


Cross-lingual transfer of knowledge in distributional language models: Experiments in Hungarian

ATTILA NOVÁK*  and BORBÁLA NOVÁK

Pázmány Péter Catholic University, Faculty of Information Technology and Bionics, Hungary

Received: April 22, 2022 • Accepted: October 1, 2022

Published online: November 22, 2022

© 2022 The Author(s)



ABSTRACT

In this paper, we argue that the very convincing performance of recent deep-neural-model-based NLP applications has demonstrated that the distributionalist approach to language description has proven to be more successful than the earlier subtle rule-based models created by the generative school. The now ubiquitous neural models can naturally handle ambiguity and achieve human-like linguistic performance with most of their training consisting only of noisy raw linguistic data without any multimodal grounding or external supervision refuting Chomsky's argument that some generic neural architecture cannot arrive at the linguistic performance exhibited by humans given the limited input available to children. In addition, we demonstrate in experiments with Hungarian as the target language that the shared internal representations in multilingually trained versions of these models make them able to transfer specific linguistic skills, including structured annotation skills, from one language to another remarkably efficiently.

KEYWORDS

distributional vs. generative models of language, zero-shot cross-lingual knowledge transfer, multilingual contextual neural language models, meaning representation parsing, named entity recognition

1. INTRODUCTION

Noam Chomsky theorized that the human brain contains a special unique mechanism he termed 'the language acquisition device' (Shatz 2007), which he thought to be quite separate and

* Corresponding author. E-mail: novak.attila@itk.ppke.hu

different from other cognitive faculties. He imagined this faculty as an innate generative universal grammar that has a number of parameters, which are set during the language acquisition process in a child's brain. Chomsky's main argument for his theory was what he perceived as a poverty of stimulus concerning linguistic input during language acquisition. Although quite influential, the idea was debated by a body of research in applied linguistics and neuroscience (Pullum & Scholz 2002). Chomsky's formulation of the problem ended up for the most part as a dogma of a school of steadily dwindling orthodox followers. While it is a trivial fact, that humans are endowed with some sort of predisposition toward language learning not exhibited by most other species,¹ *the substantive issue is whether a full description of that predisposition incorporates anything that entails specific contingent facts about natural languages*. This is what Pullum & Scholz (2002) argued against quite convincingly even without supporting the idea that domain-unspecialized algorithms for knowledge acquisition can suffice for learning natural languages given children's experience.

Predating Chomsky, originating in the work of Leonard Bloomfield and Morris Swadesh, the insight that the distribution of words or morphs is the most important source of information of all grammatical knowledge was finally formalized in Zellig Harris' work (Harris 1954). Applied as a research procedure from the 1930s, it was the task of the linguist field worker to explore distributions and categorize lexical items accordingly, which was a very labor-intensive task. Later, owing mostly to the influence of Chomsky, the generative school of linguistics largely abandoned the idea of inducing grammar from raw linguistic data. Instead, grammars were created manually.

We will attempt to demonstrate below with a very concise review of some of the key points and paradigms in the earlier and more recent history of computational linguistics and natural language processing that since the development of an appropriate type of parallel hardware made the efficient and automatic creation of rich distributional models of language feasible, these models have demonstrated remarkable linguistic abilities that surpass those ever attained by models based on the generative grammar tradition by a huge margin.

We will also cite literature showing that the internal representations of distributional models of different natural languages have been found to be highly isomorphic, i.e. these representations can be mapped to each other remarkably efficiently. Moreover, multilingual models, trained on linguistic data representing several languages, consist of shared internal representations where an aligned mapping of representations automatically emerges during the training process. These shared representations make it possible for these models to apply specific linguistic skills they attained for some language to other languages that the underlying model covers.

We present some case studies that feature Hungarian as the target language. While the feasibility of linguistic transfer for end-to-end tasks, like summarization or translation, has also been demonstrated, and is the main driving force behind these models, here we present results on structured prediction tasks where the model is expected to output some sort of linguistic annotation.

¹One can also argue that the following biological factors may be relevant to humans having exceptional cognitive faculties and being linguistic creatures: a) the human vocal tract is adapted to sophisticated and quasi-continuous vocalization, b) humans have a high-above-average neural capacity, with a large number of neurons packed into a small cerebral cortex (high neural density in the neocortex), and c) a protracted childhood with high neural plasticity lengthening the period of physiological integration and adaptation.



