A New Integrated Open-source Morphological Analyzer for Hungarian

Attila Novák[1,2], **Borbála Siklósi**[2], **Csaba Oravecz**[3]
[1]MTA-PPKE Hungarian Language Technology Research Group
Práter u. 50/a, 1083 Budapest, Hungary
[2] Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
Práter u. 50/a, 1083 Budapest, Hungary
[3]Research Institute for Linguistics, Hungarian Academy of Sciences
Benczúr u. 33, 1068 Budapest, Hungary
{novak.attila, siklosi.borbala}@itk.ppke.hu, oravecz.csaba@nytud.mta.hu

Abstract

The goal of a Hungarian research project has been to create an integrated Hungarian natural language processing framework. This infrastructure includes tools for analyzing Hungarian texts, integrated into a standardized environment. The morphological analyzer is one of the core components of the framework. The goal of this paper is to describe a fast and customizable morphological analyzer and its development framework, which synthesizes and further enriches the morphological knowledge implemented in previous tools existing for Hungarian. In addition, we present the method we applied to add semantic knowledge to the lexical database of the morphology. The method utilizes neural word embedding models and morphological and shallow syntactic knowledge.

Keywords: morphological analysis, agglutinating languages, word embedding

1. Introduction

With a joint effort of leading Hungarian research centers[1] in NLP, a project has been launched to develop a unified, integrated framework consisting of a basic language processing toolkit (BLARK) for the language. All of the components are designed to synthesize previous development efforts and to deliver the best quality and performance available, with a focus not only on language technology but also on serving the purposes of (theoretical) linguistic research. The infrastructure includes tools and resources from speech technology to syntactic analysis and is integrated into a standardized environment. All tools will be licensed open source, free to use for research and commercial purposes and will be available as plugins in the GATE architecture (Gaizauskas et al., 1996). The resulting framework being part of the national strategic research infrastructure will stand as the de facto reference NLP toolkit for the Hungarian language.

The infrastructure will implement a complete NLP pipeline including a tokenizer, a morphological analyzer, a part of speech tagger and a constituent and dependency parser. As auxiliary tools, a shallow parser (NP chunker) and a named entity recognizer will also be incorporated into the toolkit.

For highly inflectional languages like Finnish or Hungarian, a morphological analyzer is one of the core components of an NLP tool chain and its precision and coverage are critical for higher-level processing tasks. The goal of this paper is twofold: (i) to report on the development of a fast and customizable morphological analyzer, which synthesizes and further enriches the morphological knowledge implemented in previous tools existing for Hungarian; (ii) to present an algorithm based on neural word embedding models and morphological and shallow syntactic knowledge to add semantic knowledge to the lexical database of the morphology.

2. Motivation and related work

Clearly, the complex morphology of the language is a major challenge for computational processing, but standard finite-state approaches are well suited to cope with it (Beesley and Karttunen, 2003). There are a number of FST-based analyzers for a wide range of languages where the number of possible word forms would normally present a serious sparse data problem for statistical models and so the effort put into the development of a rule-based tool is rewarded in the increase of coverage and ultimately of the performance of the NLP system as a whole. It has long been argued that using the output of a high-quality morphological analyzer will reduce the error rate in many classical language processing tasks (see eg. Hajič (2000) or Müller and Schütze (2015)).

Morphological complexity presents difficulties for a precise and detailed descriptive analysis as well, therefore a concentrated effort has been made to use the necessary linguistic expertise in the design of the annotation formalism for the analysis. From an NLP perspective, almost all of the components in the infrastructure crucially depend on the output of the morphological analysis and, consequently, it has to carry the relevant information for higher-level processing tasks including fine-grained annotation for morphological, morphosyntactic and even semantic properties. At the same time, it has to fulfill the needs of linguistic research providing information on morphological segmentation and non-standard variants for example. To comply with international standards, a mapping to the Universal Dependencies (Marneffe et al., 2014) formalism is also developed.

