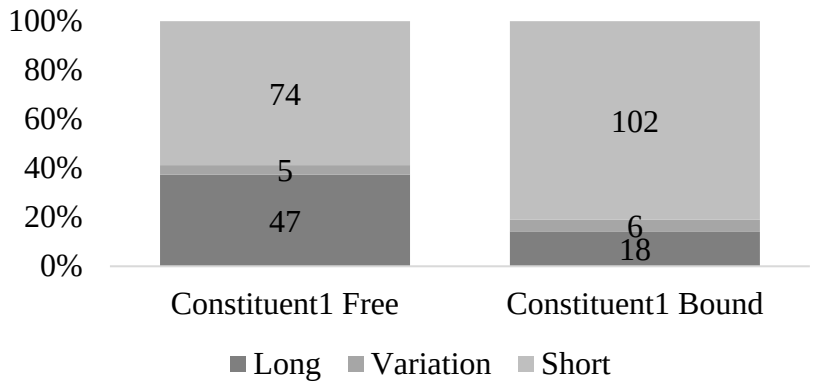


Figure 3. Vowels found in the environment for Trisyllabic Shortening depending of the free or bound nature of the first constituent

As can be seen in Figure 3, there is a clear difference between words for which the first constituent is free and those for which it is bound: if the first constituent is free (e.g. *Adamson* ↔ *Adam*, *Davidson* ↔ *David*, *Maryland* ↔ *Mary*), then there is a higher proportion of words which may have a long vowel (41%) than if it is bound (19%). Even when the words with a free first constituent are taken out, the proportion remains over twice as high as that reported by Deschamps (1994). Both rates can be interpreted as evidence of morphological complexity. In words in which the first constituent is free, the rate is five times that observed in simplex words, which constitutes stronger evidence for complexity, especially considering the source of the difference: those results once again suggest that words with a free first constituent are indeed related to the corresponding freestanding words.

5.4. Vowel reduction in the final constituent

Vowel reduction is a complex and multifactorial process (see Dabouis and Fournier submitted a) and, as it is not the main focus of this paper, the analyses presented here will probably be better seen as exploratory and will have to be dealt with in more depth in future research. Following Szigetvári (2018), I will assume that /ə, ɪ, ʊ/ and the corresponding three diphthongs /əw, ɪj, ʉw/ may be reduced (which he calls "unstressed"). As all of those vowels may also be full vowels, it will be assumed that they are reduced only if it is /ɪ/ representing an orthographic vowel that is not <i> or if they are represented in Wells (2008) as /ə/ or /i/, which can be taken to represent reduced (or "stressless", in Szigetvári's terms) occurrences of STRUT and FLEECE, respectively. Therefore, cases such as *-gate* /-gɪt/, *-ledge* /-lɪdʒ/ or *-ness* /-nɪs/ are treated as reduced. Cases where <i> or <y> are realised as /ɪ/ are treated as uninterpretable.

A first striking observation is that many second constituents have a fixed behaviour: their vowel is either always full, always reduced or it displays systematic variation. This is shown in Table 7.

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

Behaviour	Number of types	Number of words	Examples
Uninterpretable	17	327	*-bridge, -cliffe, -gill, -hill, -kin, -kins, -ling, -mills, -minster, -ridge, -smith, -ville, -wich, -wick, -win, -wych, -wyn*
Systematically reduced	23	955	*-berry, -burgh, -borough, -bury, -by, -folk, -ford, -ley, -sham, -son, -stable, -staple, -try, -wark …*
Systematically full	83	738	*-bolt, -bourn, -castle, -church, -dale, -field, -fleet, -kirk, -port, -shaw, -side, -town, -water, -wood …*
Systematic variation	1	36	*-shire*
Instability	46	1523	*-bell, -chester, -cote, -gate, -haven, -land, -ledge, -man, -more, -mouth, -stone, -wall, -wold …*

Table 7. Vowel reduction behaviour of different second constituents

However, as can be seen in Table 7, there are 46 final constituents whose behaviour is 'unstable', i.e. from one word to another, they can be full or reduced. However, that covers a range of highly different behaviours, as can be seen in Figure 4, which shows the behaviour of the 27 second constituents which appear in at least ten different words.

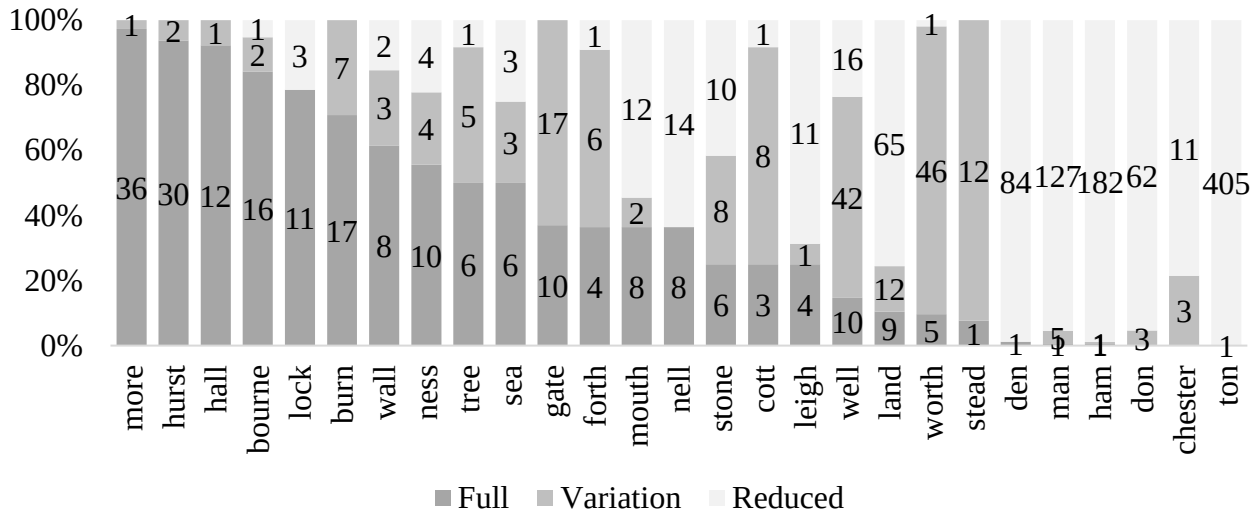


Figure 4. Reduction behaviour in the 27 second constituents which appear in at least ten different words

In order to explore what the determining factors of reduction might be, I excluded the uninterpretable cases and coded reduction as a binary variable (FULL vs REDUCED) using the main pronunciation. We saw the different reduction-blocking contexts for the final syllable in (12), and so words were coded as having or not having such a context. Words whose second constituent is disyllabic have to be taken out, which leaves 3014 words. Another factor which could be important is the free or bound nature of the constituent, as we saw that it was relevant for certain properties of the first constituent. Finally, Fidelholtz (1975) found that more frequent words tend to show more reduction than less frequent items, and this has been confirmed in recent studies of vowel reduction (Dabouis *et al.* 2020; Dabouis and Fournier submitted a). Here, this will be tested through the recurrence of constituents in the dataset. I ran a binary logistic regression using all three factors, and

only the recurrence of the second constituent and the presence of a reduction-blocking context were found to be significant predictors of vowel reduction in the second constituent. The results of the regression analysis are shown in Table 8.

Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.33	0.26 – 0.40	**<0.001**
RECURCONSTITUENT2	1.03	1.03 – 1.04	**<0.001**
BLOCKREDCTXT-YES	0.09	0.07 – 0.11	**<0.001**
AIC	1611.486		

Table 8. Binary logistic regression analysis for the presence of reduced vowels in the second constituent

Reduced vowels are more common if there is no reduction-blocking context, and if the constituent occurs in many different words.[22] The difference between items which have a reduction-blocking context and those which do not is shown in Figure 5.

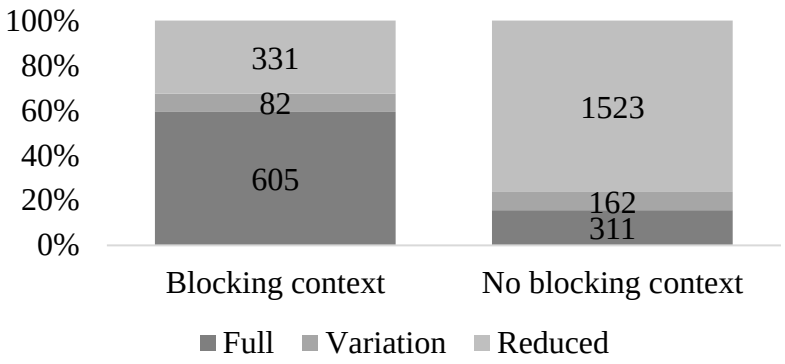


Figure 5. Vowel reduction in the second constituent depending on the presence of a reduction-blocking context

One limitation to these results is that the factors considered are characteristics of constituents, not of whole words, and so we cannot account for the kind of intra-constituent variation represented in Figure 5 at this stage. Future research may include word frequencies into account, although, in the case of place-names, there may exist strong inter-speaker differences in frequencies as the name may have a high frequency only for local speakers. Therefore, it remains to be seen whether or not a global frequency measure can be relevant to account for reduction in place-names. In any case, Wells (2008) gives several local pronunciations which indeed suggest that a high local frequency may lead to further reductions, as shown by the examples in (23). Note that those pronunciations are only introduced by "locally also", which cannot inform us on how widespread that pronunciation is locally.

(23)	**Place-name**	**Main pronunciation**	**Local pronunciation**
	Cudworth	/ˈkədwəθ/ ~ /ˈkədwəːθ/	/ˈkədəθ/
	Hunstanton	/hənˈstantən/	/ˈhənstən/
	Rothersthorpe	/ˈɹɔðəzθoːp/	/ˈɹɔðəzθɹəp/
	Rothwell	/ˈɹɔθwɛl/ ~ /ˈɹɔθwəl/	/ˈɹəwəl/
	Rutherglen	/ˈɹəðəglɛn/	/ˈɹəðəglən/

As for whether reduction provides any cues to morphological complexity, the results presented here are inconclusive. We saw in §3.2 that a suffix tends to reduce while the second

constituent of a compound tends not to. That variability between different sorts of complex structures and the fact that there are clear effects of the same reduction-blocking contexts as those observed in simplex words make it impossible to draw any conclusions based on reduction alone. However, given the results of the previous sections, the reduction of the second constituent may be taken as an indicator of whether those words, if they are analysed as complex words, should be analysed as compounds or as suffixed words.

6. Discussion

6.1. Interpretation of the results

We have seen that the phonological behaviour of the names in the dataset is overall more consistent with the assumption that these words are complex, even though we have seen there is variability in the data. Assuming that all characteristics except vowel reduction provide evidence of morphological complexity, one question may be that of the proportion of the words in the dataset which do not have at least one of these characteristics. If we exclude words which:

- do not have penultimate stress even if they have a closed penult,
- may have preantepenultimate stress,
- have an illegal medial consonant cluster,
- may have a preconsonantal stressless /ɪj/,
- have a prevocalic /ŋ/,
- have an illegal superheavy syllable,
- have a stressed vowel which does not obey Trisyllabic Shortening,

then, we are left with 1697 words. That represents a bit less than half of the data.[23] As there is variation among simplex words for some of those characteristics, having one such characteristic pointing to morphologically complexity may not necessarily entail that most speakers will assume those words to be morphologically complex. It is possible that several are needed for complexity to be perceivable, or that some characteristics are more significant than others.[24] Conversely, having none of those characteristics may not necessarily entail that such words will not be perceived as complex. For instance, many of those 1697 words contain second constituents that occur in dozens or even hundreds of words, such as *-ton, -ley, -son, -ham* or *-ford*. Distributional recurrence has been suggested to be a key factor in morphological recognition for opaque prefixed words, as discussed in §2 (Dabouis and Fournier submitted b; Forster and Azuma 2000; Fournier 1996; Taft 1994), and Köhnlein (2015) develops a similar idea for place-names in Dutch:

> [T]he repeated occurrence of toponym-endings like *-drecht, -dam*, or *-en*, in combination with the accompanying phonological characteristics, leads the learner to postulate a corresponding morpheme.

Proper names therefore form a network of various combinations of first and second elements, the densest part of which could look as in Table 9, although final constituents are the most recurrent and therefore probably play a greater role in the recognition of the internal structure of such names, even when they are not ‘free’.[25]

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

	-bury	-field	-ham	-ley	-ford	-well	-wood
ash-		*Ashfield*			*Ashford*	*Ashwell*	
black-				*Blackley*	*Blackford*	*Blackwell*	*Blackwood*
brad-	*Bradbury*	*Bradfield*		*Bradley*	*Bradford*	*Bradwell*	
bur-	*Burbury*			*Burley*	*Burford*	*Burwell*	
fair-		*Fairfield*		*Fairley*	*Fairford*		
green-		*Greenfield*	*Greenham*		*Greenford*	*Greenwell*	*Greenwood*
han-	*Hanbury*			*Hanley*		*Hanwell*	
har-				*Harley*	*Harford*	*Harwell*	*Harwood*
kings-	*Kingsbury*			*Kingsley*	*Kingsford*		*Kingswood*
mar-			*Marham*	*Marley*			
new-	*Newbury*		*Newham*				
south-			*Southham*			*Southwell*	
stan-	*Stanbury*	*Stanfield*		*Stanley*	*Stanford*	*Stanwell*	
stock-				*Stockley*		*Stockwell*	*Stockwood*
wal-			*Walham*		*Walford*		
water-	*Waterbury*						
west-	*Westbury*	*Westfield*					*Westwood*

Table 9. A part of the network of first and second constituents found in proper names

The existence of this network, and especially the recurrence of final constituents is probably sufficient for English language users to assume word-internal complexity in such names. If that is the case, then one might wonder what type of structure should be posited. As we have seen in §3.2, only the reduction of the final constituent may tease apart compounds from words with neutral suffixes, although certain suffixes do not reduce, and certain final constituents of compounds do. If we take a simplifying position, we might assume that names whose final constituent is reduced are interpreted as suffixed while those for which it is not are interpreted as compounds. Then, one question that arises is how constituents which show variable reduction (such as those seen in Figure 4) should be interpreted. One possible answer is that the source of the variation lies in different interpretations of those constituents: they are sometimes interpreted as suffixes and sometimes as stems. As certain elements show systematic diatopic differences, it is possible that different communities perceive certain elements differently (e.g. *-ham* is perceived as a suffix by British language users but as a stem by American language users).

