Ariel Drozd, Corpus linguistics, winter 2023/24

Creating a test corpus for developing and validating new rules regarding automatic processing of compound words relating to numbers and/or quantity. Part of preliminary corpora exploration.

Conventions of notation

Lexemes are represented as their base form written in small caps, eg: TRÓJSTRONNY, TRILATERAL.

Non-regex strings (including morphemes and bases), are written in italics, eg: *trzy-*, *-stronny,* .

Regexes are written in a monospace font, eg: `tr(i|y)`.

Definitions of the lexemes, if needed (eg. for disambiguation or when a corresponding lexeme is not easily found in English), are written in single quotation marks, eg: ‘definition’.

n (unlike in mathematical convention, where it usually represents a natural number!) represents any quantity, including fractions or undefined values. On rare occasions where an actual natural-only number is meant, the letter *m* is used instead of *n*.

Context of the study

This is a preliminary research that aims to prepare the author for writing their MA thesis. In general, the focus of the thesis are compound words containing segments related to numerals (cardinal or ordinal) and non-numeral quantifying expressions. Specifically, the MA thesis will aim to propose a set of improved rules of automatic morphological analysis and sense disambiguation of

"homologue words" (to be explained below) in Polish language, as well as to explore possibilities of automatically creating and displaying their lexicographical representations (definitions, exemples, flexion etc.) on-demand in digital dictionaries, based on the morphological composition of a particular query – replacing the need for manually creating separate dictionary entries for every possible realization of a recurring pattern. Ideally, a new lexical resource (compatible with existing lexical databases and grammatical analysers) would be created during the author's third cycle of studies, refining and implementing theoretical propositions made in the MA thesis.

All this requires i.al. creating a test corpus with samples containing some of the words in focus, as well as the ones only resembling them superficially, and potentially borderline examples of those two categories. Such corpus would allow for more convenient testing of the modifications made in the morphosyntactic tagger: observing practical consequences and efficiency of various approaches is the easiest in an environment where the targeted words are grossly overrepresented. Before the test corpus is created – and before the scope of the MA thesis is ultimately narrowed-down, clearly defined and sealed – it is necessary to explore the (so far) vaguely named "compound words relating to numbers and/or quantity" in existing corpora, dictionaries and other resources.

The present paper is the first part of the corpora exploration, with 3 types of documentation intended: tables containing the research queries, tables gathering examples found in the reference corpus, and a part of the target test corpus created on the basis of those examples. All of it will be used to refine assumptions and methods before moving on to conducting research for the actual thesis.

Certain terms need to already be established now as working definitions, but they are likely to be adjusted before the thesis is written.

Key definitions and concepts

Segment is used, for lack of a better word, in reference to a string containing one or more morphemes which has a global lexical meaning, but is not necessarily a whole word on its own. Segment can be described as "**linked**" if it is not

separated from other segments with a whitespace on both sides. An example of a well established grammatical class that could be referred to as a "linked segment" is the "adja" (ad-adjectival adjective) class from NKJP tagset. Segments from this class, such as CZARNO 'black', have practically the same meaning and function as corresponding regular adjectives, but only appear if linked by a hyphen to other adjectives (eg. CZARNO-BIAŁY, Eng, BLACK-AND-WHITE). While processed with NLP tools, each of those segments can be assigned its own morphonsyntactic and semantic interpretation which – together with the position of one relative to another – contributes to the global interpretation of the whole compound word. Thanks to that, there is no need to manually account for all the possible color or language combinations in order to process regular hyphenated compound adjectives such as POLSKO-UKRAIŃSKI (Eng. POLISH-TO-UKRAINIAN or 'happening between or related to both Poland and Ukraine').

While writing their thesis, the author is likely to propose a new class that would encompass quantifying and ordinal expressions, but only those forms which need to be directly linked to another segment (similarly to the "adja" class reserved for bound-only forms of adjectival segments), barring spelling errors. The main morphological difference between the intended new class and the existing "adja" class would be the lack of hyphen linking the segment: it would be bound directly and seamlessly to the neighboring token. This approach may be controversial (usually every string in between two whitespaces is considered to be one word / segment, however morphologically complex it may be), but it has its precedents, such as "agglutinate BYĆ"[1] class from the NKJP tagset.

Members of this hypothetical new class are referred to as "segments" throughout the paper, in contrast with "words" or "expressions" (in case of morphologically independent forms of quantifiers or ordinals). A general word "segment" has been chosen, because their possible classification as "(complex) morphemes", "roots",

[1] In NKJP tagset, past forms such as *siedziałem* (1st person singular, past tense of TO SIT) are interpreted as two segments: *siedział*|*em*: l-participle of SIEDZIEĆ (TO SIT) and agglutinate BYĆ (TO BE).

"preffixoids"[2], "affixes" etc. is controversial, especially considering that in some cases, inclusion of interfixes *-o-* and *-u-* in the segment may be necessary for technical reasons. It is worth noting that they are regarded as individual "segments" in the context of this paper, however they are not treated as such by existing linguistic resources.

As of today, those quantifying or ordinal segments are mostly being processed as one entity together with the segments they are linked to. In some cases, additional rules have already been added to Morfeusz morphological analyser to better deal with certain regularly composed words not included in SGJP (Grammatical Dictionary of Polish). However, this accounts only for the morphosyntactic processing (lemmatization, grammatical interpretation, syntax parsing etc.), and has been done for certain groups of words only. As for word sense disambiguation, a key lexical resource for Polish – Słowosieć (Polish WordNet, plWN in short) – currently requires a manual input for every single word and every single relation[3] in order both to display a dictionary entry to its "human users", and to map a token in a text with its semantic representation in the lexical database. By contrast, it is possible to semantically disambiguate hyphenated compound adjectives without having all of their possible combinations entered into the database – precisely thanks to the fact that they are treated as two separate segments.

Quantifying segment, or **q-segment** for short, is proposed as a term for 'a linked segment referring to a certain quantity', such as *trzy-, trój-, tri-, try-* denoting a quantity of 3, or *dwudziestotrzy-* denoting a quantity of 23. Quantity may be understood here broadly as any magnitude (size) or multiplicity (amount). Quantifying segments do not necessarily point to a defined numerical value, eg. *wielo-*, *multi-*, *poli-* denote simply the concept of "many".

---

[2] As proposed by Jadacka H. in chapter *2.2. Złożenia* of *Kultura języka polskiego. Fleksja, słowotwórstwo, składnia.* Wydawnictwo Naukowe PWN, Warszawa 2013. ISBN 978-83-01-14398-5.

