Creation of a corpus with semantic role labels for Hungarian

Attila Novák[1,2], **László János Laki**[1,2], **Borbála Novák**[1,2]
Andrea Dömötör[1,3], **Noémi Ligeti-Nagy**[1,3], **Ágnes Kalivoda**[1,3]
[1]MTA-PPKE Hungarian Language Technology Research Group,
[2]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
Práter u. 50/a, 1083 Budapest, Hungary
[3]Pázmány Péter Catholic University, Faculty of Humanities and Social Sciences
Egyetem u. 1, 2087 Piliscsaba, Hungary
{surname.firstname}@itk.ppke.hu

Abstract

In this article, an ongoing research is presented, the immediate goal of which is to create a corpus annotated with semantic role labels for Hungarian that can be used to train a parser-based system capable of formulating relevant questions about the text it processes. We briefly describe the objectives of our research, our efforts at eliminating errors in the Hungarian Universal Dependencies corpus, which we use as the base of our annotation effort, at creating a Hungarian verbal argument database annotated with thematic roles, at classifying adjuncts, and at matching verbal argument frames to specific occurrences of verbs and participles in the corpus.

1 Introduction

Recently, state-of-the-art performance in most NLP related tasks has been achieved by end-to-end systems based on neural deep learning networks (see e.g. BERT (Devlin et al., 2018) or GPT-2 (Radford et al., 2019)) surpassing the performance of previous systems employing some sort of grammatical analysis. This has raised doubts as to whether it makes sense to deal with grammatical analysis at all. At the same time, the training of end-to-end systems usually requires a great amount of training material, which is not available in most languages. Therefore, we think it may still make sense put an effort into the implementation of a grammatical analysis framework as long as the output of the system can be directly used to perform tasks relevant to everyday users.

However, we cannot be satisfied with an analysis that relies on completely abstract categories that cannot be clearly translated into terms that can be linked to what that text means in a manner that can also be understood by ordinary people. An essential element of reading comprehension is that we are able to ask meaningful questions about the given text, and this ability is closely related to the ability to answer questions. Therefore, our aim is to create a system that is actually capable of formulating relevant questions about the text it processes. To do this, many distinctions need to be made that are not present in syntactic annotation currently available for Hungarian. This article presents the first phase of this work, which aims to create an annotated corpus where the annotation contains all the features needed to generate questions concerning the text.

2 Shortcomings of the traditional analysis

Since our goal is to create a system that can generate meaningful questions, we have decided that when determining what distinctions need to be made in the annotation should be basically determined by what questions can be asked concerning the particular grammatical construction. For example, in order to be able to formulate questions concerning **noun phrases**, the *Who/What?* distinction is indispensable, so the system must be able to clearly distinguish persons from things. At the same time, we ask *who* or *what* questions concerning NP's that refer to groups or organizations depending on the role they play in the given sentence. For example, a bank is referred to linguistically as a person when sending an invoice letter, but as a thing when it is liquidated. In addition, a more detailed classification is required to generate questions about nominal predicates. Concerning the predicate in the sentence *John is a doctor*, the question *Who is John?* is not very sophisticated. *What is John's occupation?* is a question matching the predicate in the sentence much more precisely. Classifying concepts as occupations, animals, tools, behaviors, etc. also makes for the system possible to generate more specific

Proceedings of the 13th Linguistic Annotation Workshop, pages 220–229
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

questions related to non-predicative occurrences of noun phrases: e.g. *What animal have you seen in the garden?* vs. *What did you see in the garden?* This is particularly important in the case of coordinated phrases where one can only identify which conjunct is meant in the question if the question is specific enough.

To formulate questions concerning adverbials even at the most basic level, we also need a much more detailed system of distinctions than what is provided by the syntactic annotation present in currently available tree banks. Hungarian NP's headed by a word in inessive case or corresponding English PP's headed by the preposition *in* can have quite a number of different grammatical functions. Thus we ask different questions concerning them: (1) *szeptemberben* 'in September': *mikor?* 'when?', (2) *Londonban* 'in London': *hol?* 'where?', (3) *fájdalmában (felüvöltött)* '(he screamed) in pain': *mitől?* 'what made (him scream)?', (4) *magában (hisz)* '(he believes) in himself': *kiben?* 'in whom?', (5) *bajban* 'in trouble': *milyen helyzetben?* 'in what situation?', (6) *életben (marad)* '(stay) alive' lit. '(stay) in life': no question in general, this is part of a light verb construction.

Generating questions concerning not only nominal but also verbal predicates requires information not provided by currently available annotation for Hungarian. How a question concerning a verbal predicate should be formulated using specific arguments as anchors depends on the thematic roles the arguments play. *What did John do to Frank?* is an adequate question if John is an agent and Frank is a patient. In the same situation, *What happened to Frank?* and *What did John do?* are likewise adequate questions.