One experiment demonstrates the feasibility of zero-shot parsing with a dependency-based meaning representation output. In that experiment, we find that shared typological features are needed for the model to generate certain types of annotation: e.g. a model trained on a language not featuring certain types of zero elements (e.g. subject pro drop) cannot properly parse constructions containing such elements in another language.

In other experiments, we perform zero-shot transfer of named entity recognition, i.e. identification of names and the type of their referents.

2. GENERATIVE GRAMMAR VS. DISTRIBUTIONALISM – THE END OF AN ERA

Chomsky's ideas of grammar and especially syntax also influenced computational linguistics and language technology for quite some time. However, even during the era when rule-based approaches dominated natural language processing, it was not Chomsky-style transformational, minimalist, etc. grammars that actually worked. Nevertheless, the more-or-less well functioning alternatives, like HPSG (Pollard & Sag 1994) or LFG (Dalrymple 2001), that had working implementations (e.g. the Alpino parser for Dutch (Bouma et al. 2000), based on an HPSG grammar, or the LFG-based Xerox Linguistic Environment (XLE, Kaplan et al. 2004)) can be considered members of a family of generative grammars both in the sense that they are formally explicit models and that they are characterized by a recursive productivity.

Chomsky mostly ignored semantics, but computational grammars were not feasible without some implementation of semantics that could be used to tackle the problem of immense ambiguities posed by possible alternative grammatically feasible analyses. However, an effective computational semantics seemed for a long time to be a Holy Grail.

2.1. The advent of statistical machine learning paradigms

As computing power and available memory of average computers increased, the application of machine learning paradigms based on statistical methods became dominant partly supplementing,² but in most cases rather displacing rule-based solutions, and, while some sort of grammatical annotation was still an ingredient of many of these models, the annotation did not come from some hand-crafted grammar, but it was learned during model training from gold-standard expert annotation of treebanks. Statistical parsers (Klein & Manning 2003) and other statistical NLP models, like those applied in the statistical machine translation paradigm (Koehn 2009), were also much less fragile than rule-based ones: while the latter failed to produce any output for a significant portion of actually attested linguistic data (and generated too many possible outputs for the rest), the former naturally came up with a 'best' or 'most probable' solution for any input (which, however, was not necessarily correct or even well-formed). These systems, like humans, were no longer overwhelmed by ambiguity.

²E.g. in the Dutch Alpino parser, statistics from the Alpino treebank were used to rank and select output from the parser providing information about 'what makes sense'.



Meanwhile the Chomskyan tradition also lost ground to alternative formalisms owing to the relative lack of expressive power of hardcore Chomskyan syntax: functional relations of the age-old dependency syntax tradition seemed to be more usable in computational models than phrase-structure-based constituency. While the latter were almost exclusively used for decades, the parsing scene later became dominated by dependency parsers partly due to the success of the unified effort of the Universal Dependencies project.³

2.2. Neural models of machine learning

Technical evolution reached a point at the end of the 2000s where artificial neural networks (ANNs) started to achieve state-of-the-art results in some areas previously dominated by other machine learning paradigms. With the foundations set between the 1940s and the 60s (Hebb 1949; Rosenblatt 1958), functioning computational implementations of ANNs had been available since the 1970s (Werbos 1994), but they had limited practical use for a long time. The main reason for this was that training of and inference by these models could only be emulated on by today's standards snail-pace sequential hardware that also had very serious memory limitations. Another reason was the vanishing gradient problem,⁴ and a lack of massive amounts of training data needed to train complex networks.

The basic idea, however, remained the same since the beginnings: artificial neural networks consist of units attached to each other via connections of variable strength or weight, the units are activated by some non-linear activation function of their input, and it is the connection weights that largely determine the functioning of the network. Neural networks are trained in a supervised manner: connection weights are updated during training by backpropagating prediction error of the network that can be calculated as the difference of the network's prediction and the expected output.⁵

Appearance and increasing availability of powerful parallel computing architectures (Graphical Processing Units, GPUs),⁶ a constant increase in computer power and the amount of data available for training combined with a new neural training paradigm, which became known as 'deep learning' that could feasibly tackle the problem of error backpropagation during the training of deep multi-layer neural networks turned out to be a game changer not only in image processing, speech and language technology, but in practically all areas of information technology. Like a few times before, speech technology pioneered the first deep neural networks, but soon

³<https://universaldependencies.org/>.