[3] In fact, plWordNet's inconsistent and incomplete entries describing "various entities of an age expressed in years" were a direct inspiration for this research (for example, PÓŁTORAROCZNIAK has separate meanings for a 1,5-year-old human child and a 1,5-year-old animal, while JEDNOLATEK has separate entries for a 1-year-old human, animal and, additionally, a 1-year-old plant; PIĘĆDZIESIĘCIOPAROLATEK is missing from the resource despite PIĘĆDZIESIĘCIOLATEK CZTERDZIETOPAROLATEK and CZTERDZIESTOLATEK all being included).

Ordinal segment, or **o-segment** for short, is proposed for 'a linked segment referring to a place in an ordered series', such as *trzecio-* or *tert-* denoting the third place. At the moment of writing, the author can provide only one example of an ordinal segment that is not related to a specific numeric value: suffixoid *ostatnio-* 'last' as used in OSTATNIOLIGOWY 'of the last ligue'.

Numeral segment, or **n-segment** for short, is proposed as a name for 'a linked segment related to any numerical value, including fractions'. It may be quantifying (*trzy-*) or ordinal (*trzecio-*): they are both broadly related to the number 3. As of this time, the author is not able to imagine a non-ordinal, non-quantifying numeral segment, but a possibility of such segments existing is taken into account due to the preliminary nature of the study.

Q/o segment is used as a more concise equivalent of "quantifying or ordinal segment". It encompasses collectively all of the three types of segments described above (unless the study leads to discovering numeral segments that are neither quantifying nor ordinal).

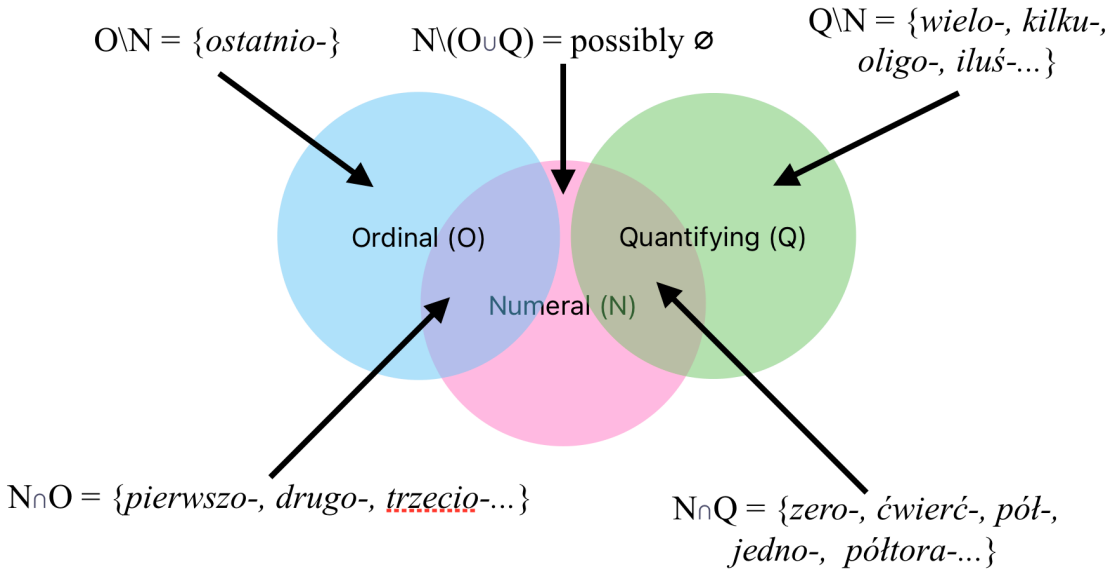

Fig. 1 Venn diagram describing logical relations between sets of numeral, quantifying and ordinal segments, together with exemplary elements of their intersections and differences.

Homologous series (of words) is proposed as a term describing a group of words in which:

- Each word has an immutable part that holds a constant meaning and form (barring flexion and phonetic alternations), which is composed of at least the (derivative) **base of the series**, and can contain additional morpheme(s),
- Each word has a mutable part, which consists of an interchangeable **q/o segment**,
- The only difference in meaning between any two words of the series is a direct result of one q/o segment being substituted with another, effectively changing the quantity or ordinal referenced.

An example of such a series is: {JEDNOLATEK, DWULATEK, DWUIPÓŁLATEK, TRZYLATEK...} (Eng. {ONE-YEAR-OLD, TWO-YEAR-OLD, TWO-AND-A-HALF-YEAR-OLD, THREE-YEAR-OLD…}). In this case the series can be represented simply as *n*-LATEK (Eng. *n*-YEAR-OLD), where *n* represents the mutable q-segment. The series may be assigned a global definition of, for example, 'someone who is *n* years old'[4].

The immutable part of the series does not need to be continuous. For many bases, a **value modifier** may be attached before the q/o segment to derive one series from another, eg: *n*-METROWY 'that has *n* meters' (a series with a continuous immutable part) can be modified with the prefix *ponad-,* creating new series PONAD-*n*-METROWY = {PONADPÓŁMETROWY, PONADJEDNOMETROWY, PONADDWUMETROWY…} 'that has more than *n* meters' (a series with incontinuous immutable part). Technically speaking, value modifiers should be regarded as a part of q-segments, as they are contributing to the global quantitative interpretation: they change the referenced value from "exactly 2" to "(2, +∞)". However, at this stage it seems more practical to consider *n*-METROWY and PONAD-*n*-METROWY as two separate series, assigning the modifier *ponad-* as a component of the immutable part together with the base, simply because it produces a more straightforward model of substitution. It is possible that this decision will be revoked in the later stages

[4] Deciding if 'someone who is *n* years old' should constitute a separate meaning (i.e. a separate homonymous lexeme) from 'an animal which is *n* years old' is not as straightforward as it may seem. For now, any proposed ad-hoc definition is for illustrative purposes only – this particular definition is simply complying with existing sense granulation in plWN, where "human-related" and "animal-related" senses usually constitute two separate lexemes.

of the study. Either way, assuming a possible discontinuity of the immutable part may be helpful to detect other unpredicted series[5].

For purposes of the thesis, the most important are the **infinite homologous series**, which can be defined as ones where the mutable part can be filled by an infinite number of different q/o segments.