Identification of thematic roles of verbal argument slots is also needed in order to be able to distinguish oblique arguments from semantically compositional relations (e.g. locative and oblique uses of *in*: *believe in something* vs. *be somewhere*). We also need to distinguish parts of idioms and light verb constructions from compositional verb-to-argument relations. It is a joke to ask a question concerning a non-compositionally related constituent:

What are you holding? — A meeting.

3 The corpus

As a starting point, we chose the Hungarian subcorpus (Vincze et al., 2017) of the Universal Dependencies (UD) corpus (Nivre et al., 2016) consisting of 1800 sentences (42000 tokens) of mainly newswire text in order to put the annotation schema we propose in a context that can be interpreted at an international level. The UD corpus contains texts in many languages annotated with morphosyntactic and dependency-based syntactic analysis using unified principles and categories. Our original plan was to supplement or refine the annotation in the Hungarian UD corpus with the information needed to formulate questions. However, it turned out that the annotation in the Hungarian sub-corpus does not correspond to the currently valid UD specification in many respects, and contains many random annotation errors, so fixing these errors turned out to be an inevitable part of our task.

According to the UD 2.0 specification[1], the internal structure of multiword expressions is to be annotated using the `flat, fixed` or `compound` dependency relations. The `fixed` relation is used exclusively to annotate fully lexicalized function-word-like structures. In many languages, such as English, multiword names are generally considered to be flat exocentric structures, and the use of `flat` is suggested to annotate the internal structure of these names with all words of the name directly attached to the first word of the name. On the other hand, the UD 2.0 annotation specification explicitly excludes the use of this type of analysis in cases where the name has a regular syntactic structure (eg. in the case of book or movie titles or a large part of names of organizations). Here the generic syntactic dependency structures are to be used. Similarly, endocentric structures should be annotated using the `compound` relation or one of its sub-types[2] (see e.g. Kahane et al. (2017) on the contradiction of applying the flat annotation to languages where names are endocentric). Hungarian noun phrases are always right-headed endocentric structures, so in the case of names that do not have a regular structure and compositional meaning, the `compound` relation is to be used. This ensures,

[1] http://universaldependencies.org/guidelines.html

[2] https://universaldependencies.org/u/overview/specific-syntax.html#multiword-expressions

for example, that case endings always attached to the head of the NP are directly accessible. E.g. the head of an object NP is always in the accusative case in Hungarian. The current annotation for names completely obscures this fact (see e.g. *az Egyesült Államokat* 'the united States[Acc]' in Figure 1)). Therefore, as one of the preprocessing steps, all multiword names, originally erroneously annotated in the corpus as `flat` structures, were automatically converted into `compound` structures (Figure 1). For the time being, the identification and further reannotation of names with a completely regular structure has not been done, as this requires manual intervention.

Structures like *Angela Merkel német kancellár* 'German Chancellor Angela Merkel' were often erroneously annotated as appositive structures in the corpus.[3] We converted these structures introducing the `compound:title_of` relation between the name and the occupation/role.

The UD 2.0 specification prescribes the use of the `obl` relation to attach NP arguments other than subjects, objects or indirect objects even in the case of nonverbal heads. Often, some other relation was used in the corpus even for verbal arguments. We were able to automatically correct most of these annotations in the case of arguments of verbs and participles (Figure 2).

In Hungarian, like in German, verbal particles are detached from the verb in various syntactic constructions, and they are moved to some distinct syntactic position. Nevertheless, these particles are considered part of the verb lemma. The verbal argument database that we created as part of our annotation effort, also contains particle verbs in this form. In the Hungarian UD corpus, on the other hand, verb lemmas did not include the particle in such cases. This needed to be fixed (adding the particle to the verb lemma) in order for the verbal argument frames could be matched to their occurences in the corpus. Many additional lemmatization errors were fixed, and we also needed to relemmatize participles so that we can match verb argument frames against them.

In Hungarian, demonstrative predeterminers agree in case and number with the head of the NP (*azokat a kutyákat* 'those dogs$_{ACC}$'). These structures were often annotated erroneously, with the demonstrative predeterminer being attached to the head of the NP using the same dependency label the whole NP was annotated with. We corrected these errors and attached all predeterminers as `det:predet` to the head of the NP (Figure 3).

Further corrections performed automatically included using `nmod:poss` instead of `nmod:att` in possessive structures, (bottom of Figure 2), attaching all postpositions using the `case` relation, and fixing clauses where the subject and the (nominal) predicate were exchanged in the annotation by mistake due to the annotators confusing the focus construction with predication. The latter errors needed to be identified manually, the correction of the identified structures was then performed automatically (Figure 4).

