Crowdsourcing for the Slovak Morphological Lexicon

Vladimír Benko
UNESCO Chair in Plurilingual and Multicultural Communication
Comenius University in Bratislava
Šafárikovo nám. 6, SK-81499 Bratilava, Slovakia
and
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Panská 26, SK-81101 Bratislava, Slovakia

Abstract. *We present an on-going experiment aimed at improving the results of Slovak PoS tagging by means of increasing the size of morphological lexicon that is used for training the respective tagger(s). The frequency list of out-of-vocabulary (OOV) word forms along with the tags and lemmas assigned by the guesser is manually checked, corrected and classified by students in the framework of assignments, so that valid lexical items candidates for inclusion into the morphological lexicon could be identified. We expect to improve the lexicon coverage by the most frequent proper names and foreign words, as well as to create an auxiliary lexicon containing the most frequent typos.*

1 Introduction

"Crowdsourcing" is a relatively recent concept that encompasses many practices. This diversity leads to the blurring of the limits of crowdsourcing that may be identified virtually with any type of Internet-based collaborative activity, such as co-creation or user innovation [1]. In their paper, authors define eight characteristics typical for crowdsourcing as follows:

- *There is a clearly defined crowd (a)*
- *There exists a task with a clear goal (b)*
- *The recompense received by the crowd is clear (c)*
- *The crowdsourcer is clearly identified (d)*
- *The compensation to be received by the crowdsourcer is clearly defined (e)*
- *It is an online assigned process of participative type (f)*
- *It uses an open call of variable extent (g)*
- *It uses the Internet (h)*

From this perspective, language data annotation performed by students in the framework of the end-of-term assignments can well be considered "crowdsourcing", even if only some of the above characteristics apply. It is also worth noting that, according to our experience, students appreciate the feeling that their work may be useful not only as a tool for classification.

2 The Problem

Slovak belongs to languages with more than one system for morphosyntactic annotation available, with two of them being actively used in our work[1] . They have been developed (partially independently) in the framework of two different research projects.

The *Slovak National Corpus* (*SNC*) [2] is using a system based on the new Czech *MorphoDiTa* tagger [3, 4] with a custom language model and a tool for guessing lemmas for unrecognized (out-of-vocabulary – *OOV*) lexical items; while the *Aranea* Project [5, 6] is using a more traditional *TreeTagger* [7, 8] with a custom language model, yet without any functionality to guess lemmas for the *OOV* lexical items. Both systems are using the *SNC* tagset[2] [9] – a fine-grained positional tagset vaguely resembling the popular *MULTEXT-East*[3] tagset utilized for several Slavic languages.

Language models for both systems, however, have been trained on the same source data – the 1.2 M token *Manually morphologically annotated corpus*[4] and the *SNC Morphology database*[5] covering approx. 100 K lemmas, yielding some 3.2 M inflected forms. This is why that, despite the fact that both systems do not produce exactly the same output, they are (almost) identical[6] in the amount of *OOV* items, that is rather high.

As both Slovak annotation systems explicitly indicate the *OOV* status of every token within a corpus, an analysis of the situation can be conveniently performed by the corpus manager, such as NoSketch Engine[7] [10]. In the *SNC* corpora, the *OOV* status is indicated by the "*XX*" value pf the "prec" attribute – this value can be observed in 54.5 million cases of 1.37 Gigatoken *prim-8.0-pubic-sane*[8] main corpus, which is 3.98% of all tokens.

In the web-based *Araneum Slovacum Maximum*[9], where the *OOV* state is indicated by the "*0*" value of the "ztag" attribute, the situation is even worse – 135.5 million *OOV*s out of 2.96 Gigatokens, i.e., 4.57%. This can be explained by the rather "low quality" of web data that, despite all efforts in cleaning and filtering the source texts, naturally contains lots of "noise" of different kinds.

