Automatic Error Detection concerning the Definite and Indefinite Conjugation in the HunLearner Corpus

Veronika Vincze[1], János Zsibrita[2], Péter Durst[3], Martina Katalin Szabó[4]

[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence

[2]University of Szeged, Department of Informatics

[3]University of Szeged, Hungarian Studies Center

[4]University of Szeged, Hungarian Linguistics PhD Programme

E-mail: vinczev@inf.u-szeged.hu, zsibrita@inf.u-szeged.hu,

durst.peter@gmail.com, szabomartinakatalin@gmail.com

Abstract

In this paper we present the results of automatic error detection, concerning the definite and indefinite conjugation in the extended version of the HunLearner corpus, the learners' corpus of the Hungarian language. We present the most typical structures that trigger definite or indefinite conjugation in Hungarian and we also discuss the most frequent types of errors made by language learners in the corpus texts. We also illustrate the error types with sentences taken from the corpus. Our results highlight grammatical structures that might pose problems for learners of Hungarian, which can be fruitfully applied in the teaching and practicing of such constructions from the language teacher's or learners' point of view. On the other hand, these results may be exploited in extending the functionalities of a grammar checker, concerning the definiteness of the verb. Our automatic system was able to achieve perfect recall, i.e. it could find all the mismatches between the type of the object and the conjugation of the verb, which is promising for future studies in this area.

Keywords: Hungarian, language learning, conjugation, error detection, morphology

1. Introduction

In this paper we focus on automatic error detection concerning the definite and indefinite conjugation in Hungarian, based on data from the HunLearner corpus (Durst et al., forthcoming). First, we shortly describe the grammatical features of Hungarian verbal conjugation, then we present the types of definite and indefinite objects. Later, we present the extended version of HunLearner and show how conjugational errors can be automatically detected in the corpus. We also offer some statistical data on the most frequent sources of errors.

2. Definiteness in verbal conjugation

The definite verb conjugation is relatively rare in the languages therefore the acquisition of its usage usually gives rise to difficulties to the foreign learners of the Hungarian language (cf. Durst & Janurik, 2011: 20). Moreover, it is also important to emphasize that there are notable differences in what features of the definite conjugation create difficulties to students of Hungarian.

In the Hungarian language, the definite conjugation of the verb is used in the consequence of the presence of a definite object in the given structure, since the definiteness of the noun should be marked on the verb (cf. Törkenczy, 2005; Guskova, 2009: 144). So, depending on the definiteness of the object we distinguish between a definite and an indefinite paradigm in all conjugations including the present, the past, the imperative and the conditional. In the Hungarian language the definite object represents an object identified in the consciousness of the speaker and the listener to the same extent (cf. M. Korchmáros, 2006: 246). Classical examples of the Hungarian direct object are the proper noun (1a) and the noun with a definite article (1b) (cf. Moravcsik, 1975: 262; Durst, 2010a: 82–83).

1 a. Vártam *Katit*.
wait-Past-1Sg.DEF Kati-ACC
'I was waiting for Kati.'
b. Olvasta a könyvet.
read-Past-3Sg.DEF the book-ACC
'He / she read the book.'

In the Hungarian language, in most cases, the definite object occurs in third person (1a–b), but sometimes it is in second person, as well (2) (cf. Bratchikova, 2013).

(2) Könyvvel ajándékozlak meg.
book-INSTR present-1Sg.2Sg.DEF PREVERB
téged.
you-ACC
'I present you with a book.'

From the point of view of computational linguistics, the detection of direct objects may be considered problematic since the syntactic realization of the direct object is not uniform, therefore its automatic detection encounters difficulties in certain cases (the different types of the direct object are listed below). In contrast with the present project, no previous studies on the errors of definite conjugation in the Hungarian language used automatic programs with the purpose of detecting the direct objects and the errors of the usage of the definite and indefinite conjugation in the Hungarian language (cf. Langman & Bayley, 2002; Durst, 2010b; Durst &Janurik, 2011).

3. Types of definite and indefinite objects

The following examples demonstrate typical cases of the definite object and the definite verb conjugation in contrast with the indefinite form and their syntactic context.

1.a. The object is a proper noun
Ismer-em *Zoltán-t.*
know-1Sg.DEF Zoltán-ACC
'I know Zoltán.'

1.b. The object is a common noun
Ismer-ek *egy* *fiú-t.*
know-1Sg.INDEF a boy-ACC
'I know a boy.'

