The Corpus of Polish Quantificational Expressions

Jakub Szymanik[1], Witold Kieraś[2]
[1]Institute of Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands,
[2]Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland
j.k.szymanik@uva.nl, wkieras@ipipan.waw.pl

Abstract

The paper presents a manually annotated corpus of Polish quantificational expressions. The quantifier annotation was conducted on top of existing gold-standard data for Polish as its separate layer. This short paper limits itself to the presentation of the process of building the corpus, however, the resource is part of a broader ongoing project and serves as a preliminary step towards further research in the distribution of semantic properties in the Polish quantificational system.

Keywords: quantifier, quantificational expression, corpus, annotation, monotonicity

1. Overview

The paper presents a manually annotated corpus of quantificational expressions of Polish. The corpus is a new separate layer of annotation in the gold-standard 1.2 million tokens large subcorpus of the National Corpus of Polish (NKJP1M, (Przepiórkowski et al., 2012), (Przepiórkowski et al., 2012)). It is a balanced set of short samples (approx. 40-60 words long) representing different text genres and available on GNU GPL. It is the most widely used resource for Polish in standard NLP and machine learning tasks covering automatic annotation of various levels. Its annotation features adjudicated sentence- and word-level segmentation, morphosyntactic description, shallow parsing (syntactic words and groups) and named entity description, as well as limited word sense disambiguation. Thus the quantificational layer contributes to the semantic level of the gold-standard annotation of the dataset and at the same time may benefit from the existing morphological and syntactic layers.

The paper describes the process of manual annotation of the corpus regarding quantificational expressions and their features. The corpus will serve as a referantial data set as well as training data for machine learning classifier.

2. Related work

According to our knowledge, there were no previous attempts at the manual annotation of quantifiers in any language so the presented corpus is a pioneer work in the field. However, two works had an important impact on the presented project.

First of the two is an extensive two-volume survey of the quantificational expressions from the cross-linguistic perspective (Keenan and Paperno, 2012; Paperno and Keenan, 2017). In the introductory chapter *Quantifier Questionnaire* many useful distinctions and guidelines for recognizing and describing quantifiers were specified. Polish was not included among 34 languages presented in the survey. The genetically and typologically closest language considered in both volumes was Russian as the only Slavonic language.

The other important source of motivations was a pilot study by (Szymanik and Thorne, 2017) in which the authors investigate the frequencies of thirty-six most common quantifiers in English in The WaCky corpus (Baroni et al., 2009). The authors have shown that semantic complexity (Szymanik, 2016) contributes to the explanation of the differences in frequency distributions. The major limitation of this study is the restriction to the small group of English quantifiers. The corpus described in the current paper will allow, for instance, to refine the results of (Szymanik and Thorne, 2017) by counting all the quantifiers that occur in a corpus with their semantic features, including semantic complexity, and recognizing the stronger statistical patterns.

3. Quantifiers

By quantifier or quantificational expression, we understand a natural language expression indicating quantity which extensionally can be represented as a relation between two sets (properties), $Q(A, B)$. Mathematically speaking, there are other possible types of generalized quantifiers, however, quantifiers taking two properties as their arguments are definitely the most common across natural languages (Peters and Westerståhl, 2006). Actually, there is no agreement among linguists whether the more complex quantifiers are even expressed in any natural language.

The annotators' task was to identify a quantifier, describe it's four features and determine the quantifier's scope. By convention we have decided to annotate a maximal nominal phrase as a scope of D-quantifier and a full verbal form (including potential negation particles, reflexive morphemes and auxiliaries) as a scope of A-quantifier. The annotation scheme does not enforce marking scope for every quantifier as it may be omitted in the text. The scope may be also shared by more than one quantifier.

The most important part of the annotation is specifying features of each quantifier in terms of four categories described in the annotators' manual and presented briefly in the following subsections.

