Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

BUILDING AN AUTOMATIC PIPELINE FOR TYPOLOGICAL RESEARCH: A CASE-STUDY

Buis A. (annebethbuis@gmail.com)
University of Amsterdam, Amsterdam, Netherlands

Bylinina L. (bylinina@gmail.com)
Meertens Institute, Amsterdam, Netherlands

This study investigates a new method of typological research. The method is based on using raw cross-linguistic data found on the internet. We collect corpora of texts in languages of interest and use them as an input into a pipeline, which with the help of computational linguistics methods extracts the grammatical information and, as an output, gives the researcher higher-level generalizations ready for cross-linguistic comparison. We use one particular typological task as a case-study: uncovering morphosyntactic cross-linguistic variation in the domain of numeral constructions. The task is a real one—building a database characterizing such variation is one of the main research goals of the 'Language and Number' project (part of the bigger 'Knowledge and Culture' program, Leiden University / University of Amsterdam / Meertens Institute). The results that we describe in the present paper have not yet been implemented as a full-scale cross-linguistic tool and should be seen at the current stage as a proof of concept. In this paper, we show the method at work for two languages: Russian and Dutch. The main goal of the current paper is thus to show that it is practically possible to set up such a pipeline and to show some problems that arise in the process.

Key words: typology, cross-linguistic studies, web-mining, grammar extraction, natural language processing, syntactic parsing, tagging

АВТОМАТИЗАЦИЯ ЛИНГВИСТИЧЕСКОЙ ТИПОЛОГИИ: ПРИМЕР ИССЛЕДОВАНИЯ

Баус А. (annebethbuis@gmail.com)
Амстердамский университет, Амстердам, Нидерланды

Былинина Л. (bylinina@gmail.com)
Мертенс Институт, Амстердам, Нидерланды

Статья предлагает новую процедуру получения данных для типологических исследований. Эта процедура использует тексты на различных языках, доступные в интернете. Собранные в интернете тексты подаются на вход последовательности программ, которые, используя методы автоматической обработки текстов на естественном языке, дают на выходе грамматические обобщения, пригодные для прямого межъязыкового сравнения. В качестве примера типологического исследования мы используем задачу описания межъязыковой морфосинтаксической вариации конструкций с числительными. Это реальная задача, сформулированная как построение типологической базы данных в рамках проекта “Language and Number” (Лейденский университет / Университет Амстердама / Мертенс институт). Описанные здесь методы еще не были использованы в качестве инструмента полномасштабного решения этой типологической задачи, однако на примере описания русского и нидерландского языков мы стремимся показать, что такое использование принципиально — и практически — возможно.

Ключевые слова: типология, межъязыковые исследования, вебмайнинг, автоматическая обработка естественного языка, автоматическое извлечение грамматической информации, синтаксический парсинг

1. Introduction

Methods of cross-linguistic research haven›t changed in significant ways for decades. The upgrades have mainly concerned the means to reach the speakers and ways to store the information, thanks to computers and internet, but deeper changes in typological methodology haven’t taken place. Most of the work on cross-linguistic variation nowadays involves (some of) the following steps (Keenan and Paperno 2012):

- Consulting the existing grammatical descriptions of languages of interest;
- Consulting cross-linguistic or language-specific descriptions of the phenomenon of interest;
- Creating a questionnaire targeting the phenomenon and sending it off to the speakers of languages of interest;
- Consulting the speakers face-to-face using translation tasks and grammaticality judgment tasks;
- Systematizing the results manually, representing the results as tables, maps or databases.

These methods are time consuming and quite loaded theoretically. The description that the typologist produces is heavily dependent on the theoretical decisions made by the authors of the work that the typologist consults. The perfect deductive, bottom-up typological process would be free from such inherited theoretical artefacts—constructions (and, potentially, grammatical categories) would, ideally, arise as a result of theory-free generalizations over raw data. It is clear that this ideal cannot be reached by traditional methods.

We investigate another method for finding cross-linguistic grammatical information—using cross-linguistic data that is already present online. The proposed methodology does not solve the problem outlined above, but makes a step in this direction.

