# Syntax is clearer on the other side – Using parallel corpus to extract monolingual data

**Andrea Dömötör**
Pázmány Péter Catholic University / H-1088 Budapest, Szentkirályi str. 28.
MTA-PPKE Natural Language Processing Group / H-1083 Budapest, Práter str. 50/a.
domotor.andrea@itk.ppke.hu

## Abstract

This paper describes the elaboration of a training corpus containing Hungarian sentences that are labelled according to a syntactic criterion, namely the syntactic role of a very common multifunctional word *volt* 'was/had'. The labels are assigned by a rule-based algorithm that specifies the function of the target word based on the English pairs of the sentences extracted from a parallel corpus. The reasoning of this idea is that the required syntactic information is easier to retrieve in English than in Hungarian. The accuracy achieved by the algorithm was fair but still needs improvement in order to use the output as reliable training data. The obtained training corpus was tested with FastText's text classifier, the results of which showed that the targeted disambiguation problem is resolvable using neural network based text classification.

## 1 Introduction

In the past years deep learning methods have come to dominate in most of the areas of computational linguistics. A general advantage of these is their robustness and relative simplicity compared to rule-based systems. The key of success in deep learning is having a large and good set of training data, therefore corpus building has become an important field of research.

This paper describes the elaboration of a training corpus containing Hungarian sentences that are labelled according to a syntactic criterion, namely the syntactic role of a very common multifunctional word *volt* 'was/had'. The labels are assigned by a rule-based algorithm that specifies the function of the target word based on the English pairs of the sentences extracted from a parallel corpus. The reasoning of this idea is that the required syntactic information is easier to retrieve in English than in Hungarian.

### 1.1 The deep learning task

The targeted deep learning task is a word sense disambiguation problem in Hungarian, namely the automatic handling of the multifunctionality of the word *volt* 'was/had'. This token can either be a lexical verb used in locative and possessive sentences (Examples 1, 2) or a copula in case of nominal predicates (Example 3).

(1) a. *Ádám otthon* ***volt.***
Adam at_home be-PST-Sg3
'Adam was at home'

b. *Ádám otthon van.*
Adam at_home be-Sg3
'Adam is at home'

(2) a. ***Volt*** *egy macskám.*
have-PST-Sg1 a cat-Poss.Sg1
'I had a cat.'

b. *Van egy macskám.*
have-Sg1 a cat-Poss.Sg1
'I have a cat.'

(3) a. *Éva nagyon szerény* ***volt.***
Eve very humble AUX-PST-Sg3
'Eve was very humble.'

b. *Éva nagyon szerény.*
Eve very humble
'Eve is very humble.'

The main difference between these functions is that *volt* in Example 3 is omitted in present tense 3rd person while the locative and possessive verbs (Examples 1 and 2) have their present forms *van*. Based on this characteristic of the examined sentence types, this research aims to differentiate between two functions of the word *volt*. These functions will be referred as *copula* (Example 3) and *lexical verb* (Examples 1 and 2) later on. These denominations are different from the Anglo-Saxon terminology where the locative *be* is also considered a copula. However, the studies on Hungarian syntax often narrow the meaning of copula to the auxiliary verb of the nominal predicate because of its exclusive capability of having a zero form. This study follows this traditional Hungarian terminology for the same reason.

In dependency parsing a lexical verb should be considered the head of the sentence while the copula (which can be omitted at least in some persons or tenses) is a complement of the predicative nominal, according to the annotation guidelines of Universal Dependencies (Nivre, 2014). Therefore, the disambiguation of these functions is crucial for parsing. However, as seen in Examples 1 and 3, disambiguation cannot be made based on corresponding lexical items (*be* or *have*) alone because the verb of locative sentences and the auxiliary of the nominal predicate also need to be distinguished, and these are both represented by *be* in English. The disambiguation of the functions of *be* requires a deeper analysis of parse structures.