3.1. D- and A-quantifiers

The first category distinguishes D-quantifiers from A-quantifiers (Bach et al., 1995; Partee, 1995). This feature of quantifiers refers to a syntactic and predicate structures in which the quantifier occurs. In the predicate-argument structure of an utterance D-quantifiers form expressions that are predicates (nominal phrases), but A-quantifiers directly build or modify predicates. This semantic distinction is

DRAFT
This is a draft of the paper. The presented data will be updated before publication.

also reflected in purely syntactic functions of the expressions: D-quantifiers are usually nouns, adjectives or numerals, whereas A-quantifiers are verb modifiers: verbal affixes, auxiliary verbs, or adverbs. In the context of our corpus A-quantifiers are almost exclusively adverbial phrases or functionally adverbial idiomatic expressions. Among the most frequent once are mostly temporal adverbs such as *zawsze* 'always', *nigdy* 'never', *często* 'often', *czasem/czasami* 'sometimes'. There exist interpretations of some verbal prefixes as A-quantifiers in Slavonic languages (i.e. Russian, (Paperno, 2012)) which can be also applied to Polish, in practice however they are rare in texts and did not occur in our corpus despite the fact that certain examples of such prefixal quantifiers were explicitly given in the annotation manual, as in the following example of verbal prefix *na-* which has cumulative meaning:

(1) Do pokoju na=wlatywało komarów.
To room na=flew mosquitoes.GEN
'A lot of mosquitoes flew into the room.'

3.2. Universal, existential, proportional

The second category distinguishes between existential (intersective), e.g., *some*, universal (co-intersective), e.g., *all*, and proportional quantifiers, e.g. *many*. The criteria for distinguishing the three are extensional and adopted after (Paperno and Keenan, 2017). For Q being a quantifier and A, B sets, $Q(A)(B)$ is determined by $A \cap B$, that is the set of As that are Bs, then Q is existential (intersective). If $Q(A)(B)$ depends on the property $A - B$, that is the set of As that are not Bs, then Q is universal (co-intersective). If $Q(A)(B)$ depends on the proportion of As that are Bs, that is $|A \cap B|/|A|$, then Q is proportional.
Among existential quantifiers, we also distinguish a class of numeral quantifiers (unmodified numerals), e.g. *5*, which are restricted only to quantifiers expressed by a number. The motivation for the additional value of the category is purely practical and technical: numeral quantifiers are one of the most frequent in texts and relatively less interesting so marking them with a separate label provides an easy way to filter them out. So far we did not distinguish among existential quantifiers a separate class of modified numerals (e.g. *more than 5*), which would be a possible future extension. In line with (Szymanik and Thorne, 2017) complexity analysis we expect that existential and universal quantifiers will be most frequent followed by the the proportional quantifiers.

3.3. Monotonicity

The third category described for each quantifier is its left and right monotonicity. Both are tested independently for each quantifier and the category can take one of three values: increasing, decreasing and non-monotonic. A quantifier Q is upward monotone (increasing) in its left (respectively, right) argument if and only if, for any sets A, B, C and D, if A is a subset of C and B is a subset of D, then $Q(A, B)$ entails $Q(C, B)$ (respectively, $Q(A, B)$ entails $Q(A, D)$). As the value of the property might not be determined directly in the context of a corpus utterance, the annotators were provided diagnostic sentences for testing monotonicity of the quantifiers.

Right monotonicity is absolutely crucial for semantic research. First of all, (Barwise and Cooper, 1981) proposed right monotonicity as one of the semantic universals, property that every language of the world satisfies. The proposed generalization can be formulated as all simple D quantifiers are right monotone. Therefore, we expect that all (or almost all) monomorphemic D quantifiers in our corpus should be right monotone. Furthermore, there is ample psycholinguistic evidence that right downward monotone quantifiers are harder to process for humans (reasoning, comprehension, verification, and acquisition), see, e.g., (Szymanik, 2016) for an overview or (Deschamps et al., 2015) for recent experimental evidence. One possible explanation associates this extra complexity with a lower overall frequency of right downward monotone quantifiers. Our corpus will allow comparing the frequencies of downward and upward monotone quantifiers.

3.4. Comparison type

Finally, the last feature described in the annotation is the comparison type. Each quantifier can be either positive, comparative, or superlative. Modified numerals come in two, semantically equivalent, flavors: comparative, e.g., *more than*, *fewer than* and superlative, e.g., *at least*, *at most*. (Geurts et al., 2010) have provided evidence that superlative quantifiers are harder to process than comparative quantifiers. Thus, as in the case of monotonicity, it would be interesting to compare the frequency of the two types of modified numerals.