We use corpora in languages of interest as an input into a pipeline, which with methods from computational linguistics extracts the grammatical information and, as an output, gives the researcher higher-level generalizations ready for cross-linguistic comparison.

We try to see whether it is practically possible to set up such a pipeline. We use one typological task as a case-study: uncovering morphosyntactic cross-linguistic variation in the domain of numeral constructions. The task is a real one—building a database characterizing such variation is one of the research goals of the 'Language and Number' project (Leiden University / University of Amsterdam / Meertens Institute). The results that we describe in the present paper, however, have not yet been implemented as a full-scale cross-linguistic tool and should be seen at the current stage as a proof of concept. We will show the method at work for two languages: Russian and Dutch.

Section 2 introduces the typological project that we will use as a case-study, Section 3—the main section of the paper—describes the 'pipeline' methodology and the procedure for the task at hand. Section 4 sums up our results for Dutch and Russian. Section 5 concludes.

2. Typology of Number Systems

Studies on number cognition show that it consists of two subsystems—the Approximative Number System (ANS) and the Object Tracking System (OTS) (Dehaene 1997, Spelke 2011). The former works with large sets, while the latter is restricted to sets with up to three-four elements. Although the relation between these two systems and natural language is crucial for theories of number, whether grammar systematically reflects this distinction is not clear. If it does, we expect to find typological data revealing a strong split between numbers below 3–4 ('OTS domain') and above 3–4 ('ANS domain') (Carey 1998, Pica et al. 2004, 2006). The "Typology of Number Systems" (TNS) database is being designed to find out.

As a unit of description, the TNS database uses a refined notion of construction as a unique combination of the values of a large set of morphosyntactic parameters. In Russian, for example, similar meanings have to be expressed differently depending on the quantity of items involved:

(1) Odna devochka ushla
1.CARD-FEM.SG.NOM girl-SG.NOM leave-PST.SG.FEM
'One girl left'

(2) Shest' devochek ushli
6.CARD girl-PL.GEN left-PST.PL
'Six girls left'

Quantity '1' requires the numeral to agree with its noun in gender and case, the noun is in singular, and the verb agrees in singular as well. In the case of quantity '6', the numeral doesn't agree with its noun, and both the noun and the verb are in plural. The nouns are in different cases. We could formulate these differences in terms of morphosyntactic parameters. Different potential values thereof in combination give us space for a typology of numeral constructions. (1) and (2) exemplify two different numeral constructions because there are parameters that take different values in these two sentences.

	'1'	'6'	?	?	...
Gender agreement with the head noun	Yes	No	No	Yes	
Case agreement with the head noun	Yes	No	Yes	No	
Number marking on the head noun	Sg	Pl	Sg	Pl	
Number marking on the verb	Sg	Pl	Sg	Pl	

The set of parameters given here is far from complete and serves illustrative purposes. The actual list of parameters currently used in the database is ten times longer. In principle, the list doesn't have to be fixed manually, given the pipeline we discuss in the current paper. The list of parameters is potentially as long as the list of grammatical categories of all classes of lexical items that can stand in direct or indirect syntactic relation to a numeral, and can be extracted from the data.

Natural language expressions for different numerosities that have the same values of morphosyntactic parameters are said to instantiate the same numeric construction. Similarly, numbers can 'belong' to a construction. For example, '6' and '7' in Russian belong to the same construction, while '1' doesn't belong to it.

Therefore, constructions are defined for different points or intervals on the number line, which we call the numeric scope of a construction. Numeric scopes of the constructions exemplified in (1) and (2) are shown on Fig. 1.

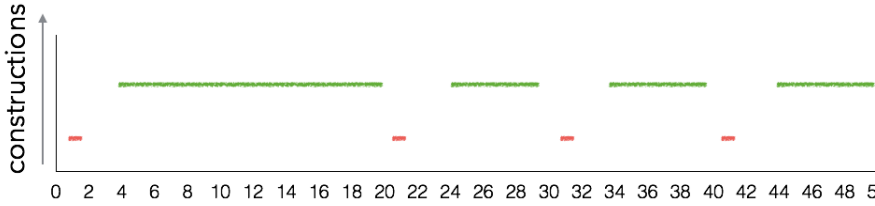

Fig. 1. Numeric scope of constructions

The descriptions of numeral constructions now contained in the database are all the result of data collection using traditional methods of typological research. The rest of the paper discusses an alternative approach to this task—automatic induction of numeral constructions.