4 The argument frame database

All stems of verbs and participles occurring in the Hungarian UD corpus were collected, and they were clustered using agglomerative clustering like in (Siklósi, 2016) based on their vector representation in a word-embedding model constructed from a morphosyntactically annotated corpus (Novák and Novák, 2018). This process effectively clustered verbs having similar distributional patterns (and argument frames). Each verb in the list was supplemented with its surface argument frames from a Hungarian verb-frame dictionary (Sass et al., 2010). Using this initial representation as a source of inspiration, we have described the possible argument frames of each verb manually. Our description contains the thematic role, the surface features (case-ending, postposition, possessive suffix, etc.), possible optionality and, if applicable, lexical/semantic constraints of each argument. Clustering helped us to streamline the process and simplified the task of annotators. Annotating verbs with similar argument frames in a batch together instead of having to process them in some random order made it possible for us to use an inheritance mechanism and improved consistency.

The main point in describing the argument frames of verbs was to provide as much information as possible to make it possible to ask the best, most accurate questions. With that in mind, our set of thematic roles is based on widely known thematic role hierarchies. However, it differs from them in minute details, just like they differ from each other. The description of verbs is intended to cover every possible meaning (argument frame).

[3] In appositive structures, like *a bátyámmal, Péterrel* 'with my brother, Peter' there is case agreement between parts of the phrase. This is not the case in these structures.

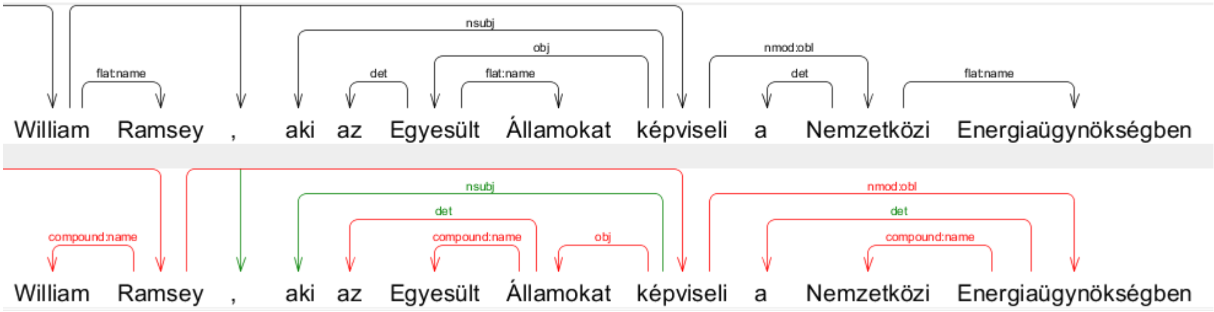
nsubj
obj
flat:name
det
flat:name
nmod:obl
det
flat:name
William Ramsey , aki az Egyesült Államokat képviseli a Nemzetközi Energiaügynökségben
compound:name
nsubj
det
compound:name
obj
nmod:obl
det
compound:name
William Ramsey , aki az Egyesült Államokat képviseli a Nemzetközi Energiaügynökségben

Figure 1: Fixing the annotation of multiword names: ‘William Ramsey representing the United States at the International Atomic Energy Agency’

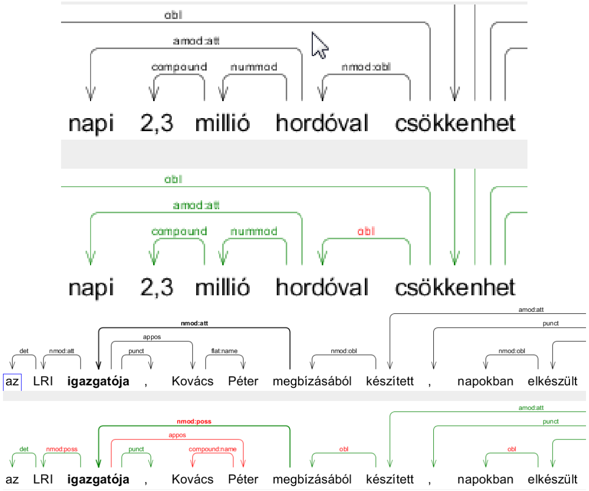
obl
amod:att
compound
nummod
nmod:obl
napi 2,3 millió hordóval csökkenhet
obl
amod:att
compound
nummod
obl
napi 2,3 millió hordóval csökkenhet
nmod:att
appos
punct
flat:name
det
nmod:att
nmod:obl
amod:att
punct
nmod:obl
az LRI igazgatója , Kovács Péter megbízásából készített , napokban elkészült
nmod:poss
appos
punct
compound:name
det
nmod:poss
obl
amod:att
punct
obl
az LRI igazgatója , Kovács Péter megbízásából készített , napokban elkészült

