LDK 2023

Validation of the Bigger Analogy Test Set Translation into Croatian, Lithuanian and Slovak

Radovan Garabík
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences,
garabik@kassiopeia.juls.savba.sk

Ana Ostroški Anić
Institute of Croatian Language and Linguistics, aostrosk@ihjj.hr

Sigita Rackevičienė, Giedrė Valūnaitė Oleškevičienė, Linas Selmistraitis
Mykolas Romeris University, {sigita.rackeviciene, gvalunaite,selmistraitis}@mruni.eu

Andrius Utka
Vytautas Magnus University, andrius.utka@vdu.lt

Abstract

This paper presents ongoing work focused on the analysis of translations of the English Bigger Analogy Test Set (BATS) dataset into three languages: Croatian, Lithuanian, and Slovak. We describe our automatic validation and further manual correction of the translations and analyse the main types of issues encountered in the dataset. The validation process involves checking the translations against morphological databases in order to uncover obvious mistakes or typos. Additionally, the translations are tested for the compliance to some of the formal guidelines for the Bigger Analogy Test Set translations, and for rudimentary grammatical correctness.

1 Introduction

1.1 Description of the the Bigger Analogy Test Set

Word embeddings are widely used in various Natural Language Processing tasks and toolkits. One of the features of the embeddings is that the vector space captures relations between the words and maps them to relations between the vectors, which leads to the word analogy based on vector arithmetic (commonly cited example is $king - man + woman = queen$) (Mikolov et al., 2013). The Bigger Analogy Test Set (BATS) was developed as a balanced analogy test set with 40 morphological and semantic relations (which yielded total 99,200 questions according to (Gladkova et al., 2016)) to draw the attention of the NLP community to word embeddings and analogical reasoning algorithms in the context of lexicographic and derivational relations (Gladkova et al., 2016). BATS includes inflectional and derivational morphology, and it also covers lexicographic and encyclopedic semantics. Each relation is represented by 10 categories, with each category containing 50 unique word pairs, e.g. $bird - feathers$ and $door - threshold$ for the relation of meronymy or $bicycle - bike$ and $loyal - faithful$ as examples representing synonymy. This layout produces 98,000 questions for the vector offset method.

The BATS bears superficial similarity to the WordNet database of semantic relations between words. While the original WordNet project (Fellbaum, 2005) covers English, numerous other WordNets and WordNet-like databases are available for many languages (Bond and Paik, 2012; Vossen et al., 2016). However, while some of the semantic relations are identical, the similarities stop there. The WordNet aims to encompass a broad range of vocabulary, ideally to cover as much of the general language as possible, and centered on the concept of sets of semantically equivalent words (*synsets*). The BATS is a specialized dataset including a pre-selected set of words and a comprehensive range of terms related to them by the given relation, incorporating highly specialized and rare lexical items. Moreover, the majority of the WordNets include only basic vocabulary or exhibit other major gaps in lexica. Nevertheless, individual language WordNets are a valuable source to consult when translating the BATS dataset.

The current study stemmed from one of the targets of the COST action *NexusLinguarum* of the creative utilization of pre-trained neural language models in order to acquire RDF relations, which form a foundation of the Linguistic Linked Open Data (LLOD) and which in turn can be used as a valuable source of curated data for Deep Learning methods. This task requires a multilingual

evaluation set of lexico-semantic relations to allow testing various potential methods for relation acquisition from neural language models across languages. Thus, the COST action started the initiative to create such a dataset by manually translating the existing English BATS dataset to as many languages as possible, by initially focusing on translating the lexico-semantic portion of the dataset. Since BATS has so far been adapted to Japanese (Karpinska et al., 2018) and Icelandic (Friðriksdóttir et al., 2022), this is indeed a large-scale initiative.

This paper presents an automated validation process developed for the purpose of assessing the translated datasets' compliance with certain formal requirements, such as spell check, basic grammar and syntax verification. It also discusses the results of validation, focusing on true and false positive results, which often indicate errors in the initial dataset or reflect deliberate decisions regarding translation equivalents.