Obviously, intransitive verbs are never used in the definite conjugation. Transitive verbs may have an indefinite object (as in 1.b.) and then they are used in the indefinite conjugation but transitive verbs that stand with a definite object (as in 1.a.) are conjugated according to the definite paradigm. Except for a few special cases, most of the grammatical objects are morphologically marked by the accusative *-t* suffix in Hungarian, making their identification easier for language learners. However, pronominal objects may be implied by the definite conjugation itself, so they may not appear explicitly. Such cases present difficulties for most language learners and they also pose challenges for computer processing.
Apart from proper names, the following structures count as definite objects when they are used in the function of a grammatical object. Where it is possible, they are presented along with the corresponding indefinite verb forms in their typical syntactic context to clearly point out the difference.

2. The object is a demonstrative pronoun
Az-t *akar-om.*
that-ACC want-1Sg.DEF
'I want that.'

3.a. The object is a noun with a definite article
A *film-et* *néz-zük.*
the film-ACC watch-1Pl.DEF
'We are watching the movie.'

3.b. The object is a noun with an indefinite article
Egy *film-et* *néz-ünk.*
a film-ACC watch-1Pl.INDEF
'We are watching a movie '

4. The object is an interrogative or a relative pronoun with the *-ik* suffix (with definitive meaning) or a noun that stands with an interrogative or a relative pronoun with the *-ik* suffix
Melyik *szobá-t* *takarít-od?*
which room-ACC clean-2Sg.DEF
'Which room are you cleaning? '

5.a. The object is a third person personal pronoun
Ismer-em *őt.*
know-1Sg.DEF him/her.
'I know him/her'

5.b. The object is a first or second person personal pronoun
Ők *ismer-nek* *engem.*
they know-3Pl.INDEF me.
'They know me.'

6. The object is a reflexive pronoun
Ismer-em *magam-at.*
know-1Sg.DEF myself-ACC
'I know myself.'

7. The object is a reciprocal pronoun
Ismer-jük *egymás-t.*
know-1Pl.DEF each other-ACC
'We know each other.'

8. The object is a noun with a possessive suffix
Róbert *könyv-é-t* *olvas-om.*
Róbert book-POSS 3Sg-ACC read-1Sg.DEF
'I am reading Róbert's book.'

9. The object is an pronoun with the meaning 'all of them'
Mind-et *lát-juk.*
all-ACC see-1Pl.DEF
'We can see all of them.'

10. The object is an objectival subordinate clause, which may be referred to by a demonstrative pronoun in the main clause
Tud-om (azt), ki vagy.
know-1Sg.DEF (that-ACC) who be-2Sg.INDEF
'I know who you are.'

Intransitive verbs do not have a definite form because they cannot take an object at all. It is interesting to note that the Hungarian definite conjugation can indicate only third person objects, which explains the difference between 5.a. and 5.b.

4. The HunLearner corpus

The HunLearner corpus contains student essays written by university students majoring in Hungarian as a foreign language (Durst et al. forthcoming). Students from Croatia wrote essays in three different topics: 'A person I like', 'Difficulties of learning Hungarian' and 'Hungarian immigrants in England'. These data have been manually corrected for grammatical errors concerning nouns and automatically annotated for the type of such errors.
Some more corpus texts have just recently been added to the data, written in the topic of 'A person I like'. This enlargement also means that now some texts are written by native speakers of other languages besides the

originally included texts written by native speakers of Croatian.

After enlargement, the HunLearner corpus currently consists of 1427 sentences and 22,000 tokens. In this bunch of texts, conjugational errors were also manually annotated by a student of linguistics, which will serve as the base of our investigations.

5. Automatic detection of mismatches in conjugation

Table 1 shows the quantitative results on mismatches in conjugation, based on gold standard data. Here we just focused on cases where the object is phonologically present in the sentence (*has object* column), so now we neglect cases when the presence of the pronominal object could be only deduced from the verbal form. We also neglect cases when the object was a subordinate clause (see Point 10 above) since subordinate clauses are not given a separate tag denoting their grammatical function by the parser, in other words, all subordinate clauses bear the same label, regardless of their grammatical function. Although it had no real effect on the results, we just mention here that for theoretical reasons, we also excluded from the experiment those verb forms that are morphologically ambiguous, so the definite and indefinite forms are the same (as in *olvastam* read-1Sg.DEF/INDEF 'I was reading') since here it cannot be decided for sure whether the language learner intended to use definite or indefinite conjugation.

Subcorpus	Verbs	Mismatch in conjugation	Has object	Unambig. verb
Difficulties	1018	11	7	7
England	564	12	8	8
A person I like	841	28	18	18
Total	2423	51	33	33

Table 1: Mismatches in conjugation.