Figure 2: Using the `obl` relation for arguments of verbs and participles: ‘... may decrease by 2.3 million barrels a day’ – ‘a recently completed [report] commissioned by Péter Kovács, director of LRI’

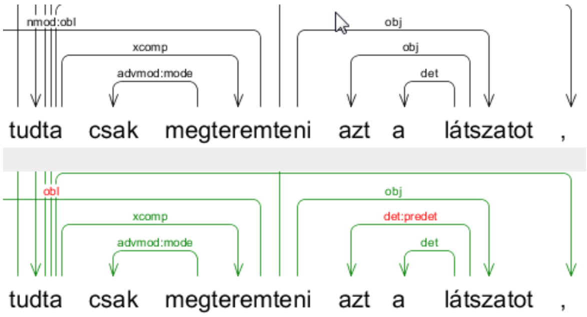
nmod:obl
xcomp
advmod:mode
obj
obj
det
tudta csak megteremteni azt a látszatot ,
obl
xcomp
advmod:mode
obj
det:predet
det
tudta csak megteremteni azt a látszatot ,

Figure 3: Correction of erroneously annotated predeterminers: ‘... was the only way he could create that impression’

Since verbs with similar meanings and argument frames were already grouped in the database, it was possible to specify common argument frames for groups of verbs. These frames are inherited automatically by verbs belonging to the same group. In addition, each verb can have its own argument frame which does not apply to the whole group.

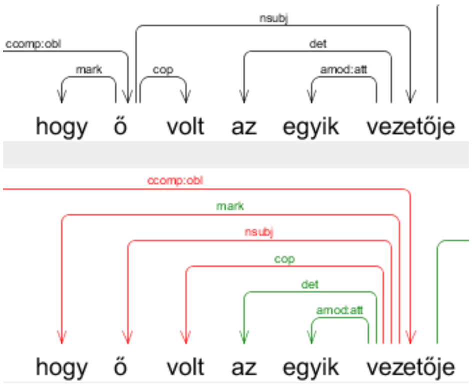

Figure 4: Correction of an exchanged subject and predicate: ‘that he (*ő*) was one of the leaders (*vezetője*) of...’

This frame can be added to the record of the specific verb.

The required and optional arguments of each verb are represented either by their thematic roles or lexically, supplemented with the required case-endings or postpositions. The identification of the thematic roles is based on the question that can be asked about the given argument or about the verb with the given argument as an anchor. For example, the question concerning the agent is *what is A doing?*, the question concerning the patient is *what is happening to P?*.

Some roles also represent a kind of semantic category, such as `CONT` which refers to the content of communication, or `ACT` which denotes an action (usually expressed as an infinitive `xcomp`). Arguments not having a specific thematic role that could not be used as an anchor when we want to ask a question about the predicate were marked using the semantically neutral theme (`TH`) thematic role.

The fixed components of idiomatic or semi-compositional verbal structures are not labeled by thematic roles but they are specified lexically.

These structures were supplemented with their own argument frame descriptions (thus interpreted as autonomous units) where this solution seemed to be justified. For example, the description of *sor kerül* ‘to take place’ (lit. ‘turn comes’) is not part of the description of the word *kerül* ‘to come’, but we have assigned an argument frame to the whole phrase as a unit. The thematic roles assigned to the verbs and verbal structures are summarized in Table 1.

Since verbs of movement imply the applicability of specific types of questions (e.g. *How did X get to Y?*), in addition to the roles listed in the table, a special annotation was applied to moving actors: for example moving agents are marked as `AGMV`. Our basic assumption was that a verb can not have more than one argument having the same thematic role. However, in some cases – where it is necessary – the co-actor is marked with the `co-` prefix. For example, *sétál valakivel* ‘to walk with someone’ is represented with `AG_coAG-vAl` (`-vAl` stands for the instrumental case-ending).

The argument frames described above could also obtain some special semantic classification which may help in the further refinement of the possible questions. The categories used for this are as follows:

perception (e.g. *to see*)
emotion (e.g. *to be glad*)
sound (e.g. *to resound*)
situation (e.g. *to be pressed for time*)
beginning (e.g. *to be established*)
cognitive (e.g. *to agree*)
communication (e.g. *to inform*)
mathematical (e.g. *to add*)
nonverbal communication (e.g. *to nod*)
self-propelled motion (e.g. *to step*)
financial (e.g. *to transfer*)
destruction (of patient, e.g. *to dry up*)
natural (e.g. *to rain*)
transformation/change (e.g. *to speed up*)
behavior (e.g. *to flirt*)
relation (e.g. *to support*)

Finally, the argument frames also have a polarity value indicating that the given event is positive, negative or neutral for the patient or experiencer.