We found for four different characteristics that there is a difference between words for which the first constituent has been categorised as ‘free’ and those categorised as ‘bound’. This suggests that, even if there is no clear semantic relationship between the constituent and the associated freestanding word (e.g. *open* and *Openshaw* (suburb in Manchester)), a link between them appears to be made. That sort of relationship is reminiscent of something that is also found in compounds, and that Jackendoff (2010) calls “strawberry morphemes”. These are “real words within compounds that play no role in the compound’s meaning” (e.g. *straw*berry, *cottage cheese*, *horse*radish, *side*kick, *airplane*). If some constituents are “strawberry morphemes”, then how should the others be analysed? Those which occur in a single word may be analysed as “cranberry morphemes” (Aronoff 1976: 10): in a word such as *cranberry*, which is a kind of berry, the constituent *cran* is never found elsewhere and has no identifiable meaning.[26] 39% of the first constituents in the dataset are bound and occur only once and could be analysed in that way. However, certain elements appear to form another category, as they occur in several forms (e.g. *somer-* has no freestanding counterpart, and yet it appears in *Somerfield*, *Somerleyton*, *Somerset*, *Somerton* and *Somerville*). In any case, all of these are analysable as roots, except final constituents which undergo reduction which may be analysed as suffixes.

We saw that the reduction of the vowel of the second constituent is related to the number of forms in which it appears, which is consistent with previous studies on vowel reduction. That result could suggest a possible path for the emergence of suffixes. It has been known for a long time that certain suffixes were originally independent words, as evoked for example by Marchand (1969: 210) regarding *-dom* and *-hood*, which "are still independent words in Old English". Recent empirical work on compounding has shown that nominal compounds are less likely to have their main prominence on the second constituent if that constituent is not informative (Bell and Plag 2012, 2013). Thus, the second constituent is deaccented if it is not very informative. Maybe an additional development could be that, if that constituent becomes more productive and is used in dozens of words, then it will be put under an increasing pressure to undergo further reductions as its frequency increases (Fidelholtz 1975; Clopper and Turnbull 2018; Bell et al. 2009). If it does indeed reduce, it may end up being interpreted as a suffix.

Before we turn to a more formal analysis, let us discuss one possible objection regarding our interpretation of the results on the relationship between a closed penult and the position of primary stress. Although this relationship is widely accepted in the generative literature, it is not so in the Guierrian tradition. Crucially, Fournier (2010a) observes that different parts of the vocabulary behave differently with regards to the position of stress in words that have a closed penult. He finds that penultimate is almost systematic only in words that are borrowed from Latinate languages such as Italian, Spanish or Portuguese (e.g. *ànacónda, extràvagánza, dìlettánte, concérto*) or from Modern Latin (e.g. *pròpagánda, enígma, meménto, alúmnus*). In the rest of the vocabulary, only a third of the relevant words have penultimate stress. Based on this observation and many others that relate to the different phonological, morphological, graphophonological and semantic properties of certain subsets of the lexicon, Dabouis and Fournier (2022) propose that English phonology is divided into different subsystems. They posit four main subsystems, §Core, §French, §Foreign and §Learned, each with its own set of specific properties, even though certain properties are shared by several or all subsystems. In their model, penultimate stress is only the rule among words with a closed penult if they belong to either §Foreign or §Learned, i.e. words that are perceived as belonging to a foreign language other than French, or as technical or scientific vocabulary. In that perspective, it could be argued that our failure to find any effects of closed penults on stress has to do with the fact that the names in the dataset do not belong to those categories of words and would be associated to the §Core subsystem. However, just as we have seen for Trisyllabic Shortening, the difference may be a probabilistic one. Indeed, we observed rates of antepenultimate stress of about 90%, while according to Fournier's (2010a) data, antepenultimate primary stress is almost never found in §Foreign and §Learned words with a closed penult, and it is found in about two thirds of simplex §Core words. Therefore, the rate of antepenultimate stress that was found in proper names is about 30% higher than what he reports for those words, and so the difference can be considered significant.

Finally, let us mention possible limitations of the present study. First, as the search for proper names was manual, some items or minor classes may have been missed. Second, we may have included words which are not etymologically complex (e.g. *London* ← Lat. *Londinium* and not < *Lon-* '?' + *-don* 'hill'). I believe that these drawbacks are not major issues for the study presented in the following sections. Indeed, the size of the dataset is large enough, so we can be confident that the generalisations that are found in it will hold should they be tested in a larger dataset. As for the inclusion of items that are not etymologically complex, we can perfectly assume that speakers that have no knowledge of their etymology analyse them as having structures that are not consistent with their etymology. As I have just argued, the recurrence of the constituents, in association with semantic and phonological properties is quite likely what makes them learnable, and so this actually would predict that such misanalyses should occur. As for the type of data used, dictionary data allows for a first study of large numbers of words, but cannot be taken to represent all the attested pronunciations of words. Thus, future studies using oral data to consolidate or challenge the present findings would be welcome.