An infinite number of possible q/o segments does not mean that any and every q/o segment has to be able to take that place. Certain bases may have restrictions for some range of numeric values or exclude certain types of segments, eg. *n*-KĄT (a series of names denoting polygons of varying amount of angles) can only take as small a value as three, and it does not accept fractional morphemes – in both cases due to the geometrical properties of a polygon. However, since the number of angles in a polygon can theoretically go up to infinity, the series is still able to to contain an infinite amount of homologues.

Homologue words, or **homologues** in short, are words that belong to series such as described above (infinite or not). The terms "homologue" and "homologous series" have been borrowed from chemistry, where they describe molecules that are very similar in their structures, only differing by the number of atoms in their carbon chain.

Impostor is proposed as a term to describe a word which, on the surface level, looks as if it could both contain a q/o segment and belong to a homologous series of words, but in fact it does not. Its surface form may be related semantically or etymologically to a quantity or ordinal (eg. PIERWSZYZNA, STOKROTKA), but without a capability of being generalized into a homologous series together with more than one other word documented in usage – either because words with similar structure do not exist, or because the difference in their meaning cannot be uncontroversially boiled down to a simple difference between values (eg. UNISEKSUALNY 'attracted to genders either similar or different than one's own, but not both at the same time', BISEKSUALNY 'attracted to genders both similar and

[5] For example, a series with reversed order of [base][q/o segment] may be discovered. So far the author can only think of the archaic SAMO-*n* = {SAMOPAS, SAMOWTÓR, SAMODWÓJ, SAMOTRZEĆ, SAMOCZWÓR} 'as a group of *n* and no more', 'alone with *n*-1 others'. This particular series is not likely to make it to the target resource for multiple reasons, but it proves the possibility of a different segment order.

different than one's own', PANSEKSUALNY 'attracted to people regardless of gender' – those definitions cannot be generalized into a pattern where the same slot is filled with either "one", "two" and "all", while all the rest remains unchanged). An impostor may also contain a string of characters only incidentally identical to a q/o segment (eg. STOŻEK) or in some cases its part (eg. GIMNASTYKA), causing it to possibly match some of the rules defined to target homologues.

Including a range of such "impostors" in the test corpus is necessary to make sure that newly proposed morphosyntactic segmentation rules will not lead to unexpected and incorrect processing of those words, such as: attempting to interpret STOKROTKA (Eng. DAISY) as 's tuple with a hundred elements' or 'a hundred tuples' based on independent meanings of segments STO, (Eng. HUNDRED) and KROTKA (Eng. TUPLE).

Pseudo-quantifying or pseudo-ordinal string is a string of characters that superficially matches an existing q/o segment, but in fact is not it (at least from the synchronic perspective). Examples of such pseudo-q/o strings are: *try-* in TRYWIALNY (matches the whole quantifying segment) and possibly *-nast-* in GIMNASTYKA (may match regexes targeted at q/o segments such as *jedenast-, dwunast-*).

Scope of interest

For the future MA thesis, the intended scope is: compound content words that contain a combination of a quantifying/ordinal segment with a non-quantifying and non-ordinal derivational base.

Most words typically classed as numerals (even morphologically compound ones such as STO DWADZIEŚCIA TRZY, KILKUNASTY) are therefore excluded from analysis, because all they denote is a value or a place in order – they do not contain any additional information such as what or who they are quantifying or ordering. There is an exception for two types of words that, at least for the Polish language, tend to be categorized as sub-classes of numerals (Jadacka, 2013): liczebniki wielorakie (Eng. "manifold" numerals) such as CZWORAKI 'that has four aspects, forms, types); liczebniki wielokrotne (Eng. iterative numerals) such as TRZYKROTNY

'that has happened, became something or been subjected to something three times'. In this paper, both of those classes will be interpreted as "normal" compound adjectives (instead of numerals) and thus included in the analysis. This decision is motivated by the semantic complexity of those words (i.e. denoting more than just a value).

A specific type of q-segments are SI prefixes such as *kilo-*, *nano-*. There exists a well codified and not too long list of them, and they are intended to be used with similarly specific and finite SI units (such as METR, GRAM). However, those prefixes may attach to many unexpected roots and still hold their meaning outside of the SI system (eg. KILOPIKSEL 'a thousand pixels'). It seems as if in this case an opposite type of series should be proposed: an immutable quantifying part + a mutable non-quantifying part (KILO-*x* 'a thousand of something' etc.). So far, due to very limited time, all of the SI prefixes are arbitrarily excluded until their properties are made sure to be adequately addressed.

For this particular paper, only the words with quantifying segments (or pseud-quantifying strings) will be taken into account. Additionally, only general (non-specialised) internet sources are going to be observed. The scope of research has been artificially narrowed down in order to make the study feasible within the imposed time frame.

After finishing this paper, the author will move on to exploring ordinal segments in the same corpus, then both types of segments in a corpus (or other collection) of specialized literature. Finally, data collected from corpora will potentially be complemented based on existing dictionary entries, as well as morphosyntactic rules that are already employed to process words with q/o segments by Morfeusz.

It is worth admitting that there is a considerable additional goal of the study: apart from research to be made in the future and to prepare for over the course of this preliminary work, this paper serves mostly a didactic purpose, as it has been created as a Corpus Linguistics course assignment.

Unresolved issues

Numeral segments will not necessarily produce an infinite series with every base they can attach to, for example *-polówka* in TRÓJPOLÓWKA (Eng. THREE-FIELD SYSTEM). One may argue the base *-polówka* produces potential homologues only in combinations with *jedno-*, *dwu-*, *trój-* or *cztero-*[6], while all the other possible combinations are meaningless. It is hard to tell where the line lies – or even if there exists a line – between a meaningful but unused (or very rarely used) word, and a non-meaningful combination of morphemes. ?SZEŚĆDZIESIĘCIOPOLÓWKA (Eng. ?SIXTY-FIELD SYSTEM) can be seen as nonsense, or as a valid term that simply happened to not have found a use in our agriculture vocabulary, but can easily be employed and understood in a sci-fi novel about extraterrestrial farmers. Current lexicographical standards[7], however, lean towards excluding theoretical (i.e. not documented in corpora) forms. For all it's worth, if we conclude that hypothetical ?SZEŚĆDZIESIĘCIOPOLÓWKA is "not a real word", we should probably constrain it from appearing in lexical resources (or at least somehow indicate its dubious or "synthetic" status to the reader).