3.5. Possible extensions

The list of categories can be obviously extended, depending on the research goals. For instance, among existential quantifiers, one may wish to distinguish value judgments, e.g., "Enough members attended to constitute a quorum" (Keenan and Paperno, 2012) and among non-monotone quantifiers, one may want to distinguish connected quantifiers, e.g., "between 5 and 7" (Chemla et al., 2019). Furthermore, building on already determined quantifier features one may want to focus, for instance, on morphosyntactically complex quantifiers, like already mentioned modifications but also boolean combinations, exception phrases (all but students), bounding phrases (twice a day), or partitive constructions (most of the) (Paperno and Keenan, 2017). The tagging system could be also extended by other quantifier properties known in the literature, like extensionality (Peters and Westerståhl, 2006). Some of those extensions may be carried out automatically or semi-automatically.

4. Annotation and tools

Since there were no attempts at the annotation of quantifiers so far and there are no specific guidelines established, we have decided to follow the general best practices in manual corpus annotation. Each sample in the corpus was annotated simultaneously by two independent annotators. Conflicts between the two were resolved by an additional adjudicator. Since the quantifier theory involves interdisciplinary research originating in logic and linguistics we have decided to recruit annotators with different backgrounds and divide

DRAFT
DRAFT
This is a draft of the paper. The presented data will be updated before publication.

them into two teams. The first team consisted of cognitive science undergraduate students. Most of them had no previous experience with linguistic annotation of any kind but had stronger background in logic. According to the recruitment process, they needed to complete at least four semesters of formal logic courses to be hired in the project. The second team consisted of four qualified linguists (graduates in Polish philology), experienced in various kinds of linguistic annotations: morphological, syntactic and semantic, but with no background in logic. Each sample was annotated by one annotator from each of the teams to diversify insights and reduce oversights in the corpus material. This approach, however, provokes many conflicts between the annotators and brings more work two the adjudicator as the vast majority of samples required adjudicator's intervention. Nevertheless, we believe that by recruiting two teams of annotators with different educational backgrounds we were able to identify all the possible quantificational expressions in the data set (sometimes even redundantly) and for that reason, any future extension of the annotation will be much faster and easier. The annotators had also access to a dedicated mailing list, where they could ask questions and discuss problems concerning their work.

The annotation was conducted in the web-based application WebAnno (Eckart de Castilho et al., 2016) designed for different types of linguistic annotation (fig. 1). WebAnno is based on Java and SQL database, so it has quite standard requirements, which makes it relatively easy to run and operate. The application allows for sharing different projects in one installation, however managing thousands of small samples is less efficient than expected.

During the process of annotation, the annotators had access to some information from other layers NKJP1M, namely: morphosyntactic tags and some selected surface syntactic groups that could be indicators of quantificational usage of an expression. The syntactic groups are limited only to adverbial groups (which could be A-quantifiers) and numeral groups (which most likely is an existential numeral quantifier). However, as we treat quantifiers primarily as semantic units annotators were not bound to those distinctions from other non-semantic layers and are even free to switch off that information from their view if they do not consider it useful.

5. Basic statistics

So far the annotation was completed by the team of linguists. The other team's work is still ongoing but has reached approx. 70% of the overall task and is expected to finish soon. For that reason, numbers provided below refer only to the annotation executed by the linguistic team but in the final version of the paper, they will be updated by the full version of the annotation including adjudication together with inter-annotator agreement coefficient. The work is expected to be concluded by the end of January 2020.

The NKJP1M corpus consists of 18,484 short samples (40-60 words long each). In 11,920 (64.5%) at least one quantifier was annotated. In total 23,165 quantificational expressions were annotated, which is 1,25 on average in each sample.