⁴The gradient (error signal) used to update the weights of connections in a multi-layer network during training decreases exponentially with the distance of a layer from the output layer, so layers close to the input may train very slowly.

⁵Neural networks are also widely used for representation learning, which is an essentially unsupervised training task. In these training scenarios no additional ground truth output annotation is provided, but the training of the network is still supervised in a strictly technical sense: the input itself is used as ground-truth supervision data to estimate prediction error both in the case of training autoencoders, which create compressed representations from which the autoencoder can more-or-less faithfully reconstruct the input, and in the case of neural language modeling, where the model is trained to reconstruct (some aspects of) the input from a manipulated/incomplete input representation.

⁶Later also special Tensor Processing Units (TPUs) were developed specifically for the training and application of deep neural networks, and integrated in supercomputers called TPU pods containing several (in an extreme case, several thousand) such units.



computer vision and natural language processing followed as domains of application. Generations of different neural network architectures followed each other resulting in new success stories at different modalities of input. Although different neural architectures turned out to be the most effective for processing visual and sequential input, this does not mean that e.g. convolutional neural networks (CNNs), most suitable for image processing, would not be successfully applicable to linguistic input: although now not state-of-the-art, they can be fair ‘cheaper’ alternatives to currently top-performing compute-and-memory-hungry transformer-based models.

2.2.1. Attention is all we need. For sequential input, a generic attention mechanism turned out to be crucial in that neural networks managed to surpass the performance of all previous machine learning paradigms. Within the NLP community, it was machine translation experts who kept coming up with breakthrough neural models, all important mile-stone models involving some new version of an attention mechanism. The introduction of attention to recurrent neural networks in 2014 was the invention that put the first nails into the coffin of the statistical machine translation paradigm (Bahdanau et al. 2014). Just three years later, the fully-attention-based transformer model (Vaswani et al. 2017) again surpassed the performance of recurrent models like long short-term memories (LSTMs) or gated recurrent units (GRUs) by a large margin.

2.2.2. An urge to predict the unknown. A generic training mechanism of simple prediction of expected or missing information also played a key role in the success story of deep (and shallow) neural networks. The currently most successful training paradigm solves the problem of limited training data for specific tasks by using a set of simple prediction tasks of reconstructing missing (masked) or identifying corrupted parts of the input based on the context to pretrain the model first. This pretraining paradigm uses only plain-text training data lacking any extra annotation. Such data is abundant because crowds of the digital population continuously generate vast amounts of text published online, which can be collected automatically by web crawler programs.

The first such successful models were obtained by training extremely simple neural networks having only a single layer of internal representation on simple and cheap sequential hardware resulting in representations, called word embeddings, constituted by the connection weights between the internal projection layer and the input or output layer of the network (Mikolov, Chen et al. 2013). These embedding models provide a representation of words in the form of real valued vectors of relatively small dimensions (compared to traditional vector space models, which relied on word co-occurrence counts). Despite their extreme simplicity, these models capture a surprising amount of semantic, syntactic, and morphological ‘knowledge’ concerning the words in the model. Embedding models are simple but efficient realizations of distributional semantics. This early result was thus also a breakthrough in computational semantics, where traditional logic-based models struggled with representing meaning at least as much as grammar-based models struggled with modeling linguistic form. The two traditional representations of form and meaning had a hard time solving the problem of constraining each other to find representations that actually make sense.

A weakness of early very simple static word embedding models was that they were not capable of handling homonymy or polysemy creating a unified representation of each word



averaging all its senses blending the sense representations in proportion to their frequency in the training corpus. This is demonstrated by a 2-dimensional rendering of nearest neighbors of the polysemous/homonymous Hungarian words *egér* ‘mouse (animal/computer peripheral)’ and *vár_V* ‘wait/expect/stand by/welcome’ vs. *vár_N* ‘castle’ in word embedding models created from a large morphologically analyzed Hungarian corpus as shown in Figure 1 (Siklósi 2018; Novák & Novák 2022b).

