ENGLISH WORD-MEDIAL MORPHONOTACTICS: A CORPUS STUDY

Paulina Zydorowicz
Katarzyna Dziubalska-Kołaczyk
Michał Jankowski
Adam Mickiewicz University, Poznań
zpaula@wa.amu.edu.pl
dkasia@wa.amu.edu.pl
mjank@wa.amu.edu.pl

ABSTRACT

The paper reports on an investigation of word-medial consonant clusters in English. Medial clusters are further subdivided into phonotactic ones, i.e. intramorphemic, and morphonotactic ones, which are morphologically complex – arising as a result of derivation or compounding. In this study we concentrate on morphonotactic clusters. We put forward the following hypothesis: since compounds may ultimately lose transparency and lexicalize, the medial clusters in compounds will tend to be relatively less marked than the medial clusters produced by derivation. In the latter, signalling a morphological boundary is a priority.

In this approach, markedness is defined on the basis of the criteria of consonant description: manner and place of articulation (MoA and PoA) as well as the sonorant / obstruent distinction (S/O) between the neighbouring elements. The verification of this hypothesis has been conducted within the Beats & Binding phonotactics, which operates with the Net Auditory Distance principle (NAD).

Keywords: phonotactics, morphonotactics, corpus study, preferences, markedness

1. INTRODUCTION

Phonotactics investigates permissible sound combinations in a language. The term *morphonotactics* was coined to refer to the interface between phonotactics and morphotactics [7]. Morphonotactics allows to specify consonant clusters which emerge as a result of the intervention of morphology. A boundary should be drawn between phonotactic clusters, which are phonologically motivated and occur within a single morpheme, e.g. /st/ in *mister* and morphonotactic clusters which arise due to concatenation, e.g. /st/ in *mis+time*.

Earlier studies of English morphonotactics investigated word-final clusters, i.e. the effect of English word-final inflection on the shape and degree of markedness of clusters [2, 3]. The aim of the present study is to investigate the influence of derivation and compounding on the emergence of complex medial clusters.

2. ENGLISH MORPHONOTACTICS

English possesses a range of derivational affixes (both prefixes ending with a consonant and suffixes beginning with a consonant) which lead to the creation of morphonotactic consonant combinations. Table 1 presents derivational affixes of English, excluding those which do not trigger morphonotactic consonant clusters [4, 9].

Table 1: The list of English derivational affixes.

prefix	arch-, circum-, cis- (on this side of), counter-, dis-, down-, en-, ex-, fore-, hyper-, mal-, il-, im-, in-, inter-, ir-, mid- mis-, non-, out-, over-, pan-, post-, step-, sub-, trans-, un-, under-, up-, vice-
suffix	-ce, -cy, -dom, -fashion, -fold, -ful, -hood, -let, -less, -like, -ling, -ly, -ment, -ness, -scape, -ship, -ster, -some, -ward(s), -way(s)

All of the morphonotactic clusters that arise due to derivation are medial (with the exception of {-ce}). Both prefixes ending in a consonant, followed by a word stem beginning with another consonant, and derivational suffixes beginning with a consonant, when added to the stem ending in a consonant lead to the rise of medial morphonotactic clusters. For the purpose of the present study, we excluded prefixes such as {inter-}, {over-}, {under-} and {fore-} as they form morphonotactic clusters in rhotic accents. Since our transcription follows the Standard British English pronunciation, the final element of the aforementioned prefixes is vocalic in nature.