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

6.2. Phonological and lexical representations for English proper names

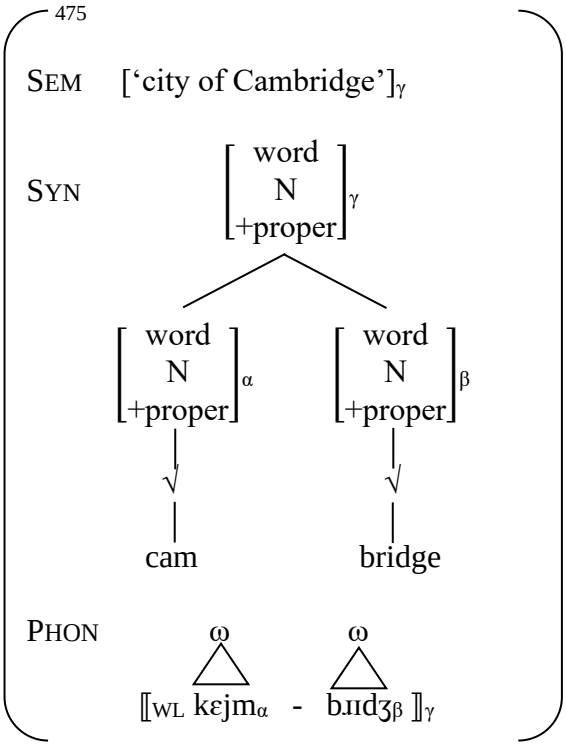
Most models of the morphology-phonology interface assume that there are phonological domains, although their nature can be either procedural (i.e. different sizes of chunks that undergo phonological computation) or representational (i.e. there are phonological constituents that partly reflect morphological structure). For example, cyclic models assume that the phonology applies to the smallest constituent of a complex expression and then it applies again at different nodes of the complex structure. Each node at which phonology applies may be called a cyclic domain. For example, Kaye (1995) assumes that, in a compound, phonology is applied to both constituents, and then to the whole compound (e.g. [[*black*][*board*]], where square brackets represent phonological domains in his analysis), or that neutral suffixes do not trigger a new application of phonology to the word (e.g. [[*dream*]*s*]).[27] Another approach which does not necessarily assumes a sequential application of phonological processes is to posit that certain morphological constituents are mapped onto phonological units, which form phonological domains. This is the approach developed in Prosodic Phonology (Nespor and Vogel 1986; Selkirk 1980). In that approach, structures similar to those posited by Kaye (1995) are assumed, but the domains that those morphological constituents map onto are assumed to be representational units, called 'prosodic' or 'phonological' words (represented using ω). Thus, a compound is commonly assumed to have the structure $((X)_{\omega^\circ}(Y)_{\omega^\circ})_{\omega'}$ (this notation is based on the assumption that phonological words can be recursive, and so the minimal phonological word projection is noted ω° while the higher projection is noted ω'; see for example Bermúdez-Otero 2011). The prosodic structure of words with neutral suffixes is more controversial. Some assume that neutral suffixes are prosodified as $((X)_{\omega^\circ}(Y)_{\sigma})_{\omega'}$, i.e. the syllable of the suffix (noted σ) attaches to a higher phonological word projection (Hammond 1999: 322-329; Szpyra 1989: 178-200), while Bermúdez-Otero (2011) assumes that they attach directly to the phonological word. Raffelsiefen (2005) makes a different proposal by claiming that there are different sources of stress-neutrality for affixes. She assumes that vowel-initial suffixes are cohering and so fuse into a single phonological word with the base to which they are attached. Consonant-initial suffixes are assumed to be non-cohering and therefore do not integrate the phonological word of the base. The source of the stress-neutrality found for some cohering suffixes is procedural in nature. It is what she calls "Paradigm Uniformity effects", which arise on an affix-specific basis. It means that affixes are assumed to be more likely than others to preserve the phonological properties of their base, stress included, and so 'stress-neutral' patterns may arise in that way. Raffelsiefen (2007) also posits that productive affixes may be recognised even when they attach to a bound root (e.g. *gormless, grateful ointment*) and that they can have the same prosodic structure as words with the same affix attached to a free stem. However, certain non-productive consonant-initial suffixes may fuse with their base, "but still deviate from canonical phonological form" (e.g. *loathsome* with /ðs/). Such deviations are then assumed to be Paradigm Uniformity effects and thus need to be licensed by a base word (e.g. in *loathsome*, /ð/ is protected from assimilation through faithfulness to *loathe*). However, in the dataset, we saw that a third of the dataset contains an illegal cluster, and a third of the words with such clusters have a bound first constituent (e.g. *Abi*/ŋd/*on*, *Hamer*/zl/*ey*, *Rotha*/mst/*ed*). Therefore, we cannot use phonological identity with an existing form to license those clusters, and so the domain solution seems preferable. As the most commonly adopted solution is that of Prosodic Phonology, I will refer to those domains as phonological words in what follows.

We can take up Raffelsiefen's idea that structures with bound constituents may be prosodically complex, if their constituents are recognised. However, I depart from her analysis as I do not assume that recognition depends on productivity in proper names but, as discussed in the previous section, on the distributional recurrence of name constituents and the identification of 'anomalous' phonological characteristics. As for the lexical representation of such names, we can assume, in line with the analyses described in §2.2, that the entries for complex names are analytically listed in the

lexicon (Bermúdez-Otero 2012). We saw that a difference between Köhnlein (2015) and Mascaró (2016) is that Köhnlein assumes that name constituents have some meaning (although they are underspecified) while Mascaró does not assume any complexity at the semantic level. It is possible that both options are actually used by speakers.

I suggest that, when the learner is faced with a name with clear phonological complexity, they can assume that it is complex morphologically but not semantically. The identified word constituents are coindexed between the phonological level and the morphological level, but there is no coindexation with semantic units that are smaller than the whole word. At this stage, the lexical entry resembles what Mascaró posits, as shown in (24) for the name *Cambridge* (the subscript Greek letters represent coindexation, and the number $_{475}$ is arbitrarily chosen to represent the index of the entry). Note that (24) does not show the whole prosodic structure that the word would eventually have, under the assumption that, within Stratal Phonology (Bermúdez-Otero 2012, 2018), this is a lexical entry specified to go through the word-level phonology (shown as $_{WL}$), i.e. a concatenation of several phonological strings.

(24) **Lexical entry for *Cambridge***

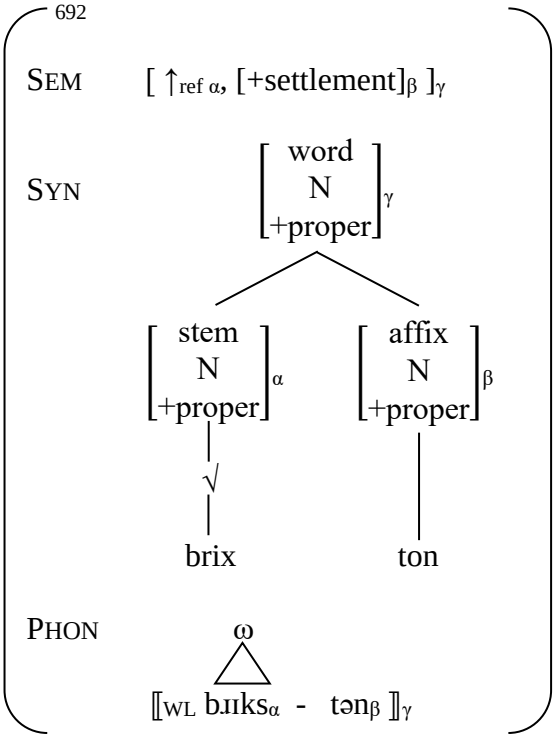


However, when faced with various city names in *-bridge*, the learner may then posit that *-bridge* is a bound root which functions as a classifier and marks the word as a settlement name. Following Köhnlein's (2015) proposal, we can posit that the first constituent is then assumed to have as its only semantic content a 'referential pointer' to a unique object in the world (represented as $\uparrow_{ref}$). Therefore, the lexical entry for *Cambridge* shown in (24) would be restructured so that its semantic representation looks like [$\uparrow_{ref\ \alpha}$, [+settlement]$_{\beta}$]$_{\gamma}$, i.e. so that there are parts of its semantic representation which are coindexed to other levels of representation.