Figure 5 shows the description of the verbs *sodródik* ‘to drift’, *hull* ‘to fall’ and *zuhan* ‘to drop/plummet’ in the argument frame database. In the first line of the extract, the `PATMV_(PATH)` frame including an optional `PATH` argument that can in turn be expanded as any combination of `SRC`, `DST` and `VIA` arguments, as well as the neutral polarity marked with `@`. refer to each verb below them. Round brackets in the descriptions indicate optionality, square brackets contain a list of examples defining a semantic category.

At the time of writing this paper, the argument frame database contains 1604 verbs with 5994 different argument frames, including the thematic role of each argument. Although frames containing optional arguments (e.g. `olvas AG_(HOW)_(PAT-t)_(REC-nAk)_(TH-rÓl)_(LOC-bAn)` ‘somebody reads (somehow) (something) (to somebody) (about something) (somewhere)’) appear as many seemingly different frames in practice, we obtained these numbers by counting the frames containing optional arguments and possible thematic role variants only once.

5 Identifying the role of adjuncts

An important task is to provide a fine-grained description regarding the role of nominals with case-ending, traditionally referred to as adjuncts. If we approach the question from the case-endings, we could say that the nominal having an inessive case-ending indicates some kind of location and answers the question *Where?*. However, if the question is e.g. *Where did Mary graduate?*, it is a joke to say *In her dream.* The case-endings answering the questions *Hol?* ‘Where?’, *Hová?* ‘Where to?’ and *Honnan?* ‘From where?’[4] are not always used to specify the location, the source or the destination. Depending on the lemma, the suffixed forms may express various temporal relations, modality, etc. For most lexical items that can refer to locations, only one set of the suffixes (e.g. only the inessive *-bAn* ‘in’, illative *-bA* ‘into’, and the elative *-bÓl* ‘from inside’ can be used to express location, source and destination), the rest of the suffixes can only be used as markers of specific oblique arguments of verbs. E.g. while for settlements outside Hungary, the locative relation is always expressed using inessive (e.g. *Londonban* ‘in London’), for most Hungarian settlements, the superessive is used (e.g. *Budapesten* ‘in Budapest’,

[4] In Hungarian, locative/lative/delative case-endings are as follows: the inessive *-bAn* ‘in’, the adessive *-nÁl* ‘at’, the superessive *-On* ‘on’; the illative *-bA* ‘into’, the allative *-hOz* ‘to’, the sublative *-rA* ‘onto’; the elative *-bÓl* ‘from inside’, the ablative *-tÓl* ‘from’, the delative *-rÓl* ‘from the top of’.

```
PATMV_(PATH)
@.
sodródik[IGE] +CHAR_ár-vAl ‘drift[V] +CHAR_tide~with’
hull[IGE] +AG_térd-rA_(CHAR~előtt) +hó +PAT~[haj|könny]-A +PAT@-pusztulás
    ‘fall[V] +AG~to~one’s~knees +snow +PAT’s~hair|tears (die:)+PAT@-decay’
zuhan[IGE] +EXP_álom-bA@.biotünet ‘drop/plummet[V] (fall asleep:)+EXP_into~dream’
```

Figure 5: An extract from the argument frame database.

Table 1: Thematic roles used in the description of argument frames

Annotation	Name	Question regarding the verb	Example
AG	agent	What is AG doing?	**John** has climbed the tree.
CHAR	characterized	What is characteristic of CHAR?	**Expertise** is an advantage.
ATTR	attribute	–	Expertise is **an advantage**.
EXP	experiencer	How does EXP feel? What has EXP perceived?	**John** loves Mary. **John** has seen a swallow.
PAT	patient	What happened to PAT?	John kissed **Mary**.
PATDST	patient-destination	What happened to PATDST? Where did PAT get to?	He painted **the wall** green.
TH	theme	–	John relies **on his intuition**.
ST	stimulus	What effect has ST (on EXP)?	John loves **Mary**. John got frightened of **his shadow**.
CONT	information content	–	John presented **the plan** to Joe.
REC	recipient	–	John presented the plan **to Joe**. **Mary** received a letter.
RES	result	How did RES come into being?	Mary baked **a cake**.
INS	instrument	What is AG using INS for?	John travels to work **by scooter**.
CAU	causer	What did CAU cause? What was the consequence of CAU?	John was late **because of an accident**.
MOT	motivation	–	John is studying to be **an engineer**.
LOC	location	What happened in/at/on... LOC?	John kissed Mary **in the cinema**.
SRC	source, starting point	–	John came **out of the room**. Mary received a letter **from John**.
DST	destination	How did AG/PAT get to DST?	John went **into the room**.
HOW	mode	–	John **deftly** climbed the tree.
ASPECT	aspect	–	John is doing well **financially**.
ACT	action	–	John wants **to work** from home.