3 The Task

The *OOV* lexical items observed in our corpora are of different nature. Besides the "true neologisms", i.e., words qualifying for inclusion even into the traditional dictionary, proper nouns (such as personal and geographical names) and their derivates, we can find also items traditionally not considered as "words" – various abbreviations, acronyms and symbols, URLs or e-mail addresses, parts of foreign language quotations and – above all – all sorts of "typos" and "errors". Inflected word forms apply to almost all previously mentioned categories, which makes the whole picture even more complex.

[1] We are aware of (at least) two more systems for morphosyntactic annotation of Slovak data that have been independently developed at Masaryk University in Brno and Charles University in Prague, respectively. These two systems, however, were not available for our work at the time of writing this paper.

[2] https://korpus.sk/morpho_en.html
[3] http://nl.ijs.si/ME/V4/
[4] https://korpus.sk/ver_r(2d)mak.html
[5] https://korpus.sk/morphology_database.html
[6] The differences are mainly caused by the fact that the *TreeTagger*-based system is also using word forms from the training corpus that were not present in the morphological database (mostly proper nouns) to ammend the morphological lexicon,
[7] https://nlp.fi.muni.cz/trac/noske
[8] https://korpus.sk/prim(2d)8(2e)0.html
[9] http://aranea.juls.savba.sk/aranea_about

In the following text we present an experiment aimed at amending the morphological lexicon used for training the language model(s) by a manually validated list of most frequent *OOV* items derived from an annotated web corpus. The annotation is to be performed by graduate students of foreign languages, in the framework of end-of-term assignment for the "Introduction to Corpus Linguistics" subject.

Having only limited "human power" (two groups with 46 students in total) at hand, we decided to follow the minimal two-fold setup (i.e., each item to be annotated by only two independent annotators) and make the task as simple as possible. This is why the annotators were not expected to check all the morphological categories provided by the respective tags, and they were asked to decide only on two parameters – lemma and word class (part of speech).

4 The Data

In the first step, we used data from the *Araneum Slovacum Maximum 17.09* web corpus of approx. 3 Gigatokens that has been independently tagged both by the *SNC MorphoDiTa* and the *Aranea TreeTagger* pipelines, and subsequently merged into a single vertical file. Then, we converted the original *SNC* morphological tags to "PoS-only" tags and produced a frequency list of all lexical items indicated as *OOV* by both taggers. This list has been further filtered to exclude word forms contained in the *Czech* morphological lexicon[10]. After deleting the unused parameters, the resulting lists contained the frequency, word form, lemma assigned by the *SNC* guesser and PoS information derived from the tag assigned by *TreeTagger* (*aTag*, using the *AUT*[11] notation). This decision has been motivated by an observation that *TreeTagger* is typically more successful in assigning morphological categories for unknown words than *MorphoDiTa*.

As we naturally could expect to be able to process only the rather small part of the list, after some experimenting with various thresholds, we decided to pass into annotation only items appearing 50 or more times, yielding to 77,169 items. This meant that each annotator would process approximately 3,300 items.

The example of source data (after discarding the frequency information and adding a unique Id) is shown in Table 1.

We can observe several phenomena here. The same lexical item is in some cases tagged as "foreign", while as "noun" or "adjective" in the others, and lemma form as well as its capitalization is sometimes guessed correctly, while sometimes not. It can be also seen, that many table items will in fact have to be merged after correcting the annotation, producing less total of correct lines.

The overall task for the annotators was to produce correct data for all lines in the table. To minimize the number of necessary keystrokes and to keep track of the changes, the data have been further modified to contain two newly added columns – *Lemmb* used as a template for correcting the value for *Lemma* (it is expected that most modifications will occur at the end of the respective string only) and *bTag* (to be filled only in case of wrong PoS assignment).