3. The pipeline

Using a pipeline means building a tool that uses a start-to-end approach that is general for each language entering the process. The number and the type of steps taken, in principle, does not depend on which language is being studied.

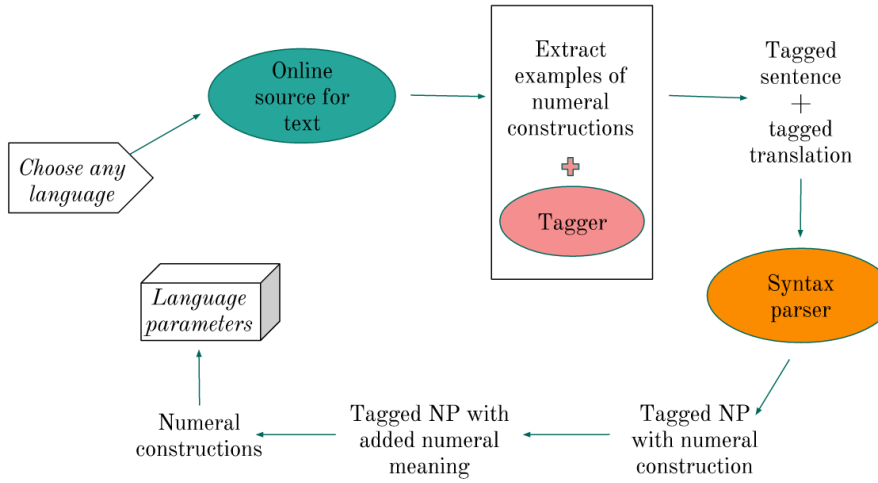

Fig. 2. The pipeline

Figure 2 is a representation of the general outline of the pipeline for extracting numeral constructions from text. As a first step the goal language is selected. In the next step, text in this goal language will be retrieved from an online source. From this data the sentences that contain numeral constructions are selected by using part-of-speech tagging (POS tagging), i.e. assigning a part of speech to each word, and ignoring any sentences that do not contain a numeral. A syntactic analysis of the accepted sentences provides us with information about with which noun the numeral is associated.

In these numeral constructions the numeral is often written in a textual representation (such as: “two” in English). To be able to check the scope of the constructions, each numeral should be converted to a digit representation. The description above gives a short overview of necessary steps for extracting information on numeral constructions. In this case-study we implemented the pipeline for cardinal numeral constructions in Dutch and Russian. Let us now describe the procedure in more detail.

3.1. Collecting a corpus

The goal of the pipeline is to supply data for a typological project. The source for construction of the corpus it is based on should therefore be available in as many languages as possible. Two possible approaches could be considered for getting text from an online source:

- Using an existing corpus of texts
- Creating a collection of texts (by scraping the web)

Under the first approach, one can use language specific corpora (Russian National Corpus for Russian, NEGRA for German, etc.), documents from the United Nations or, for example, the Universal Declaration of Human Rights. This approach has the disadvantage of being more language dependent and therefore less useful for studying a variety of languages. The second approach is to extract the text from websites that contain text in the goal language. A big advantage is that some form of online text will often be available for a lot of languages. Obtaining relevant samples of text without too many errors is, however, not an easy task.

Also available online are versions of the Bible in many different languages. These are a mix between the first and second approach; the Bible is an existing collection of text (a corpus), but it is often necessary to scrape the Bible texts (web-approach). The Parallel Bible Corpus (Mayer and Cysouw 2014) collects online versions of the Bible and currently contains information on 1169 languages. For the pipeline we opted for working with the Bible, selecting full versions (Old and New Testament) for both Russian and Dutch (we will discuss some disadvantages of this decision below). The URL's for the respective Bible versions were located using the Parallel Bible Corpus and the corpus was constructed by scraping.