As for final constituents whose vowel is reduced, the difference may be taken to manifest at the morphological level, where that constituent is identified as a suffix, and at the phonological level, where that constituent does not have its own phonological word. After, the identification of the final constituent as a toponymic or patronymic suffix, a lexical entry would look as in (25).

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

(25) **Lexical entry for *Brixton***



Finally, we saw in §5.2 that certain words have exceptional word-medial clusters that cannot be syllabified as a coda and an onset (even assuming that both are to be treated as being at word edges), in cases in which there is a medial <s>, realised as /s/ or /z/. We also saw in §4.2 that certain constituents are internally complex. In those cases, it is necessary to assume more layers of internal structure on the morphological and phonological level (e.g. *Ronaldsway* (((rɔnəld)$_{ω∘}$z)$_{ω'}$(wɛj)$_{ω∘}$)$_{ω''}$; *Bedfordshire* (((bɛd)$_{ω∘}$fəd)$_{ω'}$ʃə))$_{ω''}$), assuming that those consonants are treated like a form of neutral affix.

With those lexical representations, we can account for the phonological behaviours observed in the dataset, which I interpret as sign of morphological complexity, without necessarily assuming that the semantic representation needs to refer to the semantics of the internal constituents when they are free, or even complex at all. We also saw that close to half of the words in the dataset do not present any phonological signs of morphological complexity (at least based on the characteristics investigated in this paper). It is quite likely that, if a learner has not been exposed to many different names that contain the same final constituent, they will not posit lexical representations of the kind of (24-25) and will assume simplex structures at all levels of representation.

7. Conclusions

In this paper, I have presented the first large-scale study on the phonological characteristics of complex proper names in English. The findings overall point to morphological complexity in those words, even in the absence of clearly identifiable semantics in the constituents of names, although we saw that there is variability in the data, and I have suggested that some words in the data may be interpreted as truly simplex. I have taken over analyses from Köhnlein (2015) and Mascaró (2016) regarding how such names should be represented in the lexicon using analytic listing and coindexation. Crucially, in the proposed analysis, the subparts of a lexical entry need not correlate at all levels of representation.

We can identify three significant contributions made by this paper. First, in several of the phonological characteristics studied in this paper, we found that items for which the first constituent

has a homographic freestanding word ('free' constituents) behave differently from those for which there is no such homograph ('bound' constituents). This suggests that there can exist morphological relationships in the absence of a clear semantic relationship. Such an observation is actually not really new, as such relationships have been found among bound roots in opaque prefixed words (see §2.1), in irregularly inflected words (Aronoff 1976: 14; e.g. *understand - understood* cp. *stand - stood*) and in compounds with "strawberry morphemes" (see §6.1).[28] Allen (1980) also says that *-man* (e.g. in *chairman*), that she analyses as a suffix "must still be 'linked' with the lexical word *man* in such a way as to retain all the morphological irregularities associated with the lexical word *man*, in particular, its irregular plural" (*chairman* is inflected as *chairmen* and not **chairmans*). Second, based on the observation that more recurrent second constituents in proper names more often have a reduced vowel than less recurrent ones, and on the assumption that final constituents with a full vowel are analysed as roots while those with a reduced vowel are analysed as suffixes, I have proposed a possible path on how suffixes are created from independent words: such constituents would be deaccented and reduced as their informativity decreases and their productivity increases. Finally, contrary to what Raffelsiefen (2007) assumes, I proposed that the proper names studied in this paper provide evidence that it is possible for a word formed from non-productive bound constituents to have a complex prosodic structure.

However, a number of questions remain open for further research on English complex names. Let us mention a few:

- Can some of the phonological characteristics be related to word frequency? As we have seen in (23), it looks like there can be local frequency effects for place-names, leading to greater reductions (and maybe to what Raffelsiefen (2007) calls "High Frequency Fusion"), and so they may become more similar to simplex words.
- Can we identify effects of headedness in proper names? If the compound is left-headed or right-headed, are there phonological differences?
- Can graphotactics be used as an additional indicator of complexity? For example, there can be word-internal <VCe> structures in which V is realised as a diphthong (e.g. *Gracechurch, Lakeland, Roseberry, Bateson, Pateley*) just as it would be at the end of a word (cp. *rat ~ rate, sit ~ site, not ~ note*).[29] There may also be unusual sequences of letters that function as orthographic boundary signals, even if there is nothing unusual phonologically.
- Can we find psycholinguistic evidence of morphological relatedness in lexical decision tasks using masked priming, as is the case for opaque prefixed words (Forster and Azuma 2000; Pastizzo and Feldman 2004)?
- Can we find phonetic evidence regarding how the segments of each constituent syllabify of the kind discussed by Raffelsiefen (2005)? For example, do intervocalic consonants behave differently if they are constituent-initial or final?
- I have argued, in line with previous works, that the recurrence of elements is what makes them learnable, but how can we know how many occurrences are enough? And are initial or final elements more important for the identification of morphological structure in names?

Hopefully, this paper will have paved the way for research on the phonological characteristics of English proper names, and much remains to be done.

References

Allen, Margaret R. "Semantic and Phonological Consequences of Boundaries: A Morphological Analysis of Compounds." *Juncture: A Collection of Original Papers*. Ed. Mark Aronoff and Mary-Louise Kean. Saratoga: Anma Libri, 1980. 9–27.

Arndt-Lappe, Sabine, and Quentin Dabouis. "Secondary Stress and Morphological Structure: New Evidence from Dictionary and Speech Data." Submitted.

Aronoff, Mark. *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press, 1976.

Aronoff, Mark, and Andrea D. Sims. "The Relational Nature of Morphology." *Linguistic Morphology*

in the Mind and Brain. Ed. David Crepaldi. Abingdon & New York: Routledge, 2023. 7–25.

Bell, Alan et al. "Predictability Effects on Durations of Content and Function Words in Conversational English." *Journal of Memory and Language* 60 (2009): 92–111.

Bell, Melanie, and Ingo Plag. "Informativeness Is a Determinant of Compound Stress in English." *Journal of Linguistics* 48 (2012): 485–520.

Bell, Melanie, and Ingo Plag. "Informativity and Analogy in English Compound Stress." *Word Structure* 6.2 (2013): 129–155.

Bermúdez-Otero, Ricardo. "Cyclicity." *The Blackwell Companion to Phonology (Vol. 4: Phonological Interfaces)*. Ed. Marc van Oostendorp et al. Malden, MA: Wiley-Blackwell, 2011. 2019–2048.

