CLASSIFICATION OF THE LEXICON OF MODERN POLISH ACCORDING TO THE STRUCTURE OF CONSONANT CLUSTERS

Katarzyna Dziubalska-Kołaczyk, Michał Jankowski & Piotr Wierzchoń

Adam Mickiewicz University, Poznań, Poland
dkasia@ifa.amu.edu.pl; mjank@amu.edu.pl; wierzch@amu.edu.pl

ABSTRACT

The aim of the project this paper reports on is to identify and analyze the correlation between Polish consonant clusters and the semantic categories of the words that contain a given cluster. The material studied is a large corpus of Polish newspaper text. A common characteristic of sets of words is whether or not they contain an identical consonant cluster, which is understood here as a contiguous string of consonants not interrupted by a vowel. The aim of the study is to find other common characteristics of the words, that is the identification and analysis of a correlation between the cluster and the semantic categories that the words containing the cluster belong to. The semantic categories considered in the model pertain to the following properties of the words featuring a common cluster: derivation, part of speech, inflection, and morphotactics.

Keywords: consonant cluster, frequency, ontology, Polish language, phono-morpho-semantics

1. CURRENT STATE OF RESEARCH

What we deal with here is grouping of the lexicon into sets of words called *derivational nests*, which are related in terms of their morphology (as well as their semantics, as a consequence). Nest analysis of Slavonic languages has been reported in [23], whereas [13] and [21], among others, report studies of Polish within nest theory. The idea of the nest view for Polish was first put forward in the 50s by Wierzchowski [24], who studied what he called *proportional groups* of words.

The algorithmization of morphological research on English pertains to two issues: the algorithms for recognizing the intermorpheme boundaries and the stemming algorithms. An example of the former is Harris' algorithm described in his article *From Phoneme to Morpheme* [10]. Harris' idea was picked up in [19], and used to automatize the production of an inventory of morphemes. The issue of affix productivity in English was reported in [2, 3, 11]. Stemming algorithms are for English what the nest approach is in the study of Slavonic languages. Stemming, which makes it possible to collect derivatives of words into sets first started with the work of Harwood and Wright (reported in [18]), who, based on the analyses of Harris [9] and Lovins [17], produced a series of such algorithms (cf. [8, 12, 20]). In terms of morpheme position, the status of prefixes and suffixes for some African languages was studied by Alemayehu and Willett [1]. In the area of phonotactics and morphotactics (cf. [4, 22]) consonant clusters have been studied in terms of their position in a word, length, morphotactic boundaries and their systemic adequacy analysed in the context of the *Optimal Net Auditory Distance (NAD) Principle* (cf. [6, 7] and references given there); and as reported in [5] on German verbs.

2. TERMINOLOGY

This study is phonetic, statistical and semantic in character. Its core is the analysis of the correlation between consonant clusters and the semantic categories of the words which contain these clusters. The categories will be discussed in the section titled **The Semantic Model**. The aim of the project is the formulation of more detailed tacto-phonic and language-sensitive rules within what Levelt [16] termed the *formulator* component of his model of speech processing. The material studied was a word list extracted from a corpus of raw newspaper text rather than words in their dictionary form (lemmas).

2.1. Research assumptions

The research hypothesis is as follows: there seems to be a relationship between elements of a set of consonant clusters and elements of a set of semantic categories. Furthermore, there is a typology of these relationships. That is to say that these correlations can be classified in a certain way, according to properties relevant to the model used, such as the category of words which belong to one part of speech, the category of words which belong to one derivational nest, etc. Obviously, not all consonant clusters will denote a class of words, members of one semantic category. For example, the Polish cluster -st- is present in 26,590 words of the corpus studied. Classifying over 26K words into semantic groups seems pointless, because it is a task impossible to carry out in practice and also not worthwhile ontologically. On the other hand, the cluster -tɕstf- maps a single category of words, such as *ekouchodźstwa* (“eco-emmigration” gen. sg.), *uchodźstwo* (“refugee status” n), *wychodźstwem* (“emmigration” inst), all of which belong to the derivational nest of the words *uchodzić* (“escape” v) and *wychodzić* (“leave” v) – verbs derived from *chód* (“stride, walk” n). The cluster -ksxw- maps exactly one word in the corpus: *ekschłopaka* (“ex-boyfriend” gen.sg). And one more example: another 4-phoneme cluster -rʃʧk- maps just two words: *zmarszczka* (“wrinkle” n) and *przeciwzmarszczkowy* (“anti-wrinkle” adj), which belong to the semantic nest of the word *zmarszczka* (“wrinkle” n). No other words belong to this category.

