Studia Anglica Posnaniensia 56s1 (2021): 465–487
doi: 10.2478/stap-2021-0006

ON THE SHAPES OF THE POLISH WORD: PHONOTACTIC COMPLEXITY AND DIVERSITY

PAULINA ZYDOROWICZ, MICHAŁ JANKOWSKI, KATARZYNA DZIUBALSKA-KOŁACZYK[1]

ABSTRACT

The aim of this contribution is to identify the dominant shapes of the Polish word with reference to three criteria: cluster complexity (i.e., cluster size), saturation (the number of clusters in a word), and diversity (in terms of features of consonant description). The dominant word shape is understood as the most *frequent* or *typical* skeletal pattern, expressed by means of alternations or groupings of Cs (consonants) and Vs (vowels), e.g., CVCCV etc., or by means of specific features (of place, manner, voice, and the sonorant/obstruent distinction). Our work focuses on 2 aspects of Polish phonotactics: (1) the relation between cluster complexity and saturation of words with clusters, (2) the degrees of diversity in features of place, manner, and voice within clusters. Using corpus data, we have established that only 4.17% of word shapes have no clusters. The dominant word shape for a one-cluster word is CVCCVCV. The most frequent scenario for a word shape is to contain two clusters, of which 67% are a combination of a word initial and a word medial cluster. We have found that: (1) cluster length is inversely proportional to the number of clusters in a word; (2) nearly 73% of word types contain clusters of the same size, e.g., two CCs or two CCCs (Polish words prefer saturation over complexity); (3) MOA is more diversified than POA across clusters and words.

Keywords: Phonotactic complexity; phonotactic diversity; corpora; word shape.

1. Introduction

1.1. Polish phonotactics

Polish is a phonotactically rich language. An intra-morphemic consonant cluster can be up to 5 consonants long, while cluster complexity grows across morphological boundaries (up to 6 consonants) and across word boundaries

[1] Adam Mickiewicz University in Poznań; Faculty of English, Kolegium Heliodora Święcickiego (Collegium Heliodori Święcicki), ul. Grunwaldzka 6, 60–780 Poznań, Poland. zpaula@amu.edu.pl, mjank@amu.edu.pl, dkasia@amu.edu.pl

(up to 11 elements). Moreover, a single word may contain several clusters, and on average it actually has two clusters.

Polish phonotactics has been of interest to linguists over the last 70 years. The first works devoted to the description of Polish phonotactics date back to the middle of the 20th century (Bargiełówna 1950; Kuryłowicz 1952), and were devoted to identifying possible cluster types and formulating general observations about consonant organisation within clusters. Subsequent works on Polish phonotactics focused on cluster frequency in corpora (Dukiewicz 1985; Dobrogowska 1990, 1992; Orzechowska 2009). The complexity and the intricacies of the Polish phonotactic inventory have also been explained by theoretical approaches to phonology, e.g., generative phonology (Bethin 1992; Szpyra 1992, 1995), Lexical Phonology (Rubach & Booij 1990), Government Phonology (Gussmann & Cyran 1998; Cyran & Gussmann 1999; Cyran 2006; Gussmann 2007), Optimality Theory (Rochoń 2000) or, more recently, Onset Prominence (Schwartz 2019). Finally, selected aspects of phonotactics have been studied empirically in first language acquisition (Jarosz 2010; Zydorowicz 2010, 2019), second language acquisition (Dziubalska-Kołaczyk & Zydorowicz 2014) and psycholinguistics (Wiese et al. 2017). Corpus-based as well as empirical studies of Polish phonotactics in terms of phonetic features were conducted by Orzechowska (2019).

The approach adopted by the authors of the present contribution is that of Natural Phonology. Earlier research in the spirit of this framework focused on the exploration of various written corpora (Dziubalska-Kołaczyk 2014, 2019; Zydorowicz et al. 2016; Zydorowicz & Orzechowska 2017), the behaviour of clusters in spontaneous speech of adult language users and in the process of phonological development by children (Zydorowicz & Dziubalska-Kołaczyk 2017; Zydorowicz 2019). The aforementioned corpus studies involved obtaining statistical information on the number of cluster types, word types, and word tokens, analysing clusters in terms of cluster goodness (as defined by markedness or preferability measures), examining the morphological structure of Polish clusters and relations between such variables as cluster size vs markedness, markedness vs morphological composition, markedness vs frequency. The studies revealed the following findings: approximately 2400 cluster types have been identified. Clusters are best tolerated word-medially (which manifests itself in the greatest number of existing types), then word-initially and least welcome word-finally (the lowest number of clusters). The maximal number of elements in a word-initial cluster is 4, e.g., /pstr-/ in *pstrykać* ‘to flick’ (5 consonants are present in an exceptional river name *Strwiąż*, namely /strfj-/). Word-medially 4 consonants are permissible within a morpheme, e.g., /-kstr-/ in *ekstra* ‘great’ and 6 intermorphemically, e.g., /ntʂ+zvj/ *wewnątrzzwiązkowy* ‘union-internal’, The right edge of the word may contain up to 5 consonants in a morphologically complex cluster, e.g., /-mpstf/ *przestępstw* 'crime'-gen.pl and 3 consonants within a morpheme, e.g., /kst/ as in *tekst* 'text'. The size of a cluster predicts the presence of a morphological boundary (the longer a cluster is, the higher is the probability of an intervening morphological boundary). A higher proportion of dispreferred (or marked) clusters is found in the group of morphologically complex clusters, although it must be stressed that phonologically motivated clusters can be highly dispreferred as well, cf. /t͡ʂtɕ-/ in *czcić* 'to worship', /kt-/ *kto* 'who' or /rv-/ in *rwać* 'to tear' (these tautomorphemic examples demonstrate that sonority plateaus and reversals are attested, and although they are statistically rarer, they are not exceptional and can be very frequent in language use). Cluster markedness does not correlate with token frequency (Orzechowska & Zydorowicz 2019). There are numerous clusters which are considered universally dispreferred and are still very common in language use in Polish, e.g., /kt-/ in *który* 'which' or *kto* 'who' and /ɡd-/ in *gdy* 'when'.