Some problems may arise in the analysis of certain English lexical items. Firstly, *cranberry morphemes* deserve a special treatment [13]. The name stems from the most representative case, namely *cranberry* as contrasted with *raspberry* or *huckleberry* where *berry* is the root and {cran-}, {rasp-}, and {huckle-} are obviously bound morphemes as they cannot stand on their own. Secondly, literature finds examples of so called *marginal morphemes*. To provide an example, *deceive* / *perceive* / *receive* share the common phonological string {-ceive}, whereas it is the apparent prefixes which are replaced. However, the apparent prefixes {de-}, {per-}, and {re-} do not express the traditional meanings associated with them, i.e. {re-} in *receive* does not mean *to ceive again*. Similarly, {-ceive} does not mean anything on its own though it possesses interesting properties, namely it changes into {-cept} in forms such as *deception*, *perception* and *reception*. An analogical change occurs in words ending with {-mit}, e.g. *permit* vs *permission*, *admit* vs *admission*. Thus {-ceive} and {-mit} are considered to be morphemes of a marginal type.

The last source of morphonotactic clusters are those which arise due to compounding. examples of morphonotactic clusters arising as a result of compounding are /ndb/ in *handbag* (when unassimilated and unreduced), /lpr/ in *foolproof*, /fst/ in *beefsteak*, /tkr/ in *gatecrasher*, /th/ in *sweetheart*, /lθk/ in *healthcare*, or /stpr/ in *dustproof* . The shape of clusters resulting from compounding is rather liberal as far as their phonological make-up is concerned, including the emergence of geminate clusters which are impossible in monomorphemic words, e.g. *bookcase* vs *better*.

"A compound is usually defined (somewhat paradoxically) as a word that is made up of two other words" [5: 719]. It may be seen as a construction type or a lexical unit of certain characteristics [5]. There are several generally accepted criteria used to describe compounds: orthographic, phonological, morphological, syntactic and semantic. However, the criteria do not provide a satisfactory classification of compounds.

The orthographic criterion may be legitimate in a corpus study since it helps to identify compounds as being written as one word, e.g. *blackbird*, *bluebird* or spelled with a hyphen, e.g. *break-in*. In the present study both types will be taken into consideration in the analysis of the data.[i]

3. FRAMEWORK

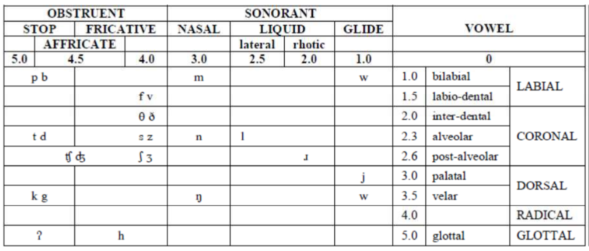
The theoretical framework for measuring cluster markedness is that of Beats-and-Binding phonotactics [1]. It specifies phonotactic preferences as well as the way to evaluate clusters according to them. The rationale behind this model of phonotactics is to counteract the preference for CV. Since CV is a preferred phonological structure and clusters of consonants tend to be avoided across languages and in performance, there must be a phonological means to let them function in the lexicon relatively naturally. This is achieved by auditory contrast and its proper distribution across the word. It is believed that auditory (perceptual) distance can be expressed by respective combinations of articulatory features which eventually bring about the auditory effect.

Any cluster in a structure which is more complex than CV is susceptible to change leading to CV, e.g. via cluster reduction (consonant deletion), or vowel epenthesis or at least vowel prothesis. A way to counteract this tendency is to increase the perceptual distance between the consonants (CC of the CCV) to counterbalance the distance between the C and the V (CV of the CCV). The distance will be expressed by the Net Auditory Distance. Besides, cluster size remains a straightforward measure of cluster complexity: longer clusters are unanimously more complex than the shorter ones.

The Net Auditory Distance (NAD) is a measure of distance between two neighbouring elements of a cluster in terms of differences in MOA (manner of articulation) and POA (place of articulation) as well as the sonorant / obstruent distinction (S/O) between the neighbouring elements. A general NAD table includes MOAs and POAs, in which manners refer to the most generally acknowledged version of the so-called sonority scale, while places are taken from Ladefoged [11]. For particular languages, more detailed tables can be devised, reflecting the differences between systems as well as including more detailed MOA and POA scales, as in the table for English (see Table 2). The numbers in the table are arbitrary. The numbers for the MOAs are based on the sonority scale which assumes equal 'distances' between members starting with STOP through VOWEL. These are expressed by the distance of 1. Affricates and liquids receive special treatment due to their phonetic characteristics. Similarly, the numbers for POAs reflect arbitrarily the distances between sounds. Again, the judgments refer to their phonetic characteristics.