[1] Research Institute for Linguistics, Hungarian Academy of Sciences; MTA-PPKE Hungarian Language Technology Research Group, Pázmány Péter Catholic University; Research Group on Artificial Intelligence, University of Szeged

3. Existing computational morphologies for Hungarian

There are a number of morphological analyzer tools for Hungarian: *Humor* (Novák, 2003; Prószéky and Kis, 1999), the Hungarian *Xerox* analyzer, *hunmorph-foma*, and a family of tools based on the *Hunmorph/morphdb.hu* (Trón et al., 2006) resource. The quality and the availability of these tools differs significantly. Each of them is able to handle a different set of morphological constructions and the coverage of their stem databases is also different. Moreover, various formalisms were used in their implementations. The resources also greatly differ concerning the availability, readability and maintainability of the sources and the extensibility of their stem database. There are also differences in the availability of the original developers, some of them being completely unavailable, while others are willing to provide some help in understanding or modifying their code.

Due to unavailability of both the sources and the developer, the use of the Xerox Hungarian analyzer as a source was out of the question. The readability of the hunmorph-foma analyzer, its maintainability, and the coverage of its vocabulary is far inferior to that of Humor and resources based on morphdb.hu. The description implemented in the hunmorph-foma analyzer is not grammar-based, thus it can only be extended by analogy, i.e. for a new word the description of an existing one with the same morphological behavior must be copied. Moreover, the modification of erroneous paradigm descriptions is also very difficult. These considerations and the fact that the source of the hunmorph-foma analyzer has recently become unavailable forced us not to use this resource as one of the sources of the new morphological analyzer.

Two Hungarian morphological analyzer implementations, *ocamorph* and *jmorph* (Trón et al., 2005) depend on the results of a single open- source Hungarian morphology development effort, the final outcome of which was the Hunmorph/morphdb.hu (Trón et al., 2006) morphological database.

A clear advantage of the hunmorph (ocamorph) and jmorph analyzers is that they are open-source and that the morphdb.hu morphological database is based on a morphological grammar. Thus it is easy to extend, correct and understand by reading the source itself. The morphdb.hu resource is based on a morphological description and lexicon compiler tool, *Hunlex*. A disadvantage of both the morphdb.hu database and of Hunlex, however, is that these haven't been developed for years now and their documentation is quite deficient. Moreover, the Hunlex tool was implemented in OCaml, and we have not been able to find a competent OCaml programmer. The ocamorph analyzer is also implemented in OCaml. In contrast, jmorph is implemented in Java. The analyzer algorithm is different in the two implementations. The way they handle compounds differs quite significantly. Yet, these differences are not documented and can only be traced by examining the output or the source code of the tools.

Another computational morphology for Hungarian, the Humor (Novák, 2003; Prószéky and Kis, 1999) analyzer and morphological database was developed independently, and, in contrast to the abandoned morphdb.hu project, has been continuously maintained.

The Humor analyzer performs a classical 'item-and-arrangement' (IA)-style analysis, analyzing the input word as a sequence of morphs. A feature-based local compatibility check between adjacent morphs and the conformance of the word structure to a finite-state word grammar automaton is performed during lookup.

The Humor database is generated using a feature-based morphological grammar from a non-redundant lexical database in a similar manner to the one implemented in morphdb.hu/hunlex.

The morphological database of the Humor analyzer had not been freely available before the project described in this paper started and the analyzer itself is also closed-source. Nonetheless, the descriptions in this resource also use a grammar making it easy to extend and correct. Its documentation is quite complete, compared to the other tools. Moreover, this is the only tool the developer of which has been available and the maintenance of the source has always been continuous. The characteristics of the Humor morphological database are shown in Table 1.

4. Evaluation of the coverage of the analyzers

As the first step, the detailed evaluation and critical comparison of the tools potentially appropriate for further development (Humor, ocamorph, jmorph) was carried out. The first step was to evaluate and compare coverage of the tools on a 35-million-word frequency list of words retrieved from a 3-billion-word uncleaned mostly web-crawled text corpus.

When enabling the productive compound analyzer function of the ocamorph tool, it ran into infinite loops for many input words, thus this functionality was turned off. Moreover, while the word grammar in Humor describes productive Hungarian compound constructions quite accurately, the description of compounds in the morphdb.hu grammar is extremely simple: it allows an arbitrary number and type of nominal stems (nouns, adjectives or numerals) to occur in a compound in an arbitrary order (accepting for example the non-existing *tevepiroshatvannegyven*, 'camelredsixtyforty' as a correct compound), this tool returned much more nonsense analyses with the productive compounding function on than the other two analyzers.

The quantitative results of the evaluation of each analyzer are shown in Table 2. As it can be seen in Figure 1, for words with a frequency between 500,000 and 1,000,000, ocamorph had the highest coverage, while in the other regions, the Humor analyzer performed best.