1.2. Aims and crucial terminology

In this paper, we extend our previous analysis and investigate the phonotactic potential of the Polish word as a whole and not just single clusters. We are interested in how clusters are organised within words both qualitatively and quantitatively. Our research has been motivated by the questions concerning the relationship between cluster complexity and composition, and the number of clusters a word can bear (referred to as cluster saturation). Crucially, we have also been interested in the positions of clusters within a word (initial, medial, or final) and their mutual coexistence. We have also looked for potential co-occurrence restrictions in cluster composition in the word domain, or, in other words, a degree of compositional diversity allowed in a single word.

We begin by introducing crucial terminology which will be employed in our analysis.

Cluster type: a cluster with unique composition, e.g., /tr/
Word form: all existing forms of a lemma, e.g., *trawa* 'grass', *trawy* 'grass'-pl., *trawą* 'grass'-instr. (these examples are counted as three word forms)
Word token: a repetition of a word form in the corpus, e.g., *trawa* 'grass' has 638 repetitions in the SUBTLEX-PL corpus, thus 638 tokens
Word shape: the word's skeleton / template where sounds are generalised to Cs (consonants) and Vs (vowels), e.g., *siostra* 'sister' /ɕɔstra/ CVCCCV or to features (place of articulation, manner of articulation, and voice), e.g., the cluster in *sto* 'a hundred' can be described as dental + dental, fricative + plosive, voiceless + voiceless
(Cluster) complexity: cluster size/length (which are used interchangeably)

Saturation: the number of clusters in a single word, e.g., *przetransportowaliśmy* /pʂɛtranspɔrtɔvaliɕmɨ/ 'we transported' (which is saturated with 5 clusters) vs *niezarejestrowany* /ɲɛzarɛjɛstrɔvanɨ/ 'unregistered' (saturated with 1 cluster)
Diversity: the difference in terms of phonetic features between the member consonants of a cluster; 3 features are considered, namely, place of articulation (POA), manner of articulation (MOA), and voicing. To illustrate diversity with an example, the cluster /br/ in the word *brat* 'brother' is diversified in terms of MOA and POA, but not in terms of voicing. Sequence /pr-/ in *prać* 'to wash' is diversified with respect to all three criteria. Diversity can be defined as the opposite of feature identity.

2. NAD model of phonotactics (Dziubalska-Kołaczyk 2014, 2019) and the interaction between phonotactics and morphonotactics

In the introduction we have specified our theoretical framework as that of natural phonology. To study clusters we have referred to the syllable-less Beats-and-Binding phonology embedded in the natural framework. Consequently, we examine clusters in terms of their position in a word (initial, medial, and final) and not within the domain of the syllable. The Net Auditory Distance model of phonotactics is derived from the B&B phonology. In the NAD model, the composition of a cluster is expressed by NAD – net auditory distance. Every articulatory gesture produces an auditory effect; every segment/phoneme is auditorily distant from its neighbours as a result of their differing articulatory set-ups. Therefore, auditory distances in the NAD model are measured in terms of articulatory features which brought them about. Optimal combinations of distances generate the most preferred clusters.

The NAD principle defines cluster preferability with respect to a given cluster's position in the word: different clusters are preferred in word initial, word medial, and word final position. The NAD principle is not just about a distance between two segments – it is about a relation of distances in a cluster, specified accordingly to its position in the word. Three major articulatory parameters have been selected for NAD measurement: manner of articulation (MOA), place of articulation (POA), and the distinction between sonorant (S) and obstruent (O) sounds (S/O) NAD is the sum of distances calculated in terms of the three parameters, i.e., NAD = |MOA| + |POA| + S/O.

The formulas for cluster preferability (=preferences) operate within the domain of a word. The idea behind them is to define the optimal distribution of distances in a cluster to guarantee its survival. For example, in a word initial double cluster (i.e., a cluster consisting of two consonants) #C1C2V, the distances need to be so distributed as to counteract the CV preference. This is captured by the inequality below which reads: NAD between C1 and C2 should be greater than (or equal to) NAD between C2 and V: NAD (C1,C2) ≥ NAD (C2,V) The NAD calculator has been designed to perform the calculations of the NAD automatically. (The tool is accessible online at http://wa.amu.edu.pl/nadcalc/).

Phonological criteria are not enough to investigate clusters. Morphological operations of inflection, word-formation, and compounding also contribute to the creation of consonant clusters. The area of interaction between morphotactics and phonotactics is called morphonotactics (Dressler & Dziubalska-Kołaczyk 2006). A general hypothesis regarding the outcome of this interaction predicts the morphonotactic clusters to be relatively more marked (lower on the scale of preferability) than the purely phonotactic ones.

In the present study, we do not use NAD for the analysis of clusters. However, we do analyze the relevance of the NAD criteria of MOA, POA, and voice as well as sonorant/obstruent distinction for the composition of clusters in Polish. Consequently, the present results provide us with a basis to assign relative weights to the NAD criteria with reference to their predictability force. In other words, we will be able to estimate the contribution of the particular criteria to the overall NAD product.

3. Corpus study

3.1 Methodology

The corpus data used in this study was adopted from SUBTLEX-PL (Mandera et al. 2014), which is a list of word frequency estimates derived from film and television subtitles, containing approximately 450K types (over 101M tokens). The choice of this resource was motivated by several factors. First of all, corpora of film subtitles have gained popularity among linguists over the last 10 years. Currently, subtitle-based resources are available for a number of languages, including American English SUBTLEX-US (Brysbaert, New & Keuleers 2012), British English (SUBTLEX-UK), German (SUBTLEX-DE, Brysbaert et al. 2011), Portuguese (SUBTLEX-PT, Soares et al. 2015) or Polish (SUBTLEX-PL, Mandera et al. 2014). Numerous studies have explored the functionality and reliability of these resources, confirming their validity. For instance, frequencies based on SUBTLEX corpora explain more variance in lexical decision times than other available word frequency measures (cf. frequency counts according to SUBTLEX-DE vs CELEX in German, SUBTLEX-UK vs the British National Corpus in British English, or SUBTLEX-PT vs P-PAL in Portuguese). Following these findings, the Polish version of the corpus was considered a reliable, valid and sufficient resource in the study of phonotactic patterns.

From among the various versions of the frequency list we chose the one based on at least 3 sources. After eliminating several groups of entries which it was thought were unsuitable for the study, such as those tagged as unrecognized by the spellchecker used, entries with unclear part-of-speech tags, and entries with non-Polish or non-alphabetic characters, we obtained a list of 267 191 "clean" Polish words (100 598 516 tokens, 50 186 lemmas). This list was transcribed with the use of software and rules originally designed and described in Jankowski (1994), and successfully used in a recent consonant-cluster-related project (Zydorowicz et al. 2016).