As it was expected, the most numerous group among the quantifiers are unmodified numerals constituting 30.5% of all units. D-quantifiers are ten times more frequent than A-quantifiers (21063 to 2101). Existential quantifiers (8003) are more frequent than universal ones (4866) which are again more frequent than proportional (3199) which is consistent with the semantic complexity prediction mentioned in the introduction (Szymanik and Thorne, 2017). Comparative and superlative comparison types occur only marginally in the corpus with 229 and 106 occurrences respectively, again comparatives are more frequent than superlatives as expected (Geurts et al., 2010).

As our experience shows, the monotonicity property is the hardest to annotate and brings the most errors to the data, which will need to be corrected in the process of adjudication. However, a preliminary review of the annotation regarding monotonicity confirms the general expectations. Right non-monotone quantifiers are almost absent in the data and those that occur are either false positives (annotation errors) or very specific quantifiers such as *kilka*, *kilkanaście*, *kilkadziesiąt*, *kilkaset* meaning 'more than X and less than Y' (e.g. *kilkanaście* means 'between ten and twenty') and their A-type counterpart (*kilkakrotnie*, *kilkunastokrotnie*). Furthermore, the preliminary data shows that the number of downward monotone quantifiers is comparable to the number of upward monotone quantifiers suggesting that the psycholinguistic effect of monotonicity may not be due to the relative frequencies as is often conjectured in the literature (Degen and Tanenhaus, 2019).

6. Conclusions and future work

The first step in future work with the annotated corpus is an insightful analysis of the annotated units with respect to possible extensions of the quantifier description. The four categories considered in our project are by no means exhaustive and there are many other possible features of quantifiers that could be added in the extended annotation, see section 3.5.

The manually annotated data will also serve as a training corpus for a machine learning classifier aimed at the automatic annotation of quantifiers in large corpora. Based on the annotation we are planning to carry an extensive corpus-based analysis of quantifiers distribution in Polish. One natural direction would be repeating the research on semantic complexity conducted by (Szymanik and Thorne, 2017). The biggest weakness of their analysis was the restriction to 36 most common quantifiers. Using our corpus we could have much broader coverage, approximating all quantifier expressions in Polish, and therefore any statistical generalization about the influence of various semantic factors on linguistic distribution would be more robust. Also, such analysis would be based on a significantly typologically different language than English. Furthermore, an additional factor of text genre could be taken into account.

The corpus will be available on the web both as a separate layer of annotation together with the whole NKJP1M indexed in the corpus search engine and as an XML source tarball for processing in other projects. The corpus will be indexed in MTAS (Brouwer et al., 2017), a multi-tier annotation search engine which allows for indexing multiple layers of annotation. The quantifier layer will be accessible from the Corpus Query Language together with other layers of annotation, which will enable searching for alignments

DRAFT
DRAFT
This is a draft of the paper. The presented data will be updated before publication.