The jmorph and ocamorph tools were probably developed using different versions of the same lexical database. The rather significant difference in their performance, however, is partly due to the different algorithms implemented in them. Items left unanalyzed by all tools in the most frequent regions are due to tokenization, punctuation and spelling errors in the corpus.

stem lexicons		**lemmas/lexemes**	**allomorphs**
generic vocabulary		**95811**	**141718**
	original lexicon extended	75132	105473
	closed-class stems (pronouns, numerals, etc.)	744	3675
	from dictionaries and corpora	19935	32570
terminological lexicons		**110129**	**178324**
	Geographical and human names	40262	
	Nuclear technology	911	
	Financial/Administration	4736	
	English	1920	
	Medical	40813	
	Defense	21487	
all		**205940**	**320042**
	of that compounds	89415	126728
	polymorphemic/suffixed		7720
suffix lexicon		**lexemes**	**allomorphs**
	all	**283**	**12041**
	polymorphemic		10959
rule files	**operations**	**rules**	**lines**
stem rules file			
	45 declarations	520 rules	2074 lines
	596 allomorph generating operations	220 stem allomorphy rules	
suffix rules file		50 rules	233 lines
	86 allomorph generating operations	34 allomorphy rules	
word grm	**states**	**transitions**	**flags**
word grammar automaton			
	47 states	602 transitions	20 flags
	categories	**properties**	
encoding definition of features and word grammar categories			
	102 word grammar categories	102 vector-encoded properties	
		187 matrix-encoded properties	

Table 1: Components of the Hungarian morphological description

Analyzer	Number of words left unanalyzed
Humor	13 754 680
ocamorph	23 248 165
jmorph	17 152 815

Table 2: The number of unknown words for each tool in the list of 35 million word forms.

5. The new morphological analyzer system

The new morphological analyzer system is composed of three layers. The first one is the source database, i.e. the stem lexicon and the morpho(phono)logical grammar, which is readable for linguists. The second layer is a source database converter, which transforms data from the first layer to resources for the third layer. The third layer then is the morphological analyzer framework itself.

The morphological databases used by the ocamorph,

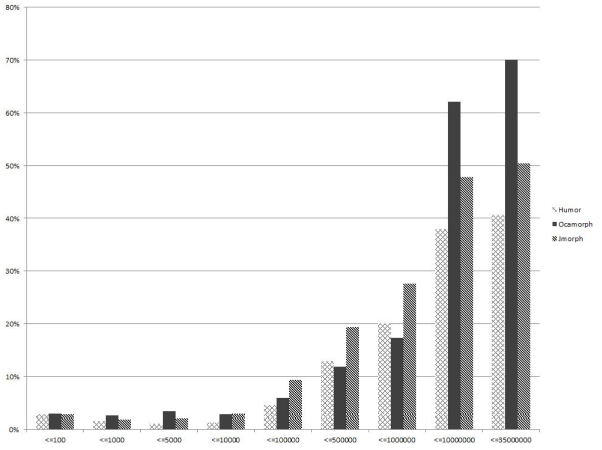

Figure 1: The ratio of unknown words for each tool in various token frequency regions.

jmorph and Humor analyzers had also been created using a similar three-layer architecture. In the current project, the morphological database of the analyzer is created from the Humor grammar with the tool described in (Novák, 2014). This tool converts the morphological grammar to a finite-state representation.

5.1. The stem database

The source database was created as the synthesis of the stem databases and the morphological rule systems of the Humor and the morphdb.hu morphologies. The storage of pragmatic, dialectal, semantic and morphological features is supported. Pragmatic features are used to assign information concerning style and non-standard orthography. In addition, frequency information is added to the entries. A similar combination of features was not present in any of the former existing databases. Moreover, the source database has been enriched with semantic features and ontological classification.

The set of morphological features was obtained by merging the category systems of the Humor and the morphdb.hu lexicons and resolving cases where the two descriptions contradict. Even though the Humor database originally contained semantic category tags, which were not displayed in the output of the analyses, the set of these tags was extended and checked using distributional semantic models.