1.2 Analysed Languages

The Slovak language belongs to the West Slavic group of Slavic languages. It is the official and main language in Slovakia, spoken by about 5 million native speakers (conservative estimate based on the 2011 census data). It can be characterized as a medium-level inflected, subject-verb-object language with three grammatical genders, seven cases[1], two grammatical numbers, three tenses and two verbal aspects. Adjectives are inflected for gender, number and case and agree with the noun in these categories. These features are shared with most Slavic languages.

Being in the group of the Western South-Slavic languages, Croatian is typologically very similar to Slovak, with which it shares many grammatical features, e.g. the level of inflectional complexity, three grammatical genders, two grammatical numbers, and agreement between nouns and adjectives. It also has seven cases, three simple and three compound tenses, three moods, and four participles (Tadić, 2007). Its standardized variety is the official language of the Republic of Croatia, and is spoken by about 7 million native speakers around the world (Eberhard et al., 2023).

The Lithuanian language is one of two living languages of the Baltic branch of the Indo-European language family (the other living Baltic language is Latvian). It is the official state language of the Republic of Lithuania and has about 2.67 million speakers in Lithuania and about 0.6 million speakers abroad (VLE, 2023). Lithuanian is a highly inflected language. Notional parts of speech are inflected by cases (nouns, pronouns, adjectives, participles, numerals), by person (verbs) or are uninflected (adverbs). The parts of speech inflected by cases have two or three grammatical genders (nouns have two, while the other parts of speech have three), two grammatical numbers (some pronouns have, in addition, the dual number), and the declension system comprised of case paradigms, the number of which varies across the parts of speech. Nouns and adjectives agree in gender, number and case. Verbs have three grammatical persons, two grammatical numbers, four tenses, four moods and two voices. The only uninflected notional part of speech is adverb, but many adverbs still have the morphological category of degrees of comparison (Ambrazas et al., 2006).

Slovak, Croatian and Lithuanian thus share several grammatical features that make them quite compatible for the cross-linguistic comparison and this analysis. All three languages are synthetic, SVO with a relatively free word order, with medium to high level inflection, and in general they have two grammatical numbers and three genders. All have noun-adjective agreement in gender, number and case, and – not less relevant – all three have adverbs as the only uninflected part of speech that appears in the lexico-semantic part of the BATS dataset.

The remainder of the paper is structured as follows: the guidelines for translating the BATS dataset are briefly presented in the next section. In section 3, the morphological databases of Croatian, Lithuanian and Slovak are described, which were used for the validation process, explained in section 4. The results of validation are discussed in detail in section 5, from the point of view of each language.

2 Description of the BATS Translation Process

We begin by introducing several expressions that will be used throughout the article. We use the term *source word* to indicate the word from which

[1]The number of cases and genders depends on the level of abstraction of morphological analysis and on inclusion of marginal features; thus sometimes we encounter six cases and four genders

the semantic relation originates. Conversely, we refer to the word related by the given semantic relation (i.e. the second member of the related pair of words), as the *target word*. The term *word* encompasses both single words and multiword expressions in this context. It is important to note that these terms are not related to the notion of the 'source' or 'target' language. If we take meronyms as an example, in the English original dataset *roof* is the source word, while *shingles, tiles, wood, metal* are the target words in the meronymic relation. Similarly, in the Slovak translation, *strecha* is the source word, while *škridle, dlaždice, drevo, kov* are the target words.

By *entry*, we understand one source word, accompanied by all the target words, and all corresponding translations in the given language. We call a single source word with the corresponding translation (or multiple translations) an *item*. An *entry* is thus composed of the list of *items*.