Bermúdez-Otero, Ricardo. "Stratal Phonology." *The Routledge Handbook of Phonological Theory*. Ed. S J Hannahs and Anna R K Bosch. Abingdon, OX: Routledge, 2018. 100–134.

Bermúdez-Otero, Ricardo. "The Architecture of Grammar and the Division of Labour in Exponence." *The Phonology and Morphology of Exponence - the State of the Art*. Ed. Jochen Trommer. Oxford: OUP, 2012. 8–83.

Burzio, Luigi. *Principles of English Stress*. New York: Cambridge University Press, 1994.

Chomsky, Noam, and Morris Halle. *The Sound Pattern of English*. New York: Harper & Row, 1968.

Clopper, C. G., and Rory Turnbull. *Exploring Variation in Phonetic Reduction: Linguistic, Social, and Cognitive Factors*. Ed. Francesco Cangemi et al. Berlin, New York: de Gruyter Mouton, 2018.

Cruttenden, Alan. *Gimson's Pronunciation of English*. 8th editio. Oxon & New York: Routledge, 2014.

Dabouis, Quentin. "English Phonology and the Literate Speaker: Some Implications for Lexical Stress." *New Perspectives on English Word Stress*. Ed. Nicolas Ballier et al. Edinburgh: Edinburgh University Press 2023. 117–153.

Dabouis, Quentin, Jean-Michel Fournier, et al. "English Word Stress and the Guierrian School." *New Perspectives on English Word Stress*. Ed. Nicolas Ballier et al. Edinburgh: Edinburgh University Press, 2023. 53–82.

Dabouis, Quentin, Guillaume Enguehard, et al. "The English 'Arab Rule' without Feet." *Acta Linguistica Academica* 1.67 (2020): 121–134.

Dabouis, Quentin, and Jean-Michel Fournier. "An Empirical Study of Vowel Reduction and Preservation in British English". Submitted a.

Dabouis, Quentin, and Jean-Michel Fournier. "Opaque Morphology and Phonology: Historical Prefixes in English". *Journal of Linguistics* (2024).

Dabouis, Quentin, and Jean-Michel Fournier. "The Stress Patterns of English Verbs: Syllable Weight or Morphology?" *New Perspectives on English Word Stress*. Ed. Nicolas Ballier et al. Edinburgh: Edinburgh University Press, 2023. 154–191.

Dabouis, Quentin, and Pierre Fournier. "English PhonologiES?" *Modèles et Modélisation En Linguistique / Models and Modelisation in Linguistics*. Ed. Viviane Arigne and Christiane Rocq-Migette. Bruxelles, Belgique: Peter Lang, 2022. 215–258.

Deschamps, Alain. *De l'écrit à l'oral et de l'oral à l'écrit*. Paris: Ophrys, 1994.

Deschamps, Alain. *English Phonology and Graphophonemics*. Paris: Ophrys, 2004.

Fidelholtz, J. "Word Frequency and Vowel Reduction in English." *Chicago Linguistic Society* 11 (1975): 200–213.

Forster, Kenneth I., and Tamiko Azuma. "Masked Priming for Prefixed Words with Bound Stems: Does Submit Prime Permit?" *Language and Cognitive Processes* 15.4–5 (2000): 539–561.

Fournier, Jean-Michel. "Accentuation Lexicale et Poids Syllabique En Anglais : L'analyse Erronée de Chomsky et Halle." *Paper presented at the 8th meeting of the French Phonology Network held at the Université d'Orléans on 1–3 July* (2010).

Fournier, Jean-Michel. "From a Latin Syllable-Driven Stress System to a Romance versus Germanic

Morphology-Driven Dynamics: In Honour of Lionel Guierre." *Language Sciences* 29 (2007): 218–236.
Fournier, Jean-Michel. "La Reconnaissance Morphologique." *8ème Colloque d'Avril Sur l'anglais Oral*. Villetaneuse: Université de Paris-Nord, CELDA, diffusion APLV, 1996. 45–75.
Fournier, Jean-Michel. *Manuel d'anglais Oral*. Paris: Ophrys, 2010.
Giegerich, Heinz J. *Lexical Strata in English: Morphological Causes, Phonological Effects*. Cambridge: Cambridge University Press, 1999.
Guierre, Lionel. "Essai Sur l'accentuation En Anglais Contemporain : Eléments Pour Une Synthèse." Ph.D. dissertation, Université Paris-VII, 1979.
Guierre, Lionel. "Mots Composés Anglais et Agrégats Consonantiques." *5ème Colloque d'Avril Sur l'anglais Oral*. Villetaneuse: Université de Paris-Nord, CELDA, diffusion APLV, 1990. 59–72.
Halle, Morris, and Samuel Keyser. *English Stress: Its Form, Its Growth, and Its Role in Verse*. New York: Harper & Row, 1971.
Halle, Morris, and Karuvannur Puthanveettil Mohanan. "Segmental Phonology of Modern English." *Linguistic Inquiry* 16 (1985): 57–116.
Halle, Morris, and Jean-Roger Vergnaud. *An Essay on Stress*. Cambridge, MA: MIT, 1987.
Hammond, Michael. "Frequency, Cyclicity, and Optimality." 2003. URL: http://www.u.arizona.edu/~hammond/kslides.pdf
Hammond, Michael. *The Phonology of English: A Prosodic Optimality-Theoretic Approach*. Ed. Jacques Durand. Oxford: Oxford University Press, 1999.
Harris, John. *English Sound Structure*. Oxford: Blackwell, 1994.
Harris, John, and Edmund Gussmann. "Final Codas: Why the West Was Wrong." *Structure and Interpretation. Studies in Phonology*. Ed. Eugeniuc Cyran. Lublin: Folium, 1998. 139–162.
Hayes, Bruce. *A Metrical Theory of Stress Rules*. Ph.D. dissertation, Yale University: N.p., 1980.
Hayes, Bruce. "Compensatory Lengthening in Moraic Phonology." *Linguistic Inquiry* 20 (1989): 253–306.
Hayes, Bruce. "Extrametricality and English Stress." *Linguistic Inquiry* 13.2 (1982): 227–276.
Hempl, George. "The Stress of German and English Compound Geographical Names." *Modern Language Notes* 11.4 (1896).
Herment, Sophie. "The Pedagogical Implications of Variability in Transcription, the Case of [i] and [u]." *English Pronunciation : Issues and Practices (EPIP). Proceedings of the First International Conference*. Ed. Alice Henderson. Presses universitaires Savoie Mont Blanc, 2010. 177–188.
Hyman, Larry. *A Theory of Phonological Weight*. Dordrecht: Foris, 1985.
Jackendoff, Ray. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press, 2002.
Jackendoff, Ray. *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press, 1997.
Jackendoff, Ray. "The Ecology of English Noun-Noun Compounds." *Meaning and the Lexicon*. Oxford: Oxford University Press, 2010. 413–445.
Jones, Daniel. *Cambridge English Pronouncing Dictionary*. 17th ed. Cambridge: Cambridge University Press, 2006.
Kaye, Jonathan. "Derivations and Interfaces." *Frontiers of Phonology*. Ed. Jaques Durand and Francis Katamba. London & New York: Longman, 1995. 289–332.
Köhnlein, Bjorn. "The Morphological Structure of Complex Place Names: The Case of Dutch." *Journal of Comparative Germanic Linguistics* 3.18 (2015): 183–212.
Liberman, Mark, and Alan Prince. "On Stress and Linguistic Rhythm." *Linguistic inquiry* 8.2 (1977): 249–336.
Lindsey, Geoff. *English after RP: Standard British Pronunciation Today*. Cham: Palgrave Macmillan, 2019.
Marchand, Hans. "The Categories and Types of Present-Day English Word-Formation." 1969: 545.

