An MLU Estimation Method for Hungarian Transcripts

György Orosz[1,2] and Kinga Mátyus[3]

[1] Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
50/a Práter street, 1083 Budapest, Hungary
[2] MTA-PPKE Hungarian Language Technology Research Group
50/a Práter street, 1083 Budapest, Hungary
oroszgy@itk.ppke.hu
[3] MTA Research Institute for Linguistics
33. Benczúr street, 1068 Budapest
matyus.kinga@nytud.mta.hu

Abstract. Mean length of utterance (MLU) is an important indicator for measuring complexity in child language. A generally employed method for calculating MLU is to use the CLAN toolkit, which includes modules that enable the measurement of utterance length in morphemes. However, these methods are based on rules which are only available for just a few languages not involving Hungarian. Therefore, in order to automatically analyze and measure Hungarian transcripts adequate methods need to be developed. In this paper we describe a new toolkit which is able to estimate MLU counts (in morphemes) while providing morphosyntactic tagging as well. Its components are based on existing resources; however, many of them were adapted to the language of the transcripts. The toolchain performs the annotation task with a high precision and its MLU estimates are correlated with that of human experts.

1 Introduction

Mean length of utterance (MLU) is a metric that has been widely used for measuring linguistic productivity of children for almost a hundred years. Utterance lengths are usually calculated in morphemes for morphologically complex languages, while in the case of analytical languages counting only words is also a feasible solution. Concerning the calculation of MLU in morphemes (MLUm), several tools are available which employ natural language processing methods and could be used to boost this labor-intensive task. Nonetheless, none of them is able to deal with Hungarian transcripts. However, existing resources with slight modifications can be used to estimate MLUm for transcripts of Hungarian child language.

2 Background

Ever since the complexity of child language has been measured, several methods have been developed. While manual counting prevailed for decades, automatic counting tools have been sought for in the past years.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 173–180, 2014.

2.1 Measuring Morphosyntactic Complexity

Several studies (e.g. [3]) showed that MLUm indicates language development for normal children, especially at very early stages. In contrast, MLUw was shown to be highly correlating [7,16] with MLUm in the case of analytical languages such as English or Irish. Therefore, some studies concur that MLUw is a reliable measure as opposed to MLUm, where researchers often need to make ad hoc decisions on what (not) to count [4].

However, Crystal also points out [4] that MLUm is a good way to measure morphologically complex languages (see e.g. [2]). Hungarian is an agglutinative language, thus MLUm can be considered to be a more reliable indicator of language development than MLUw (similarly to Turkish [20]). Moreover, previous studies which measured language development in Hungarian manually [18,23] mostly employed MLUm as a metric.thus we used their work as bases.

In the case of corpora which follow the CHAT guidelines [9], MLU values (including MLUm) can be calculated with the CLAN [8] toolkit. This system is widely used, since it contains components that perform the necessary preprocessing steps. One of its modules is MOR, which is a morphological analyzer specially designed for spoken language corpora. A subsequent component is POST, which does the morphological disambiguation for such texts. With these components, the number of morphemes can be calculated in a corpus of transcribed spoken data in a number of languages. However, they lack rules for Hungarian, thus none of them can be used for processing such transcripts.

2.2 Tagging Spoken Language

One of the pioneers in tagging spoken language was Eeg-Oloffson [6], who used manually annotated transcripts to train a statistical tagger. Others employ and adapt statistics that derive from written language corpora [11,13,15]. Furthermore, building domain specific rules could also lead to satisfactory taggers (e.g. [12]), while combination of such systems with stochastic tools [1] yields effective algorithms as well.

Based on the studies above, a proper morphological annotation system aiming to process transcripts must be able to handle the following types of difficulties: *i*) existence of new morphosyntactic tags which are missing from the tagset of the training data, *ii*) occurrence of tokens with non-standard orthography in the texts, *iii*) the number of words unknown to a statistical tagger are increased compared to written language corpora, *iv*) in the case of stochastic taggers, if probability estimates are derived from a written language training corpus, the model learnt can become non-representative (e.g. the distribution of PoS tags may significantly differ in written and spoken language).