The morphdb.hu database contains about 13,000 lemmas which are not covered by the 200,000-word Humor stem database. However, these 13,000 lemmas include numerous orthographically substandard or simply mistyped word forms. Although a part of these was included intentionally and was marked as such, the rest had to be checked and marked or removed manually. Mistyped words had to be identified and mapped to their correct forms. This process was supported by generating possible correction candidates for words in the list. First, we created a list of correct forms and multiword expressions from the stem database of Humor and the most comprehensive orthographic dictionary of Hungarian (Laczkó and Mártonfi, 2004). Then, using the A* algorithm (Huldén, 2009), and a confusion matrix as the error model, a ranked list of correction candidates was generated for each misspelled word. Finally, the correction candidates were manually validated.

5.2. The category system

In addition to the main part-of-speech categories, subcategories have also been included in the new database. Some of these subcategories (i.e. types of proper names, types of adverbs etc.) had been included in the source of the Humor or morphdb.hu grammars, but they were not present in the representations output by the run-time analyzers, as the resource compilers were not configured to transfer this information as explicit tags into the compiled lexicons. In addition to features already present in at least one of the original lexicon sources, we introduced other relevant semantic and morphosyntactic categories. The usual morphological features: inflections, derivations, compound boundaries and their types have also been derived by unification of the representations in the two former systems. Phonological features are also part of the database. This includes pronunciations that cannot be automatically derived from the orthographic forms or are alternative pronunciations, dialectal or sociolectal variants. The description also includes the CV-(consonant-vowel)-skeleton of words.

In order to perform the unification of the grammatical descriptions, we reviewed the source of the morphdb.hu grammar, implemented in the Hunlex description language, and identified differences from the Humor description. We modified the Humor source or added missing information where it was necessary. In the morphdb.hu database, adverbs are categorized into several subclasses. We validated and extended this categorization with distributional models. We used neural word embedding models using the `word2vec` tool in a manner described in Section 6..

The results revealed several errors in the categorization of adverbs in the morphdb.hu database. The identification and manual correction of these categorization errors was facilitated by performing a hierarchical clustering of words based on their semantic vector representations. Groups of similar words can be categorized at once, and erroneously categorized ones clearly stand out in the company of similar words with different category tags.

It is interesting to note that when using different distance metrics and different models (lemmatized and analyzed vs. raw, see Section 6.) clearly different aspects dominate in the vector-space-model-based classification of words. This also revealed some problematic cases in the Humor database concerning words having both an unsegmented adverb analysis and a more detailed otherwise equivalent analysis revealing the internal structure of the word. When clustering the words annotated as adverbs in the morphdb.hu database based on their representation built from the disambiguated and lemmatized corpus, a large, grammatically quite heterogeneous group appeared among the resulting clusters. This cluster consisted of elements that had been reduced by a deeper analysis to a lemma different from the original word, but some occurrences of these words in the corpus received an adverbial annotation (due to the inconsistent annotation of the training corpus used to build the model of the tagger/lemmatizer). These rare analyses, however, did not result in reliable continuous vector representations, thus these words appeared as elements in a single odd cluster created from these representations. The common feature of these words was indeed only this inconsistency, i.e. the adverbial analyses being rare and noise-like compared to their more detailed analyses.

The distributional analysis revealed other types of annotation anomalies in the corpus. For instance, words unknown to the Humor analyzer and receiving specific types of erroneous lemmas and PoS tags during automatic morphological annotation were grouped together. Lexical gaps in the morphological analyzer and misspellings may lead to cases where the guesser in the tagger erroneously tags and lemmatizes words. Lemmas resulting from similar errors were grouped together by the model. E.g. the 'lemmas' *pufidzsek(i)* 'puffy jacket', *rövidnac(i)* 'shorts', *napszemcs(i)* 'sunglasses', *szemcs(i)* 'glasses', *szmöty(i)* 'gunk' etc. all lack a word-final *-i*. They result from the lemmatizer guesser erroneously cutting the ending *-it* from the accusative form of these words.

The classification of conjunctions in the Humor and morphdb.hu databases also differs to some extent. In the Humor database, only words that have a clearly conjunction-like distribution in clauses (i.e. they must be clause-initial) are categorized as conjunctions. Discourse markers that have clause linking pragmatic function but have an adverb-like distribution within the clause (i.e. they are free to move within the clause) were classified consistently as adverbs. In contrast, some of these words are also classified as conjunctions in the morphdb.hu database. In the new, unified categorization system, these are classified as a subclass of adverbs. Relative pronouns, which have the same distribution as conjunctions, were also not categorized as conjunctions, but as relative pronouns keeping their main PoS category.