### 1.2 The aimed solution

Copular, locative and possessive sentences have clear distinctive structural characteristics, however, a rule-based method is not effective for Hungarian. One source of difficulty is that in Hungarian the word order does not define the syntactic role of the words. Other characteristic that complicates the automatic handling of Hungarian is that it is a so-called pro-drop language, which means that the subject of the sentence is not necessarily overt. Both mentioned characteristics of Hungarian syntax obstacle the detection of predicative nominals to such an extent that the specification of the sentence types listed above would need an in-depth analysis (morphology and NP-chunking). It seems more advantageous to solve this problem with a deep learning method, like neural network based sentence classification.

For this approach a large amount of labelled data is required. This study focuses on the acquisition of training data for a sentence classifier. The obtained data was tested with FastText's text classifier (Joulin et al. (2016), Bojanowski et al. (2016)).

### 1.3 Baseline results

The results will be compared to the performance of the `e-magyar` toolset which is an integrated text processing pipeline for Hungarian (Váradi et al. (2018)). The system has 8 modules that cover the most common NLP tasks (tokenizer, morphological analyzer, lemmatizer, POS tagger, dependency parser, constituent parser, NP chunker, NER tagger). For the specific task of this paper I used the dependency parser module (which obviously uses the analyses of the modules of lower levels). A test set of 1000 sentences was parsed and classified according to the parser's analyses. If there was a word in `PRED` relation with *volt* the sentence was assigned a `copular` tag, otherwise it received a `lexical` tag. The tags were reviewed manually. The results are displayed in Table 1.

| | |
|---|---|
| Erroneous labels | 186 |
| Accuracy | 81,4% |

Table 1: Results of the evaluation of the `e-magyar` tool on 1000 sentences

As the achieved accuracy result shows, the monolingual pipeline analysis struggles with the ambiguity of *volt*.

## 2 Method

A neural network based sentence classifier that could solve the problem described in Section 1.1 needs training data with sentences that are annotated with the corresponding function (verb or copula) of the target word. As manual labelling is time-consuming, it was inevitable to find a method for automatic

labelling. The basic idea of this method is to use an English-Hungarian parallel corpus. Contrary to Hungarian, English has a restricted word order and no pro-drop, which characteristics allow to make syntactic decisions based on local information. That means that the English pairs of the Hungarian sentences can help to define the function of the word *volt*, by applying fewer and simpler rules as if we used the Hungarian part only.

### 2.1 The parallel corpus

For data extraction I used an English-Hungarian lemmatized, morphologically analyzed and disambiguated, word-aligned corpus (Novák et al., 2019). This research did not contribute to the creation of this corpus.

The base of the corpus is OPUS Opensubtitles (Lison and Tiedemann, 2016) which contains 644,5 million tokens of aligned sentences. As first step, both sides of the corpus were morphologically analyzed and disambiguated. The English side was lemmatized with the morpha tool (Minnen et al., 2001) and tagged with Stanford tagger (Toutanova et al., 2003). On the Hungarian side the lemmatization and disambiguation was made with PurePos (Orosz and Novák, 2013) which uses the analyses of the Humor analyzer (Novák, 2014). The analyzed texts were transformed on both sides so that every original token is represented by two tokens: (1) the lemma and its main POS-tag and (2) other morphosyntactic tags belonging to the token.

Example 4 shows a pair of preprocessed sentences (*Szeretlek, kedvesem. – I love you, dear*).

(4) a. `szeret[IGE] [Ie1] ,[PUNCT] kedves[FN] [PSe1][NOM]`

b. `I#PRP love#VB [P] you#PRP ,#, dear#RB`

The preprocessed sentences were word aligned with the fast align programme (Dyer et al., 2013). The alignments of Example 4 are displayed in Table 2.

| | |
|---|---|
| szeret[IGE] | love#VB |
| [Ie1] | I#PRP, you#PRP |
| ,[PUNCT] | ,#, |
| kedves[FN] | dear#RB |
| [PSe1][NOM] | dear#RB |

Table 2: The alignments of Example 4

### 2.2 The labelling algorithm

Having the prepared parallel corpus, the first step was to extract the sentences that contained a form of the target word (*volt*) on the Hungarian side. These sentences were labelled according to the syntactic role (copula or lexical verb) of the target word with a rule-based algorithm implemented in Python3.