3.2. POS tagging

Part-of-speech tagging is "the process of [statistically] assigning a part of speech or other syntactic class marker to each word in a corpus" (Jurafsky and Martin, 2009). There are a lot of available tagging tools, but for this project the most important requirement is selecting a POS tagger that tags numerals correctly. The tagging of numerals by a POS tagger does not depend on the implementation of the tagger itself, but on the corpus that was used for the training of the tagger. POS taggers for English, for example, are often based on the same training corpus. As a result of this is most English POS taggers interpret ordinal numerals as adjectives.

A second criterion for selecting a POS tagger for this project is the level of detail in the tagging. The information produced in the pipeline should be valuable for typological research. Thus, a level of specificity is needed in the tagging, in particular singular vs. plural and case-marking on nouns. The tagset POS taggers use also depends on the original training corpus for the tagger. As a result POS taggers for different taggers often use different tagsets. Mapping between different tagsets is not a trivial problem, since the relations between tagsets for different languages are not one-to-one.

For a comparison of available taggers Asmussen's survey of POS taggers (2015) was consulted. We preferred to use a non-language specific tagger in this pipeline project, to make it directly compatible with as many languages as possible. Out of the list of non-language specific taggers, as made up by Asmussen, TreeTagger (Schmid 1995) includes the biggest range of languages (19).

3.3. Syntactic parsing

At this stage a syntactic structure was assigned to the sentences from the corpus allowing us to find the noun the numeral is syntactically connected with. Syntactic parsers have been developed for a range of languages (English, Dutch, etc.), but are much less common than POS taggers. A syntax counterpart of TreeTagger (for a big set of languages) is not available yet. Since this case study focuses on Dutch and Russian, a parser was implemented for each of these languages.

Alpino (Noord, 2006) is a Dutch dependency parser based on the Prolog programming language. The accuracy lies around 88% and the parser is available online. Alpino comes with a graphical interface, but it can also be easily adapted to fit other projects, in particular our pipeline project. In 2012 a contest for Russian dependency parsers was held (Gareyshina et al.). The Russian dependency parser for the Malt Parser (Sharoff & Nivre 2011) won third place with an accuracy of 83% and was the only accessible syntax parser for Russian at the moment of creating this project (the authors thank Svetlana Toldova and Max Ionov for help in setting up the parser). It is important to note that Alpino and the Russian Malt Parser are—for Dutch and Russian respectively—the only parsers that could fit within this project. Although the parsers use a slightly different approach, we will therefore not compare their technical implementation in this paper.

3.4. Assigning number value

The noun phrases obtained by syntactic parsing contain a (complex) numeral and a noun. As is the case in most written text, most numerals in the Bible are represented as words and not as digits. For the purpose of the pipeline it is, however, much more interesting to use a digit representation of the numeral. In that way the data from the pipeline can be useful for identifying splits in numeric scope. We used a naïve approach to convert written numerals in Russian or Dutch to their digit forms, based on NLP module Pattern (De Smedt & Daelemans 2012).

First the numeral is identified. In the case of complex numerals this amounts to all words tagged as numeral and any words that have a coordinating function. This set is passed to a translation service which translates it to English. Only for English a few tools for converting written numerals into digits are available, and they are not very advanced yet. The converter tries to return a digit corresponding to the input—in our case, the input is a translation of the numeral set from Russian or Dutch into English.

Naturally, the result depends greatly on the quality of the translation and the complexity of the number. For Russian, it is successful in 55,4% of the cases. The score for Dutch is much higher, with 94,0% found digits. This can probably be explained by the fact that Russian numerals might be harder to translate due to their morphological complexity. Dutch cardinal numerals do not change and are therefore easy to translate. It might be the case that the most problems occur with really complex numerals with very high numeric value. If it is so, this is not a big obstacle for the task at hand.

4. Results

Let us now turn to the results obtained by processing the Bible via our pipeline approach. We discuss results for Dutch and Russian in turn.

4.1. Dutch

For Dutch a total of 1838 sentences with cardinal numerals were extracted from the Bible data. Dutch does not have a case system and nouns can only bear a plural marker (versus overt marking for singulars). Therefore the only feature that can be obtained from the data is the marking for singular versus plural. We looked at the relationship between the numeral and the noun, by comparing the numerals that occur with singular nouns with the numerals that occur with plural nouns.