The presence of the S/O distinction is signalled by 1, the absence of it by 0. For instance, S/O (C1C2) when both consonants are sonorants or obstruents equals 0, and when one is a sonorant and the other obstruent - 1.

Table 2: The values of MOA and POA of English consonants.



OBSTRUENT			SONORANT				VOWEL		
STOP		FRICATIVE	NASAL	LIQUID		GLIDE			
	AFFRICATE			lateral	rhotic				
5.0	4.5	4.0	3.0	2.5	2.0	1.0	0		
p b			m			w	1.0	bilabial	LABIAL
		f v					1.5	labio-dental	
		θ ð					2.0	inter-dental	CORONAL
t d		s z	n	l			2.3	alveolar	
	ʧ ʤ	ʃ ʒ			ɹ		2.6	post-alveolar	
						j	3.0	palatal	DORSAL
k g			ŋ			w	3.5	velar	
							4.0		RADICAL
ʔ		h					5.0	glottal	GLOTTAL

The NAD Principle evaluates cluster markedness with reference to universal phonotactic preferences. The preferences for word-medial clusters are presented below.

The condition for a double medial (V1C1C2V2)

1) NAD (V1,C1) NAD (C1,C2) < NAD (C2,V2)

The condition reads:
For a word-medial double cluster, the NAD between the two consonants should be less than between each of the consonants and its respective neighbouring beat, and it may be equal to the NAD between the first consonant and the beat preceding it.

To illustrate with an example, the medial cluster /st/ in the word *mister* would be analysed in the following way:

Example of a medial CC
NAD V1, C1 = |MOA V1 - MOA C1| + |S/O (V1C1)|
NAD C1, C2 = |MOA C1 - MOA C2| + |POA C1 - POA C2| + |S/O (C1C2)|
NAD C2, V2 = |MOA C2 - MOA V2| + |S/O (C2V2)|

/st/ in *mister*
NAD V1, C1 = |0 - 4| + |1 - 0 | = 4 + 1 = 5
NAD C1, C2 = |4 - 5| + |2.3 - 2.3| + |0-0| = 1 + 0 + 0 = 1
NAD C2, V2 = |5 - 0| + |0 - 1| = 5 + 1 = 6

The condition is fulfilled as 5 1 < 6

The condition for a triple medial (V1C1C2C3V2)

(2) NAD (V,C1) NAD (C1,C2)
& NAD (C2,C3) < (C3,V2)

Example of a medial CCC
NAD V1, C1 = |MOA V1 - MOA C1| + |S/O (V1C1)|
NAD C1, C2 = |MOA C1 - MOA C2| + |POA C1 - POA C2| + |S/O (C1C2)|
NAD C2, C3 = |MOA C2 - MOA C3| + |POA C2 - POA C3| + |S/O (C2C3)|
NAD C3, V2 = |MOA C3 - MOA V2| + |S/O (C3V2)|

/stl/ in *firstly*

NAD V1, C1 = |0 - 4| + |1 - 0| = 4 + 1 = 5
NAD C1, C2 = |4 - 5| + |2.3 - 2.3| + |0 - 0| = 1 + 0 + 0 = 1
NAD C2, C3 = |5 - 2.5| + |2.3 - 2.3| + |0 - 1| = 2.5 + 0 + 1 =3.5
NAD C3, V2 = |2.5 - 0| + |1 - 1| = 2.5 - 0 = 2.5

The left-hand condition is fulfilled as 5≥ 1
The right-hand condition is NOT fulfilled 3.5 > 2.5

4. CORPUS STUDY

4.1. Resources

The resources used for the purpose of the study include a list of inflectional forms based on a well-established dictionary and a word frequency list based on a large, well-balanced corpus.