Today’s much more complex deep neural networks (typically transformers) build contextual language models during the training, in which the representations pertaining to individual elements of the input reflect the actual sense of the item in the very specific linguistic context. As mentioned before, the most widespread training paradigm consists of pre-training deep neural models using large raw corpora to perform simple tasks like the prediction of masked items in the input based on words in the context (masked language modeling, MLM), and subsequently fine-tuning these models for specific tasks like sentiment analysis, question answering, natural language inference, etc. Pre-training the models on raw corpora makes it possible for the models to learn ‘language’ from more data rather than having to pick it up from the much more limited amount of annotated training data available for the specific tasks. The first transformer-based model using masked language model pre-training was BERT (Devlin et al. 2019). Later similar models varied details of the architecture and/or the training procedure resulting in various improvements. XLNet (Yang et al. 2019) uses permutations rather than masking. ELECTRA (Clark et al. 2020) is trained using a separate network that corrupts the input used for training the model by replacing some tokens with other tokens having a similar distribution, and the task of the pre-trained network is just to identify the corrupted tokens rather than to reconstruct them. This simplified training task results in significantly faster convergence of the model during training.

When these pre-trained deep neural models are probed for the presence of specific linguistic capacities, they show remarkable abilities of making subtle distinctions concerning various aspects of the input. All this in spite of the fact that these models are only exposed to plain (and to some degree dirty) linguistic data during training without any multimodal contextual clues a child is exposed to when learning language.

2.2.3. The emergence of linguistic competence. Although these artificial neural models cannot be claimed to mimic the operations of natural neural networks in the human brain (despite biological analogies inspiring the original neuron model and some of the network architectures, development of the models was later driven more by a motivation to optimize performance rather than by faithfulness to biological analogies), the success of this training paradigm may also be taken as an argument against Chomsky’s hypothesis that a poverty of stimulus would block a generic learning hardware from picking up language. These models are not pre-wired to learn ‘language’, nevertheless, they succeed without formal supervised training involving negative evidence, without being exposed to multimodal clues, driven only by a generic ‘motivation’ to learn to predict phenomena observable in the environment, which in their case is limited to raw unanchored linguistic data.

Prediction is a primary capacity and function of a nervous system. Self-correction of prediction errors concerning raw language data alone drives these systems to converge to internal representation states that can be used to perform diverse specific tasks concerning language.