2.2. The studied material

The set of Polish consonants is the following (cf. [15]): b, ʦ, ʧ, ʨ, d, ʤ, f, g, ɟ, x, j, (j̃), k, c, l, m, n, ŋ, ɲ, p, r, s, ʃ, ɕ, t, v, w, (w̃), z, ʣ, ʑ, ʥ, ʒ (31+2 phonemes). The set of Polish vowels is the following: a, e, i, o, u, ɨ. A Polish word is a sequence of characters that are members of either of the two sets. A consonant cluster in a Polish word is a contiguous string of consonants not interrupted by a vowel.

2.3. The semantic model

The semantic model used in the study assumes four categories: (a) derivation, (b) part of speech, (c) inflection, (d) morphotactics. In other words, the Polish lexicon (in the form of a corpus of newspaper text) is studied in terms of application of the four above mentioned criteria to the form of consonant clusters.

2.3.1. The derivation submodel

The derivation submodel features the category *derivation nest* (derivation paradigm) [13, 21, 23], which is a set of words which are dependent synchronically on a non-derivative word, i.e., they are formed by derivation mechanisms available in a particular language (eg., affixation, negative derivation, etc). For example, the list: features a set of derivatives of the non-derivative word *Afryka* (“Africa”), and is a slighlty modified textbook example (cf. [13]). A similar list from the corpus: *grejpfrut* (“grapefruit” nom.sg), *grejpfruta* (“grapefruit” gen.sg), *grejpfrutach*, *grejpfruty* (“grapefruit” nom.pl), *grejpfrutów* (“grapefruit” gen.pl), *grejpfrutówce* (“grapefruit vodka” dat.sg) features words which all contain the cluster -jpfr-, which does not occur in any other words in the corpus. Thus we can say that the cluster -jpfr- maps the category “the derivation nest GREJPFRUT”.

AFRYKA	**“**Africa” n fem
Afrykanin	“African” n masc
Afrykanka	“African” n fem
afrykański	“African” adj
afrykanistyka	“African studies” n fem
afrykanista	“African studies scholar” n masc
afrykanistka	“African studies scholar” n fem
afrykanizacja	“africanization” n fem
afrykanizować	“africanize” v

2.3.2. The part-of-speech submodel

The part-of-speech submodel features the category *part of speech*, which is a category of words which can be identified because of their specific (semantic, morphological, syntactic) characteristics. Traditionally, Polish features the following parts of speech: Noun, Verb, Adjective, Adverb, Pronoun, Numeral, Preposition, Conjunction, Particle, Exclamation, a division that goes back to the Greco-Roman tradition. Example: w̃stc = *anty-cząstki* (“anti-particle” n.pl), *chrząstki* (cartilage pl), *cząstki* (“particle” n.pl), *piąstki* (“fist” n.pl dim). In the corpus studied the cluster -w̃stc- maps the category “POS Noun”. Example: ntʃr = *wewnątrzredakcyjna* (“internal to the editorial team” adj), *wewnątrzresortowa* (“interdepartmental” nom.sg), *wewnątrzresortowych* (“interdepartment-al” gen.pl), *wewnątrzrosyjska* (“inter-Russian” nom.sg fem), *wewnątrzrosyjskie* (“inter-Russian” nom.sg neut). In the corpus studied the cluster -ntʃr- maps the category “POS Adjective”.

2.3.3. The inflection submodel

The inflection submodel features the category: *forms of a lexeme which belong to one inflectional paradigm*. An inflectional paradigm is a set of all words which are in a relation of inflectional opposition to each other, i.e., they differ solely in terms of their inflectional parameters, specific to a given part of speech (cf. also [7]). For example, in the case of the Polish Noun these are Number and Case. In the case of the Polish Adjective these are Number, Case and Gender as well as Degree. For Adverb, it is only Degree (eg. *silnie*, *silniej*, *najsilniej* (“strongly, more strongly, most strongly”)). As can be seen, the present model preserves the distinction between derivation and inflection. Example: -tfst- = *przedwstępna*, *przedwstępne*, *przedwstępnej*, *przedwstępny*, *przed-*

wstępnych, *przedwstępną* (various case forms of "preliminary" adj). In the corpus studied the cluster -tfst- maps the category of "the inflectional paradigm PRZEDWSTĘPNY". Naturally, it is not necessary for a given cluster to map all the possible inflectional forms of a paradigm. In reality, only a part of the paradigm is actually recorded.