Certain bases will almost certainly form non-infinite homologuous series (for example, the ones derived from uncountable nouns which are attachable to non-numeric quantifying segments only). Are they worth examining and describing as well? On one hand, it is probably much faster to create 3 or 4 manual entries in the lexical resources instead of coming up with elaborate rules to generate them automatically. However, with a series of 10 or 20 homologues, automating their processing may begin to be worth it, especially if certain processes (such as creating lexical relations), due to a possible strong analogies between them, may target multiple series at once.

Luckily, at this (preliminary) stage those questions do not need to be accurately answered, as they carry very little practical consequences. Even if certain bases

---

[6] This base combined with higher numbers may produce meaningful and useful words, such as SZEŚCIOPOLÓWKA (https://polska-org.pl/509926,Gola_Swidnicka,Kosciol_filialny_sw_Marcina.html, DOA: 6.12.2023), but its meaning will not be homologue to TRÓJPOLÓWKA understood as THREE-FIELD SYSTEM.

[7] As witnessed by the author during their work as editor of Polish WordNet (precise rules on word selection are gathered in internal documentation and therefore uncitable).

or lexemes later turn out to be uninteresting or irrelevant, they will serve as good negative examples. These issues are signaled already, but they are purposefully left to be resolved after the data is gathered.

Why use a reference corpus to explore q/o segments and the homologous series?

The language system allows for practically unconstrained formation of hypothetical words with q/o segments. We could, in theory, define the rules of that formation, apply them systematically to every possible base and create a list of every possible homologous series. For example, by analogy to a series *n*-OSOBOWY = {JEDNOOSOBOWY, DWUOSOBOWY, TRZYOSOBOWY...} 'able to fit *n* persons', we could easily come up with a series such as: *n*-OŚMIORNICOWY = {JEDNOOŚMIORNICOWY, DWUOŚMIORNICOWY, TZYOŚMIORNICOWY...} 'able to fit *n* octopi'. While preparing dictionary entries, words from that series could as easily be employed in meaningful sentences: *Zoo kupiło stuośmiornicowe akwarium*. (Eng. *The Zoo has bought a tank that is able to fit a hundred octopi.*) However, words from this hypothetical series may have never been used before and may never will, and both in NLP and in lexicography, we try to direct our attention towards the words most likely to be encountered in real-life usage.

Thanks to the existence of corpora, instead of relying only on systemic rules and/or the author's personal language intuition and habits, it is possible to explore a large sample of texts and find objectively salient examples.

One may argue that the exploration of q/o segments (and the words they appear in) should start with consulting the existing descriptive resources anyway. In fact, out of necessity, some didactic and encyclopedic resources (for example, Wikipedia entries containing names of chemical compounds) were used as starting points in preparation for searching the corpus. Nonetheless, the author tried to avoid transcribing q-segments from words gathered in dictionaries or glossaries, especially those already employed in NLP: this research should ultimately result in a novel lexical resource, and not a transformation of the existing ones. Lexical resources may be used later to account for the author's own

blind spots, but using them in the very beginning would pose a risk transposing all the biases from previous researcher's work into this one, leaving no way to balance it out. On the contrary, exploring the corpora first cannot influence the already documented content of dictionaries and tools, leaving them as a convenient way to fill in the blanks at the end of the exploratory phase.

Methodology of exploring the reference corpus

Choosing the corpus

For the purpose of exploration, **Polish Web 2012 (plTenTen12, RFTagger)** has been chosen as the reference corpus due to the following reasons:

1. Its language is Polish, which is the focus of the study;
2. It's big (9,387,142,186 tokens – considerably more compared with the 5,216,428,620 tokens of its 2019's counterpart);
3. It's contemporary – we don't expect substantial revolution in the use q/o segments over the course of less than one and a half decade;
4. It's available in the SketchEngine platform – making it both convenient to work with and suitable for learning to use this particular tool;
5. It contains internet sources which are likely to be representative of the contemporary colloquial and general language.

Searching the corpus and transferring data

Lexemes that likely belong to a homologous series should be discoverable by searching all the words containing any q-segment and checking the top of the frequency list. As numeral segments are, by definition, infinite, an arbitrary list of queries needs to be chosen. In doing so, the following line of reasoning has been employed:

- Q/o segments do not necessarily appear at the beginning of the word, although it is presumed to be the majority of cases;
- For o-segments, the closer the value is to one, the more likely the designate is likely to exist: *n*th place implies the existence of (*n*-1)th place etc., but

not the other way round. And the more something is likely to exist, the more (at least in theory) we are likely to talk about it; additionally, the closer something is to being the 1st, the more important (or worth talking about) it likely is;

- However, in the case of q-segments, it is near to impossible to theorize which values are most likely to appear. Certain important bases may "prefer" particular quantities for extralinguistic reasons (such as *-nożny*, Eng. *-legged*, is probably co-occuring the most often with values of 2 and 4, but not as often with 3, as the most talked-about beings have either two or four legs); guessing "favorite" value ranges for every homologous series is a task that defeats the purpose of corpus research;
- It is thus necessary to explore a possibly broad range of numeral q-segments, i.e instead of just searching for low values such as 2, allow for segments referring to 12, 22 … 102 etc. to be found as well – the same goes for fractions;
- For some bases – especially ones containing uncountable nominative roots or ones who have broad and evenly spread range of "favorite" values – the highest likelihood to find an occurrence may be to look for non-numerical q-segments or q-segments relating to ranges of values as well;
- There is at least one modifier (*ponad-*) that is likely to be found on the left side of a q-segment – it is worth examining words starting with it as well, instead of only looking for specific q-segments.

Taking all that into account, the author has prepared the list of search queries starting from non-numeral and zero quantifying segments, moving on to segments related to positive integers, then to ranges, to fractions, and finally modifiers (eventually, only one linked modifier has been defined). In each case, a list of exemplary values / value modifications is defined first (e.g. "all", "10 + m for $0 < m < 10$", "m + 0.5", "more than"). For every value, as many corresponding strings of characters as possible are listed. Interfix `(u|o)` is omitted in most cases. The strings are then generalized into regexes more convenient for searching – eg. strings *poly-* or *poli-* into `pol(y|i)`. Interfixes `(u|o)` may be

added to a regex to make it less ambiguous, especially when it comes to very short strings.

To come up with a possibly broad range of strings, the author supported their own linguistic intuition with additional sources such as a morphology textbook (Jadacka, 2013), Wikipedia articles or internet searches. The list of strings identified as relevant can be found in Appendix 1. It is by no means exhaustive, but sufficiently long to provide enough data for analysis.