lit. ‘on Budapest’). On the other hand, as oblique arguments of the verbs *hisz* ‘believe’ and *múlik* ‘depend’, all nouns take the inessive ‘in’ and the superessive ‘on’ suffixes, respectively. Lemmas can thus be be classified concerning what functional/semantic relation is expressed by the combination of the lemma and each case ending. We identified such classes and defined templates that describe the semantic role of each suffixed form in the template. For all words (lemmas) belonging to the specific class, the template yields the semantics of each suffixed form.

The task can also be formulated as a classification of adverbs. There are, of course, adverbs of place such as *a sarkon* ‘at the corner’ or *bankban* ‘in a bank’, and adverbs of time such as *télen* ‘in winter’, *decemberben* ‘in December’. However, we also find adverbs of duration, e.g. *5 hónapra* ‘for five months’ or a category that we could term as ‘adverbs of garment’ such as *kabátban* ‘wearing a coat’. 31 main categories have been identified, some of which can be divided into several subcategories. Together with the subcategories, we have divided the adjuncts having locative case-endings into 51 classes. To illustrate some of the subcategories, in Table 2, we present lemmas which, when combined with a subset of the locative suffixes, function as adverbs of place. When combined with other locative suffixes, they cannot function as heads of adjunct phrases. In these cases, they can only depend on some head word selecting that specific suffix as the marker of a specific oblique argument.

The first two columns of the table show the main category (in this case, *loc*) and its subcategories (*all*, *ine*, *city-sup*, etc.). This is followed by an

category		example	-bAn (inessive)	-nÁl (adessive)	-On (superessive)
loc	all	*szekrény* ‘wardrobe’	where	where	where
loc	ade	*Microsoft*	in what	where	on what
loc	ine	*állam* ‘state’	where	at what	on what
loc	sup	*címoldal* ‘title page’	in what	at what	where
loc	ine-sup	*könyv* ‘book’	where	at what	where
loc	city-ine	*Altenkirchen*	where	where	on which city
loc	city-sup	*Budapest*	in which city	where	where
loc	country	*Afganisztán* ‘Afghanistan’	where	where	on which country

Table 2: Examples of lexical items that function as heads of locative adverbial phrases when combined with a specific subset of locative case suffixes (cells marked with *where*). With other suffixes, they can only function as oblique arguments of some predicate.

example lemma belonging to the given subcategory, and the best applicable questions for each of the case-endings *-bAn*, *-nÁl* and *-On*, respectively. The questions indicate what role the suffixed word form plays in a sentence.

6 Automatic identification of semi-compositional structures

When identifying idiomatic and semi-compositional verbal constructions, we focused on the behavior of phrases with regard to the relevant question that can be asked about the given phrase. In the case of *döntést hoz* ‘to make a decision’ (lit. ‘to bring a decision’), *What does A bring?* is not an acceptable question. Similarly, *Where does A bring P?* is incorrect regarding the phrase *szóba hoz* ‘to mention’ (lit. ‘to bring into word’).

We have implemented an algorithm for collecting such phrases from a parallel corpus. First, we generated word alignments in the English-Hungarian parallel subcorpus of the OpenSubtitles corpus consisting of 644.5 million tokens (Lison and Tiedemann, 2016) using the fast align tool (Dyer et al., 2013). To alleviate data sparseness problems due to the rich morphology of Hungarian, to improve alignment quality and to facilitate the subsequent light verb construction and idiom identification process, we used a morphosyntactically annotated version of the corpus. The English side was annotated using Stanford tagger (Toutanova et al., 2003) and the morpha lemmatizer (Minnen et al., 2001), while the Hungarian side using the PurePos tagger (Orosz and Novák, 2013) and the Hungarian Humor morphological analyzer (Novák et al., 2016). We further post-processed the output of the tagger/lemmatizer tool combos, to generate an annotation in which each word is represented by one or two tokens on the Hungarian as well as on the English side. The first token is the lemma with the main POS tag attached to it, while the other optional token consists of possible extra morphosyntactic tags (such as tense, case, etc.) if present. We extracted at most 7-token-long parallel phrases from the word-aligned corpus using the phrase extraction algorithm using the grow-diag-final heuristic implemented in the Moses SMT toolkit (Koehn et al., 2007). Of the pairs of phrases extracted, we only kept pairs containing exactly one verb both on the English and on the Hungarian side. For each Hungarian verb from these phrase pairs, we collected all the nouns on the Hungarian side that were aligned with the verb on the English side. For example, in *döntést hoz* ‘to make a decision’ (lit. ‘to bring a decision’), the Hungarian verb is *hoz* ‘to bring’. If it is aligned with the verb *decide* on the English side, the noun *döntés* ‘decision’ is also aligned with this verb, as it does not appear as a separate word on the English side. Note that even if the English translation is *make a decision*, *döntés* on the Hungarian side is also usually aligned with *make* as well as with *decision*, because *make* only corresponds to *hoz* ‘bring’ only in this and a few other similar light verb constructions. In contrast, e.g. in the case of the compositional *táskát hoz* phrase ‘to bring a bag’, *bring* and *bag* are both present on the English side, so these are only aligned with their Hungarian equivalents. Finally, we have normalized and sorted the list of nouns collected for the Hungarian verbs, based on their frequency and their homogeneity regarding the given verb. We cut off the end of the resulting list (where only phrases having a compositional meaning were gathered). The algorithm generated 6531 candidate expressions for 309 verbs.