The next step of data preparation was to determine the skeletal structure of Polish words or, in other words, the number of different word shapes (n = 2 968). Word shapes are word patterns whose phonological content is generalized to Cs and Vs. To illustrate with an example, the word /brat/ 'brother' represents a CCVC pattern. Scripts were written to convert transcription of word forms into word shapes, mechanically. Thus, in the present work, we will refer to three levels of analysis, namely, word shapes, word forms, and word tokens. Subsequently, all consonants were tagged for place of articulation POA, manner of articulation MOA, and voice.

The following hypotheses and supporting research questions have been formulated.

3.2. Hypotheses

Hypothesis 1: saturation and complexity

The more clusters in a word, the shorter the clusters are. It is assumed that in order to ensure articulatory balance in a word, words with numerous heavy clusters will be avoided. This hypothesis is also accompanied by a related research question: How do clusters coexist in terms of length?

Hypothesis 2: feature diversity

Manner of articulation features will be more diversified in clusters than place features. This assumption is derived from several linguistic facts and empirical findings. Firstly, Baumann & Wissing (2018: 77) demonstrated that "large differences in the manner of articulation between segments contribute to a cluster's success in acquisition and diachrony" while the place of articulation has an impeding effect. Furthermore, preliminary results from the acquisition of Polish phonotactics reveals that children modifying clusters through substitution preserve manner features better than place features. In other words, in segment substitutions within clusters, the place of articulation is changed more frequently than manner of articulation (Zydorowicz 2019: 193). This points to the greater stability, and to some extent, a greater role/importance of manner features in preserving contrasts. Thirdly, the principle responsible for the organisation of sounds within words and syllables is sonority, which is based on the degree of aperture of the vocal tract. The building blocks of the sonority hierarchy are consonant classes ordered according to the manner of articulation (plosives – affricates – fricatives – nasals – liquids – semi-vowels).

The predictions regarding voicing consider the position of a cluster in a word:
a) clusters at word edges are expected to be varied in terms of the laryngeal activity.
b) clusters in the medial position are likely to be less constrained in terms of voicing or are likely to be voiced.

These hypotheses stem from the following observations. Typologically speaking, universally unmarked clusters are those composed of obstruents and sonorants (to be more precise obstruents and sonorants word initially, and their reversals word-finally). According to some sonority scales (i.e., those which include voicing as a relevant criterion), voiceless plosives + sonorants ensure a larger contrast than voiced plosives + sonorants. Voiced plosive + sonorant sequences are also rarer cross-linguistically (for example, Polish possesses both /fl-/ and /vl-/ or /sn-/ and /zn-/, whereas English possesses only the first cluster of each pair). It may be surmised that the universal preferences (for a voiceless obstruent followed by a sonorant) will be reflected in the structure of Polish words. Therefore, diversified voice combinations within clusters are expected. It is assumed that clusters in the medial position will be less constrained by articulatory or auditory demands, thus a certain degree of freedom is expected. However, the neighbourhood of (voiced) vowels may be conducive to the prevalence of voiced sequences.

3.3. Results

We begin our analysis with general observations pertaining to the structure of the Polish word. The corpus data revealed an interesting proportion between words with and without clusters. It transpired that a Polish word by default contains at least one cluster (Table 1).

Table 1 shows that only 4.18% of word shapes and 17.90% of word forms are devoid of clusters. However, such words are in frequent use and constitute 55% of the corpus. They provided 126 word shapes. Words with clusters generated 2 842 word shapes, which constitutes 95.82% of all word shapes. The data split into various cluster lengths is presented in Table 2 where type and token frequency are given.

Table 1. The numbers of words with and without clusters.

	word shapes	word forms	word tokens
without clusters	126 4.2%	47 870 17.9%	55 100 315 54.7%
with clusters	2 842 95.7%	219 550 82.1%	45 592 700 45.2%
total	2 968	267 420	100 693 015

Table 2a. The distribution of clusters in terms of clusters size (all positions separate).

	cluster types			clusters in word forms			clusters in word tokens		
cl size	initial	medial	final	initial	medial	final	initial	medial	final
2	236	448	171	62729	214261	6513	18568468	29869562	3773429
3	227	792	64	12111	33366	208	1915648	3638392	178878
4	38	204	8	635	4191	44	55167	388898	6768
5	1	22	1	1	164	5	4	17711	811
6		1			4			392	
total	502	1467	244	75476	251986	6770	20539287	33914955	3959886

Table 2b. The distribution of clusters in terms of clusters size (all positions combined).

cl size	cluster types	clusters in word forms	clusters in word tokens
2	855	283503	52211459
3	1083	45685	5732918
4	250	4870	450833
5	24	170	18526
6	1	4	392
total	2213	334 232	58 414 128

The data shows that as the number of elements in a cluster increases, fewer combinatorial possibilities are exploited. This observation is compatible with Dukiewicz (1985) who points out that having 28 consonantal phonemes in Polish (established after Paulsson 1969) one can obtain 756 various doubles, 19 656 triples, and 491 400 4-member clusters. A similar statement had been made with respect to English phonotactics: in English the number of consonants is 22, which gives us the number of 484 different consonant combinations in a double cluster. However, only 28 possibilities exist in English. In the case of triples the potential is 10 648, out of which only 8 occur (Greenberg 1978, Universal 1). Computationally, the bigger the cluster is, the more possible combinations can be generated. In practice, however, as the length of the cluster increases, fewer possibilities are utilized. In our data, this statement is true for word forms, word tokens, and, to a large extent, cluster types, with the exception of triple clusters which outnumber 2-member sequences.

The most frequent word shape (according to word form frequency) is CVCCVCV as exemplified by the word *dziękuję* ‘thank you’ /ʥɛŋkujɛ/. We obtained the following saturation values (numbers of clusters in a word shape and word form):

Table 3. Saturation of Polish words with clusters.

number of clusters in a word	word shapes	word forms	example
1 cluster	648	122 897	/ʥɛŋkujɛ/ *dziękuję* ‘thank you’
2 clusters	1 057	79 772	/napravdɛ/ *naprawdę* ‘in reality’
3 clusters	847	15 702	/vwaɕʨivjɛ/ *właściwie* ‘indeed’
4 clusters	264	1 152	/pʂɨvjɛʑliɕmɨ/ *przywieźliśmy* ‘we brought’
5 clusters	28	41	/pʂɛtɕɛmbjɔrt͡ʂɔɕʨi/ *przedsiębiorczości* ‘entrepreneurship’ dat.