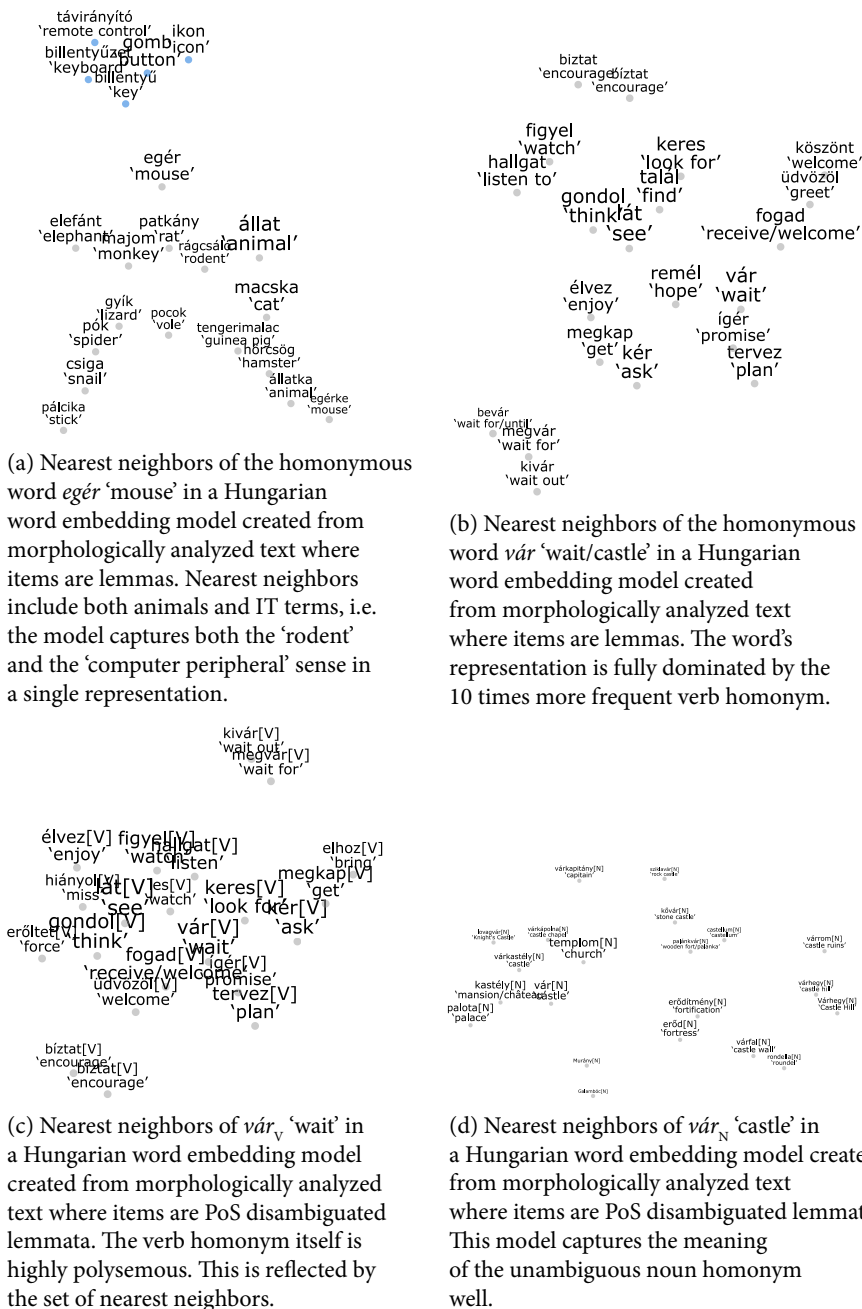


Figure 1. 2-dimensional projections of nearest neighbors of homonymous/polysemous Hungarian words in various Hungarian word embedding models. The images were generated using the t-SNE algorithm from 300-dimensional representations



Although the models need to be separately fine-tuned to actually perform tasks different from simple prediction of what is missing or corrupted, often a quite limited amount of data is enough to get them perform surprisingly well, even if most of the model is frozen (not allowed to learn) during fine-tuning.

One may still argue that the *amount* of raw linguistic training data a child is typically exposed to would not be enough for these artificial neural models to learn a representation performing on par with a child. This is probably true. And specialized neural subnetworks underlie the human linguistic capacity⁷ similarly to the manner some brain areas like the fusiform face area in the fusiform gyrus in the temporal lobe or the occipital face area in the inferior occipital gyrus apparently play a crucial role in human face perception (Haxby et al. 2000; Zhen et al. 2013). Humans, like other primates, also have pre-wired spider and snake recognition subnetworks (Le et al. 2013; Rakison & Derringer 2008), as these are essential for tropical primates' survival. For modern humans living in environments lacking these specific types of natural threats, these are just a less useful part of our biological heritage resulting in bothering phobias rather than providing a skill essential for survival. These biological facts are, however, immaterial to the generativism vs. connectionism debate, the context of which the original Chomskian argument was formulated in. Even if the artificial neural networks currently used to handle face recognition or language-related tasks have a structure quite different from the ones present in the biological networks involved in performing these tasks and even if these artificial networks take significantly more data to train than their specialized biological counterparts, the indispensability or utility of a specific generative theory of human grammars does not follow from these facts more than that of a generative theory of face or spider or snake images.

2.2.4. What do these models 'know' about language? As an example of emerging linguistic representations, it was shown that syntactic dependency relations between tokens in text can be extracted from these models simply by mapping representations of tokens at specific layers of the model using a linear transformation (i.e. projecting their representation to a 32–256 dimensional 'syntax hyperplane') and assuming a dependency link between nearest neighbors in the mapped representation (Hewitt & Manning 2019). Gold standard dependency annotation of dependency relations is needed to find the optimal mapping (i.e. for finding the hyperplane to look at), but the representation itself is already there without explicit training for syntactic analysis, and it is fairly accurate albeit not perfect. Note that the model does not only quite efficiently predict nearest neighbors, but distances between all word pairs in the sentence.

Figure 2 shows dependency relations within an English sentence extracted from a contextual model using a linear 'syntax hyperplane' structural probe (here a linear transformation projecting the representation to a hyperplane of 128 dimensions) based on representations in layer 16 of the English contextual language model BERT-Large. Note that this simple projection model does not identify the direction of the attachment. One metric that can be used to evaluate the 'parses' the models yields is UAS (Undirected Unlabeled Attachment Score), which was found to be 0.817 for layer 16 in the BERT-Large English language model.