Table 1. Source Data

Id	*Word*	*Lemma*	*aTag*
sk_11184	dvojťažiek	dvojťažka	Nn
sk_11185	dvojťažiek	dvojťažky	Nn
sk_11186	dvojťažka	dvojťažka	Nn
sk_11187	Dvojťažka	dvojťažka	Nn
sk_11188	Dvojťažka	Dvojťažka	Nn
sk_11189	Dvojťažka	dvojťažka	Yx
sk_11190	Dvojťažka	Dvojťažka	Yx
sk_11191	dvojťažkách	dvojťažke	Nn
sk_11192	dvojťažke	dvojťažka	Nn
sk_11193	dvojťažkou	dvojťažka	Nn
sk_11194	dvojťažku	dvojťažka	Nn
sk_11195	dvojťažky	dvojťažka	Nn
sk_11196	dvojťažky	dvojťažky	Av
sk_11197	dvojťažky	dvojťažky	Nn
sk_11198	dvojtisícovku	dvojtisícovka	Nn
sk_11199	dvojtlačidlo	dvojtlačidlo	Nn
sk_11200	dvojtraktovú	dvojtraktový	Aj
sk_11201	dvojumývadlom	dvojumývadlom	Nn
sk_11202	dvojumývadlom	dvojumývadlom	Yx
sk_11203	dvojzákrutovej	dvojzákrutovej	Aj
sk_11204	dvojzákrutovej	dvojzákrutovej	Yx
sk_11205	dvojzápasovú	dvojzápasový	Aj
sk_11206	dvojzónovú	dvojzónový	Aj
sk_11207	dvolezite	dvolezite	Nn
sk_11208	dvolezite	dvolezite	Yx
sk_11209	Dvonča	Dvonča	Nn
sk_11210	Dvonča	Dvonč	Nn
sk_11211	Dvončom	Dvonča	Nn
sk_11212	Dvončom	Dvonč	Nn

As has been already mentioned, each item (line of the table) has to be annotated by two independent annotators. We decided, however, not to split the data in a straightforward way, but to assign each alphabetical segment of the data to *three* annotators using a rule as follows: each triple of lines will be split into three tuples containing first and second, first and third and second and third lines, respectively. Moreover, the whole lot of data has been split to three parts, so that each annotator could get three different sections of the alphabet in his or her data.

By applying this fairly "sophisticated" assignment scheme, we expected to improve the overall uniformity and quality of the output, as well as to prevent "collaboration" among students, as no two assigned lots were identical.

An excerpt of the data from Table 1 assigned to a single annotator is shown in Table 2.

Table 2. Data to Annotate

Id	*Word*	*Lemma*	*Lemmb*	*bTag*	*aTag*
sk_11184	dvojťažiek	dvojťažka	dvojťažka		Nn
sk_11185	dvojťažiek	dvojťažky	dvojťažky		Nn
sk_11187	Dvojťažka	dvojťažka	dvojťažka		Nn
sk_11188	Dvojťažka	Dvojťažka	Dvojťažka		Nn
sk_11190	Dvojťažka	Dvojťažka	Dvojťažka		Yx
sk_11191	dvojťažkách	dvojťažke	dvojťažke		Nn
sk_11193	dvojťažkou	dvojťažka	dvojťažka		Nn
sk_11194	dvojťažku	dvojťažka	dvojťažka		Nn
sk_11196	dvojťažky	dvojťažky	dvojťažky		Av
sk_11197	dvojťažky	dvojťažky	dvojťažky		Nn

Note that the "missing" every third Id results from the assignment scheme.

5 The Crowd Annotation

The split data has been uploaded as excel spreadsheets to a shared Google disk and assigned randomly to the respective annotators. The task has been assigned in the middle of

[10] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1836

[11] http://aranea.juls.savba.sk/aranea_about/aut.html

the semester, after the students already got acquainted with the basic concepts of corpus morphosyntactic annotation and acquired the elementary querying skills.

The instructions for annotating the data were as follows.

(A) Only *Lemmb* and *bTag* columns may be modified.

(B) If both *Lemma* and *aTag* values are correct, nothing has to be done.

(C) If *aTag* value is wrong, the correct value should be inserted in bTag.

(D) If *Lemma* value is wrong, it should be corrected in Lemmb.

(E) If the word form is obvious typo (missing or superfluous letter, exchanged letters), or the word does not contain the necessary diacritics, the correct lemma marked by an asterisk should entered in *Lemmb*.

(F) If the correct word form cannot be reconstructed by simple editing operations, i.e., cannot be recognized (e.g., part of the word as a result of hyphenation), the value of *bTag* will be "*Er*" (error).