Singular: 2–20, 22–25, 28, 30, 32, 36, 37, 40, 42, 43, 47, 50, 52, 54, 56, 60, 62, 65, 66, 72, 73, 74, 80, 92, 95, 98, 100, 112, 120, 138, 150
Plural: 2–20, 22–25, 28, 30, 32–35, 40, 42, 45, 48, 50, 60, 65, 70, 80, 82, 90, 99, 100, 110, 120, 123, 127, 130, 137, 140, 150, 180

The division in Dutch between singular and plural actually lies at > 1—as in many languages, but this is not reflected in the data. Further inspection of the output leads us to conclude that the numbers showing up with singular are either parsing errors or examples of the following kind:

(3) zeventien jaar
17 year
“17 years”

The words “jaar” (year), “keer” (time; as in: “two times”) and a few time- or unit-denoting nouns in Dutch can appear in singular with numerals other than ‘1’. Lexical subcategorisation issues also have to do with genre. The Bible contains words that a POS tagger might mis-tag—say, “cherubim” (a type of angel). Both “cherubim” (singular or plural) and “cherubims” (only plural) exist in Dutch, which makes this word hard to classify. The latter type of error is hard to avoid, but excluding words like “jaar”, “keer” and measure nouns could greatly improve the Dutch results.

By leaving out the nouns “jaar” (year), “maal” (time, as in “two times”) and “man” (man) when marked as singular, the following results can be produced:

Singular: 2–7, 12, 13, 15, 20, 22, 30, 32, 36, 42, 47, 54, 65, 100
Plural: 2–20, 22–25, 28, 30, 32–35, 40, 42, 45, 48, 50, 60, 65, 70, 80, 82, 90, 99, 100, 110, 120, 123, 127, 130, 137, 140, 150, 180

The list of numerals higher than one occurring with a singular is shorter, as expected, but has not completely disappeared. The remaining numerals are results of parsing errors or less common nouns that are also exceptions.

Importantly, the list of numbers that occurred with a singular noun, "1" is missing. In Dutch the same word is used for the numeral 1 ("een") as for the indefinite article. We suspect that all numerals 1 have been parsed as indefinite articles. However, this is not necessarily an issue of the pipeline. The information—or the lack thereof—we extracted here sides with research that has been done on Dutch and in Dutch there is no clear distinction between the numeral 1 and the indefinite article (Barbiers 2007).

4.2. Russian

After processing we obtained 3477 sentences with numerals for Russian. This number is considerably higher than the number of sentences we extracted for Dutch (1838). The difference can partly be explained by a difference in maximum sentence length between the syntax parsers. The parser for Russian can process sentences that are (approximately) 10 words longer than the maximum length sentence in Dutch. As a result, more of the sentences from the Russian corpus will have been parsed.

Russian is clearly a harder case for our pipeline—Russian has rich morphology and free word order, which raises the number of potential morphosyntactic constructions. Massive grammatical ambiguity causes trouble for syntactic parsing. The results we obtained for Russian were less clean than the Dutch ones, so we had to think of heuristics and potential future ways of getting a more reliable result.

The Russian version of the Bible contained a higher number of numerals that were already written in their digit form. In these cases the missing morphological information from the numeral often caused the noun to be tagged incorrectly. In order to deal with other classes of parsing errors that are harder to classify, we filtered out the cases when a numeral occurred in combination with noun in singular/plural only once.