⁷Broca's and Wernicke's areas were identified already in the 19th century, but their exact function, structure and the additional brain structures relevant to language understanding and production are still largely unexplored.



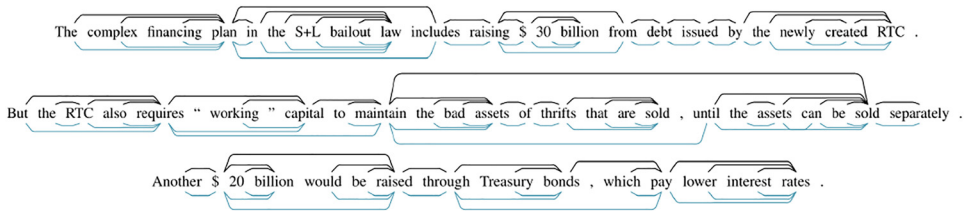


Figure 2. Dependency relations extracted from layer 16 representations of English sentences from the Penn Treebank in BERT-Large using a linear ‘syntax hyperplane’ nearest neighbors structural probe. Gold dependencies at the top, system output at the bottom. The results above are from [Hewitt & Manning \(2019\)](#)

A recent critique of the method ([Maudslay & Cotterell 2021](#)) has demonstrated that performance of the transformer models perceived by the syntactic probe degrades significantly when tested on ‘Jabberwocky’ input, i.e. on text where *all content words* are replaced by plausible non-words with matching part of speech,⁸ which the authors interpret as the models relying heavily on distributional properties of content words rather than embodying a strong model of syntax.⁹

2-D visualization of vector representation of tokens at middle layers of transformer language models along with their gold dependency labels in multilingual models trained on text in many different languages has showed a fairly consistent mapping of nodes with identical labels across languages ([Chi et al. 2020](#)).¹⁰

In one research ([Glavas & Vulic 2020](#)), explicit syntactic knowledge from human-curated treebanks was injected into pre-trained transformer models applying intermediate parsing training (IPT) before further finetuning them for language understanding (sequence classification and multiple choice classification) tasks. However, this knowledge was found to be redundant w.r.t. the structural language knowledge transformers obtain through masked language model pretraining: further training on gold standard parse trees was not found to consistently improve their performance on tasks where earlier top performing systems relied on parsers trained in a supervised manner. Another research ([Sachan et al. 2021](#)) found improvements in semantic role labeling (SRL) and relation extraction performance only when combining contextual language models with gold parses (also at inference time), i.e. this can be considered another negative result. [Bai et al. \(2021\)](#), in contrast, report consistent improvements in performance on linguistic benchmarks tasks in GLUE ([Wang et al. 2018](#)) after tweaking attention heads in transformer models based on parser output in their Syntax-BERT models as part of intermediate pre-training.

⁸Plausible non-words: word forms that conform English phonotactic and morphotactic constraints.

⁹Note, that in Carroll’s *Jabberwocky*, far from all content words are non-words, and many of the ones that are, are onomatopoeic or resemble something that makes sense.

¹⁰The multilingual UD corpus provides fairly consistent dependency annotation across languages: this made such an evaluation possible.



2.3. Multilingual and language-agnostic representations

For simple word embedding models, a number of methods were shown to be effective to map the representations across languages or to create cross-lingual word embeddings (CLWEs). A CLWE is a shared cross-lingual word vector space where words with similar meanings obtain similar vectors regardless of their actual language. Most mapping methods are projection-based. The projection is achieved by learning a piecewise linear transformation based on a seed dictionary (in earlier implementations containing several thousand items), through which a monolingual WE space can be mapped to another monolingual space (Mikolov, Le & Sutskever 2013). The transformation maps each word vector in the source language space to a point in the vicinity of the vector of its translation in the target language space. Later implementations worked with just a few hundred cognate seed pairs, identical strings or simply numerals (weak supervision). Then even the need for a seed dictionary was dropped using adversarial training. Although a seed dictionary needs to be induced also in the fully unsupervised mapping methods, this is achieved based on a heuristic that translations have similar similarity distributions across languages.