Originally, we planned to evaluate the algorithm using a list of light verb constructions[5] (LVC’s) (Vincze, 2011) created from the syntacti-

[5] The list contains 1524 items.

cally annotated monolingual Szeged Dependency Treebank (Vincze et al., 2010) and the English-Hungarian SzegedParalell corpus. However, we found that it would be a mistake to consider the original list gold standard. Only 83.6% of the expressions on the Szeged list met our criteria. For the rest, we found that there is nothing odd about asking a question concerning the nominal part of the supposed light verb construction. Finally, we manually evaluated the union of the original list and the entries returned by our algorithm (7538 items altogether). The original list had a recall of 32.2% of the true positives on the union of lists. The precision and recall values for our algorithm turned out to be P=28.6%, and R=84.2%. As a result, we managed to extend the list of Hungarian LVC's and verbal idioms significantly compared to the original Szeged list. We extended our lexicon of argument frames with LVC's and idioms on the list we obtained by manually adding other arguments along with their thematic roles.

7 Aligning argument frames to their occurrences in the corpus

The first step of the algorithm aligning the argument frames to their occurrences in the UD corpus is reading the source lexicon files containing the argument frame descriptions and checking them for syntactic errors. It generates the full argument frame description for each verb by applying an inheritance mechanism adding the argument frames belonging to the verb group to those pertaining only to the given verb.

Explicit and implicit constraints on the form of arguments (suffixes, postpositions etc.) implied by the thematic roles in argument frame descriptions are converted into constraints on features and dependency relations applied to the morphological and syntactic annotations in the UD corpus, respectively. We align the argument frames to the verbs and the heads of phrases attached to them in the corpus using these constraints. The thematic roles *location* (LOC), *destination* (DST) and *source* (SRC) cover noun phrases the head of which is marked with the suffixes and postpositions used to denote location, direction and source (*Where?, Where to?, From where?*) and adverbs having such meaning. The argument frames of many verbs contain the thematic role PATH, which can be replaced by any combination of destination, source and location passed by (VIA). For the sake of readability, case suffixes are represented by their underlying phonological form in the argument frame database. The alignment algorithm converts these descriptions into feature constraints that match the morphosyntactic descriptions in the UD corpus.

Hungarian is a 'pro drop' language, i.e. subject and singular object pronouns generally have an explicit phonological representation in sentences only if they are stressed (e.g. focused or in contrastive topic position). Subjects and objects having no surface realization are recovered in the alignment algorithm by introducing implicit pronouns and assigning the corresponding thematic role to them if the argument frame contains such arguments and there is no explicit subject or object in the given clause. For infinitives, gerunds and participles, verbal argument frames are matched by implicitly binding subjects and objects depending on the type of the construction, while the rest of the arguments are matched in the regular manner. Since objects (NP's marked with accusative case) and infinitives can only occur as arguments, not as adjuncts, frames that do not contain an object/infinitive are discarded if an explicit object/infinitive is attached to the actual verb instance.

The lexically bound nominal element of some of the light verb constructions is an apparently possessive form (annotated as the head of the phrase like in other possessive structures), e.g. *szomszédja nyakára küldte az adóhatóságot* 'he set the tax authority to check up on his neighbour' (lit. 'he sent the tax authority onto his neighbour's neck'). In these constructions, the actual argument that is to be assigned a thematic role is not this semantically empty word, which rather functions like an oblique case ending or postposition, but the possessor (i.e. the neighbor in the previous example). These structures are converted into a form similar to postpositional phrases. The real argument (*his neighbour*) becomes the head in the modified structure, and thus the appropriate thematic role can be directly assigned to it.

When multiple frames match the specific verb instance, the most specific frame is selected: matching light verb or idiomatic constructions are ranked high, otherwise match candidates are ranked by the length of the matching argument list.