A Polish word may contain between 0 and 5 clusters (zero is not included in Table 3 as words devoid of clusters have been filtered out). If a word contains one cluster, it is most likely to be medial, then initial, and lastly final. The most frequent scenario for a word shape is to contain 2 clusters, whereas the most frequent scenario for a word form is to contain one cluster.

The next step of the analysis was to examine the coexistence of clusters in different word-positions, which is presented in Figure 1.

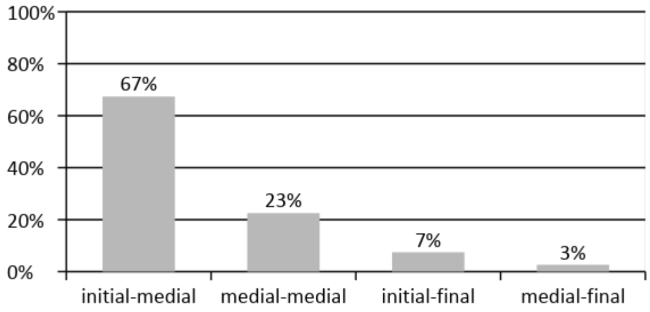

Figure 1. Coexistence of clusters in terms of all word positions.

In the most typical scenario, medial clusters co-exist with initials.

Having provided general information, about cluster frequency and distribution in a word, we move on to the discussion of the hypotheses.

3.3.1. Hypothesis 1: saturation vs complexity

The first hypothesis concerned the relation between the number of clusters in a word (saturation) and cluster length (complexity). Word forms containing clusters were first split into groups based on the number of clusters they have. Then within those subgroups word forms with clusters of specific length were counted. Table 4 presents the results of this analysis. For example, among the words with 2 clusters, there are 78 299 word forms with double clusters, 19 967 with triples, 2170 with quadruples etc. The results can be generalized that cluster length is inversely proportional to the number of clusters in a word. This means that if a word contains several clusters, they tend to be shorter.

Table 4. The number of word forms with regard to cluster length and saturation.

	number of clusters in a word			
length	2 clusters	3 clusters	4 clusters	5 clusters
CC	78 299	15 685	1 152	41
CCC	19 967	4 443	467	29
CCCC	2170	597	63	1
CCCCC	89	26	3	–
CCCCCC	3	1	–	–
total	100 528	20 752	1 685	71

The explanation for those results may come from the fact that long clusters (4–6 consonant members) code specific morpho-semantic information (Dziubalska-Kołaczyk, Jankowski & Wierzchoń 2011). The demand for these functions is limited. In contrast, double and triple clusters have a universal function. Therefore, the number of 4- and 5- member clusters does not grow alongside pattern length and are independent of saturation.

Hypothesis 1 was accompanied by a follow-up research question how clusters coexist in terms of length in a word. The results are presented in Table 5. The data was divided into three categories: IS (identical size of all clusters), S+1 (neighbourhood of a cluster larger by 1 consonant) and S+2 (neighbourhood of a cluster larger by 2 consonants).

Table 5. Cluster coexistence in words in terms of cluster size.

category	word shapes	word form frequency	token frequency
IS	7	70 423 72.85%	8 871 953 75.01%
S+1	18	23 631 24.45%	2 712 820 22.93%
S+2	42	2 613 2.70%	243 530 2.06%
total	67	96 667 100%	11 828 303 100%

The results show that Polish words tend to contain clusters of the same size, which constitutes 72.85% of all word types. 24.45% of word-types contain a cluster which coexists with another cluster longer by 1 segment. Co-occurrence of clusters which differ in length by 2 segments is negligible. A close inspection of the data revealed that three double clusters are preferred over two triple ones. We conclude that Polish prefers saturation (more clusters) over complexity (length).

3.2.2. Hypothesis 2: Feature diversity

We have investigated the diversification of clusters in terms of their (1) place of articulation, (2) manner of articulation, and (3) voice. We looked at homogeneity vs heterogeneity in place and manner features in a cluster. For instance, word initial /st-/ in *sto* ‘hundred’ is homogenous in place and heterogenous in manner. The percentages show how often POA and MOA agree in identity.

Table 6. The degree of feature identity in clusters (as exemplified by CCs).

position	size	POA	MOA
initial	CC	9.29%	6.14%
	CCC	0.22%	0.98%
medial	CC	29.21%	7.17%
	CCC	5.72%	0.01%
final	CC	73.13%	1.48%
	CCC	1.49%	–

Table 6 demonstrates how often a given place or manner repeats itself in a double and triple cluster. In double clusters, place of articulation is less diversified than manner as percentages expressing feature identity for POA are substantially higher. The second observation is that place and manner features behave differently across word positions: feature identity in POA is strongly preferred word finally and the least welcome word initially; manner-wise, feature identity is most dispreferred word-finally and best tolerated in the word-medial position. Triple clusters show much less identity in POA and MOA. In other words, POA and MOA feature identity decreases alongside cluster length, which means that a feature is unlikely to repeat itself in longer clusters.

The third feature under scrutiny was voice. We distinguished three kinds of clusters: voiced, voiceless, and mixed. As demonstrated in Figure 2, voiceless clusters prevail in the initial position. Word-medially, voiced and mixed clusters are balanced. The word-final position is most diversified with mixed clusters prevailing.