Figure 3 shows nearest neighbors of the projections of some lexical items presented in Section 2.2.2 *vár_V* ‘wait/expect/stand by/welcome’ and *vár_N* ‘castle’ in an English word embedding model created from morphologically analyzed text. The mapping between the models was performed following the method of Mikolov, Le & Sutskever (2013) as described in Novák & Novák (2018). Comparing these with snapshots of the Hungarian model in Figure 1b shows that the two models are quasi-isomorphic: i.e. they have similar lexical items in the neighborhood of these lexemes.

There is a significant body of research literature describing work concerning cross-lingual transfer using deep-neural-network-based models. The machine translation community pioneered the first multilingual models (Firat et al. 2016; Johnson et al. 2017). It was

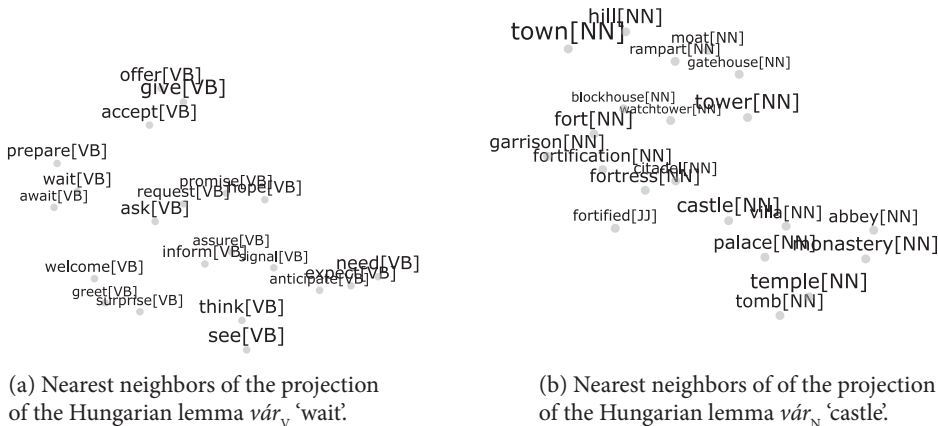


Figure 3. Nearest neighbors of the projection of Hungarian words in an English word embedding model created from morphologically analyzed text where items are PoS disambiguated lemmata. The images were generated using the t-SNE algorithm



discovered that, with neural machine translation models, it is possible to improve performance in specific lower-resource languages and language-pairs, and even to translate between language pairs for which the model had not been trained at all by training the encoder and the decoder of the model in a shared manner on multiple languages. This resource-sharing also made direct translation between all of the represented languages possible, and resulted in savings in resources concerning both training, storage and inference, i.e. using the model in production.

Multilingual training turned out to be fruitful not only in the domain of machine translation. Multilingual pre-training of contextual language models like multilingual BERT (Devlin et al. 2019) and XLM-RoBERTa (Conneau, Khandelwal et al. 2020) made cross-lingual knowledge transfer efficient for other NLP tasks as well. These models were trained using the same training algorithm as their English or other monolingual counterparts but on massively multilingual corpora consisting of more than 100 languages. These models have been used to train massively multilingual syntactic dependency parsers (Kondratyuk & Straka 2019), zero-shot named entity recognizers (Wu et al. 2020), etc., with even specific multilingual benchmarks prepared for testing the cross-lingual generalization capability of models on various tasks such as sentence-pair classification, structured prediction (POS tagging, NER), question answering, natural language inference and sentence retrieval (Hu et al. 2020).

Conneau, Wu et al. (2020) and Dufter & Schütze (2020) examined the conditions necessary for the emergence of multilingual shared representations in contextual language models, and found that, contrary to former expectations, neither shared vocabulary nor a high similarity of domain is necessary for an effective cross-lingual transfer: parameter sharing in the top layers leads to the emergence of multilingual representations, i.e. the model must receive enough data during training to begin to share parameters across the languages represented. Conneau, Wu et al. (2020) also showed that monolingual contextual representations can be aligned to each other quite effectively, like static word embeddings.

3. CROSS-LINGUAL TRANSFER OF ANNOTATION MODELS: TARGETING HUNGARIAN

In this and the following section, we present two experiments where we used multilingual models to perform zero-shot structured prediction tasks for Hungarian. As mentioned in Section 2.3, the underlying multilingual models were trained on massively multilingual corpora, which included Hungarian among the more than 100 languages they cover.