(G) If the word form is obvious foreign word, the value of *bTag* will be "*Yx*".

(H) It is not necessary to evaluate whether the word form is "literary" – words of "lower" registers (such as slang) also have "correct" lemmas.

The annotators were also instructed to check all "non-obvious" items by querying the corpus and analyzing the respective contexts. The initial training was performed during one teaching lesson in a computer lab, so that possibly all frequent problems could be explained.

6 First Results and Problems

Out of 46 students, 43 managed to complete the assignments in time. Table 3 shows an example of the correctly annotated data.

Table 3. Annotated Data

Id	*Word*	*Lemma*	*Lemmb*	*bTag*	*aTag*
sk_11184	dvojťažiek	dvojťažka	dvojťažka		Nn
sk_11185	dvojťažiek	dvojťažky	dvojťažka		Nn
sk_11187	Dvojťažka	dvojťažka	dvojťažka		Nn
sk_11188	Dvojťažka	Dvojťažka	dvojťažka		Nn
sk_11190	Dvojťažka	Dvojťažka	dvojťažka	Nn	Yx
sk_11191	dvojťažkách	dvojťažke	dvojťažka		Nn
sk_11193	dvojťažkou	dvojťažka	dvojťažka		Nn
sk_11194	dvojťažku	dvojťažka	dvojťažka		Nn
sk_11196	dvojťažky	dvojťažky	dvojťažka	Nn	Av
sk_11197	dvojťažky	dvojťažky	dvojťažka		Nn
sk_11199	dvojtlačidlo	dvojtlačidlo	dvojtlačidlo		Nn
sk_11200	dvojtraktovú	dvojtraktový	dvojtraktový		Aj
sk_11202	dvojumývadlom	dvojumývadlom	dvojumývadlo	Nn	Yx
sk_11203	dvojzákrutovej	dvojzákrutovej	dvojzákrutový		Aj
sk_11205	dvojzápasovú	dvojzápasový	dvojzápasový		Aj
sk_11206	dvojzónovú	dvojzónový	dvojzónový		Aj
sk_11208	dvolezite	dvolezite	dôležitý*	Aj	Yx
sk_11209	Dvonča	Dvonča	Dvonč		Nn
sk_11211	Dvončom	Dvonča	Dvonč		Nn
sk_11212	Dvončom	Dvonč	Dvonč		Nn

We can see that PoS information was corrected in four cases, lemma form in nine cases and its capitalization in two cases. One lexical item was marked as "error", as it lacked all diacritics and used nonstandard spelling.

The quick analysis, however, revealed that the annotation is much below the expected quality. We will discuss some of the issues. The basic statistics is shown in Table 4.

Table 4. Results of Annotation

	Count	**%**	**%**
Assigned lines	77,169	100.00	
Lines annotated at least once	76,413	99.02	
Lines annotated twice	60,048	77.81	100.00
Lines agreed on lemma	39,469	51.15	65.73
Lines agreed on lemma and PoS	33,371	43.24	55.57

The rather low values of the raw inter-annotator agreement suggests that the resulting data has to be analyzed thoroughly before the procedure can be used within a similar larger-scale annotation attempt in the future.

The quick analysis revealed some frequent issues – different treatment of (prototypically) proper names written in lowercase, assigning PoS information to symbols and foreign words, incoherent use of asterisks, etc. Some of these issues can be solved by an automated procedure but some will require more detailed instruction so that a correct annotation could be obtained.

After merging the duplicate "fully agreed" items from the pevious table, 27,135 unique lines were obtained. Table 5 shows the word class distribution of the resulting data.

Table 5. Annotated Data PoS Distribution

PoS	**Count**	**%**
Nn	20,043	73.86
Aj	5174	19.07
Pn	46	0.17
Nm	27	0.10
Vb	464	1.71
Av	261	0.96
Pp	8	0.03
Cj	10	0.04
Ij	42	0.15
Pt	24	0.09
Ab	185	0.68
Xy	1	0.00
Yx	490	1.81
Er	343	1.26
?	17	0.06
	27,135	100.00

The values in the table basically follow our expectations: most unrecognized items belong to main convent word classes – nouns and adjectives. Moreover, out of the 20,043 words tagged as nouns, 14,190 (70.80%) begin with uppercase letter, i.e., they are most likely proper nouns.