Detailed translation guidelines to be used as internal for the *Use Case 4.1.3 – Acquiring RDF Relations with Neural Language Models* were drafted by the task coordinator specifically for the task of translating the BATS dataset into 19 European languages. However, translation processes did not all start at the same time, and they are currently at various stages. The guidelines prescribed manual translation as they were intended to focus on possible issues in finding equivalents for the original English examples strictly. In particular, machine translation and post-editing is strictly prohibited. Apart from the expected common semantic phenomena, such as polysemy and synonymy, English examples contained a large number of culturally specific words, which were deemed as potentially too language specific, and for which finding appropriate equivalents proved to be challenging. For this reason, as well as in order to achieve a high level of validation, all translations were to be carried out manually. For each English word, the most common or the most frequent equivalent in the target language was chosen. Translation equivalents could be tested with a quick Google search to compare frequencies or by consulting dictionaries, word embeddings, online resources, etc., and choosing the most relevant translation. There was a possibility to add other equivalents commonly used on the line below the final target word, not aligned with a specific target word. In order to identify duplicates, i.e. two or more words in the target language that are used for one word in the original dataset, the label `DUPLICATE` was to be used. Similarly, in cases where there was no appropriate equivalent word in the translation, the label `NO_TRANSLATION` was used. In order to allow for replicability and comparison of the English data and the translated files, the guidelines strictly forbade changing anything in the original English dataset, including obvious errors and the duplication of words in certain pairs.

In the Slovak translation of the dataset, we decided to keep the translations blank in such instances, as it was frequently impossible to find an adequate number of valid and distinct target words. This approach differs from the use of the `NO_TRANSLATION` keyword. In the latter case, it indicates the existence of either a genuine lexical lacuna or a situation where the target word's concept is too regional and does not have a direct (loanword) equivalent in the target language.

In Table 1 we summarize the categories, identified by prefixes of the individual files. We will use these identifiers to refer to the categories and their translations.

category ID	relation
L01	hypernyms – animals
L02	hypernyms – misc
L03	hyponyms – misc
L04	meronyms – substance
L05	meronyms – member
L06	meronyms – part
L07	synonyms – intensity
L08	synonyms – exact
L09	antonyms – gradable
L10	antonyms – binary

Table 1: List of lexical categories

3 Morphological Databases

In the validation, we use morphological databases, i.e. triplets of *lemma, word, morphosyntactic description (MSD) tag* for some validation steps. We briefly describe the databases for our analysed languages.

3.1 Croatian

The Inflectional lexicon hrLex 1.3 (Ljubešić, 2019) is an inflectional lexicon of the Croatian language in which each entry consists of a word form, lemma, MSD, MSD features, UPOS, morphological features, frequency, and per-million frequency. The wordform, lemma, and MSD frequencies are

calculated on the hrWaC v2.2 corpus. The process of compiling the initial lexicon is described in (Ljubešić et al., 2016). The database met all the validation requirements, but minor issues in initial lemmatization (e.g. that participles are lemmatized as verbs) led to creating false positives in the validation process.

3.2 Lithuanian

The Lithuanian Morphological Database was specially designed for the validation of Lithuanian BATS translation. The database contains all types and lemmas for nouns, adjectives, verbs, and conjunctions extracted from the Joint Corpora of Lithuanian, as well as their morphological analyses. The wordlist of types, which is the base of the Lithuanian Morphological Database, is freely accessible from the CLARIN-LT repository (Dadurkevičius, 2020). The database includes more that 1.43 million unique word forms (types). Since the database includes only 4 parts of speech, our validation generated errors for translation including the missing parts of speech, i.e. numerals, adverbs, prepositions, and pronouns.

3.3 Slovak

The Slovak Morphological Database is a database of lemmas and their inflected word forms. The database includes 114,634 lemmas, selected from various Slovak dictionaries and supplemented with the most frequent words from the Slovak National Corpus. Each lemma is provided with a full paradigm along with morphological tags representing grammatical information. The database currently holds about 1.3 million unique word forms, for a total of 3.8 million entries (including homonyms). The database is used for automatic lemmatization and tagging of texts in the Slovak National Corpus and other Slovak corpora (Garabík and Mitana, 2022).