The labelling programme first checks the English tokens aligned to *volt*. If *volt* is aligned to a non-auxiliar *have* or an expletive *there*, the sentence is labelled as lexical. If the target word is aligned to a form of *be*, the sentence can either be copular or locative, therefore further rules are required to make the decision. In other cases, the sentence is dismissed because if none of the above listed tokens is aligned to *volt*, the English pair of the sentence can not be used for labelling reliably.

In case of *volt* aligned to *be*, the algorithm selects a "keyword" on the English side, the Hungarian alignments of which define the label of the sentence. The keyword is supposed to represent a (part of a) nominal predicate or a non-nominative argument. Therefore, the algorithm searches for the canonical position of these in English sentences.

For keyword selection the programme first specifies whether the sentence is interrogative. If the sentence is declarative the keyword is the first token following *be* that is not an NP-modifier (*very, more* etc.) or a word of negation (Example 5). If the sentence is a yes-no question or its question word is *what*, *who*, *whose*, *which*, *how* or *why*, the programme follows the same principles as with declaratives but skips one more word due to the inversion of word order (Example 6). If the sentence has another question word (*where*, *when* etc.), the sentence is labelled as lexical.

(5) a. *Régen ez egy kvalitás volt.*

It used to be **a** quality.

b. *Nem volt otthon.*

He was not **at** home.

(6) a. *Mi volt ez a zaj?*

What was that **noise**?

b. *Miről volt szó?*

What was it **about**?

The algorithm then checks the morphological tags aligned to the keyword and labels the sentence based on these. The sentence is assigned a `lexical` label if the aligned morphological tag is a non-nominative case marker. If the keyword is aligned to a determiner or a nominative nominal the sentence is labelled `copular`. The tags listed in Table 3 cover all the morphological tags that are aligned to a keyword in the corpus.

| **lexical** | | **copula** | |
|---|---|---|---|
| HA | adverb | DET | determiner 'the, a an' |
| HA\|NM | adverbial pronoun | DET\|NM | determinative pronoun |
| NU | nominal postposition | MN | adjective |
| INE | inessive 'in' | MN\|NM | adjectival pronoun |
| SUP | superessive 'on' | FOK | comparative adjective |
| ELA | elative 'from inside' | FF | superlative adjective |
| ADE | adessive 'at (place)' | SZN | numeral |
| ESSMOD | modal essive '-ly' | SZN\|NM | numeral pronoun |
| ILL | illative 'into' | FN | noun |
| ALL | allative 'onto' | FN\|NM | nominal pronoun |
| SUB | sublative 'to (somewhere)' | PS | possessive nominal |
| CAU | causative 'for (reason)' | OKEP | '-ing' |
| ABL | ablative 'of' | MIB | past participle (adjectival) |
| HIN | past participle (passive constructions) | | |
| INS | instrumental 'with' | | |
| DEL | delative 'about' | | |
| DAT | dative 'to (someone)' | | |
| TER | terminative 'until' | | |
| TEM | temporal 'at (time), during' | | |
| ESSNUM | numeral essive '(three) of us' | | |

Table 3: The morphological tags aligned to keywords and the assigned labels

The algorithm also applies some special lexical rules where the morphological tags would be misleading. First, we should mention a special construction that Kádár (2011) calls *environmental copula construction*. These are NP + VAN 'be' constructions that comprise weather, ambient or environmental conditions. Environmental copula constructions do not behave as "other" copular constructions: they do not omit the copula in present tense third person. This means they should be labelled as sentences with a lexical verb, but the keyword-based part of the algorithm would obviously tag them as copular (see Example 7).

(7) a. *Sötét volt és köd.*
dark be-PST-Sg3 and fog

'It was dark and foggy.'

b. It was dark and foggy.