There are however deeper problems for our procedure when applied to Russian. As known independently, number marking in Russian cardinal constructions depends not only on the numeric value of the cardinal, but also on other grammatical properties of the construction. If we simply follow the same procedure as for Dutch, the results will be uninformative—there will be a large overlap between numbers that go with singular and numbers that go with plural, hiding the interaction between grammatical characteristics of the construction. What we can do instead is run the same procedure for subsets of our examples. Each subset will fix the potentially relevant grammatical parameters. As an example, we divided our sample into sub-samples according to case marking on the noun—this allows us to uncover the dependency between case and number marking in cardinal constructions. Here are the results:

NOMINATIVE
Singular: 2, 5, 7, 8, 10, 12–14, 16, 18, 20, 25, 30, 36, 40, 50, 66, 70, 75, 80, 85, 120, 200, 212, 250, 300, 360, 400, 450, 500, 600, 690, 700, 760
Plural: 2, 3

ACCUSATIVE
Singular: 2, 3, 5, 7, 10, 12, 20, 77, 120, 300, 832
Plural: 2, 3, 5, 6, 7, 8, 9, 10, 12, 16, 30, 50, 70, 200, 500

GENITIVE
Singular: 0.5, 2–4, 7, 12, 22–24, 26, 30, 32–34, 42, 52, 84, 162, 182, 242, 284, 300, 403, 600, 782, 962
Plural: 2–20, 23, 25–30, 32, 35, 36, 38, 40, 45, 48, 50, 60, 62, 65, 66, 70, 72, 75, 77, 80, 85–87, 90, 98–100, 105, 110, 112, 119, 120, 128, 130, 140, 144, 150, 160, 174, 180, 187, 200, 202, 205, 207–209, 218, 220, 250, 260, 290, 300, 350, 364, 365, 390, 400, 402, 410, 420, 430, 435, 445, 450, 453, 454, 464, 468, 500, 504, 532, 540, 541, 546, 550, 560, 576, 595, 600, 604, 645, 674, 675, 700, 703, 736, 745, 752, 757, 773, 777, 800, 807, 815, 820, 830, 837, 840, 895, 900, 905, 910, 912, 930, 950, 969, 1331

DATIVE
Singular: 40, 107
Plural: 2, 4, 7, 9–12, 20, 32

INSTRUMENTAL
Singular: 4, 10
Plural: 2–7, 10, 12, 20, 22, 50, 70

LOCATIVE
Singular: 2
Plural: 2–5, 7, 10, 12

These results are not very clean, due to the errors of the parser we used and mistakes in the input (typos/mistakes in the Bible text that we used). We deleted the examples with parser errors manually to be able to judge the cleaner result of the procedure and assess its feasibility. This is what we got:

NOMINATIVE
Singular: 2
Plural:

ACCUSATIVE
Singular: 2–3, 832
Plural: 2, 3, 5, 6, 7, 8, 9, 10, 12, 16, 30, 50, 70, 200, 500

GENITIVE
Singular: 0.5, 2–4 and numbers higher than 20 ending with 2–4
Plural: 2 and higher

DATIVE / INSTRUMENTAL / LOCATIVE
Singular:
Plural: 2 and higher

There are several things to note here. One, the sample seems too small for the task at hand—partitioning the sample further shows gaps (particularly visible in the nominative case here). Second, the number 1 is systematically absent (the parser doesn't tag it as a numeral). Three, there is a split between two groups of cases: nominative/accusative/genitive allows for low numbers to combine with singular nouns, while this is not so for dative/instrumental/locative. This is a genuine distinction in the Russian grammar. Finally, there are overlaps in accusative and genitive—low numbers can combine both with singular and plural nouns. These are two points where other grammatical factors determine number marking—in particular, it is animacy and case marking on the numeral. We could repeat our procedure of sub-sampling to reveal these factors. Importantly, we don't have to know in advance which grammatical categories affect number marking. In principle, we could manipulate all available categories, running the procedure for subsets with fixed values of the parameters until all the overlaps between numeric scopes are gone (if we exhaust all available parameters and still have overlaps, this might mean we are dealing with optional marking).

It is clear how to improve the result by manipulating the input text and the size of the sample together with the cut-off point for filtering out parser artefacts. However, at this point of the project this was not implemented.