The resulting 33 cases were analyzed in detail, concerning the type of the object. It was revealed that the most frequent source of errors was when the object is a demonstrative pronoun (Point 2 above): it triggers definite conjugation but in 25% of the errors, it co-occurred with an indefinite verb. Other frequent errors are a bare common noun (i.e. without an article) or a relative pronoun as the object: in 15-15% of the errors, they do not co-occur with the required type of conjugation. Together with the errors induced by common noun with a definite article (Point 3.a above), these types altogether are responsible for two third of the mismatches in conjugation, so they should be paid special attention in language teaching and learning.

Our results also show that the definite object + indefinite conjugation (55%) is a more frequent phenomenon than the opposite, i.e. indefinite object + definite conjugation.

The texts of HunLearner were POS-tagged and dependency parsed by *magyarlanc*, a linguistic preprocessing toolkit of Hungarian (Zsibrita et al., 2013). On the basis of the syntactic and morphological analysis we were able to define rules for the object-verb agreement, which made it possible to automatically collect those sentences where there was a mismatch between the definiteness of the object and the verbal conjugational pattern. An example for such a rule: we checked whether the object noun has any article. If it has a definite article, then the verb it is attached to must be used in the definite form.

We then evaluated the performance of our rule-based system on the gold standard data with the metrics precision, recall and F-measure interpreted on the mismatches. The system achieved perfect recall, that is, it was able to identify all the problematic cases, however, its precision was lower with a score of 32.67, and so, the overall F-score was 49.62. However, we think that in an automatic system that seeks to help language learners the main task is to identify all of the possible errors and the fact that our method achieves perfect recall even at this early stage of research can be considered promising.

Some errors in performance were due to errors in morphological or syntactic parsing. We evaluated the accuracy of POS-tagging on the corpus, and *magyarlanc* was able to obtain an accuracy of 90.96% (including all the erroneously chosen or misspelled words written by the language learners)[1]. An interesting source of error for POS-tagging was that learners of Hungarian seem to have problems with the correct use of accents, which might have influenced the results of our system since in some cases, the accent is a distinctive marker of definite or indefinite conjugation, such as in *olvassak* (read-IMP-1Sg.INDEF) 'I should read' or *olvassák* (read-IMP-3Pl.DEF or read- 3Pl.DEF) 'they should read it' or 'they are reading it'. Moreover, there are cases in the verbal paradigm where all the other morphological features are the same except for definiteness like in *festene* (paint-COND.3Sg.INDEF) 'he would paint' vs. *festené* (paint-COND.3Sg.DEF) 'he would paint it'. Thus, if the accents are not used properly, it might be interpreted as a conjugational error.

6. Typical errors

In this section we illustrate the most typical problematic cases with samples from the corpus. First, we give the sentences in their original form, and then we also provide a flawless version of the same sentence in parentheses, where all types of errors concerning word order, syntax, morphology, accents and other errors have been corrected.

[1] 6.53% of the tokens are misspelled or used erroneously in the corpus, which strongly influences POS-tagging: neglecting them, *magyarlanc* achieves an accuracy of 97.3%, which is similar to the POS-tagging results obtained on standard Hungarian texts.

The object is a proper noun:
Mindenki nagyon ***szeret*** ***Magyarországot.***
everybody very like-3Sg.INDEF Hungary
'Everybody likes Hungary very much.'
(*Mindenki nagyon szereti Magyarországot.)*

The object is a noun with a definite article:
A lányok akik ***néztek*** ***a filmet****, az*
the girls who watch-Past-3Pl.INDEF the film-ACC the
egesz filmig táncoltak és
whole film-TER dance-Past-3Pl.INDEF and
buliztak.
party-Past-3Pl.INDEF
'The girls who were watching the film were dancing and partying during the whole film.'
(*A lányok, akik a filmet nézték, az egész film alatt táncoltak és buliztak*.)

Indefinite object:
Gondoltam, hogy most könnyebb
think-Past-1Sg.DEF that now easier
fogom találni valamilyen más munkát*, de*
will-1Sg.DEF to.find some other job-ACC but
nem volt így.
not was so
'I thought that now it will be easier for me to find another job but it was not so.'
(*Azt hittem, hogy most könnyebben fogok másik munkát találni, de nem így lett.*)

És néha ***látom*** ***nagyon erős***
and sometimes see-1Sg.DEF very strong
nátionaliszmusot*.*
nationalism-ACC
'And sometimes I can see a very strong nationalism.'
(*És néha nagyon erős nacionalizmust látok*.)