Mascaró, Joan. "Morphological Exceptions to Vowel Reduction in Central Catalan and the Problem of the Missing Base." *Catalan Journal of Linguistics* 15 (2016): 27–51.

McCarthy, John J., and Alan S Prince. "Prosodic Morphology 1986." *Linguistics Department Faculty Publication Series* 1996.

Moore-Cantwell, Claire. "Weight and Final Vowels in the English Stress System." *Phonology* 37.4 (2020): 657–695.

Nespor, Marina, and Irene Vogel. *Prosodic Phonology*. Foris: Dordrecht, 1986.

Newell, Heather. "Deriving Level 1 / Level 2 Affix Classes in English : Floating Phonology, Cyclic Syntax." *Acta Linguistica Academica* 68.1–2 (2021): 31–76.

Pastizzo, Matthew J., and Laurie B. Feldman. "Morphological Processing: A Comparison between Free and Bound Stem Facilitation." *Brain and Language* 90.1–3 (2004): 31–39.

Pater, Joe. "Non-Uniformity in English Secondary Stress." *Phonology* 17.2 (2000): 237–274.

Plag, Ingo, and Laura Winther Balling. "Derivational Morphology: An Integrative Perspective on Some Fundamental Questions." *Word Knowledge and Word Usage: A Cross-Disciplinary Guide to the Mental Lexicon*. Ed. Vito Pirrelli, Ingo Plag, and Wolfgang U Dressler. Berlin & Boston: de Gruyter Mouton, 2020. 295–335.

R Core Team. "R: A Language and Environment for Statistical Computing." 2023.

Raffelsiefen, Renate. "Morphological Word Structure in English and Swedish: The Evidence from Prosody." *Fifth Mediterranean Morphology Meeting* (2007): 209–268.

Raffelsiefen, Renate. "Paradigm Uniformity Effects versus Boundary Effects." *Paradigms in Phonological Theory*. Ed. Laura J. Downing, T. Alan Hall, and Renate Raffelsiefen. Oxford: Oxford University Press, 2005. 211–262.

Scheer, Tobias. *A Guide to Morphosyntax-Phonology Interface Theories. How Extra-Phonological Information Is Treated in Phonology since Trubetzkoy's Grenzsignale*. Berlin: Mouton de Gruyter, 2011.

Selkirk, Elisabeth O. *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge: MIT Press, 1984.

Selkirk, Elisabeth O. "The Role of Prosodic Categories in English Word Stress." *Linguistic inquiry* 11.3 (1980): 563–605.

Sonderegger, Morgan, and Partha Niyogi. "Variation and Change in English Noun/Verb Pair Stress: Data and Dynamical Systems Models." *Origins of Sound Change*. Ed. A.C.L. Yu. Oxford: Oxford University Press, 2013. 262–284. Web.

Szigetvári, Péter. "Stressed Schwa in English." *The Even Yearbook* 13 (2018): 81–95.

Szigetvári, Péter. "Syncope in English." *The Even Yearbook* 5 (2002): 139–149.

Szpyra, Jolanta. *The Morphology-Phonology Interface: Cycles, Levels and Words*. London: Routledge, 1989.

Taft, Marcus. "Interactive-Activation as a Framework for Understanding Morphological Processing." *Language and Cognitive Processes* 9.3 (1994): 271–294. Web.

Trevian, Ives. *English Suffixes: Stress-Assignment Properties, Productivity, Selection and Combinatorial Processes*. Bern: Peter Lang, 2015.

Troubetzkoï, Nikolaï S. "Die Phonologischen Grenzsignale." *Proceedings of the 2nd International Congress of the Phonetic Sciences*. Cambridge: Cambridge University Press, 1936. 45–49.

Vennemann, Theo. *Preference Laws for Syllable Structure and the Explanation of Sound Change*. Berlin, New York: Mouton de Gruyter, 1988.

Watts, Victor. *The Cambridge Dictionary of English Place-Names: Based on the Collections of the English Place-Name Society*. Cambridge: Cambridge University Press, 2004.

Wells, J.C. *Longman Pronunciation Dictionary*. 3rd ed. London: Longman, 2008.

Wennerstrom, Ann. "Focus on the Prefix: Evidence for Word-Internal Prosodic Words." *Phonology* 10 (1993): 309–324.

Published in *Anglophonia* [online], 36, 2024.
URL: https://doi.org/10.4000/11qbh

[1] I would like to thank Björn Köhnlein, the audiences of the ALOES/ALAES workshop at the 61st SAES congress and of the 20th Old World Conference in Phonology and two anonymous reviewers for their remarks and suggestions. All errors are mine alone. This research has benefited from funding from the ANR (ANR-21-FRAL-0001-01) and the DFG (AR 676/3-1) for the ERSaF project (English Root Stress across Frameworks).

[2] Following the common practice of the English-speaking literature, stresses are indicated using diacritic symbols. An acute accent thus indicates primary stress and a grave accent indicates secondary stress (note that this is different from what is often done in the French tradition, and so here *èntertáin* is equivalent to what some would note ˌ*enterˈtain*).

[3] All the transcriptions given in this paper are phonemic transcriptions adapted from Wells (2008). Adaptations include some of the main recent changes that standard British English has undergone, and which are discussed by Lindsey (2019). Almost all vowels are affected: the symbol /æ/ for the TRAP lexical set will be replaced with /a/, the symbol /e/ for DRESS will be replaced with /ɛ/, diphthongs like FACE or GOAT are analysed as vowel-glide sequences (e.g. /ɛj/, /əw/) and the centring diphthongs used by Wells are replaced with long monophthongs. Three pairs of symbols used by Wells are merged in Lindsey's analysis: /ə/ and /ʌ/ are transcribed as /ə/, /iː/ and /i/ as /ɪj/ and /uː/ and /u/ as /ʉw/, though the symbols /ə/, /i/ and /u/ will be taken to represent stressless, reduced realizations of those vowels. Syllable boundaries (marked with spaces) are also not taken over from the dictionary. Finally, the symbol /r/ used by Wells will be replaced by /ɹ/.

[4] Dabouis and Fournier (submitted b) also review psycholinguistic evidence which shows that speakers indeed analyse such words as complex units rather than as simple words.