After having reviewed the morphdb.hu grammar, we were able to automatically transfer the features of lexical elements from the morphdb.hu lexicon to the Humor lexicon. Still, this process also required manual checking, because some of the features were not valid or missing.

In addition to different lexical and grammatical coverage, existing Hungarian computational morphologies also differ in the tagsets they use in the generated annotation. Humor has its own set of tags, morphdb.hu uses another notation known as KR (Kornai et al., 2004), and the other analyzers also use their own tagsets. The Hungarian version of the MSD tagset (Erjavec, 2010) is also widely used, in spite of the fact that it is not the native tagset of any Hungarian morphological analyzer, and that it does not cover derivation. In the current project, a new set of morphological tags was defined for Hungarian based on the tag formalism proposed in the Leipzig glossing rules[2] but adapting them to the obvious practical requirements imposed by the fact that we want to tag a language that has an orthography (which e.g. excludes the possibility of using the hyphen as a tag separator). We also defined mappings to all traditionally used Hungarian morphosyntactic tagsets.

5.3. The framework

The resulting source database is readable and editable for an expert linguist easily, using a plain text editor. The lexicon contains only unpredictable features of lexical items: their lemma (including segmentation of compounds), their PoS, and optionally any unpredictable features such as irregular pronunciation, membership in a closed stem alternation class, irregular morphological features, irregular allomorphs and semantic or pragmatic features. Moreover, an application is also being developed that makes the extension or the modification of the stem database even simpler for those lacking the expertise of a computational linguist by proposing likely features and generating model paradigms. This layer also contains rules that infer predictable features and generates allomorphs. The description also includes a finite-state word grammar describing possible morphological constructions.

The second layer of the system, which is the source database converter, transforms the stem database to an allomorph-based representation, which is implemented as a *lexc* (Beesley and Karttunen, 2003) lexicon. Word grammar constructions are represented in the lexicon by flag diacritics, which can optionally be eliminated from the compiled transducer. The compiled lexc lexicon is used by the third layer, which is implemented in HFST (Lindén et al., 2009) using the Divvun/Giellatekno infrastructure (Moshagen et al., 2013).

The new framework provides the possibility of creating customized domain-specific and register-specific analyzers based on the source database using xfst-style regex filters. The native morphological tag system can be mapped to any of the three annotation schemes generally used for Hungarian (KR, Humor, MSD), and so the compatibility with already existing annotated corpora is ensured.

6. Enlarging the database and enriching it with semantic knowledge

Even though the lexicon of the new analyzer, created by unifying the Humor and morphdb.hu lexicons, is larger than any of the former ones, there is still room to improve the lexical coverage of the system. In addition, we wanted to enrich the database with semantic features that constrain morphological constructions, are referred to by orthographic rules, or play an important role in determining syntactic distribution, such as color terms, mass nouns, names of ethnic groups and languages, professions etc. We expected that manual addition of these semantic features to the new lexical database would be a very labor-intensive task. Thus, an automated method seemed to be reasonable to apply.

Many current natural language processing techniques use word embedding representations to acquire semantic information from raw corpora. In these systems lexical items are represented as points in a real vector space. It has been shown that similar words tend to be closer to each other in this high-dimensional space and that these similarities can of both semantic and (morpho)syntactic type (Mikolov et al., 2013b; Mikolov et al., 2013a).

Thus, we built two types of models using the `word2vec`[3] tool, the widely-used framework for creating word embedding representations. In the first model, we used a nearly 3-billion-word raw web-crawled corpus of Hungarian (applying boilerplate removal). In the second model, we applied POS-tagging and morphological analysis to the same corpus. Then, each word was segmented into a lemma and a morphosyntactic tag. Thus, these morphological features were also considered when building the feature vectors for words, and the terms were lemmas instead of surface word forms. Both models were built using a skip-gram model with 300-dimensional feature vectors and a window radius of 5. The details of the models and the clustering algorithm can be found in (Novák and Siklósi, 2016).

We used the two models to extract coherent semantic groups from the corpus. Since our goal in this task was to organize words along their semantic similarity, rather than their syntactic behavior, we used the model built from the lemmatized version of the corpus only. We created a web application to aid the exploration and visualization of the

[2] https://www.eva.mpg.de/lingua/resources/glossing-rules.php

[3] https://code.google.com/p/word2vec/

models and the retrieval of semantically restricted vocabulary.