2.3 Tagging (Spoken) Hungarian

Beside Pos tagging, finding the roots of words is an indispensable task for estimating MLUm. There are numerous studies investigating the tagging performance of machine learning methods on Hungarian. Only two taggers are freely available and perform lemmatization as well. These tools are usually trained on the Szeged Corpus [5] (SZC), since it is the only linguistic resource for Hungarian that is manually annotated.

PurePos [14] is an open-source full morphological disambiguator system which incorporates the Humor [17] morphological analyzer (MA). The tool is based on statistical trigram-tagging algorithms, but it is extended to employ rule-based components effectively.

magyarlanc [24] is a freely available language processing chain for parsing Hungarian. Its morphological tagging component is based on the Stanford tagger [21] and incorporates a MA which relies on morphdb.hu [22].

We are not aware of any research investigating the tagging of spoken Hungarian. Moreover, there is no study aiming to calculate MLUm for Hungarian transcripts automatically. Therefore this article investigates methods for analyzing transcripts of Hungairan child speech.

3 The HUKILC Corpus

The contemporary Hungarian Kindergarten Language Corpus (HUKILC) [10] was selected as a base of our research, since there were no morphosyntactically annotated transcripts available. This corpus has been compiled predominantly for child language variation studies. It contains 62 interviews with 4.5–5.5 year-old kindergarten children from Budapest, recorded in spring 2012. The interviews are 20–30 minutes long, and consist of different types of story-telling tasks. Its transcription was carried out using the Child Language Data Exchange System (CHILDES) [9], following its guidelines. The corpus consists of about 39,000 utterances with 140,000 words.

As for morphological annotation of the corpus, general tagging principles were established first. We chose the morphosyntactic labels and lemmata of the Humor analyzer to represent morphological analyses. Next, an annotation manual was developed for human annotators to guide their work in the morphological disambiguation of the corpus. Finally, 6 interviews with about 1,000 utterances were labelled manually. This gold standard corpus was split into two sets of equal sizes: a development and a test set containing 509 and 449 lines respectively.

4 Tagging the Transcripts

The morphological tagging algorithm employed here is a hybrid one, composed of a morphological analyzer, a stochastic tagger tool and domain-specific disambiguation rules as well. Since the tagset of Humor was chosen to be used for the annotation, a plausible solution was to employ this analyzer. Further on, PurePos was used to disambiguate between the morphological tags, and Szeged Corpus was employed as a training corpus for the tagger.

In order to apply a MA prepared to analyze written texts, its analyses had to be adjusted for the transcripts. Thus, rules adapting its analyses – based on regular expressions and domain-specific word lists – were developed. Their formulation could be done with high confidence, since most of the transcripts contained controlled conversation covering only a few topics.

As a first step, morphological analyses of about 40 words typical of spoken language were created manually. These tokens were mostly interjections not used in written language (such as *hűha* 'wow'), while some adverbs were regarded as utterance words[4] in the corpus (e.g. *komolyan* 'seriously'). Furthermore, those tokens that are written in one word in transcripts but are spelled as two words in formal texts were also added to the lexicon. An example is *légyszíves* 'please' which is written formally as *légy szíves*. Finally, diminutive analyses were also provided where it was necessary. E.g. *kutyus* 'doggy' was also analysed as n.dim with the lemma *kutya* 'dog' beside the old label n and the *kutyus* 'doggy' root. This process was carried out by investigating the lemmata produced by Humor: if the deletion of the derivational affix resulted in a root that was enumerated in a domain-specific list, a new diminutive analysis was created as well.

Concerning the disambiguation process, PurePos was extended with rules in order to customize its knowledge to the target domain. First, the tagger was forced to assign diminutive analyses when a related label had previously been selected by the disambiguator. Then further enhancements were carried out by investigating the common mistakes of the chain on the development dataset.