The regexes from the Appendix 1 are searched for in the reference corpus using **Wordlist → find → lemmas → starting with**. Note that certain regexes start with `.*`, which means they are matched even if the lemma does not begin with them (i.e. `.*nast` matches with such lemmas as NASTOLETNI, but also KILKUNASTOMETROWY).

The results have been gathered in a Google Sheets document. While transferring lemmas from Sketch Engine results to the spreadsheet, the following rule were employed:

1. From 1st to 10th place, all the results have been placed in the table.
2. After 10th place, only lemmas likely belonging to a homologous series are placed in the table.
3. Search results are examined in order from the most frequent until the 100th place, or when the number of their occurrences drops below 2000 (if this has not happened within the first 100 places). Those limits were decided arbitrarily after conducting a few test searches and reaching a compromise between the time/cognitive constraints and a likelihood of encountering an interesting word.
4. At least 3 impostors for each query should be placed in the table (unless, of course, the results contain less than 3 of them all together).
5. More impostors may be added if they seem particularly interesting[8].

[8] For example, if the same base attaches to more than one pseudo-q string and thus even more convincingly creates an illusion of belonging to a homologous series, as it is the case with DWUKROPEK and WIELOKROPEK

6. "Lemmas" produced entirely by software mistakes[9] are excluded from the results. To verify if a suspicious lemma indeed results from a mistake, its concordances are checked. All the word occurrences gathered under an erroneous "lemma" are simply ignored, instead of being counted as additional occurrences of the actual lexeme they belong to – they may be a collection of word forms belonging to multiple separate lexemes, and spending time on sorting them out is not justifiable by the small number of cases they add up to.
7. All independent numerals (cardinal and ordinal) are excluded from the results, even if they are found in the top 10 most frequent words for a given query.
8. Word sense disambiguation and part-of-speech distinction are not taken into account. This may cause some classifying errors later on, but the preliminary stage of the study is aimed at collecting as much potentially useful data as possible – at the risk of including some residual meanings or other irregularities.

Additional N-gram search

To check for potential blind spots of the q-segment-oriented queries – and to account for possible disjunctive spelling of words starting with *ponad-* – the N-gram function is used for performing and advanced search with the following parameters (only modification of the default settings are described):

- Case insensitive;
- Length: 2-3 tokens;
- Attribute: lemma;
- Starting with `(ponad|około|przeszło|gdzieś|przynajmniej)`

Quantity modifiers included in the query are likely to precede words relating to quantity, including complex words with q-segments (for example, *około dwuletni* 'about 2 years old', *przeszło dwuletni* 'of more than 2 years'). Searching for N-grams starting with quantity modifiers is a way to look for quantifying

[9] For example, if an adjective or a noun is lemmatised to a surface form different from its actual nominative case.

expressions without requesting a specific value of the quantifier, therefore making it possible to find words potentially overlooked with the first approach.

Organizing data

The examples of words with q-segments (or pseudo-q strings) found in the reference corpus are gathered in tables using Google Sheets. For clarity, there are separate sheets for words containing segments/strings related to:

1. Non-numerals and 0, eg. WIELOLATEK (value of "much" or "many"), ZEROLATEK (value of 0);
2. Integers, eg. OŚMIOLATEK (value of 8);
3. Ranges, eg. ILUŚNASTOLATEK (value from the approximate range of 11–19), KILKULATEK (value from the approximate range of 2–9);
4. Fractions, eg. OŚMIOIPÓŁLATEK (value of 8.5).

Each lexeme is introduced in a separate row containing:

1. Its full lemma (canonical form of the lexeme);
2. Total number of lexeme occurrences in the corpus;
3. Extracted q-segment (or pseudo-q string);
4. Extracted non-quantifying part from the left side (if it exists);
5. Extracted non-quantifying part from the right side (if it exists);
6. General formula of the series that the lexeme (supposedly) belongs to;
7. Series' supposed (provisional) class.

Words and series are still not disambiguated between possible multiple senses or meanings: all the word forms lemmatized to a certain base form are treated as belonging to the same single lexeme. Lemmatization relies entirely on the Sketch Engine built-in tools, and only the obvious errors are excluded.

The lemma and its number of occurrences are copied directly from Sketch Engine in their original tabular format, thus lowering the risk of typing mistakes. Then, q-segments / pseudo-q strings are manually typed in the table. A formula =PRAWY(A_*row*#;DŁ(A_*row*#) – D_*row*#)[10] is used to automatically extract the

[10] All the letters indicate a column in accordance with the table organization in the table gathering the results, ie. A is the ID of the column where the full lemmas are kept, and D is the

immutable right-side part from most lemmas. If the lemma contains an immutable part on the left side of the q-segment as well, it needs to be typed into correct cell manually, and the formula for right-side extraction is adjusted into =PRAWY(A_*row*#;DŁ(A_*row*#) – D_*row*# – C_*row*#).

All samples are automatically generalized into a formula of a so-far theoretical homologous series. To generate the formula, one of the two functions is used:

- =JOIN("-"; "n";E_*row*#) if the immutable part is located only on the right side of the q-segment / pseudo-q string;
- =JOIN("-"; C_*row*#;"n";E_*row*#) for discontinuous immutable parts.

Provisional classes of the homologous series (chosen out of those proposed in Table 1) need to be assigned manually based on intuition.

If a particular lemma, or a whole supposed series, causes doubts due to its ambiguity, it receives a "questionable" class. If assigning a class poses problems unrelated to the word-sense ambiguity, a reverse search is performed (looking for words ending with the base of the supposed series). Series is classified as "infinite" if segments related to values above 10 are discovered, "finite" if at least 3 possible homologues are discovered, but none of them relates to a value above 10, and "impostor" if only 1-2 possible homologues are discovered. For example, *n*-SPADOWY has the highest value of 8, and should therefore be coded as finite. Meanwhile *n*-RDZENIOWY includes such lemmas as DWUNASTORDZENIOWY or STURDZENIOWY and in consequence is coded as a (potentially) infinite series. This coding is likely to be adjusted once data from other corpora is gathered.

If a lexeme is recognised as an "impostor", the series' theoretical formula is kept anyway, as it may lead to discovering classifying errors once the tables are sorted (i.e. revealing an unforeseen regularity between 3 or more supposed impostors).

Class	Meaning
Infinite	The lexeme seems to belong to a series with an infinite number of possible q-segments.

ID of the column containing q-segment / pseudo-q string. "_row#" is a placeholder for the row number of the entry in question.