2.3.4. The morphotactics submodel

The morphotactics submodel features the category *words with a morphological boundary within a cluster*. Some words in a language are viewed as non-derivatives, in which case it is assumed that they are indivisible morphologically, e.g. *kształt* ("shape" n). Some such words may form derivative units by binding with certain morphemes, eg. *bezkształt* ("shapelessness"), *odkształcić* ("dis-shape"), *przekształcić* ("transform"), *kształtność* ("shapliness"). As a result, a morphological boundary within a word appears, eg. *bez-kształt*, *od-kształcić*, *prze-kształcić*, *kształtn-ość*. Some of these boundaries cut through consonant clusters[1]:

s|kʃ = *bez-kształt*
t|kʃ = *od-kształcić*

A morphological boundary may thus appear within a consonant cluster which maps a specific set of words, as well as within each of the words mapped that way. Example: z|vzgl = bez|względna ("absolute" adj nom.fem), bez|względnego ("absolute" adj gen.masc), bez|względnej ("absolute" adj gen.fem), bez|względnie ("absolute" adv), bez|względność ("absoluteness" n), bez|względnymi ("absolute" adj instr.pl). In the corpus studied the cluster -zvzgl- maps the category of words (all of) which feature a morphological boundary, i.e. the prefix *bez-* ("without, no, non-") and the basic form *wzgląd* ("consideration" n) (in various part-of-speech forms). A morphological boundary may run between lexical morphemes, as does -s|ɕtɕ- in *bezściółkowy* ("not involving mulch" adj), or between a lexical and a grammatical morpheme, as does m|k in *domku* ("house" dim loc.sg).

3. PREPARATION OF MATERIAL

The corpus studied in this project is a collection of newspaper articles spanning two years (over 100,000 files and approx. 48.5 million running words). The motivation for the choice of a nation wide newspaper was as follows: (a) the text is easy to access and process (plain ascii text), (b) subject scope (politics, culture, sports, hobbies, etc) guarantees a wide lexical spectrum. The corpus features over 600,000 types (unique words) transcribed using tools extended from those reported in [14].

4. QUANTITATIVE RESULTS

Apart form identifying semantic categories in a text corpus by examining the properties of the clusters contained in the words, it is possible to make additional quantitative observations by taking advantage of the frequency data available. For example, based on the data in the following format (unique word and its frequency in the corpus) one

domki 65	domknięciu 1	domknąć 5
domkiem 12	domknięto 1	domku 86
domknięcie 4	domknięty 1	domków 90[2]
domknięciem 1		

observes that the cluster -mk- maps 10 types and -mkn- maps 6 types. Further, it is easy to obtain a total number of words (tokens) mapped by a cluster. For the data shown here: -mk- = 266 (65 + 12 + 4 + 1 + 1 + 1 + 1 + 5 + 86 + 90), -mkn- = 8 (4 + 1 + 1 + 1 + 1). With data available for tokens and types, it is possible to calculate a "quantitative cluster efficiency" index (QCE): -mk- = 266/10 = 26.6, -mkn- = 8/6 = 1.3. For example, the QCE of -jpfr- (as in *grejpfrut*) is 1.5 (9/6). In all, more than 1700 clusters were extracted from the corpus and the distribution observed was as shown in Table 1, where D, I, M, P stand for "derivation", "inflection", "morphotactics", and "part of speech" classes, respectively, and N denotes "no category", based on an analysis of each group of clusters. The %N column shows a percentage of the "no category" clusters in each length group.

Table 1: Unique clusters according to length and category.

cluster length	number of clusters	D	I	M	P	N	% N
2	487	22	2	5	12	451	92,61
3	970	168	76	164	206	300	30,93
4	224	63	39	73	82	28	12,50
5	47	1	4	22	23	0	0
6	3			2	3	0	0
Total	1731						

It is clear for example that the longer the cluster, the more likely it is for it to be associated with only one or just a few tokens, types and lexical categories. This can, in fact, be deduced from quantitative data before going into qualitative category analysis. One quick statistic that can be very useful in this preliminary analysis is what we call the *max* of a cluster. *Max* is simply the number of occurrences in the corpus of the most frequent word that contains a given cluster. The longer clusters seem to appear in fewer words and among the fewer words there seems to be one with a relatively high frequency (the *max*), which we identified as a case of the "strong default". For example the cluster -ntpl- (10817 tokens, 32 types) is represented 6888 times by *wątpliwości* ("doubt" n.pl) (63,7%), whereas the second most frequent example *niewątpliwie* ("doubtless" adv) appears 1910 times (17%). Needless to say, all the examples for -ntpl- belong to one lexical category grouped around the stem -*wątpl*- which appears in such words as *wątpliwość* ("doubt" n), *wątpliwy* ("doubtful"), *niewątpliwie* ("doubtless" adv), etc. This is what we like to call a "superdefault". The ratio of *max* to cluster frequency seems to confirm the observation that longer clusters map fewer categories, fewer types, and fewer tokens. It is easy to confirm this by looking at the average values of max/token and type/token ratios (which we might call measures of *lexical load*) for clusters depending on cluster length, as shown in the following table.