Therefore, these constructions are handled lexically, based on a list of nominals that usually form a part of an environmental copular construction.

There are other cases where keyword selection fails and these could be called consistent translational differences. This means that some English copular clauses are consistently translated to Hungarian with a lexical verb.

The most common case of this is the translation of "being right". As seen in Example 8, in Hungarian "being right" is literally expressed as "having the truth" which is, syntactically, a possessive structure but the algorithm labels it as copular based on its English pair. The case of "being lucky" is similar (see Example 9), however, this expression also has a copular version in Hungarian.

(8) a. *Igazad volt.*
truth-Poss.Sg2 have-PST-Sg3
'You were right.'
b. You were right.

(9) a. *Neki volt szerencséje.*
he-DAT have-PST-Sg3 luck-Poss.Sg3
'He had luck.'
b. He was lucky.

The algorithm handles these cases (and two further similar ones: "being necessary" and "being ready") with exceptional lexical rules.

The labelling algorithm is summarized in Table 4.

| | |
|---|---|
| **Step 1**: Check aligns of *volt* | |
| *have* | `lexical` |
| *there* | `lexical` |
| *be* | go to Step 2 |
| other | dismiss sentence |
| **Step 2**: Special lexical rules | |
| environmental copular construction | `lexical` |
| *right, lucky, necessary, ready* | `lexical` |
| other | go to Step 3 |
| **Step 3**: Keyword selection | |
| declarative sentence | token following *be* |
| yes-no question | *be* + 2 tokens |
| *what, who, whose, which, how, why* | *be* + 2 tokens |
| other wh-question | `lexical` |
| **Step 4**: Assign label according to keyword | |

Table 4: Summary of the labelling algorithm

### 2.3 Sentence classification

The obtained labelled corpus was used as training data for FastText's text classifier. I prepared two versions of the training corpus: one contains the original sentences while in the other the sentences are represented with the POS-tags of their words only. Both corpora were trained for the same classification task.

## 3 Results

The output of the labelling script was 791130 labelled sentences, 458270 of which was tagged as copular and 332860 as containing a lexical verb. These numbers show that the target word - as expected - is extremely common which allows to build a reasonably big corpus for our specific task.

The performance of the algorithm was evaluated on a random sample of 1000 sentences, 598 of which is copular and 402 contains a lexical *volt*. (The same sentences were used for the baseline test described in Section 1.3.) The labels that the algorithm gave on this sample were reviewed manually, and also corrected so that FastText could use the same sample as gold standard test data. The results are displayed in Table 5.

| | |
|---|---|
| Erroneous labels | 108 |
| Accuracy | 89,2% |

Table 5: Results of the evaluation of the labelling algorithm on 1000 sentences

The labelling algorithm overperformed the baseline result (81,4%) significantly, however the achieved accuracy is still far from a gold standard training corpus. The obtained labelled corpus was subject to the neural network based classification experiment anyways.

The accuracy results of FastText classifier are displayed in Table 6. As seen, the classifier works well despite the deficiencies of the training corpus.

| | |
|---|---|
| Original sentences | 89,6% |
| POS-tags | 91,5% |

Table 6: Results of sentence classification (FastText)

## 4 Discussion

As seen in Section 3 both the labelling algorithm and the sentence classifier achieved significantly higher accuracy than the baseline, however, the quality of the training corpus still needs to be improved. This section reviews the labelling algorithm's most common reasons of failure and the possibilities to avoid them.

### 4.1 Translational differences

The error analysis of the labelling algorithm revealed that the major part of errors does not originate from the algorithm itself. There are labelling mistakes that can be considered "extraneous", because they are caused by erroneous POS-tagging or alignment. Other very common sources of errors are the occasional differences between the English sentences and their Hungarian translations. The algorithm attempts to avoid this problem by disregarding those sentences where *volt* is not aligned to either *be* or *have*. But this constraint still allows a considerable number of sentences where the inconsistent structural, or sometimes also semantic differences of the paired sentences cause difficulties to the labelling algorithm. In Example 10 the Hungarian sentence (10a) is copular but in its English pair (10b) the verb (aligned to *volt*) is *have*, therefore the algorithm assigned a lexical label to the sentence. Example (11a) is a locative sentence but the programme considered it copular based on its English version (11b), which is indeed copular.