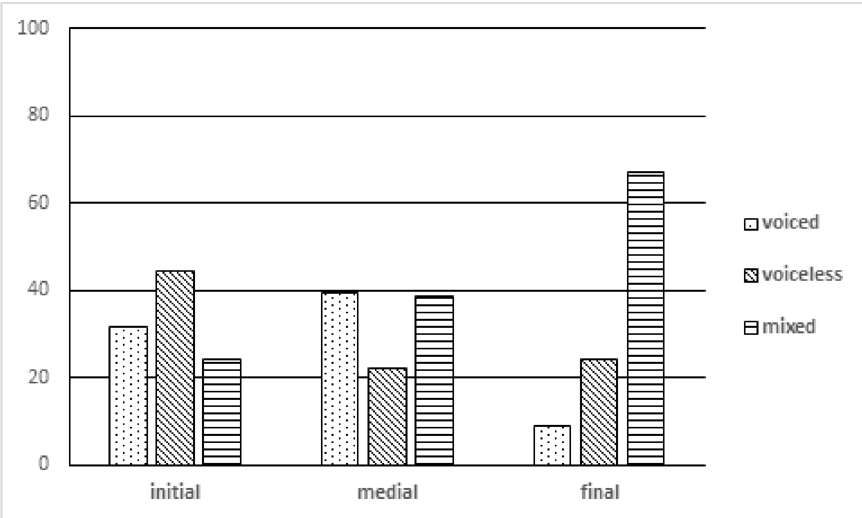

Figure 2. Voice distribution in CC clusters across word positions.

Since the predictions concerning the prevalence of voicing patterns in different word positions were based on the universal preferences pertaining to combinations of sonorant and obstruent consonants, we also provide the distribution of all combinatorial possibilities and their popularity in all word positions. Apart from indicating the frequency of the four profiles (SS, OO, SO, OS), we also indicate the voicing variant for obstruents ('+v' stands for voiced, '–v' stands for voiceless). Sonorants are voiced by default.

Table 7. Sonorant / obstruent distribution in word-initial double clusters.

sonorant / obstruent feature	voice pattern	word form frequency	%	token frequency	%	example CC	example word
–vO–vO	– –	27773	44%	7131082	38%	fʂ	*wszystko* 'everything'
–vOS	–+	15185	24%	3716664	20%	pr	*proszę* 'please'
+vOS	++	12534	20%	5078628	27%	dl	*dla* 'for'
+vO+vO	++	5549	9%	1174134	6%	gʥ	*gdzie* 'where'
SS	++	1473	2%	1455759	8%	mɲ	*mnie* 'me'
S+vO	++	162	0.26%	8280	0.04%	wz	*łzy* 'tears'
S-vO	+–	53	0.08%	3921	0.02%	rt	*rtęć* 'mercury'
		62 729		18 568 468			

Table 8. Sonorant / obstruent distribution in word-medial double clusters.

sonorant / obstruent feature	voice pattern	word form frequency	%	token frequency	%	example CC	example word
–vO–vO	– –	47319	22%	6650986	22%	st	*jestem* 'I am'
S-vO	+–	42545	20%	4257000	14%	lk	*tylko* 'only'
–vOS	–+	39703	19%	5234428	18%	ɕl	*jeśli* 'if'
+vOS	++	36606	17%	6542297	22%	bj	*ciebie* 'you'
SS	++	22647	11%	2729499	9%	mj	*rozumiem* 'I see'
S+vO	++	17728	8%	2723677	9%	ɳʥ	*będzie* 'it will be'
+vO+vO	++	7692	4%	1730933	6%	bʐ	*dobrze* 'well'
total		214261		29869562			

Table 9. Sonorant / obstruent distribution in word-final double clusters.

word-final clusters							
sonorant / obstruent feature	voice pattern	word form frequency	%	token frequency	%	example CC	example word
S–vO	+–	4149	64%	1422320	38%	nʦ	*więc* 'so'
–vO–vO	– –	1570	24%	2070750	55%	st	*jest* 'is'
+vOS	++	400	6%	144525	4%	gw	*mógł* 'he could'
–vOS	–+	222	3%	85855	2%	sw	*pomysł* 'idea'
SS	++	172	3%	49979	1%	lm	*film* 'film'

With respect to the hypotheses formulated in Section 3.2, it was predicted that clusters at word edges are likely to be mixed in terms of voicing. This assumption was confirmed only for word-final clusters, whose majority are mixed (67%). However, word-initially, voiceless clusters prevail (44%). This fact can be explained by a high proportion of obstruent clusters, which, in turn, can partly be attributed to morphology. Polish possesses several non-syllabic consonantal prefixes: {s}, {z}, {w} {ws}, {wz}, which when attached to the stem beginning with a consonant, may form morphologically complex structures. The prefixes are obstruents themselves and they may combine with obstruent consonants initial in a word, e.g., /s/+*topić* 'to melt'-perf., *z+gubić* 'to lose'-perf., *w+stać* 'to stand up', *w+pisać* 'to write in', *ws+chodzić* 'to rise', *wz+bogacić (się)* 'to get rich'. The last two form triple clusters of the type OOO, but the prefixes themselves have the OO structure. In fact, further analysis revealed that the top 4 examples include /pʂ/ and /s/ + stop clusters (/sp/, /st/, /sk/). In fact, /pʂ/

constitutes 49% (!) of the word-initial OO clusters. The cluster owes its high type frequency to the productive prefix {przed} which generates a considerable number of derivatives. The /s/ + stop clusters constitute, respectively, 9, 8 and 7% of the OO cluster subset. The expectations pertaining to medials were largely met: on the one hand, word-medial position is the most tolerant of clusters, and on the other hand, the voicing of the neighbouring voiced vowels may favour the voicing of the cluster. Therefore, the majority of medial CCs tend to be either voiced (40%) or mixed (39%).

Having established that the predominant word shape in Polish has two double clusters, one initial and one medial and no other clusters, further analysis concentrated on shapes that meet this general criterium, i.e., words that begin with a double cluster, may or may not have one or more consonants before a double medial cluster, and then may or may not feature one or more consonants following the middle cluster. To facilitate programmatic analysis of the data, a "cluster length" shape was used in place of the previously used "CVCV" shape. A cluster length (CL) shape is a word shape where digits 1-6 represent lengths of consonant groups (including single C's) in a word. It is an alternative shape convention to the one using 'CV' symbols, where groups of 'C' symbols are substituted by a digit corresponding to the number of consonants in a particular group. For example, '221' is equivalent to 'CCVCCVC' (which, for example, is the shape for *człowiek* 'human being'). In *człowieka* in the table below we have an initial double cluster (2), one medial double (2) and one single consonant (1). The zero in the shape '2210' means "no final consonant(s) present". If a word starts with a vowel, the CL shape has a '0' in the first position. Table 10 below lists the first ten CL shapes for words where there are exactly two clusters of length 2, one initial and one medial. The items are ordered by the number of words that fit a specific shape. The total number of word types that fit these shapes is 92 024.