5. Conclusion

We described a pipeline for automatic extraction of cross-linguistic grammatical information. We restricted our study in several ways: 1) we focussed on constructions with numerals; 2) we only described Dutch and Russian; 3) as a parameter, we only looked at case marking on the head noun. The pipeline proved to be implementable, although with the growth of morphological complexity (in Russian, compared to Dutch) the output decreased in reliability. The major factors obscuring the output are parsing errors, lexical subcategorization issues and interdependencies between grammatical categories. Dealing with lexical subcategorization can be a part of the pipeline procedure: we know which classes of nouns tend to have special number marking properties cross-linguistically (mostly, these are measure nouns). We can identify these nouns in the sample by translation, and we can also identify nouns that are guaranteed not to have these problems (that would be animate nouns)—and then study these two groups separately in a language of interest. If there turns out to be no difference between the two groups, the lexical class of measure nouns will be treated as not grammatically relevant in a language. Similarly, when—as in Russian—number marking depends on other grammatical properties of the construction, such as case or animacy, we can resort to further stratification of the sample (for example, looking separately at examples with different case marking on the noun). We do not need to know in advance that the dependency is there and what the dependency is—in principle,

nothing prevents the procedure from running several times on groups of sub-samples organised according to different grammatical categories of either of the two words in the pair. Finally, parsing errors cause a serious problem for this approach. An obvious way to filter out random errors is to set up a cut-off point for an item to enter the construction. This will, obviously, increase precision but reduce recall: if the corpus is not big enough, one would get artificial gaps in the numeric scope of numeral constructions. However, these issues can be fixed by increasing corpus size.

The numeral "one" was not present in either of the samples, since in Dutch it is tagged as an indefinite article and in Russian it also caused problems for the tagger. In further research we should come up with a solution for this, because of the fact that for many languages "one" is important for the division between singular and plural. However, for this project, running the pipeline on the Dutch data made clear that a next important step is subcategorisation of nouns. Running the pipeline on the Russian data taught us that another step to improve the pipeline approach is to find interdependencies between different grammatical categories.

References

1. *Asmussen, J.* (2015) Survey of POS taggers. Technical report, DK-CLARIN WP 2.1, http:korpus.dsl.dkclarin/corpus-docpos-survey.pdf.
2. *Barbiers, S.* (2007) Indefinite numerals ONE and MANY and the cause of ordinal suppletion. Lingua, 117(5), 859–880.
3. *Carey, S.* (1998) Knowledge of number: its evolution and ontogeny. Science. 282 no. 5389 pp. 641–642.
4. *Dehaene, S.* (1997) The number sense. Oxford: Oxford University Press.
5. *De Smedt, T. & Daelemans, W.* (2012) Pattern for Python. Journal of Machine Learning Research, 13: 2031–2035.
6. *Gareyshina, A., Ionov, M., Lyashevskaya, O., Privoznov, D., Sokolova, E. G., Toldova, S.* (2012) RU-EVAL-2012: Evaluating Dependency Parsers for Russian. In: Proceedings of COLING 2012, pp. 349–360.
7. *Jurafsky, D. and J. H. Martin.* (2009) Syntactic Parsing In Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.
8. *Keenan, Edward L. & Denis Paperno (eds.)* (2012) Handbook of quantifiers in natural language (Studies in linguistics and philosophy v. 90). Dordrecht, New York: Springer.
9. *Mayer, Thomas & Michael Cysouw* (2014) Creating a massively parallel Bible corpus. In: Proceedings of LREC 2014, 3158–3163.
10. *Noord, van Gertjan.* (2006) At Last Parsing Is Now Operational. In: TALN 2006, pp. 20–42
11. *Pica, P, C. Lemer, V. Izard & S. Dehaene.* (2004) Exact and approximate arithmetic in an Amazonian indigene group. Science, 306(5695), 499–503.
12. *Pica, P., S. Dehaene, V. Izard & E. Spelke.* (2006) Core knowledge of Geometry in an Amazonian Indigene Group. Science n° 311, 381–384.

13. *Schmid, H.* (1995) Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.
14. *Sharoff, S., Nivre, J.* (2011) The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Proc. Dialogue 2011, Russian Conference on Computational Linguistics.
15. *Spelke, E.* (2011) Natural Number and Natural Geometry. In: S. Dehaene & E. Brannon (eds.), Space, Time and Number in the brain. Searching for the Foundations of Mathematical Thought. Attention and Performance XXIV, Chapter 18. Academic Press, 287–317.