For each category an initial word was selected and the top 200 most similar words were retrieved from the model. Then, the top 200 most similar words were retrieved for items selected by a simple mouse click (taken from the bottom of the previous list). This step was repeated about 10 times. Repeated occurrences were filtered out when retrieving the subsequent lists. The result lists were then merged. Moreover, it was also checked by quick inspection whether the lists did in fact contain mostly relevant items. Those that did not, were deleted by a single click. Throwing these words away, the algorithm was applied again resulting in purer lists. Thus, starting from one word for each category, hundreds or thousands of related words could be retrieved semi-automatically with minimal human interaction that could hardly have been done manually.

6.1. Results

We evaluated the task of semantic categorization by manually counting the number of correct and incorrect words in the given category. However, in order to be able to perform this validation efficiently, the result lists were clustered automatically so that these groups could be reviewed at once. Due to the clustering applied to the sets of words, the validation of the results became very efficient and easy. The results of the manual evaluation for the following semantic categories: languages, occupations, materials and within that textiles, colors, vehicles, greetings and interjections, and units of measure are shown in Table 3. We categorized the words (or clusters, if they were homogenous) as correct, erroneous or related.

	Correct		**Errorneous**		**Related**		**All**
Lng	755	60.69%	98	7.88%	391	31.43%	1244
Occ	2387	93.32%	134	5.24%	37	1.45%	2558
Mat	1139	84.06%	162	11.96%	54	3.99%	1355
Tex	120	51.28%	114	48.72%	0	0.00%	234
Col	870	63.18%	392	28.47%	115	8.35%	1377
Vhc	1141	72.26%	239	15.14%	199	12.60%	1579
G&I	334	24.85%	261	19.42%	749	55.73%	1344
UoM	1457	60.08%	909	37.48%	59	2.43%	2425

Table 3: The results of semantic categorization. Lng: languages, Occ: Occupations, Mat: materials, Tex: textiles, Col: colors, Vhc: vehicles, G&I greetings and interjections, UoM: units of measure

6.2. Error analysis

In order to further evaluate our method, we performed a detailed analysis on the task of identifying names of languages. We got the following results by manually evaluating the 1244 suggestions of the system. Table 4 shows numerical distribution of different types of language names retrieved from the corpus. These types are the following. The first group includes languages, language types, etc:

- Standard language name: the official name of a language in standard orthographic form
- Fictitious language name: the name of a language invented by the author of some literary work.
- Name of a group or family of languages: e.g. *uráli* ‘Uralic’
- Ethnic term but not the name of a language: e.g. *zsidó* ‘Jewish’. These terms are often informally used as if they were the names of languages.
- Name of a script: e.g. *dévanágari* ‘Devanagari’, *cirill* ‘Cyrillic’ These behave similarly to names of languages concerning the grammatical constructions used with them.
- Language type: e.g. *kreol* ‘Creole’, *patois* ‘Patois’, *pidzsin* ‘Pidgin’ (final part of names of languages of this type)

The second group includes attributes of languages:

- Attribute of geographic location: attribute part of the name of a language, dialect or language group, which itself cannot be used as the name of a language e.g. *iraki* ‘Iraqi (Arabic)’, *mezopotámiai* ‘Mesopotamian (languages)’
- Other (non-geographical) attribute: *rabbinikus* ‘Rabbinic (Hebrew)’

The third group includes orthographic variants, synonyms and misspellings:

- Synonym: another (e.g. historical) name of a language, e.g. *tót* for *szlovák* ‘Slovak’, *hellén* for *görög* ‘Greek’.
- Orthographic variant (of a language, group or dialect name): arachic form, phonetic variant, or one with Latinate orthography, e.g. *franczia* for *francia* ‘French’, *bulgár* for *bolgár* ‘Bulgarian’, *szittya* or *scytha* for *szkíta* ‘Scythian’
- Grave misspelling: name of a language, dialect or group with letters missing or swapped

These three groups include words that can be considered as languages from the aspect of the original task, i.e. enhancing the database of the morphological analyzer. These covered 92.12% of the 1244 candidate words. The rest (i.e. 7.88%) were non-language words. This figure includes language pairs, like *magyar-angol* ‘Hungarian-English’, where the languages do not form a proper language group, but not e.g. *bajor-osztrák* ‘Bavarian-Austrian’, where the two form a coherent dialect group.