The word types belonging to this set were further analysed with respect to the features of the consonants present in their clusters. The word types were ordered according to the combinations of features occurring most commonly.

Table 11 shows ten most frequent feature combinations for the initial clusters in words where there are exactly two clusters of length 2, one initial and one medial.

Table 10. Top examples for the "two clusters of length 2, one initial and one medial" shape.

CL_shape	word form freq	most frequent example
2210	10710	*człowieka*
221	9396	*więcej*
2211	9148	*przepraszam*
2120	7125	*pieniądze*
2121	7008	*znalazłem*
220	6926	*prawda*
22110	6772	*prawdziwego*
21210	5220	*zrozumiano*
21120	4356	*wiadomości*
22111	4350	*zjednoczonych*

Table 11. Ten most frequent feature combinations for initial clusters.

word form frequency	CC	POA	MOA	voice	most frequent example
5736	pʂ	bilabial alveolar	plosive fricative	–V –V	*przepraszam*
1297	pr	bilabial alveolar	plosive trill	–V +V	*prawda*
1155	sp	dental bilabial	fricative plosive	–V –V	*spokojnie*
1000	st	dental dental	fricative plosive	–V –V	*statku*
734	sk	dental velar	fricative plosive	–V –V	*skończyć*
677	kr	velar alveolar	plosive trill	–V +V	*kręci*
667	tr	dental alveolar	plosive trill	–V +V	*trudno*
652	vj	labial palatal	fricative glide	+V +V	*więcej*
495	zd	dental dental	fricative plosive	+V +V	*zdejmij*
472	pj	bilabial palatal	plosive glide	–V +V	*pieniądze*

Table 12 below shows ten most frequent feature combinations for the medial clusters in words where there are exactly two clusters of length 2, one initial and one medial.

Table 12. Ten most frequent feature combinations for medial clusters.

type fr	CC	POA	MOA	Voice	Most frequent example
1403	nʦ	dental dental	nasal affricate	+V –V	*więcej*
837	st	dental dental	fricative plosive	–V –V	*prostu*
606	sk	dental velar	fricative plosive	–V –V	*blisko*
559	nt	dental dental	nasal plosive	+V –V	*prezydenta*
547	ɕʨ	alv-pal alv-pal	fricative affricate	–V –V	*mieście*
523	tk	dental velar	plosive plosive	–V –V	*środku*
501	ɕm	alv-pal bilabial	fricative nasal	–V +V	*mieliśmy*
472	ŋk	velar velar	nasal plosive	+V –V	*drinka*
460	vj	labial palatal	fricative glottal	+V +V	*człowiek*
454	mj	bilabial palatal	nasal glottal	+V +V	*zrozumieć*

The words where there are exactly two clusters of length 2, one initial and one medial have four pairs of cluster feature combinations in common. These four pairs are:

initial	vj	labial palatal	fricative glide	+V +V	lpfg11
medial	nc	dental dental	nasal affricate	+V –V	ddna10

initial	pʂ	bilabial alveolar	plosive fricative	–V –V	bapf00
medial	pr	bilabial alveolar	plosive trill	–V +V	bapt01

initial	zr	dental alveolar	fricative trill	+V +V	daft11
medial	bj	bilabial palatal	plosive glottal	+V +V	bppg11

initial	zm	dental bilabial	fricative nasal	+V +V	dbfn11
medial	ntş	dental alveolar	nasal affricate	+V –V	dana10

where the clusters in column two are the most frequent among each group. The sequences in column six, where the single-character symbols correspond to the features listed in columns three, four, and five, were used in the programmatic analysis as what might be called "cluster feature shapes".

Table 13 below shows the most frequent words representing each pair of feature combinations along with the numbers of word types whose initial and medial clusters represent the above pairs.

Table 13. Most frequent words with each pair of feature combinations.

most frequent example	initial_type_fr	medial_type_fr	initial_CC	medial_CC
więcej	652	1403	vj	nc
przepraszam	5736	258	pş	pr
zrobię	302	303	zr	bj
zmęczony	299	359	zm	ntş

In order to find out which feature combination pairs were represented by the largest number of word types, a further analysis was performed which revealed the following ranking.

Table 14. A ranking of feature combination pairs.

most frequent word type	initial feature shape	medial featureshape	total number of word types
przepraszam	bapf00	bapt01	218
więcej	lpfg11	ddna10	43
zmęczony	dbfn11	dana10	30
zrobię	daft11	bppg11	17

which appears to suggest that the word *przepraszam* ('I'm sorry') may be the most representative word type of the most common word shape in Polish, as the set it represents not only is the largest in size but also outnumbers the other sets by more than two factors.

Table 15 below shows first ten word forms (sorted by token frequency) of each set represented by the words in the ranking above.

Table 15. A ranking of most frequent words representing each feature combination set.

bapf00 + bapt01		lpfg11 + ddna10		dbfn11 + dana10		daft11 + bppg11	
przepraszam	135069	*więcej*	86170	*zmęczony*	4659	*zrobię*	27457
przeprosić	3843	*wierzący*	241	*zmęczona*	3376	*zrobione*	3821
przeprosiny	2368	*wierzących*	82	*zmęczeni*	667	*zrobienia*	3769
przeprowadzić	2270	*wierzącym*	66	*zmęczonego*	335	*zrobią*	3680
przepraszamy	2199	*wierząca*	63	*zmęczenie*	285	*zrobiony*	1048
przyprowadź	1850	*wiodący*	39	*zmęczone*	224	*zrobienie*	833
przyprowadzić	1183	*wiodącym*	38	*zmęczoną*	224	*zrobiono*	794
przepraszać	1171	*wiedzący*	35	*zmęczenia*	178	*zrobiona*	570
przeproś	834	*wiodącą*	29	*zmęczyło*	138	*zrobieniu*	227
przyprowadził	766	*wiejący*	29	*zmęczyłem*	136	*zrobieniem*	114
word types	218		43		30		17
total token frequency	170 145		87 141		11 160		42 761

The high token frequency of the word *przepraszam* in the table above may further confirm the status of this word as the most representative word type in Polish.