4 Validation Description

4.1 Validation Levels

The automated validation process assesses the translated dataset compliance with formal requirements, which encompasses the syntax of the files, spell-check, and a simple grammar check of multiword terms. During this validation, we recognize three degrees of significance:

- `ERR` is a hard error, either a formatting error, or a duplicate translation. Issues labeled as `ERR` have high probability of being true positives
- `WARN` is a less serious issue, including spelling mistakes or unusual characters in the terms. These issues are quite often false positives.
- `NOTE` is just a notice. This is used to indicate missing translations.

4.2 Validation Steps

The first step involves the initial validation of the formal format following the BATS translation guidelines. This step focuses on a limited set of checks to allow for progress to the subsequent validation stages. The syntactical checks, in the sense of the formal syntax of the entries, include the following criteria: the translation must not be empty, multiword expressions should use the underscore character as the word separator instead of spaces, and all-capitals entries longer than one character should only consist of the strings `DUPLICATE` or `NO_TRANSLATION` as their values.

The second step involves validating the orthography and grammar of the entries. We compare the entries against a morphological database that includes lemmas and inflected words. Since we assume single-word translations to be lemmas, the validation fails if a translation is not present in the list of lemmas from the morphological database.

In the case of two-word translations, where the first word is an adjective or a participle and the second word is a noun, the second word must be included in the list of lemmas (specifically, nominative singular in almost all cases[2]) to pass the validation, and the first word has to agree with the noun in gender, case and number – or to be more precise, since the intra-lexeme homonymy is significant in all the three languages, at least one of the possible triplets of *gender, case, number* should agree with the noun.

If the translation consists of more than two words, or two words that are not an adjective (or a participle) and a noun, the validation passes if all the words are present in the list of possible word forms, and they do not need to be in the basic form. These multiword translations are mostly noun phrases, and as such they usually consist of variously inflected words: nouns, adjectives and

[2]With the exception of pluralia tantum and some defective nouns lacking the nominative.

prepositions. However, a small portion of multiword units are also verb phrases.

These validation steps ensure basic correctness of the translations. However, many of the original English words are in plural (for various reasons, mostly due to usage or the common perception of concepts, e.g. *claws, pebbles, whiskers*), and the translations follow them rather faithfully. Although we could have easily added the plurals to the list of lemmas, we decided to include such translations in the list of warnings, lest we overlook easily visible errors.

The third step checks for duplicate translations (identically translated target words) within one entry. We consider the duplicates in the English original to be errors of the original dataset, and ignore them in this step. Overall, there are 154 duplicates in the original English dataset out of 5866 target words, comprising about 2.6% of the data.

5 Validation Results

		first run			final run		
category	en	hr	lt	sk	hr	lt	sk
L01	828	825	967	821	835	965	821
L02	876	838	845	796	848	844	796
L03	1507	1474	1799	1700	1474	1786	1685
L04	198	199	251	199	203	250	199
L05	113	119	152	125	119	151	125
L06	834	835	852	914	835	852	909
L07	254	263	303	287	263	303	287
L08	186	211	272	213	211	273	213
L09	881	869	865	1004	869	865	994
L10	190	203	207	192	203	205	192

Table 2: Translated target words per language and category. Note that there can be more translations than the original items in the English dataset (denoted by *en* in the table)

.

In the following Tables 3 and 4, the originally translated data (before validation) is called the *initial run*; data where the issues identified by the validation are fixed is called the *final run*. In Table 3, we show the number of issues found in the first version of the translations, per language and per category. Note that the issues with the NOTE level (i.e. untranslated words) are not comparable between languages – the Slovak dataset often leaves the translation empty by design; the Croatian dataset has not been completely translated by the time of writing this article. Table 4 shows the results after manual corrections. The last row shows the amount of corrected issues as a percentage of the difference from Table 3. Although the percentage appears to be small in some cases, the remaining issues are (confirmed by further proofreading) predominantly 8false positives, thus these corrections eliminated practically all the mistakes of these types. Notably, we eliminated all the ERRs and significantly reduced other issues (mostly related to typos and spelling mistakes). The increase of Slovak NOTEs is caused by deleting some of the duplicates, thus moving those ERRs into NOTEs.