Table 2: Correlation between cluster length and lexical load.

cluster length	max/token	type/token
2	23.4	6.7
3	29.5	19.5
4	35.7	23.6
5	60.8	45.9

A more important observation can be made about the form of clusters with respect to the category they belong to. Clusters classified as **D** (such as, eg., -brvj- as in *brwiowy*, *brwiach* ("eye-brow") etc.) tend to cover a substantial part of the lexical stem of the word. **P** clusters (such as, eg., -fʨʨ- as in *sprawdźcie* ("check" v.imp.pl), *krzywdźcie* ("hurt" v.imp.pl) tend to cover a substantial part of the grammatical ending. On the other hand, **I** clusters such as -ʦkɲ- (as in *ocknie* ("come to" v.3.sg), *ocknięciu* ("coming to" n.loc.sg), *ockniemy* ("come to" v.1.pl) seem to map groups of 3-5 (usually quite frequent) word forms. And, finally, clusters classified as both **D** and **P**, such as -ɲʧm-, as in *skończmy*, *zakończmy*, *dokończmy*, *kończmy* (variants of "let's finish"), play an important role in shaping two significators of the words: lexical and grammatical.

5. FUTURE RESEARCH

Both the semantic categories and the quantitative data can be used in further research in such diverse areas as: analysis of correlation between cluster categories and the NAD [7], automatic recognition of neologisms, selection of example material for phonetics courses.

6. REFERENCES

[1] Alemayehu, N., Willett, P. 2002. Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing* 17(1), 1-17.

[2] Baayen, H. 1994. Productivity in language production. *Language and Cognitive Processes* 9, 447-469.

[3] Baayen, H., Lieber, R. 1991. Productivity and English derivation: A corpus based study. *Linguistics* 2, 801-843.

[4] Bargiełówna, M. 1950. Grupy fonemów spółgłoskowych współczesnej polszczyzny kulturalnej. *Biuletyn Polskiego Towarzystwa Językoznawczego* 10, 1-25.

[5] Beedham, C. 2005 Eine phonotaktische verbindung zwischen starken verben und grammatischen wörtern der deutschen gegenwartssprache. *Deutsch als Fremdsprache* 42(3), 167-172.

[6] Dressler, W.U., Dziubalska-Kołaczyk, K. 2006. Proposing morphonotactics. *Wiener Linguistische Gazette* 73, 1-19.

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

[8] Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27, 153-198.

[9] Harris, Z. 1951. *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

[10] Harris, Z. 1955. From phoneme to morpheme. *Language* 31(2), 190-222.

[11] Hay, J. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39, 1041-1070.

[12] Jacquemin, C., 1997. Guessing morphology from terms and corpora. *Actes, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* Philadelphia, 156-165.

[13] Jadacka, H. 1995. *Rzeczownik Polski Jako Baza Derywacyjna*. Warszawa: Wydawnictwo Naukowe PWN.

[14] Jankowski, M. 1994. Practical automatic phonemic transcription systems. *Studia Anglica Posnaniensia* 28, 143-150.

[15] Jassem, W. 2003. Polish. *Journal of the International Phonetic Association: Illustrations of the IPA* 33(1), 103-107.

[16] Levelt, W. 1989. *Speaking: From Intention to Articulation*. Cambridge, Massachusets: The MIT Press.

[17] Lovins, J.B. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Ling.* 11, 22-31.

[18] Nagórko, A. 1974. Stan i perspektywy badań ilościowych w słowotwórstwie opisowym. *Poradnik Językowy* 1, 1-13.

[19] Neuvel, S., Fulop, S. 2002. Unsupervised learning of morphology without morphemes. *Proceedings of the Workshop on Morphological and Phonological Learning of ACL/SIGPHON-2002*, 31-40.

[20] Porter, M.F. 1980. An algorithm for suffix stripping. *Program* 14, 130-137.

[21] Skarżyński, M. 1981. *Tworzenie Wyrazów w Języku Polskim*. Kielce: IKN ODN.

[22] Śledziński, D. 2010. Analiza struktury grup spółgłoskowych w nagłosie oraz w wygłosie wyrazów w języku polskim. *Kwartalnik Językoznawczy* 3-4, 62-83.

[23] Tichonov A. 1985. *Slovoobrazovatel'nyj slovar' russkogo jazyka*. Moskwa.

[24] Wierzchowski, J. 1959. Uwagi słowotwórczo-leksykalne. *Biuletyn Polskiego Towarzystwa Językoznawczego* XVIII, 223-229.

[1] Morpheme boundaries are marked with a vertical bar and only when it is relevant to our morphotactics submodel.

[2] *domki*, *domkiem*, *domku*, *domków* are forms of *domek* ("house" dim) and *domknięto*, *domknięty*, *domknięcie*, *domknięciem*, *domknięciu*, *domknąć* are all members of the POS DOMKNĄĆ ("close shut") category.