Figure 6 shows how the original annotation of a sentence in the original UD corpus was corrected

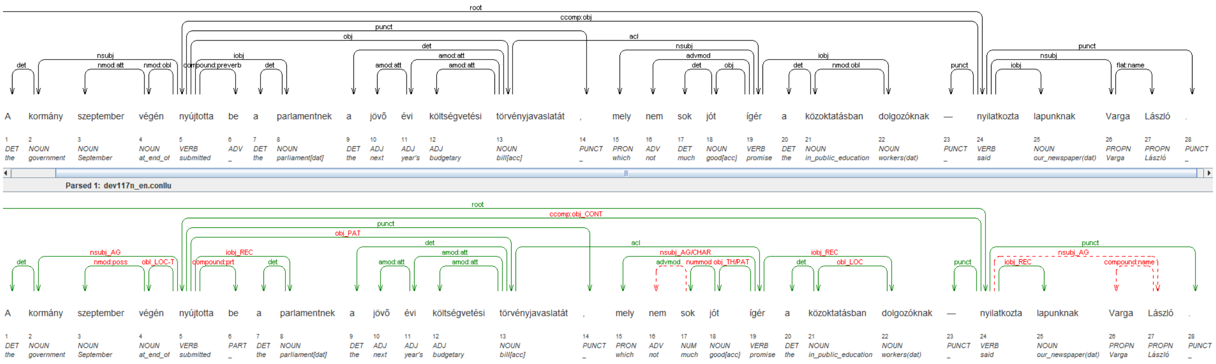

Figure 6: The result of automatic correction and assignment of thematic roles to heads of phrases and clauses attached to verbs in a sentence in the Hungarian UD corpus: 'The government submitted its bill for next year's budget to the parliament at the end of September: it bodes no good for those working in the public education sector, László Varga told our newspaper'

and extended with thematic role labels by the the adjunct and verbal argument frame matching algorithm.

8 Conclusion

Within the scope of the ongoing research presented in this article, we have created a semantically rich corpus annotation for Hungarian using the Hungarian UD subcorpus as a starting point. Our future tasks include integrating the argument frames of nominal predicates and manual checking of the generated thematic role annotation. We may also need to finetune the interaction of lexically-driven automatic adjunct annotation and verbal argument frame alignment, and the ranking of matching argument frames. Furthermore, arguments annotated with thematic roles will need to be semantically further subcategorized to be able to generate the right questions. We have done this, but we have not yet integrated this information with the rest of the annotation. We will also need to extend our corpus (converting and correcting parts of the Szeged Dependency Treebank not included in the Hungarian UD corpus) to provide enough training material for a semantic parser.

Acknowledgments

This research has been implemented with support provided by grants FK 125217 and PD 125216 of the National Research, Development and Innovation Office of Hungary financed under the FK 17 and PD 17 funding schemes as well as by grant ÚNKP-18–3-III-PPKE-26 of the New National Excellence Program of the Ministry of Human Resources. We would like to thank Kinga Fegyó and Ivett Bognár for their participation in creating verbal argument frames and assigning thematic roles.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Sylvain Kahane, Marine Courtin, and Kim Gerdes. 2017. Multi-word annotation in syntactic treebanks - propositions for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 181–189, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*. The Association for Computer Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Guido Minnen, John Carroll, and Darren Pearce. 2001.

Applied morphological processing of English. *Nat. Lang. Eng.*, 7(3):207–223.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Attila Novák and Borbála Novák. 2018. POS, ANA and LEM: Word embeddings built from annotated corpora perform better. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2018*, Hanoi, Vietnam. Springer International Publishing, Cham.

Attila Novák, Borbála Siklósi, and Csaba Oravecz. 2016. A new integrated open-source morphological analyzer for Hungarian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Bálint Sass, Tamás Váradi, Júlia Pajzs, and Margit Kiss. 2010. *Magyar igei szerkezetek: a leggyakoribb vonzatok és szókapcsolatok szótára [Hungarian verbal constructions: a dictionary of the most frequent arguments and phrases]*. A magyar nyelv kézikönyvei. Tinta Könyvkiadó.

Borbála Siklósi. 2016. Using embedding models for lexical categorization in morphologically rich languages. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016*, pages 115–126, Konya, Turkey. Springer International Publishing, Cham.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Morristown, NJ, USA. Association for Computational Linguistics.

Veronika Vincze. 2011. *Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses*. Ph.D. thesis.

Veronika Vincze, Richárd Farkas, Zsolt Szántó, and Katalin Ilona Simkó. 2017. Universal Dependencies and morphology for Hungarian – and on the price of universality. In *Proceedings of EACL 2017*, pages 356–365.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).