4. Discussion and conclusion

In this study we searched for the dominant shape of the Polish word. First, we hypothesised that cluster complexity is related to word saturation with clusters. It was expected that the more clusters a word contains, the shorter they are. Indeed, we have found that cluster length is inversely proportional to the number of clusters in a word. Second, we pondered about the co-occurrent patterns of clusters of different lengths. The results demonstrated that in the majority of cases clusters of the same size tend to co-occur (72.8%). Additionally, we observed that Polish words prefer saturation over complexity: three doubles in a word are preferred over two triples. Third, we predicted higher degree of diversification of MOA over POA within clusters. The prediction was grounded in the fact that manners of articulation are reflected in the sonority scale which is generally acknowledged to be responsible for the distinctive perception of sound sequences. Diversification by means of places of articulation is less functional, and may even be dysfunctional when assimilation of place occurs. As predicted, POA is significantly less diversified than MOA in all word positions, with as much as 73.13% POA identity word finally in double clusters. The degree of diversification of both POA and MOA rises with cluster size. This finding may serve as feedback to the NAD principle which is based on several criteria, including POA and MOA features. The results obtained in the present study sanction the future introduction of weights of these features in preferability calculations.

Fourth, we examined voice agreement in double clusters of all positions. According to the natural typological tendencies, initial clusters were expected to contain a voiceless obstruent and a sonorant. Similarly, word-finally, mixed clusters were presumed to occur, both due to the natural universal (or typological) preference of sonorants and obstruents. Moreover, the language-specific process of *Auslatverhärtung* leads to the pronunciation of a voiceless obstruent. Medially, a greater tolerance or freedom in the voice constraint was expected, but because of the intervocalic position, predominance of voiced clusters was expected. The results confirmed the prediction in two contexts: word-medial clusters turned out to be voiced or mixed, and word final clusters turned out to be mixed. Word-initially, voiceless clusters prevailed, which may partially be attributed to the role of morphology, i.e., the phonological shape of prefixes.

In conclusion, we have identified several characteristics of the representative Polish word. The statistics based on word forms indicate that Polish words with one cluster prevail, in which case the cluster is most likely to be medial. The next most common pattern is a word with two clusters which tend to be of equal size; in this case medials co-exist with initials. Moreover, three double clusters are preferred over two triple ones.

Having established that the dominant word shape is that containing two clusters, one word-initially and one word-medially, we searched for prevalent feature combinations in these positions. In the word-initial position, the dominant articulatory set-up is a voiceless bilabial plosive (as C1), followed by a voiceless alveolar fricative (as C2), which is represented by cluster /pʂ/. Word-medially, the winner is a voiced dental nasal followed by a voiceless dental affricate, i.e., /nʦ/. We did not find a word that would fulfill both conditions simultaneously, however, when we allowed some flexibility and extended the ranking to 10 most favoured feature arrangements, we found that 4 feature pairs qualify and match one of the 10 preferences in each word position. These include: (1) a voiceless bilabial plosive followed by a voiceless alveolar fricative + a voiceless bilabial plosive followed by a voiced alveolar trill in *przepraszam* 'I'm sorry', (2) a voiced labiodental fricative followed by a voiced palatal approximant + a voiced dental nasal followed by a voiceless dental affricate in *więcej*, (3) a voiced dental fricative followed by a voiced alveolar trill + voiced bilabial plosive followed by a voiced palatal approximant in *zrobię* 'I will do', and (4) voiced dental fricative followed by a voiced bilabial nasal + voiced dental nasal followed by a voiceless alveolar affricate in *zmęczony* 'tired'.

In the present contribution, we explored the entire SUBTLEX-PL corpus in search of phonotactic generalisations pertaining to the formulated hypotheses. However, one must bear in mind that in all of the calculations performed so far, the role of morphology has not been acknowledged, revealed, or controlled. Morphology constitutes another potential effect on feature distribution and the phonological layout of the clusters. It should be verified whether the impact of a given feature (MOA, POA, voice, sonorant/obstruent) on the composition of a cluster is regulated by morphology. Therefore, there is a dire need to continue this line of research taking into account the morphological criterion. The first step to gain an insight into the internal structure of clusters has already been taken and preliminary results, albeit obtained from word-initial double cluster types, revealed significant differences between feature distribution in clusters which are purely lexical, purely morphonotactic and mixed.

REFERENCES

Bargiełówna, Maria. 1950. Grupy fonemów spółgłoskowych współczesnej polszczyzny kulturalnej. *Biuletyn Polskiego Towarzystwa Językoznawczego* 10. 1–25.

Baumann, Andreas & Daan Wissing. 2018. Stabilising determinants in the transmission of phonotactic systems: Diachrony and acquisition coda clusters in Dutch and Afrikaans. *Stellenbosch Papers in Linguistics Plus* 55. 77–107. DOI: 10.5842/55-0-780

Bethin, Christina Y. 1992. *Polish syllables: The role of prosody in phonology and morphology*. Slavica Publishers.

Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41. 977–990. DOI: 10.3758/BRM.41.4.977

Brysbaert, Marc, Boris New & Emmanuel Keuleers. 2012. Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods* 44. 991–997. DOI: 10.3758/s13428-012-0190-4

Brysbaert, Marc, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte & Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology* 58(5). 412–424. DOI: 10.1027/1618-3169/a000123

Cyran, Eugeniusz. 2006. Polish and English syllable structures. How different are they? *Zeszyty Wszechnicy Świętokrzyskiej. Filologia Angielska* 1(23). 151–160.

Cyran, Eugeniusz & Edmund Gussmann. 1999. Consonant sequences and governing relations: Polish initial consonant sequences. In Harry van der Hulst & Nancy Ritter (eds.), *The syllable: Views and facts*, De Gruyter Mouton. 219–247. DOI: 10.1515/9783110806793.219

Dobrogowska, Katarzyna. 1984. Śródgłosowe grupy spółgłosek w polskich tekstach popularnonaukowych [Word-internal clusters in popular science texts]. *Polonica* X. 5–34.

Dobrogowska, Katarzyna. 1990. Word internal consonant clusters in Polish artistic prose. *Studia Phonetica Posnaniensia* 2. 43–67.

Dobrogowska, Katarzyna. 1992. Word initial and word final consonant clusters in Polish popular science text and in artistic prose. *Studia Phonetica Posnaniensia* 3. 47–121.

Dressler, Wolfgang U. & Katarzyna Dziubalska-Kołaczyk. 2006. Proposing morphonotactics. *Rivista di Linguistica* 18(2). 249–266.