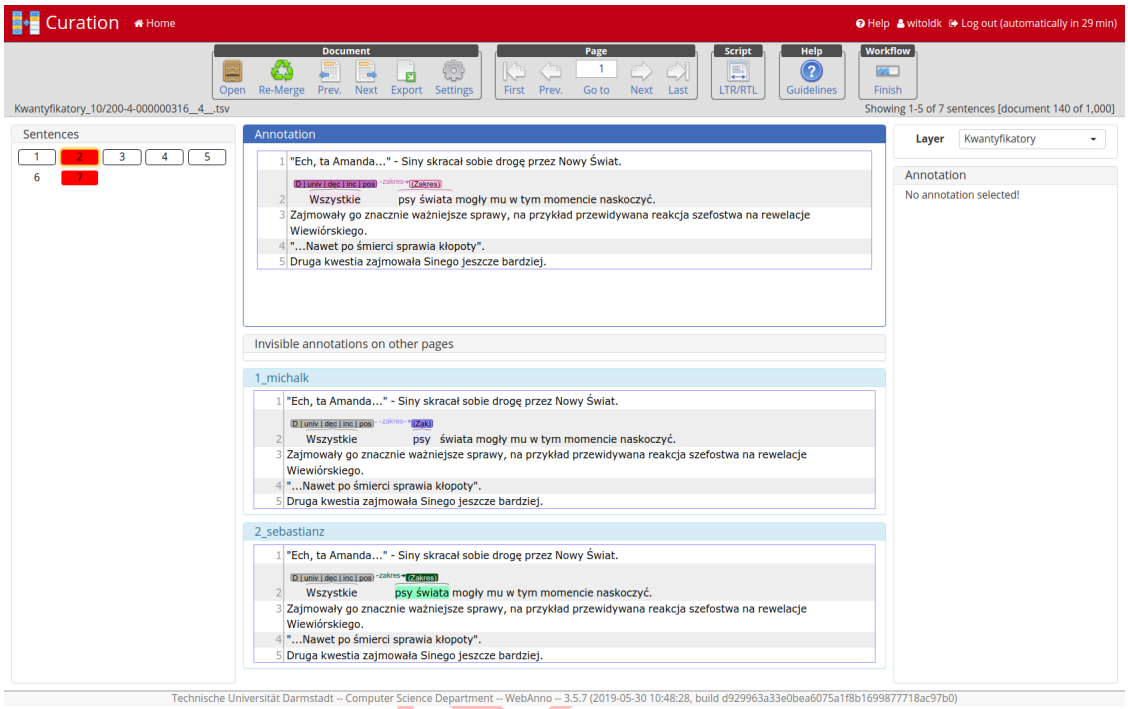

Figure 1: An example of annotated sample with one quantifier *wszystkie* 'all' in WebAnno web application. The two annotators agree with each other with respect to quantifier and its features, the conflict concerns only the scope of the quantifier. The second annotator marks full nominal phase, as required in the manual.

between grammatical and quantificational layers.

7. Acknowledgements

The work being reported was financed by National Science Centre, Poland grant 2017/25/B/HS1/02911.

8. Bibliographical References

E. Bach, et al., editors. (1995). *Quantification in Natural Languages*, volume 54 of *Studies in Linguistics and Philosophy*. Springer.

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4:159–219.

Brouwer, M., Brugman, H., and Kemps-Snijders, M. (2017). MTAS: A Solr/Lucene based Multi Tier Annotation Search solution. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, number 136, pages 19–37. Linköping University Electronic Press, Linköpings universitet.

Chemla, E., Buccola, B., and Dautriche, I. (2019). Connecting content and logical words. *Journal of Semantics*, 36(3):531–547.

Degen, J. and Tanenhaus, M. K. (2019). Constraint-based pragmatic processing. In *The Oxford Handbook of Experimental Semantics and Pragmatics*.

Deschamps, I., Agmon, G., Loewenstein, Y., and Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, 143:115 – 128.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Geurts, B., Katsos, N., Cummins, C., Moons, J., and Noordman, L. (2010). Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes*, 25(1):130–148.

Keenan, E. and Paperno, D. (2012). *Handbook of Quantifiers in Natural Language*. Studies in Linguistics and Philosophy. Springer Netherlands.

Paperno, D. and Keenan, E. (2017). *Handbook of Quantifiers in Natural Language*. Number t. 2 in Studies in Linguistics and Philosophy. Springer International Publishing.

Paperno, D. (2012). Quantification in Standard Russian. In *Handbook of Quantifiers in Natural Language* (Keenan and Paperno, 2012).

Partee, B. H., (1995). *Quantificational Structures and Compositionality*, pages 541–601. Springer Netherlands, Dordrecht.

Peters, S. and Westerståhl, D. (2006). *Quantifiers in Language and Logic*. Clarendon Press, Oxford.

A. Przepiórkowski, et al., editors. (2012). *Narodowy Korpus Języka Polskiego*. Warszawa.

Szymanik, J. and Thorne, C. (2017). Exploring the relation of semantic complexity and quantifier distribution in

DRAFT
This is a draft of the paper. The presented data will be updated before publication.

large corpora. *Language Sciences*.
Szymanik, J. (2016). *Quantifiers and Cognition. Logical and Computational Perspectives*. Studies in Linguistics and Philosophy. Springer.

9. Language Resource References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.
Przepiórkowski, A. and Bańko, M. and Górski, R. L. and Lewandowska-Tomaszczyk, B. (2012). *NKJP1M*. `http://clip.ipipan.waw.pl/NationalCorpusOfPolish`.

DRAFT
DRAFT
This is a draft of the paper. The presented data will be updated before publication.