	hr			lt			sk		
	N	W	E	N	W	E	N	W	E
L01	41	120	8	0	240	42	7	128	10
L02	85	8	7	0	97	22	1	23	3
L03	1226	20	0	1	226	32	162	293	45
L04	0	39	4	0	44	4	0	34	1
L05	0	0	0	0	6	1	3	1	0
L06	695	19	2	6	84	85	88	136	35
L07	97	20	0	0	97	10	14	17	0
L08	0	27	1	0	31	5	74	21	0
L09	597	29	2	0	162	26	226	90	16
L10	3	16	3	1	67	4	69	6	1
Σ	2744	298	27	8	1054	231	644	749	111

Table 3: Number of NOTEs (N), WARNs (W) and ERRs (E) per language and category, initial run.

	hr			lt			sk		
	N	W	E	N	W	E	N	W	E
L01	4	115	0	0	152	0	10	109	0
L02	0	11	0	0	46	0	1	21	0
L03	1226	20	0	0	200	0	165	281	0
L04	0	39	0	0	49	0	0	34	0
L05	0	0	0	0	4	0	3	1	0
L06	695	18	0	0	79	0	89	134	0
L07	95	19	0	0	76	0	14	16	0
L08	0	26	0	0	28	0	74	18	0
L09	597	29	0	0	156	0	226	82	0
L10	0	9	0	0	64	0	69	3	0
Σ	2617	286	0	0	854	0	651	699	0
$-\Delta\Sigma/\Sigma$ [%]	4.6	4.0	100	100	20.0	100	-1.1	6.7	100

Table 4: Number of NOTEs (N), WARNs (W) and ERRs (E) per language and category, final run.

	hr			lt			sk		
	s	d	t	s	d	t	s	d	t
L01	1	7	0	36	6	0	0	8	2
L02	0	7	0	16	6	0	1	2	0
L03	0	0	0	10	15	7	3	42	0
L04	2	2	0	3	1	0	1	0	0
L05	0	0	0	0	1	0	0	0	0
L06	0	2	0	43	30	12	7	28	0
L07	0	0	0	3	7	0	0	0	0
L08	0	1	0	5	0	0	0	0	0
L09	0	2	0	11	13	2	0	16	0
L10	0	3	0	1	3	0	0	1	0
Σ	3	24	0	128	82	21	12	97	2

Table 5: Number of ERR types, initial run.

In Table 5, we analyse the types of the errors (is-

sues with the ERR severity). We use these codes:

- *s* means there is a space in the translated item, instead of the correct underscore
- *d* means the item is a duplicate of an already existing translation within one entry
- *t* stands for a typo in the value that should have been DUPLICATE (e.g. DULICATE, DUBLICATE etc.) or NO_TRANSLATION (however, there were no misspelled NO_TRANSLATION items found)

6 Discussion of False Positive Warnings

The warnings produced by the automated validation process are of three different types: agreement, spelling, capitalisation. They include false positive cases, the number of which depends on the design of each morphological database used for validation.

6.1 False Positive Warnings in Slovak

Slovak stands out with very few false positive warnings. Somewhat surprisingly, the adjective+noun orthographic/grammar check resulted in only two warnings in the Slovak translations, in L09 *cobwebby* → *pokrytý_pavučinami* (covered-NOM-MSC-SG cobwebs-INS-FEM-PL, i.e. 'covered by cobwebbs') and *doddering* → *upadajúci_vekom* (declining-NOM-MSC-SG age-INS-MSC-SG, i.e. 'declining because of age'), both false positives.

6.2 False Positive Warnings in Croatian

There were no agreement warnings for the Croatian data. False positives in the Croatian data mostly referred to participles, which are lemmatized in the inflectional lexicon as verbs. Common warnings referred to adjectives when they had been translated in their definite form, instead of using a canonical indefinite form commonly appearing in traditional dictionaries of Croatian, e.g. *besmrtni, uzlazni, završni* instead of the indefinite forms *besmrtan, uzlazan, završan*, 'immortal, rising, final'. However, this also depends on the type of an adjective, e.g. relational adjectives are always used in their definite form, while possessive adjectives always appear in the indefinite form.