A frequent error of the chain was the mistagging of *akkor* 'when' and *azért* 'in order to'. These words are pronouns and can be categorized as either adverbial, noun phrase level or demonstrative ones, and can also behave as pronomial adjectives. Generally, when *akkor* is followed by *amikor* 'when' (as in *Akkor érkezett meg, amikor mentem* 'He arrived, when I left') and when *azért* is followed by *mert* 'because' (as in the sentence *Azért eszik, mert éhes* 'He eats, because he is hungry') these pronouns are demonstrative ones. Furthermore, such co-occurances are more frequent in the transcripts than in the Szeged Corpus, since they are usually used for reasoning or telling a story. As these long-term dependencies could not be learnt by the trigram tagger used, rules were employed to tag these tokens correctly.

The next issue was the case of the word *utána* 'afterwards, then; after him/her/it'. It can either be used as an adverb of time (as in the sentence *Utána elindultunk* 'Then we left') and as a postpositional phrase meaning 'after him/her/it, following him/her/it' (as in *Elindultunk utána* 'We went after him'). The former usage is much more common in spoken language: when this word is directly followed by conjunctions such as *meg* 'and' or *pedig* 'however', it is always an adverb. Therefore *utána* was tagged as an adverb in the transcripts when it is followed by one of these trigger words.

The last rule introduced deals with *meg*, which may function as a verbal prefix or as a conjunction. Moreover, it is commonly used as an expletive in spoken language. Therefore, the conjunctive label was assigned to the word when there was not any verb in its two token window.

5 MLU Estimation

As a first step, general principles of counting morphemes were established. This was based on the work of Brown [3], Retherford [19], Wéber [23] and Réger [18], with

[4] Annotation schemes for Hungarian distinguish utterance words and interjections. An utterance word forms a sentence or an utterance alone by interrupting or managing the communication. In contrast, interjections are either onomatopoeic or used to indicate emotions.

some necessary modifications. The basic principles were: *1*) only meaningful words were analyzed, thus fillers (filled pauses such as *öööö* 'er'), punctuation marks and repetitions are not counted in the utterances; *2*) phatic expressions (e.g. *igen* 'yes, mhm') serving to maintain communication and not conveying meaning were omitted; *3*) inflectional suffixes and lemmata were each counted as one unit; *4*) derivational morphemes (including diminutives) were not counted as separate ones, *5*) reciprocal and indefinite pronouns (e.g. *minden#ki* 'everybody') and compound words (such as *kosár#labda* 'basketball') were counted as one morpheme.

In a language with such a rich derivational system as Hungarian, it is often very complicated to determine the lemmata. This is even more difficult in our case, since no common methodology exists to determine the boundary of productivity in child language. Following the work of Brown [3], proper names (such as *Nagy Béla*, *Sári néni* 'Miss Sári') and lexicalized expressions (e.g. *Jó napot* 'Good morning'), which are frequently used in speech, were also considered as one unit. Their identification was based on capitalization rules and a domain specific list.

As for the automatization of rules, they were implemented relying on the morphological annotation of the corpus. First, each item on the list of fillers was eliminated. Afterwards, tagged words known to the MA were split into morphemes by the Humor analyzer. If more than one analysis was created for a word, the least complex one was chosen, since in the majority of the cases the analyses only differed in the number of derivative tags and compound markers (which we previously decided not to count). As the labels of the annotation scheme were composed of morphemic properties, the estimation for unknown words could be based on the morphosyntactic labels. This calculation was carried out by counting only the inflection markers in the guessed tags.

6 Evaluation

First of all, the morphosyntactic tagging performance of the system was investigated. For this, we calculated its accuracy following the work of Orosz et al. [14]. Therefore, full analyses – containing both the lemmata and the tag – were compared to the gold standard data, not counting punctuation marks and hesitation fillers.