The wordlist is based on the CUV2 lexicon compiled by Mitton [12] from the Oxford Advanced Learner's Dictionary of Current English [10], which comprises approximately 70.5K items including inflectional forms along with UK phonemic transcriptions. US transcriptions and an additional 13,8K items were added to the original CUV2 lexicon by Sobkowiak for his Phonetic Difficulty Index software [14]. For the present study this 84,5K lexicon was stripped of proper nouns and duplicate forms, which brought the total number of items down to approximately 66K. The transcriptions analyzed were UK ones.

Frequency data for the items studied were extracted from a frequency list based on the 410 million word Corpus of Contemporary American English (COCA) [6]. In other words, the corpus was used solely as a source of word frequency information. The list contains approximately 500,000 word forms, along with their grammar codes, number of occurrences, and number of sources in which they appear.

4.2. Methodology of morphological division

The automatic morphological parsing of English word-medial consonant clusters proceeded in several steps. Clusters in compounds were isolated by first finding words in the transcription resource which were composed of orthographic segments which themselves were individual words in the resource, and, in addition to that, the concatenated transcriptions of the segments matched the transcription of the processed word. In the second phase, if the first segment's transcription featured a final consonant or cluster and the second segment's transcription featured an initial consonant (or a cluster), that string of consonants was recorded as a cluster of specific length with the word as its source and categorized as a "compound" cluster, i. e. one that was the result of compounding. An example of such a cluster is /sb/ in *baseball* (base+ball). Derived words were investigated by analysing the affixes presented in Table 1. All words with the affixes were extracted from the database. Parsing was based on the idea that after affix stripping, the remaining stem is an existing word in the lexicon and its transcription matches that of the corresponding transcription string in the processed word. For example, the initial part of the word *helpless*, when separated from the suffix {-less}, is found in the dictionary as a separate entry and their transcriptions match. This algorithm ensures that the parsing is applied to genuine morphemes. On the other hand, pseudomorphemes or cranberry morphemes are not captured by this rule. To illustrate with an example, *distinguish* will not be divided morphologically into {dis-} + {tinguish} as the latter does not appear as a dictionary entry. One pitfall of this approach is that words such as *invent*, where both {in-} and {vent-} can be found as separate entries, will be parsed by the rule. Such cases were intercepted by the researchers and sorted out on an individual basis. Another group of items to be treated that way were clusters in words with a {-ship} ending such as *battleship* (compound) and *friendship* (suffix = derivation), which were initially accepted to both the categories and then the conflict was resolved manually.

4.3. Hypothesis

The aim of this study is to investigate English word-medial morphonotactics quantitatively.

Our hypothesis pertains to the difference between clusters arising from derivation and compounding. Since compounds may ultimately lose transparency and lexicalize, the medial clusters in compounds will tend to be relatively less marked [8][ii], i.e. closer to their phonotactic word-medial counterparts, than the medial clusters produced by derivation. In the latter, signalling a morphological boundary is a priority.

5. RESULTS

Table 3 below presents the results obtained for medial clusters of all lengths (2-5 consonant clusters). The numbers are given for cluster types and word types in the dictionary and supplemented by the frequency from the corpus (word tokens).

Table 3: Word-medial morphonotactics: quantitative data.

length	**cluster types**	**word types**	**word tokens**
2	284	5741	6761507
3	275	1819	1704795
4	55	160	89849
5	3	4	827
total	617	7724	8556978

Double and triple clusters underwent the analysis according to two criteria: morphological operation triggering the cluster (derivation vs compounding) as well as the Net Auditory Distance (preferred vs dispreferred). The results are presented in Table 4.

Table 4: Type of morphological operation vs preferability.