(10) a. *Egy rossz álom volt.*
a bad dream AUX-PST-Sg3
'It was a bad dream.'

b. You had a bad dream.

(11) a. *Ők voltak itt először.*
they be-PST-Pl3 here first
'They were here first.'

b. They were the first ones here.

These errors can hardly be avoided, however, the handling of translational differences may worth further consideration. Other possible solution could be the use of parallel corpora with "stricter" translations, like documents of the European Union. The disadvantage of this approach would be the limited domain.

### 4.2 Special cases

The error analysis also revealed some special cases that are not covered properly by the current version of the algorithm.

A recurrent problem was the handling of nominals with arguments, like "being sure about something" or "being responsible for something" (Example 12). In some of these cases the argument is omitted in the English sentence but it is present in its Hungarian pair. Therefore, the case marker of the argument on the Hungarian side is aligned to the English nominal which is often the labelling algorithm's keyword. As described in Section 2.2 a keyword aligned to a non-nominative case marker indicates that the sentence has a lexical *volt* which is not true in these cases.

(12) a. *bárki is volt érte a felelős.*
whoever ever AUX-PST-Sg3 it-CAU the responsible
'whoever was responsible for it.'

b. whoever was responsible.

The handling of these special cases needs a more detailed analysis.

## 5 Conclusions

The main idea of this paper was to retrieve syntactic information in a parallel corpus, by relying on another language in which the automatic disambiguation of the structure is easier. The described algorithm uses English sentences to define the syntactic role of a target word in the Hungarian translations. The goal was to create a labelled corpus that can be used as training data for a neural network based sentence classifier.

The results show proof of concept for the idea, although the accuracy still needs to be improved. The classifier, however, seems to deal fairly with the deficiencies of the training corpus, especially if we use the POS-tags instead of words. The cause of the difference of performance of the two kinds of training corpus may be the small size of the corpora. If only the POS-tags are used the vocabulary is significantly smaller which facilitates the creation of good embeddings. The successful classification based on POS-tags also demonstrates that the difference between copular and lexical *volt* is in great part coded in the sentence structure.

In sum, the experiments described in this paper demonstrated that parallel corpora can be useful to support syntactic analysis in any cases where the targeted structure is more explicit in an another language. On the other hand, FastText's results confirmed that neural network based text classifiers are not for sentiment or topic identification only, they can capture structural differences as well.

## References


Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *CoRR*.

Edit Kádár. 2011. Environmental Copula Constructions in Hungarian. *Acta Linguistica Hungarica*, 2011(4):417–447.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Guido Minnen, John A. Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Joakim Nivre. 2014. Nonverbal Predication and Copulas in UD v2. `http://universaldependencies.org/v2/copula.html`. Accessed: 2019-02-27.

Attila Novák. 2014. A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 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.

Attila Novák, László János Laki, and Borbála Novák. 2019. Mit hozott édesapám? Döntést – Idiomatikus és félig kompozicionális magyar igei szerkezetek azonosítása párhuzamos korpuszból [identification of Hungarian idiomatic and light verb constructions from a parallel corpus]. In *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019) [15th Hungarian Conference on Computational Linguistics]*, pages 63–71, Szeged. Szeged University.

György Orosz and Attila Novák. 2013. PurePos 2.0: a hybrid tool for morphological disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, pages 539–545, Hissar, Bulgaria. Incoma Ltd. Shoumen, Bulgaria.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tamás Váradi, Eszter Simon, Bálint Sass, Iván Mittelholcz, Attila Novák, Balázs Indig, Richárd Farkas, and Veronika Vincze. 2018. E-magyar – A Digital Language Processing System. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).