Table 1. Accuracy of the different tagging chains

Method	Tagging accuracy	
	Token	Sentence
Baseline	91.97%	68.37%
DIM	94.92%	79.96%
CONJ	95.53%	81.74%
The full chain	96.15%	83.96%

For measuring the individual advances of the enhancements presented, three different settings were evaluated on the test set. The first of them was a baseline that used the raw analyses of Humor disambiguated by PurePos. The second system (DIM) employed the extended vocabulary and handled the diminutive analyses as described in

Section 4. The next one – marked with CONJ – used further rules aiming to tag *azért* and *amikor* correctly. Finally, the last system presented contains all the enhancements detailed above. Measurements in Table 1 show that in contrast with the baseline tool which tagged erroneously 3 out of 10 sentences, the accuracy of the full chain is comparable with that of the tagging methods for written corpora [24]. Furthermore, each of the enhancements improved the overall performance significantly.

As for the MLU estimation task, two metrics were used for the evaluation. Mean relative error[5] (MRE) was calculated to show the average relative deviation of the estimated morpheme counts from the one of human annotators.

In addition, Pearson's correlation coefficient was employed to measure the correlation between the output of the processing chain and the counts of human annotators. Since both metrics required a gold dataset, MLUm was manually calculated for 300 utterances in the test set.

Table 2. Evaluation of the MLU estimation algorithm

Tagged utterances	MRE	Correlation
Gold standard	0.0279	0.9933
The full chain	0.0449	0.9901

Table 2 presents the evaluation of the morpheme count estimation algorithm. Both the gold standard data and the output of the tagging tool were used as an input for the estimator, thus enabling detailed comparison of the components. The latter values confirm that the overall performance of the methodology described in this study is satisfactory. Therefore, the estimation algorithm introduced can be used to measure the morphosyntactic complexity of Hungarian spoken language in practice.

7 Conclusion

In this study, methods for measuring the morphosyntactic complexity in a corpus of Hungarian spoken child language were investigated. First, an annotation scheme for the HUKILC corpus was created, then a morphological tagging chain was developed. Although the components of our method use resources created for analysing written corpora, its most typical errors could be located and fixed. Further on, principles for counting MLUm for the corpus were laid down and got implemented. The tool developed is suitable for the estimation task, and its morphological disambiguation performance reaches the accuracy of taggers created for written language. In addition, the pipeline architecture of the system allows its modification, thus it can be used as a basic resource in the research of child language.

The contribution of our study is threefold. First, guidelines for Hungarian transcript annotation were created. Further on, MLUm calculation principles for Hungarian were

[5] $MRE = \sum_{i=1}^{n} \frac{|a_i - p_i|/a_i}{n}$, where a_i marks the manual morpheme count of the ith utterance and p_i stands for the ith prediction.

collected, adapted for the HUKILC corpus and got automatized. Finally, a tool is presented that not just performs morphological tagging with a high accuracy, but is an adequate one to be used in practice for estimating morpheme counts. Therefore, the labor-intensive manual calculation could be replaced by the execution of a tool, radically shortening the time required for measuring MLUm.