Other false positives in the Croatian data related to spelling include adjectives in the form of participles, e.g. *natopljen* 'saturated', *pobjesnio* 'outraged', *prestrašen* 'scared', *ukočen* 'stiff', *uspaničen* 'panicky', *zarobljen* 'trapped', *zaspao* 'asleep' and a small number of proper adjectives correctly spelled, e.g. *košćat* 'bony', *majušan* 'tiny'. Adverbs were another category triggering warnings, e.g. *isprijed* 'ahead', *napolju* 'outside', and *postrani* 'aside' as well as colloquial words probably not found in the morphological database, e.g. *bajk* 'wheel', *bajs* 'cycle', *klinac* 'kid', *deran* 'tike', and *lupež* 'rascal'. As expected, plural forms were also not recognized, as previously mentioned *šape* 'paws', *oči* 'eyes', *zubi* 'teeth', and *jaja* 'eggs', as well as specialized terms such as *cementit* 'cementite', *lubanjac* 'craniate', *patkarica* 'anseriform bird', *plodvaš* 'placental', and *svitkovac* 'chordate', most of which have a place in the animal taxonomy in the category L01 hypernyms-animals.

6.3 False Positive Warnings in Lithuanian

In the Lithuanian data, 24 false positive adjective+noun agreement warnings have been produced. This is due to the limits of the Lithuanian Morphological Database, which does not include inter-lexeme homonyms, e.g. the word form of the definite adjective *baltosios* 'white' may be used as singular genitive or as plural nominative; the word forms of the adjective *lengva* 'light, not heavy' and the noun *kamera* 'camera' may be used as singular nominative or singular instrumental; however, in all these and similar cases, the database includes only one of the word forms and occasionally the included word form does not coincide with the one which has to be in the translation. E.g., in the translation, the adjective *žydra* 'bluish' has to be in singular nominative (as it agrees with the noun in singular nominative), but the database includes only the word form *žydra* tagged as singular instrumental; therefore, such a case produced an adjective+noun agreement warning.

In addition, in the Lithuanian data, many false positive spelling warnings were produced. They were of two major types: the ones related to lemmatisation and the ones related to the limits of the Lithuanian Morphological Database.

The false positive warnings related to lemmatisation were produced in the cases where the provided single-word translations were included in the database, but did not match with the lemma-forms in the database. The following categories of translations produced the false positive warnings of this type:

1) single-word translations which are definitive adjectives as they are lemmatised as indefinitive adjectives in the database, e.g. *aukštesnysis* 'eutherian' is lemmatised as *aukštas*;

2) single-word translations which are participles as they are lemmatised as infinitives in the database, e.g. *svyruojantis* 'hesitant', *dvejojantis* 'inconclusive' – lemmas *svyruoti, dvejoti*;

3) single-word translations which are nouns in plural nominative as they do not coincide with lemma-forms in the database, e.g. *plėviasparniai* 'hymenopteron', *papuošalai* 'jewellery' – lemmas *plėviasparnis, papuošalas*;

4) single-word translations which are nouns in singular genitive or plural genitive as they do not coincide with lemma-forms in the database, e.g. *placentos* 'placental', *kaukolės* 'craniate', *šunų* 'canine', *žinduolių* 'mammalian' – lemmas *placenta, kaukolė, šuo, žinduolis*;

The false positive warnings related to the limits of the Lithuanian Morphological Database were produced in the cases where the provided translations were words or comprised words that were not included in the database. The following categories of translations produced the false positive warnings of this type:

1) specialised single-word terms such as *aspidas* 'elapid', *liugeris* 'lugger' or multi-word terms that include highly specialised words such as *katinių šeimos gyvūnas* 'felid';

2) single-words which do not comply to the language norms, but were used for translation because they are frequent in the daily speech, such as *hamburgeris* 'hamburger', *fišburgeris* 'fishburger';

3) single-words of parts of speech that were not included in the database or multi-words which comprise parts of speech that were not included in the database (pronouns, adverbs, prepositions, etc.), e.g. *kažkas* 'somebody', *aukštyn* 'up', *žemyn* 'down', *virš* 'above', *po* 'under', *liūdnas ir kartu malonus* 'bittersweet', *dirbinys iš vielos* 'wirework', *išvesti iš proto* 'madden'.