The rather low value of the "Er" class can be explained by the observation that errors, despite their being frequent, rarely behave "paradigmatically", i.e., a single correct word form can produce many *different* incorrect ones.

7 Conclusions and Further Work

There were several goals to be achieved by the annotation. Firstly, we would like to produce a validated list of most frequent neologisms to be included in the morphological lexicon; in this stage, we even do not expect to generate full paradigms for those lexical items. Secondly, we wanted to get the list of the most frequent typos and other types of errors that could also be used as a supplement to that lexicon, but also as source data for a future system for data normalization. And lastly, we also wanted to obtain a list of most frequent foreign lexical items appearing in Slovak corpus data.

Although the detailed analysis of the annotated data is yet to be performed, some conclusions can be seen already. They can be summarized as follows:

(1) To minimize the consenquences of students' failed assigments, a three-fold setup would be probably better.

(2) The Annotation Guidelines must be as precise as possible, showing not only the typical problems and their solutions, but also the seemingly "easy" cases. One-page instruction, as it was in our case, is definitely not sufficient.

(3) The most common errors were associated with the treatment of proper nouns. An automatic procedure based on frequencies of lower/uppercased word forms would most likely perform better.

(4) The other common issue was the proper form of lemma for adjectives (it should be masculine and nominative singular). As the morphology of Slovak adjectives is fairly regular, a procedure to fix it automatically would be feasible.

(5) One of the fairy frequent PoS ambiguity in our data was the "Nn"/"Yx" (noun/foreign) case. The manually annotated data, however, show that the real number of "foreigns" is rather low, yet in introduces a lot of noise into the annotation process. It would therefore be reasonable to substitute all tags for "foreigns" with that of "nouns" in the future annotation.

In the near future, besides the new round of a similar annotation effort with an improved setup, we would like to combine its results with those oblained in the framework of the ensemble tagging experiment described in our other work [11].

Acknowledgment

This work has been, in part, funded by the Slovak KEGA and VEGA Grant Agencies, Project No. K-16-022-00, and 2/0017/17, respectively.

References

[1] E. Estellés-Arolas and F. González-Ladrón-de-Guevara. Towards an Integrated Crowdsourcing Definition, Journal of Information Science, 38 (2): 189–200, doi:10.1177/0165551512437638.

[2] M. Šimková and R. Garabík. Slovenský národný korpus (2002–2012): východiská, ciele a výsledky pre výskum a prax. In Jazykovedné štúdie XXXI. Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu). Ed. K. Gajdošová – A. Žáková. Bratislava: VEDA 2014, pp. 35–64.

[3] D. "johanka" Spoustová, J. Hajič, J. Raab and M. Spousta. Semi-Supervised Training for the Averaged Perceptron POS Tagger. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pp. 763–771, Athens, Greece, March. Association for Computational Linguistics.

[4] J. Straková, M. Straka and J. Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[5] V. Benko. Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014.

[6] V. Benko. Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016). – Portorož : European Language Resources Association (ELRA), 2016, pp. 4245–4248. ISBN 978-2-9517408-9-1.

[7] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester. 1994.

[8] H. Schmid. Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop, Dublin. 1995.

[9] R. Garabík and M. Šimková. Slovak Morphosyntactic Tagset. In Journal of Language Modeling. Institute of Computer Science PAS, 2012, Vol. 0, No. 1, pp. 41–63.

[10] P. Rychlý. Manatee/Bonito – A Modular Corpus Manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2007. pp. 65–70. ISBN 978-80-210-4471-5.

[11] V. Benko and R. Garabík. Ensemble Tagging Slovak Web Data. Accepted for presentation at the SlaviCorp 2018 Conference, Prague, 24–26 September, 2018. Unpublished.