References

1. Bick, E., Mello, H., Panunzi, A., Raso, T.: The annotation of the C-ORAL-BRASIL oral through the implementation of the Palavras Parser. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). pp. 3382–3386. ELRA, Istanbul, Turkey (2012)
2. Bowerman, M.: Early syntactic development: A cross-linguistic study with special reference to Finnish. Cambridge University Press (1973), http://www.google.com/books?hl=en&lr=&id=SEM4AAAAIAAJ&oi=fnd&pg=PR9&dq=Early+syntactic+development:+A+cross-linguistic+study+with+special+reference+to+Finnish&ots=_KkadLHE8X&sig=6qaazKAsEYUfaK_Bze36IVvpfvA
3. Brown, R.: A first language: The early stages. Harvard University Press (1973), http://doi.apa.org/psycinfo/1973-30971-000
4. Crystal, D.: Review of R. Brown 'A first language'. Journal of Child Language 11, 289–307 (1974)
5. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus, Lecture Notes in Computer Science, vol. 3206, pp. 19–23. Springer (2004), http://www.springerlink.com/index/EE68QPHBFD2H2ECD.pdf;http://dblp.uni-trier.de/db/conf/tsd/tsd2004.html#CsendesCG04;http://dx.doi.org/10.1007/978-3-540-30120-2_6;http://www.bibsonomy.org/bibtex/2ca56255ee5d99e2fcc5c10c8a7a17702/dblp
6. Eeg-Olofsson, M.: Probabilistic Tagging of a Corpus of Spoken English. University of Goteborg: Department of Computational Linguistics (1991)
7. Hickey, T.: Mean length of utterance and the acquisition of Irish. Journal of Child Language 18(3), 553–569 (1991), http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=4236024
8. MacWhinney, B.: The childes project: Tools for analyzing talk. Child Language Teaching and Therapy 8(2), 217–218 (1992)
9. MacWhinney, B.: CHAT manual (1996), http://childes.psy.cmu.edu/
10. Mátyus, K., Orosz, G.: MONYEK: morfológiailag egyértelműsített óvodai nyelvi korpusz. Beszédkutatás (2014 (in press))
11. Mendes, A., Amaro, R., do Nascimento, M.F.B.: Morphological tagging of a spoken Portuguese corpus using available resources. In: Branco, A., Mendes, A., Ribeiro, R. (eds.) Language technology for Portuguese: shallow processing tools and resources. pp. 47–62. Lisboa: Colibri (2004)
12. Moreno, A., Guirao, J.M.: Tagging a spontaneous speech corpus of Spanish. In: Proceedings of Recent Advances in Natural Language Processing (RANPL) 2003. pp. 292–296. Borovets, Bulgaria (2003)
13. Nivre, J., Grönqvist, L., Gustafsson, M., Lager, T., Sofkova, S.: Tagging spoken language using written language statistics. In: Proceedings of the 16th conference on Computational linguistics-Volume 2. pp. 1078–1081. Association for Computational Linguistics (1996), http://dl.acm.org/citation.cfm?id=993370

14. Orosz, G., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013). pp. 539–545. INCOMA Ltd. Shoumen, Bulgaria (2013), http://aclweb.org/anthology//R/R13/R13-1071.pdf
15. Panunzi, A., Picchi, E., Moneglia, M.: Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus: C-ORAL-ROM Italian. In: 4th Language Resource and Evaluation Conference (LREC). pp. 563–566 (2004)
16. Parker, M.D., Brorson, K.: A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). First Language 25(3), 365–376 (2005), http://fla.sagepub.com/content/25/3/365.short
17. Prószéky, G.: Industrial applications of unification morphology. In: Proceedings of the Fourth Conference on Applied Natural Language Processing. p. 213. Association for Computational Linguistics, Morristown, NJ, USA (1994)
18. Réger, Z.: Mothers' speech in different social groups in Hungary. Children's Language 7, 197–222 (1990), http://www.google.com/books?hl=en&lr=&id=2dnKnexHgHsC&oi=fnd&pg=PA197&dq=Mothers%E2%80%99+Speech+in+Different+Social+Groups+in+Hungary&ots=q6ztdWneYK&sig=sVzNK92a--l4Ba8hYA4BODAUtUw
19. Retherford, K.S.: Guide to analysis of language transcripts. Thinking Publications University (1993)
20. Saygın, A.P.: A Computational Analysis of Interaction Patterns in the Acquisition of Turkish. Research on Language and Computation 8(4), 239–253 (2010)
21. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Hearst, M., Ostendorf, M. (eds.) Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. pp. 173–180. Association for Computational Linguistics (2003), http://portal.acm.org/citation.cfm?id=1073478
22. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., Simon, E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of the Fifth conference on International Language Resources and Evaluation. pp. 1670–1673 (2006)
23. Wéber, K.: "Rejtelmes kétféleség" – A kétféle igeragozás elkülönülés a magyar nyelvben. Ph.D. thesis, University of Pécs, Pécs, Hungary (2011), http://nydi.btk.pte.hu/sites/nydi.btk.pte.hu/
24. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of Recent Advances in Natural Language Provessing 2013. pp. 763–771 (2013)