Estimating recall is a more difficult task in lack of an exhaustive list of languages in Hungarian. (If such a list existed, we could have used that in the original task.) The same applies to other categories, such as colors, professions, etc. In terms if precision, we achieved similar results for these other categories.

type	example	precision
a standard language name	*joruba* ‘yoruba’	38.39%
fictitious language name	*újbeszél* ‘Newspeak’	1.07%
name of a dialect	*Cockney*	5.14%
name of a group or family of languages	*uráli* ‘Uralic’	4.21%
orthographic variant	*scytha* for *szkíta* ‘Scythian’	9.88%
synonym	*hellén* for *görög* ‘Greek’	2.00%
		60.69%
misspelling	*ngol* ‘nglish’	7,46%
ethnic term but not the name of a language	*zsidó* ‘Jewish’	2.25%
name of a script	*cirill* ‘Cyrillic’	1.57%
language type	*kreol* ‘Creole’	0.70%
attribute of geographic location	*iraki* ‘Iraqi (Arabic)’	18.57%
other attribute	*rabbinikus* ‘Rabbinic (Hebrew)’	0.87%
		31.43%
not a language (includes language pairs)	*magyar-angol* ‘Hungarian-English’	7.88%

Table 4: Detailed analyses of the results for languages. The values are the percentage corresponding to the category in the 1244-word-long list of candidates.

7. Conclusion

We have implemented a new open-source morphological analyzer for Hungarian unifying the lexical databases and morphological grammars of two legacy systems. The new analyzer has a significantly enhanced database into which new stems and new features were also introduced. To enrich the database with further semantic and subcategorial features, we used word embedding and hierarchical clustering techniques, producing lists of words for certain semantic and syntactic categories in a semi-supervised manner resulting in high-precision lexical data with a very low demand on human expert interaction.

8. Bibliographical References

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications.

Erjavec, T. (2010). Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Gaizauskas, R., Cunningham, H., Wilks, Y., Rodgers, P., and Humphreys, K. (1996). GATE: An environment to support research and development in natural language engineering. In *Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence*, pages 58–66.

Hajič, J. (2000). Morphological tagging: Data vs. dictionaries. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 94–101, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huldén, M. (2009). Fast approximate string matching with finite automata. *Procesamiento del Lenguaje Natural*, 43:57–64.

Kornai, A., Rebrus, P., Vajda, P., Halácsy, P., Rung, A., and Trón, V. (2004). Általános célú morfológiai elemző kimeneti formalizmusa [The output formalism of a general-purpose morphological analyzer]. In *II. Magyar Számítógépes Nyelvészeti Konferencia*, pages 172–176, Szeged. SZTE.

Laczkó, K. and Mártonfi, A. (2004). *Helyesírás*. A magyar nyelv kézikönyvtára. Osiris.

Lindén, K., Silfverberg, M., and Pirinen, T. A. (2009). Hfst tools for morphology - an efficient open-source package for construction of morphological analyzers. In Cerstin Mahlow et al., editors, *SFCM*, volume 41 of *Communications in Computer and Information Science*, pages 28–47. Springer.

Marneffe, M.-C. D., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford Dependencies: A cross-linguistic typology. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). 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, June. Association for Computational Linguistics.

Moshagen, S. N., Pirinen, T. A., and Trosterud, T. (2013). Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 343–352, Oslo University, Norway.

Müller, T. and Schütze, H. (2015). Robust morphological tagging with word representations. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL*, pages 526–536, Denver, Colorado.

Novák, A. (2003). Milyen a jó Humor? [What is good Humor like?]. In *I. Magyar Számítógépes Nyelvészeti Konferencia*, pages 138–144, Szeged. SZTE.

Novák, A. (2014). A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1068–1073, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1207.

Novák, A. and Siklósi, B. (2016). Using embedding models for lexical categorization in morphologically rich languages. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3-9, 2016*. Springer International Publishing, Cham.

Prószéky, G. and Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 261–268, Stroudsburg, PA, USA. Association for Computational Linguistics.

Trón, V., Kornai, A., Gyepesi, G., Németh, L., Halácsy, P., and Varga, D. (2005). Hunmorph: Open source word analysis. In *Proceedings of the Workshop on Software*, Software '05, pages 77–85, Stroudsburg, PA, USA. Association for Computational Linguistics.

Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., and Simon, E. (2006). E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In *In: Proceedings of LREC 2006*, pages 1670–1673.