7 Conclusions

The validation process proved valuable, particularly in identifying duplicate translations and highlighting spelling mistakes.

Numerous false errors and warnings (false positives) have various causes. Some stem from incomplete morphological databases used for validation, indicating insufficient coverage in certain languages like Lithuanian. Others arise from errors and decisions made during the creation of the original dataset or reveal language-specific variations in lemmatization (e.g., indefinite vs. definite adjectives or participles lemmatized as verbs). Additionally, there may be missing highly specialized terms in domains such as biological taxonomy or nautical terminology. Given that we could not modify the original dataset, we had to find appropriate equivalents that accurately reflect the relationships found in the original. These often involved using lemmas in the plural form, colloquial or culturally specific words, etc.

However, the warnings and notices generated during validation also served as additional checks in cases where there was no existing translation. This could occur due to oversight during the translation process or the absence of a suitable equivalent. In such cases, the validation process provided an opportunity to compare these translation gaps with equivalents in other languages and potentially find effective solutions. While this paper primarily focuses on the formal aspect of translating BATS into different languages, it is worth noting that there were numerous lexical gaps specific to English-speaking regions of the world, as well as many domain-specific words or terms requiring verification in terminological resources. These translations had few or no occurrences even in very large corpora, especially within the meronym categories.

The analysis reveals that the accuracy of the initial translations varied among the languages, primarily due to differences in the effort invested in the translations, the approaches taken to the guidelines, and the resolution of problematic entries in the original dataset, rather than inherent differences between the languages.

Acknowledgements

This study is based upon work from the COST Action NexusLinguarum - European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology, https://www.cost.eu/).

References

Vytautas Ambrazas, Emma Geniušienė, Aleksas Girdenis, Nijolė Sližienė, Dalija Tekorienė, Adelė

Valeckienė, and Elena Valiulytė. 2006. *Lithuanian Grammar*. Baltos lankos.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71, Matsue.

Virginijus Dadurkevičius. 2020. Assessment Data of the Dictionary of Modern Lithuanian versus Joint Corpora. CLARIN-LT digital library in the Republic of Lithuania.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas. Twenty-sixth edition.

Christiane Fellbaum. 2005. WordNet and wordnets. In Keith Brown and et al., editors, *Encyclopedia of Language and Linguistics*, second edition edition, pages 665–670. Elsevier, Oxford.

Steinunn Rut Friðriksdóttir, Hjalti Daníelsson, Steinþór Steingrímsson, and Einar Sigurdsson. 2022. IceBATS: An Icelandic Adaptation of the Bigger Analogy Test Set. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4227–4234, Marseille, France. European Language Resources Association.

Radovan Garabík and Denis Mitana. 2022. Accuracy of Slovak Language Lemmatization and MSD Tagging – MorphoDiTa and SpaCy. In *LLOD Approaches for Language Data Research And Management, Abstract Book*, pages 93–95, Vilnius. Mykolo Romerio universitetas.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. 2018. Subcharacter Information in Japanese Embeddings: When Is It Worth It? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37, Melbourne, Australia. Association for Computational Linguistics.

Nikola Ljubešić. 2019. Inflectional lexicon hrLex 1.3. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4264–4270, Portorož, Slovenia. European Language Resources Association (ELRA).

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Marko Tadić. 2007. Building the Croatian Dependency Treebank: the Initial Stages. *Suvremena lingvistika*, pages 85–92. 63/1.

VLE. 2023. Visuotinė lietuvių enciklopedija (General Lithuanian Encyclopedia). `https://www.vle.lt/straipsnis/lietuviu-kalba/`. LNB Mokslo ir enciklopedijų leidybos centras.

Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual Global WordNet Grid. In *Proceedings of the 8th Global WordNet Conference (GWC2016)*, pages 419–426, Bucharest.