Dukiewicz, Leokadia. 1985. Nagłosowe grupy spółgłoskowe w polskich tekstach popularnonaukowych i prasowych [Onset consonant clusters in Polish popular science and press texts]. *Studia Gramatyczne* 6: 17–34.

Dziubalska-Kołaczyk, Katarzyna. 2014. Explaining phonotactics using NAD. *Language Sciences* 46. 6–17. DOI: doi.org/10.1016/j.langsci.2014.06.003

Dziubalska-Kołaczyk, Katarzyna. 2019. On the structure, survival and change of consonant clusters. *Folia Linguistica Historica* 53(s40–1). 107–128. DOI: 10.1515/flih-2019-0006

Dziubalska-Kołaczyk, Katarzyna, Michał Jankowski & Piotr Wierzchoń. 2011. Classification of the lexicon of modern Polish according to the structure of consonant clusters. In Wai Sum Lee (ed.), *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII): August 17-21, 2011*, City University of Hong Kong. 619–622.

Dziubalska-Kołaczyk, Katarzyna & Paulina Zydorowicz. 2014. The production of high-frequency clusters by native and non-native users of Polish. *Concordia Working Papers in Applied Linguistics* 5. 130–144.

Greenberg, Joseph H. 1978. Some generalisations concerning initial and final consonant clusters. In Joseph H. Greenberg (ed.), *Universals of human language. Vol. II: Phonology*, Stanford University Press. 243–279.

Gussmann, Edmund. 2007. *The phonology of Polish*. Oxford University Press.

Gussmann, Edmund & Eugeniusz Cyran. 1998. Polish consonantal sequences: A phonological testing ground. In Eugeniusz Cyran (ed.), *Structure and interpretation. Studies in phonology*, Folium. 127–138.

Heuven, Walter J. B. van, Paweł Mandera, Emmanuel Keuleers & Marc Brysbaert. 2014. Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* 67(4). 1176–1190. DOI: 10.1080/17470218.2013.850521

Jankowski, Michał. 1994. Practical automatic phonemic transcription systems. *Studia Anglica Posnaniensia* 28. 143–150.

Jarosz, Gaja. 2010. Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of Child Language* 37(3). 565–606. DOI: 10.1017/s0305000910000103

Kuryłowicz, Jerzy. 1952. Uwagi o polskich grupach spółgłoskowych. *Biuletyn Polskiego Towarzystwa Językoznawczego* 11. 54–69.

Mandera, Paweł, Emmanuel Keuleers, Zofia Wodniecka & Marc Brysbaert. 2014. Subtlex-pl: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods* 47. 471–483. DOI: 10.3758/s13428-014-0489-4

Orzechowska, Paula. 2009. *Morphonotactics in English and Polish. A dictionary- and corpus-based study of word-final consonant clusters*. Ph.D. dissertation, Adam Mickiewicz University Poznań.

Orzechowska, Paula. 2019. *Complexity in Polish phonotactics: On features, weights, rankings and preferences*. Springer Singapore.

Orzechowska, Paula & Paulina Zydorowicz. 2019. Frequency effects and markedness in phonotactics. *Poznań Studies in Contemporary Linguistics* 55(1). 157–179. DOI: 10.1515/psicl-2019-0006

Paulsson, Olof. 1969. Das Phonemsystem der modernen polnischen Literaturschprache. *Scando-Slavica* 15(1). 216–236. DOI: 10.1080/00806766908600523

Rochoń, Marzena. 2000. *Optimality in complexity: The case of Polish consonant clusters*. Akademie Verlag.

Rubach, Jerzy & Geert Booij. 1990. Syllable structure assignment in Polish. *Phonology* 7(1). 121–158. DOI: 10.1017/S0952675700001135

Soares, Ana Paula, João Machado, Ana Costa, Álvaro Iriarte, Alberto Simões, José João de Almeida, Montserrat Comesaña & Manuel Perea. 2015. On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *Quarterly Journal of Experimental Psychology* 68(4). 680–696. DOI: 10.1080/17470218.2014.964271

Schwartz, Geoffrey. 2019. *When is a cluster really a cluster?* Paper presented at the Approaches to Phonetics and Phonology conference, Lublin, 21–23 June 2019.

Szpyra, Jolanta. 1992. Ghost segments in nonlinear phonology: Polish yers. *Language* 68(2). 277–312. DOI: 10.2307/416942

Szpyra, Jolanta. 1995. *Three tiers in Polish and English phonology*. Wydawnictwo UMCS.

Wiese, Richard, Paula Orzechowska, Phillip M. Alday & Christiane Ulbrich. 2017. Structural principles or frequency of use? An ERP experiment on the learnability of consonant clusters. *Frontiers in Psychology* 7: 2005. DOI: 10.3389/fpsyg.2016.02005

Woliński, Marcin, Zygmunt Saloni, Robert Wołosz, Włodzimierz Gruszczyński, Danuta Skowrońska & Zbigniew Bronk (eds.). 2020. *Grammatical Dictionary of Polish*. (4th edn online (pre-release).) Institute of Computer Science PAS.

Zydorowicz, Paulina. 2010. Consonant clusters across morpheme boundaries: Polish morphonotactic inventory and its acquisition. *Poznań Studies in Contemporary Linguistics* 46(4). 565–588. DOI: 10.2478/v10010-010-0028-0

Zydorowicz, Paulina. 2019. *Polish (mor)phonotactics in first language acquisition, connected speech and cluster processing*. Wydawnictwo Naukowe UAM.

Zydorowicz, Paulina & Katarzyna Dziubalska-Kołaczyk. 2017. The dynamics of marked consonant clusters in Polish. In Elena Babatsouli (ed.), *Proceedings of the International Symposium on Monolingual and Bilingual Speech 2017*, ISMBS. 318–324.

Zydorowicz, Paulina & Paula Orzechowska. 2017. The study of Polish phonotactics: Measures of phonotactic preferability. *Studies in Polish Linguistics* 12(2). 97–121. DOI: 10.4467/23005920SPL.17.005.7023

Zydorowicz, Paulina, Paula Orzechowska, Michał Jankowski, Katarzyna Dziubalska-Kołaczyk, Piotr Wierzchoń & Dawid Pietrala. 2016. *Phonotactics and morphonotactics of Polish and English: Theory, description, tools and applications*. Wydawnictwo Naukowe UAM.