Finite	The lexeme seems to have at least 2 homologues, but likely not an infinite amount. It is also used when the number of homologues can theoretically reach infinity, but in practice seems to be constrained to less than 10 homologues in actual usage.
Impostor	The lexeme most probably does not belong to any valid homologous series.
Questionable	The lexeme may need semantic disambiguation before assigning a class or cannot be accurately classified for other reasons (for example, a research is needed to verify if the meaning of the word is compositional from a synchronic point of view)

Table 1. Class coding for lexemes. Meaning of each code is purposefully expressing uncertainty – data needs to be re-assessed and re-coded after it is collected from more sources.

lemma	N° of occurrences	non-quantifying part (left)	q segment / pseudo-q string	non-quantifying part (right)	general formula	class of the series
kilkulatek	23	–	kilku	latek	*n*-latek	infinite
dwulatek	10	–	dwu	latek	*n*-latek	infinite
stonoga	5	–	sto	noga	*n*-noga	impostor

Table 2. Exemplary table (filled with artificial data) illustrating data coding in a hypothetical situation where a total of 23 occurrences of KILKULATEK, 10 of DWULATEK, and 5 of STONOGA have been discovered in the corpus.

Unpredicted issues

Numbers are presented in Sketch Engine with commas as thousand separators, which causes the Google Sheets to misinterpret them as fractions or strings. To accommodate for that, the display of numbers has been temporarily set to three spaces after the comma. Once the data is gathered, the commas are removed automatically (using the "Search and replace" option) and the display of numbers is changed back to display no digits after decimal point.

Methodology of creating the target corpus

Purpose of the corpus

As it has been mentioned, the corpus is primarily intended for tool testing, and not for conducting linguistic research. It may find its use for later lexicographical

work (by facilitating the creation of definitions and examples for series it contains), but it should not be viewed as a reliable source for frequency comparisons or as a sufficient resource for isolating meanings of polysemic words.

Additionally, the process of preparing the corpus will in itself contribute to verifying classes assigned in Appendix 2. If errors are discovered (eg. meaning of a word or a segment turns out to have been misunderstood in an impactful way), they will be corrected.

Those purposes – illustrative and exploratory rather than representative – influence heavily the way the corpus is going to be composed.

Structure

The corpus created while writing this paper is actually a sub-corpus of the target test corpus for the future MA thesis, since it contains only words with the q-segments. The structure of the complete corpus should look similar to the folder tree presented in *Fig. 2*.

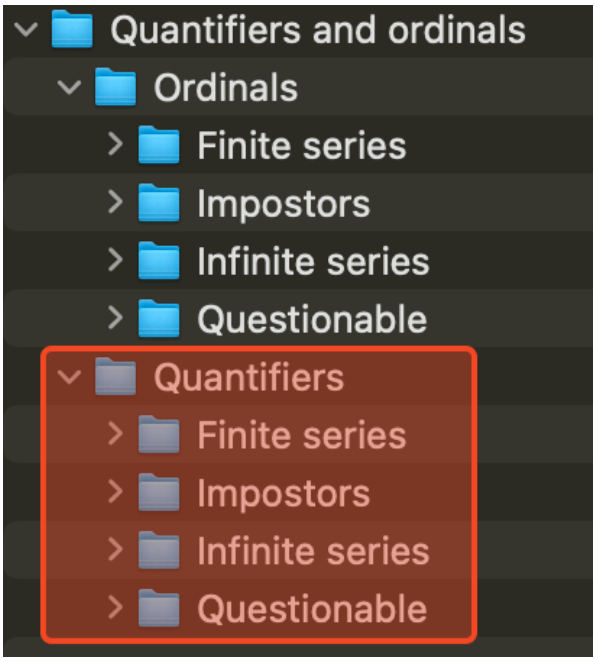

Fig. 2 Proposed structure of the full test corpus. The part related to this paper is highlighted in red. "Ordinals" refers to the texts containing words with o-segments (or pseudo-o strings), "Quantifiers" refers to the texts containing words with q-segments (or pseudo-q strings), while "Finite series", "Impostors', "Infinite series" and "Questionable" refer to the texts containing words that would be tagged as such in the Appendix 2.

This part of the corpus needs to have at least 30,000 tokens due to external guidelines set for the project. To keep it more or less balanced, each of the four sub-parts should have not less than 5,000 tokens – however, more is expected for infinite series. In order for the corpus to be functional, the texts should be very brief (around one paragraph or 500-1000 characters each) and contain at least one word with a q-segment or a pseudo-q string. If too long texts are uploaded, it will be hard to find relevant words and check how they are being processed after applying changes to the tools. On the other hand, if the texts are too short (i.e. they are single sentences or part of the sentences), the environment becomes too controlled and artificial, potentially rendering the testing unreliable and inadequate for predicting the processing in "real-life" applications.

Searching for text samples

To find authentic paragraphs containing various members of homologous series (or their impostors), word forms are searched on the Word Wide Web using Google search engine and Wikipedia search engine.

Sources of the texts are not documented – doing so would create an additional, non-essential workload to an already ambitious task of hand-selecting a few thousand short texts.

Because the samples are very brief, they will be grouped into one file with their supposed homologues or – for impostors – with other words containing the same pseudo-q string. If two or more paragraphs of the same text are merged to achieve the 500-1000 characters threshold, line breaks are removed.

Each file corresponding to a particular series should contain examples with various q-segments and various grammatical forms of the bases (in terms of number, case, gender). Series are not semantically disambiguated, so examples in one file can – but do not have to – contain mixed meanings of polysemic or homonymous bases.

Each fragment has to contain at least one instance of a relevant word (member of a series or an impostor with a given pseudo-q string). Additional members of the same or different series can appear in the same fragment as well. If a fragment

contains words from other series represented in the corpus, it cannot be repeated in the file associated to it as well – for each series, different samples need to be found.

Results

Summary

In the **Polish Web 2012** corpus, a total of 660 unique lexemes have been identified, representing 319 potential homologous series with 130 likely to be impostors.