The object is a demonstrative pronoun:
De a hétvégén keresztül olvasok
but the weekend-SUP during read-1Sg.INDEF
a magyar hireket az interneten
the Hungarian news-ACC the internet-SUP
és csak idegensítek, mert
and only get.nervous-1Sg.INDEF because
látok azt*,*
see-1Sg.INDEF that-ACC
hogy milliárdokért építtetnek
that billion-Pl-CAU build-CAUS-3Pl.INDEF
mélygarázst.
deep level garage-ACC
'During weekends, I read the Hungarian news on the internet and I only get nervous because I can see that they are having deep level garages built for billions.'
(*De hétvégente olvasom a magyar híreket az interneten, és csak idegeskedem, mert azt látom, hogy milliárdokért építtetnek mélygarázst.*)

The object is a general pronoun:
Egyszer Alfred azt mondta hogy Alma
once Alfred that-ACC say-3Sg.DEF that Alma
a nő aki ***mindent tudja***
the woman who everything-ACC know-3Sg.DEF
róla.
about.him
'Alfred said once that Alma is the woman who knows everything about him.'
(*Egyszer Alfred azt mondta, hogy Alma az a nő, aki mindent tud róla.)*

The object is a relative pronoun:
Nem kell elfelejni, hogy azt, amit
not should to.forget that that-ACC what-ACC
mondtam, csak arról az emberekről
say-Past-1Sg.DEF only that-SUB the man-Pl-SUB
lehet mondani, ***akit*** *nem* ***ismerem****.*
can to.say who-ACC not know-1Sg.DEF
'It must not be forgotten that the things I said can be said only about the men that I know.'
(*Nem szabad elfelejteni, hogy amit mondtam, csak azokról az emberekről lehet mondani, akiket ismerek*.)

7. Usability of results

Our results may be fruitfully applied in language teaching on the one hand as the statistical analysis makes it possible for the students and the teachers to concentrate on grammatical structures that seem to give rise to more difficulties. On the other hand, from a natural language processing point of view, definiteness errors in conjugation may be automatically corrected as the automatic detection of the type of the object triggers the type of conjugation. If the sentence does not contain the required form, a grammar checker may automatically propose some corrections concerning the word form of the verb.

8. Conclusions

Here we presented our approach to automatically detect conjugational errors concerning definiteness in a Hungarian learners' corpus. Our results reveal grammatical structures that might pose problems for learners of Hungarian, which can be fruitfully applied in the teaching and practicing of such constructions from the language teacher's or learners' point of view. On the other hand, these results may be exploited in extending the functionalities of a grammar checker, concerning the definiteness of the verb.
The HunLearner corpus is freely available at our website for research and educational purposes:
http://www.inf.u-szeged.hu/rgai/hunlearner.

9. Acknowledgements

This work was supported in part by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11 / 1 / KONV-2012-0013).

10. References

Bratchikova = Братчикова, Н.С. (2013). Эталонные знаки феномена "богатый" в сознании современных представителей финского лингвокультурного сообщества [http://phil.spbu.ru]

Durst, P. (2010a). Kutatásmódszertani kérdések a magyar mint idegen nyelv elsajátításában *THL2. A magyar nyelv és kultúra tanításának szakfolyóirata*. Budapest, Balassi Intézet, Budapest. pp. 82–90.

Durst, P. (2010b). A magyar mint idegen nyelv elsajátításának vizsgálata – különös tekintettel a főnévi és igei szótövekre, valamint a határozott tárgyas ragozásra. Unpublished PhD thesis. Pécs, University of Pécs.

Durst, P.; Janurik, B. (2011). The Acquisition of the Hungarian definite conjugation by learners of different first languages. *Lähivõrdlusi. Lähivertailuja* 21**.** Tallinn, Estonian Association for Applied Linguistics (EAAL). pp. 19–44.

Durst, P.; Szabó, M.; Vincze, V. ; Zsibrita, J. forthcoming. Using automatic morphological tools to process data from a learners' corpus of Hungarian.

Guskova = Гуськова, А.П. (2009). Категории определенности / неопределенности, вида, залога, in *Венгерский язык. Справочник по грамматике*. Москва, Живой язык. pp. 144–156.

Langman, J.; Bayley, R. (2002). The acquisition of verbal morphology by Chinese learners of Hungarian. *Language variation and Change* 14. pp. 55–77.

M. Korchmáros, V. (2006). *Lépésenként magyarul. Magyar nyelvtan – nem csak magyaroknak*. Szeged, Szegedi Tudományegyetem.

Törkenczy, M. (2005). *Practical Hungarian Grammar*. Budapest, Corvina Books Ltd., 2nd edn.

Zsibrita, J.; Vincze, V.; Farkas, R. (2013). magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: *Proceedings of RANLP-2013*, Hissar, Bulgaria.