	NAD	**der**	**comp**
double clusters			
type	pref	48.28%	61.46%
	dispref	51.72%	38.54%
token	pref	51.98%	63.09%
	dispref	48.02%	36.91%
triple clusters			
type	pref	67.18%	44.02%
	dispref	32.82%	55.98%
token	pref	66.49%	52.93%
	dispref	33.51%	47.07%

The results of the study confirm our hypothesis for medial double clusters: compounding generates more preferred clusters than derivation, both in the case of types and tokens. However, the results are reversed in the case of triple clusters, in which case clusters triggered by derivation tend to be more preferred.

6. CONCLUSIONS

We hypothesised that medial clusters in compounds would tend to be relatively less marked than the medial clusters produced by derivation. Medial clusters of two consonants in our data indeed supported the claim: they tended to be more preferred in compounds than in derivatives. However, the three-consonant clusters showed the opposite tendency. There may be a number of reasons for the apparently different nature of double and triple medial clusters. Before we voice them, however, further and more detailed analysis of the data is required.

7. REFERENCES

[1] Dziubalska-Kołaczyk, K. 2009. NP extension: B&B phonotactics, *PSiCL* 45 (1): 55-71.

[2] Dziubalska-Kołaczyk, K., Zydorowicz, P., Jankowski, M. 2013. English Morphonotactics: A Corpus Study. *The Phonetician. A Publication of ISPhS International Society of Phonetic Sciences* 107-108: 53-67.

[3] Zydorowicz, P., Orzechowska, P., Jankowski, M., Dziubalska-Kołaczyk, K., Wierzchoń, P., Pietrala, D. (in press). *Phonotactics and morphonotactics of Polish and English. Theory, description, tools and applications*.

[4] Bauer, L. 1983. *English word-formation*. Cambridge: CUP.

[5] Bauer, L. 2006. Compound. In: Brown, K. (ed.), *Encyclopaedia of Language and Linguistics*. Oxford: Elsevier.

[6] Davies, M. 2011. *Word frequency data from the Corpus of Contemporary American English* (COCA). Downloaded from http://www.wordfrequency.info on January 24, 2011.

[7] Dressler, W. U., Dziubalska-Kołaczyk, K.. 2006. Proposing Morphonotactics. *Rivista di Linguistica* 18.2, 249-266.

[8] Dressler, W. U., Dziubalska-Kołaczyk, K., Pestal, L. Change and Variation in Morphonotactics. *Folia Linguistica* 31. 51-67.

[9] Hatch, E., . Brown. C. 1995. *Vocabulary, semantics, and language education*. Cambridge: CUP.

[10] Hornby, A. S. 1974. *Oxford Advanced Learner's Dictionary of Current English*, Third Edition. Oxford: Oxford University Press.

[11] Ladefoged, P. 2006. *A course in phonetics*. Boston: Heinle & Heinle.

[12] Mitton, R. 1992. "A description of a computer-usable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English." A text file bundled with the resource file.

[13] O' Grady, Dobrovolsky, W. M., Aronoff, M.. 1993. *Contemporary linguistics: Introduction*. New York: St. Martin's Press.

[14] Sobkowiak, W. 2006. PDI revisited: lexical co-occurrence of phonetic difficulty codes. In:. Sobkowiak, W., Waniek-Klimczak, E. (eds.). 2006. Dydaktyka fonetyki języka obcego. Neofilologia VIII. *Zeszyty naukowe Państwowej Wyższej Szkoły Zawodowej w Płocku*. Płock: Wydawnictwo Państwowej Wyższej Szkoły Zawodowej. *Proceedings of the Fifth Phonetics in FLT Conference*, Soczewka, 25-27.4.2005. 225-238.

[i] Bauer [5] warns against drawing conclusions concerning the cohesiveness of a compound on the basis of its spelling: "Rainforest, rain-forest, and rain forest are all easily attestable" [5: 720].

[ii] Among the clusters found in the medial position, the "compound clusters" are the only ones that are truly phonotactically preferred. Thus, in Lithuanian compound formation, the rise of phonotactically marked consonant clusters appears to be disfavoured, and in this respect Lithuanian is similar to other languages which employ vowel interfixation for that purpose." [8: 61].