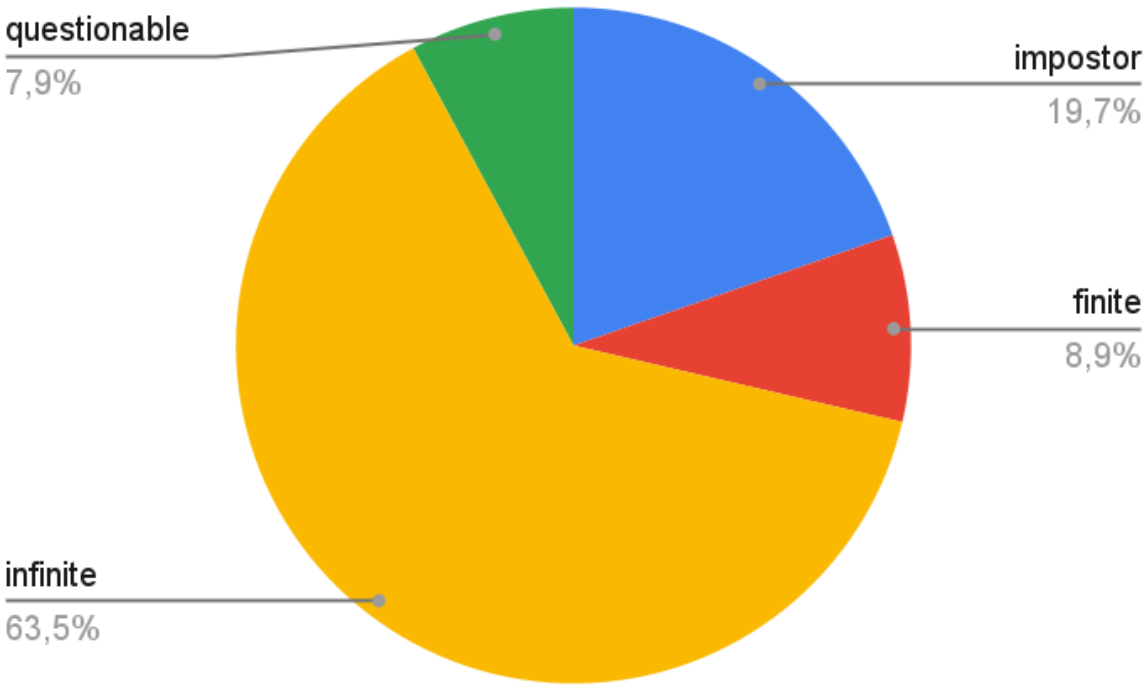

Fig. 3 A pie chart showing classification of lemmas as possibly belonging to an infinite (63.5%) or finite (8.9%) series, being impostors (19.7%) or constituting a questionable (7.9%) case.

The prevalence of lemmas classified as potentially belonging to an infinite series is conforming to the author's expectations. Out of the non-impostors, the most commonly occurring series are *n*-LETNI (39), *n*-LATEK (19), *n*-KROTNIE (16). The most unexpected series is *n*-SPADOWY '(about a roof) that has *n* slopes' with 3 different lemmas identified in the corpus.

A complete list of the results can be found in Appendix 2. The corpus containing examples found on the basis of the research results is available in Sketch Engine.

Issues and shortcomings

Not every intended query listed Appendix 1 has been used during the study. Certain series containing q-segments – such as ones related to chemical compounds' names or containing other "sophisticated" segments (such as *dodeka-*) – will likely be observable only in specialized texts and thus need additional, targeted research. Their exploration is out of scope of this particular paper, which is indicated by a "for specialized texts" note in the table. Other q-segments were left unchecked simply due to running out of time to finish the project, which is indicated by an "unexplored" note.

On the other hand, certain regexes from Appendix 1 have been tested and have not yielded any or almost any anticipated results in the reference corpus. An example of such a quantifying segment is *uni-*, which was expected to be present in at least a few short series, such as {UNILATERALNY, BILATERALNY, TRILATERALNY}. In reality, none of the actual matched strings in various lemmas from the corpus (eg. UNIWERSYTET, UNIWERSALNY, UNIA) turned out to be a valid quantifying segment, at least from the synchronic perspective. Similarly, *di-* and *bi-* have yielded collectively only one somehow relevant lemma: BINARNY (possible homologues: {UNARNY, TRYNARNY, HEKSARNY}), three "interesting" impostors: BISEKSUALNY (possible homologue: PANSEKSUALNY), DIALOG (possible homologue: MONOLOG), BICEPS (possible homologue: TRICEPS), and hundreds of completely irrelevant results. In such cases, the results are not included in the table (apart from impostors sharing a base with other results), and the string receives a "removed" note in the Appendix 1 table. However, those regexes will be re-included for searches in other resources.

Re-classification

While examining other internet sources for the purpose of creating the corpus, following errors have been discovered:

1. A lot of occurrences of TRZYKROĆ turned out to mostly refer to a proper name MATKA BOSKA TRZYKROĆ PRZEDZIWNA (one of St. Mary's titles), while its homologues such as CZĘSTOROĆ were mostly found in scientific texts, albeit in very small numbers. Due to that, the class has been changed from "infinite" to "questionable".
2. Some potential homologues of WIELONARZĄDOWY have been discovered, leading to changing its class from "impostor" to "questionable", as more research on their meaning is needed. As of today, it is suspected that *n*-NARZĄDOWY may be a finite series that belongs to specialized medical vocabulary.
3. The class of DWUMECZ has been changed from "questionable" to "impostor" after learning that the meanings of DWUMECZ and TZYMECZ are not homologous to each other.
4. More homologues of *n*-SEZONOWY have been observed (with values as high as a hundred), so it was re-classified as potentially infinite.

Additional N-gram search

The additional search has not revealed any new lemmas that would not have been already discovered using the Wordlist function. Relevant unambiguous results (sorted by the decreasing number of occurrences) include:

- PONAD DWUGODZINNY*;*
- PONAD DWULETNI*;*
- PRZYNAJMNIEJ DWUKROTNIE*;*
- PONAD DWUKROTNIE WYSOKI*;*
- PONAD TRZYKROTNIE*;*
- PONAD CZTEROKROTNIE*;*
- OKOŁO DZIESIĘĆ CENTYMETROWY*;*
- PONAD DZIESIĘCIOLETNI*;*

- PONAD DZIESIĘCIOLETNI DOŚWIADCZENIE;
- PONAD DWUDZIESTOLETNI;
- PONAD DWUDZIESTOLETNI DOŚWIADCZENIE
- PONAD TRZYDZIESTOLETNI.

For ambiguous N-grams, concordances have been checked to verify if they contain free numerals or q-segments separated from bases due to a spelling mistake. It revealed only singular relevant bundles (such as *ponad stu kilometrowy* for the PONAD STO N-gram, or *ponad trzy gwiazdkowym hotelu* for the PONAD TRZY N-gram), which were also already represented in the Appendix 2.