[5] Kaye (1995) describes these structures as having 'analytic morphology'. I call "neutral" suffixes which do not affect stress, but more generally preserve the pronunciation of their base.

[6] Herment (2010) notes that there are inconsistencies between dictionaries on that matter (e.g. Wells (2008) gives /i/ before *-ness* but Jones (2006) only gives /ɪ/ for that suffix). Cruttenden (2014) also notes that there is inter-speaker variation.

[7] That property has even been analysed as a diagnostic of the neutrality (or non-cohesiveness) of the suffix in analyses that have sought to distinguish between suffix classes (see Newell (2021) for an overview of such analyses). Such words display apparent overapplications of nasal cluster simplification (/ˈda**m**ɪŋ/ *damning*, /ˈsɑj**n**ə/ *signer*, /ˈrɔ**ŋ**ə/ *wronger*) and contrast with cases in which such clusters are maintained before stress-shifting (or cohesive) suffixes (e.g. /da**m**ˈ**n**ɛjʃən/ *damnation*, /ˈsɪ**gn**əl/ *signal*, /ˈlɔ**ŋg**ɪtjʉwd/ *longitude*).

[8] An anonymous reviewer points out that some of those items, *mainland* and *fireman*, are not exactly "non-transparent", and I would agree that this is far from being a homogenous set.

[9] The Guierrian School is French school of English phonology founded by Lionel Guierre. One of its prominent characteristics is the use of extensive datasets to study English stress and spelling-to-sound correspondences. This may seem less 'original' nowadays, but it markedly distinguished it from most generative approaches, which did not base their analyses on quantitative data until recently.

[10] http://keithbriggs.info/English_placename_element_distribution.html [Accessed on 15/02/2023].

[11] If the homographic word is marked as being a borrowing (i.e. Wells indicates a foreign pronunciation), the constituent is not treated as free.

[12] Details on the syllabification procedure adopted can be found in §5.2.4.

[13] Syncope here involves the loss of a medial /ə/ and thus the loss of a syllable (e.g. *Fredericton* /ˈfɹɛd(ə)ɹɪktən/). Compression is when a high vowel loses its syllabicity to be realized as a glide. Assuming that FLEECE and GOOSE are /ɪj/ and /ʉw/, this means that the first element of the diphthong is lost to leave only the glide, thus making it a form of syncope (see Szigetvári 2002).

[14] This word may be pronounced /ˈwɛstmɪnɪstə/, a pronunciation that Wells (2008) marks with "!!", which represents a pronunciation that is "unexpected for this spelling".

[15] Note that Wells (2008) uses two types of notations to mark optional segments: those in italics are "sounds that may optionally be elided" while those in superscript are "sounds that may optionally be inserted". The former were included in the clusters studied here, but not the latter.

[16] The effects were tested in a binary logistic regression, but only the free nature of the first constituent came out as a significant predictor of the presence of an illegal cluster. If we eliminate one variable, we can run a chi-square analysis, which shows the significant effect of that variable ($\chi^2 = 42.551$, df = 1, $p < .001$).

[17] As there can be re-spellings of words ending in /ɪj/ with productive morphology (e.g. *body – bodies – bodily*), words for which a homograph exists if a final <i> is rewritten <y> or <ie> were treated as free (e.g. *Merriman ↔ merry*).

[18] Interestingly, the variation observed above in *Eddystone* is consistent with these observations, as the variant with /ɪj/ occurs with the non-reduced variant of the second constituent.

[19] The vowels treated as long are /ɑj, ɛj, ɪj, ʉw, əw, oː, aw, oj, ɑː, əː, ɛː, ɪː, ɵː/ and those treated as short are /a, ɛ, ɪ, ɵ, ɔ, ə/. Once again, in medial clusters, the sounds indicated in superscript in Wells (2008) are not taken into consideration.

[20] As for illegal clusters, the difference between free and bound constituents is only statistically significant for the first constituent ($\chi^2 = 16.242$, df = 1, $p < .001$). Running the same test on the presence of illegal superheavy syllables, the difference is also statistically significant ($\chi^2 = 17.697$, df = 1, $p < .001$).

[21] It is also possible that a free second constituent favours the interpretation of the word as a complex word and could be associated to higher chances of having a long vowel, but no effect of the free nature of the second constituent can be found.
[22] The latter effect can be seen in Figure 4, where more recurrent constituents are more often realised with reduced vowels than less recurrent ones.
[23] That figure would probably be smaller if phonetic cues were taken into account. For example, certain medial clusters were treated as legal onsets but may be syllabified heterosyllabically to reflect morphological structure. To take but one example, sequences of a voiceless stop and a sonorant that form legal onsets would be expected to be realized phonetically with a voiceless sonorant before full vowels (e.g. *Bla*[kw̥]*all*, *Char*[tw̥]*ell*, *Shi*[pl̥]*ake*) if the cluster is syllabified as a branching onset, but the sonorant would be expected to be voiced if it is syllabified as a coda-onset sequence (see the citation from Giegerich (1999) in §3.2). As suggested in the conclusion, graphotactics may also play a role in signalling morphological structure.
[24] Similarly, in a discussion on the role of medial consonant clusters in the recognition of the internal structure of opaque prefixed words, Fournier (1996) asks if such cues to morphological complexity are sufficient.
[25] As noted by Köhnlein (2015) "[c]lassifiers and geographical suffixes may, however, become relevant for the coining of new / fictive place names". This is clearly true for English, as the constituents found in my dataset are used in books: *Bitterbridge, Tumbleton, Winterfell, Wickenden, Gulltown, Coldwater, Snakewood, Crakehall, Kingswood, Harrowlands* (in G. R. R. Martin's *Game of Thrones*), *Hobbiton, Tuckborough, Tookbank, Newbury, Stockbrook* (in J. R. R. Tolkien's *The Lord of the Rings*), *Little Hangleton*, *Hogwarts* cp. *Hogsmeade* (in J. K. Rowling's *Harry Potter*)) as well as in TV series (*Smallville* (Superman), *Sunnydale* (Buffy the Vampire Slayer), *Ambridge, Borchester, Felpersham* (The Archers), *Springfield* (Simpsons)) or video games (*Blackwater* (Red Dead Redemption), *Bullworth* (Bully, GTA IV), *Cullingtown* (Teardown)).
[26] We saw in §3.2 that this is how Giegerich (1999: 276) analyses the constituent of complex place-names.
[27] Kaye assumes that domains may end in an empty nucleus, which then can account for apparently illegal clusters (e.g. *dreams* would be [[dɹɪjmØ]zØ], where Ø stands for an empty nucleus).
[28] One possible analysis may be that such elements are analysed as part of the same lemma as the freestanding forms, i.e. a unit that has a specific morphological behaviour but can have quite different meanings (Aronoff & Sims 2023).
[29] Note that the same is true in certain words derived from bound roots (e.g. *baleful, doleful, grateful*).