However, two results: PONAD MIESIĘCZNY and PONAD GODZINNY stood out. On one hand, they do not contain any q-segment, at least on the surface level. On the other hand, they clearly point to a specified value of "more than one" (month or hour). It may be an important argument for including *ponad-* as a q-segments, but it leads to other important questions as well (see Discussion, section: ∅ q-segment).

Discussion

It is a very early stage to draw substantial conclusions about the quantifying segments. Even though a big piece of data has been gathered already, the time constraints have allowed for only a few q-segments and homologous series to be explored, and to do so in only one corpus.

Nevertheless, certain observations have already been made. Those topics can be expanded in further research.

"Immutable part" alterations

Certain alternations of the immutable part of the series have been observed, for example *n*-RAKI = {WIELORAKI, JEDNORAKI, CZWORAKI, PIĘCIORAKI…} manifests an alternation of {DWOJAKI, TROJAKI}. So far those groups of lexemes are treated as belonging to different series (by virtue of an automatic assignment of different formulas, based on their "immutable" parts). There is a need to merge them into

one series without overcomplicating its notation. Perhaps, formulas and immutable parts may simply be converted to regex before further processing.

Word vs series frequencies

Certain lemmas may have a very high frequency of use even if all their homologues are far less likely to appear, such as high-frequency WIELOZADANIOWY and its low-frequency numeral homologues[11]. So far, the lemmas have been picked and generalized into homologous series on the basis of their individual frequencies. A marker of the "importance" for a whole series may be needed in order to decide which ones should actually be focused on while implementing the results in NLP tools and resources.

∅ q-segment

Certain series seem to contain lemmas with no visible q-segment on the surface, but nevertheless relate to a value as if a q-segment was present. Such potential "∅ q-segments" could mean either "one", "some" or "a lot" depending on the series' base:

- MIESIĘCZNY has the same meaning as JEDNOMIESIĘCZNY 'of one month';
- IGŁOWY 'that has or employs (some) needles' fits into the scheme of *n*-IGŁOWY 'that has or employs *n* needles';
- KALORYCZNY is practically synonymous with WYSOKOKALORYCZNY 'that has a lot of calories'.

The same cannot be said for other series – for example removing a q-segment form *n*-POKOJOWY 'containing *n* rooms' leaves us with a word meaning 'peaceful' or 'related to a room' (and not: 'containing one/some/a lot of rooms'), which does not belong to any homologous series.

[11] For lemmas containing string *zadaniowy* and a quantifying segment, the following have been found: 12,459 occurrences of WIELOZADANIOWY, 66 occurrences of JEDNOZADANIOWY (123 adjusted for tagging errors), and only 9 of MULTIZADANIOWY, 6 of DWUZADANIOWY and 5 of BEZZADANIOWY.

No lemmas with potential "∅ q-segments" were included in the results – the queries were not designed for them to show up and their appearance was too incidental to allow for any methodic data gathering.

It is possible that the differences in meaning may be connected to particular combinations of semantic properties of the root noun in the base (expressed on such axes as: concrete/abstract or countable/uncountable). A separate study with a quantitative analysis can be proposed for examining if combinations of root's semantic properties systematically influence the particular meaning and an overall possibility of employing "∅ q-segment" in the series.

Intensity quantifiers

There emerges a potential pattern of series with qualitative adjectives as their bases, such as: NIEAUTOMATYCZNY – PÓŁAUTOMATYCZNY – AUTOMATYCZNY (Eng. MANUAL – HALF-AUTOMATIC – AUTOMATIC). At first, they do not seem like an interesting target to be analyzed as homologous series, due to their extreme shortness (we might add some other possible q-segmens such as *ćwierć-*, denoting ¼, but not many more) and due to the fact that regular antonyms created with *nie-* (Eng. *non-*) are probably already handled well within the automatic processing of Polish language. However, with a very large number of adjectives that could conform to this easily predictable pattern, there may exist an efficient way to improve processing of qualitative adjectives with segments such as *pół-*, *ćwierć-* etc. quantifying the intensity of a given property. Other potential "intensity-qualifying segments" include: *mega-*, *super-*, *giga-* in their colloquial (non-SI related) meanings.

Quantitative / ordinal ambiguity of numeral segments

A very interesting case of ambiguity has been discovered: for lemmas such as DWUDZIESTOWIECZNY, both quantifying ('that has duration of 20 centuries') and ordinal ('originating from, happening in or otherwise relating to the XXth century') semantic interpretation of the segment is possible. Although the latter is more likely, the discovery of such lemmas as DWUWIECZNY ('that has duration of 2 centuries') is a reason to keep *n*-WIECZNY in quantitative series as well.

A different but similar problem is related to segments such as DWUDZIESTOLECIE. The interpretation of the q/o segment as quantitative or ordinal may be ambivalent even in case of one, unambiguous meaning of the whole word: it can be defined as '20th anniversary' (ordinal interpretation of the numeric segment) or 'the moment when 20 years has passed' (quantitative interpretation of the numeric segment). The existence of TRZYLECIE and not TRZECIOLECIE points to a qualitative meaning of this segment, even though the ordinal interpretation seems more "natural" as a definition.

It is unclear how many bases have the potential to create such ambiguous lemmas: in most cases, the meaning of the base itself points to a more probable interpretation of the q/o segment, but, as it has been demonstrated, it may not always be the case.

"Finite" in theory and in practice

Certain homologous series have been classed as "finite", because their length is limited due to the semantic properties of the base (such as having an uncountable noun as a root, which makes them generally unsuitable for linking with numeral segments – non-numeral segments are a more or less closed class).

Uncountableness is not an exclusive reason for the base to produce finite series – some may be finite because there are some non-linguistic limitations of their use. However, some series were most likely classified as "finite" due to pure chance: they most likely allow for an infinite number of meaningful homologues, but their occurrences in the reference corpus were too scarce to justify classifying them as infinite. It is expected that many series will be re-evaluated in this regard after observing more data.

Conclusion

The process of researching compound words with q/o segments is still in its early phase. The author can confidently say that their knowledge on the use of tools has been expanded in a way that will facilitate further gathering of the linguistic data. And although they did not resolve any particular problems yet, they have

managed to identify the questions that may be worth asking and answering in their MA thesis.

Bibliography

Jadacka H., *Złożenia of Kultura języka polskiego. Fleksja, słowotwórstwo, składnia*. Wydawnictwo Naukowe PWN, Warszawa 2013. ISBN 978-83-01-14398-5.