Before presenting these experiments, we show how syntactic structures emerge in multilingual models using the linear ‘syntax hyperplane’ structural probe described in Section 2.2.4. Figure 4 shows dependency relations in English sentences extracted from layer 6 of the multilingual BERT model. The results are somewhat inferior to those obtained by Hewitt & Manning (2019) from layer 16 of the larger monolingual English BERT-Large model (UAS: 0.783 vs. 0.817). Here a more coarse-grained 32-dimensional syntax hyperplane was used, and the more restricted capacity of the multilingual BERT model (12 layers, 110 million vs. 24 layers and 340 million parameters of BERT-Large) seems also to have been overwhelmed by the 104 languages it was trained on. As we will see later, another multilingual model, XML-RoBERTa performs significantly better in many tasks than multilingual BERT.





Figure 4. Dependency relations extracted from layer 6 representations of English sentences from Universal Dependencies English Web Treebank 2.0 in multilingual BERT using a 32-rank linear ‘syntax hyperplane’ nearest neighbors structural probe. Gold dependencies at the top, system output at the bottom, mismatches in red

In Figure 5, we show dependency relations assigned by the model to Hungarian sentences from the Szeged Dependency Treebank. Here we also probed layer 6 of multilingual BERT, however, the 32-dimensional syntax hyperplane was identified using data concatenated from 11 UD corpora not including Hungarian: the training portion of one UD corpus each for Arabic, Czech, German, English, Spanish, Farsi, Finnish, French, Indonesian, Latvian and Chinese following the work presented in Chi et al. (2020). The UAS obtained by this model for Hungarian is 0.693.





Figure 5. Dependency relations extracted from layer 6 representations of Hungarian sentences from UD version of the Szeged Dependency Treebank in multilingual BERT using a 32-rank linear ‘syntax hyperplane’ nearest neighbors structural probe. Gold dependencies at the top, system output at the bottom, mismatches in red

3.1. Dependency-based meaning representation parsing

Although the line of research on computational dependency grammars was for a long time limited to tree representations due to implementation concerns,¹¹ many deeper meaning representations have been introduced that use acyclic structures (and usually also variables).

¹¹This is no longer so since enhanced dependencies have been introduced, although for most languages covered by the Universal Dependencies project, no, or only a very limited and approximate machine-generated enhanced annotation exists.

Annotated resources (meaning banks) have been created for many of these formalisms with English being the most prominent object language, although for many formalisms, other languages have also been covered.

Some of these meaning banks contain rich syntactic-semantic annotation based on general theories of grammar. With certain simplifications, semantic annotations in these resources can be converted to graph representations, which generic parsing algorithms can be trained to generate.

One such resource, Elementary Dependency Structures (EDS), is based on English Resource Grammar (Flickinger et al. 2017) aka English Resource Semantics (ERS) (Flickinger et al. 2014) annotation. The underlying linguistic theory is Head-Driven Phrase Structure Grammar (HPSG, Pollard & Sag 1994) with Minimal Recursion Semantics (MRS, Copestake et al. 2005). In EDS, ERS representations are turned into variable-free semantic dependency graphs consisting of labeled graph nodes representing logical predications and edges representing labeled argument positions. The conversion from ERS to EDS discards information on semantic scope. The nodes are anchored to spans of the input string. Figure 6 shows the EDS representation of an English sentence.

Abstract Meaning Representation (AMR, Banarescu et al. 2013) features graphs comparable to EDS, but with more abstract predication labels due to application of extensive lexical decomposition and normalization towards verbal senses, e.g. representing *similar* as the verbal sense ‘resemble’. In contrast to EDS (and some other models), AMR nodes are not explicitly anchored to spans of the surface form, because when semantic decomposition and normalization is applied, anchoring of individual component nodes becomes more than non-trivial. Figure 7 shows the AMR representation of the same English sentence. Normalization to verbal senses may also result in the inversion of canonical argument relations, e.g. *researcher* is represented as the person who is ARG0 of the verb *research*. The same applies to the subject of the relative clause in the example sentence. Negation is represented as the feature *polarity* -.

Less ambitious in the coverage of aspects of meaning, Universal Conceptual Cognitive Annotation (UCCA, Abend & Rappoport 2013) is an abstract annotation featuring only purely semantic categories and structure. The foundational layer of UCCA consists of a very basic set of semantic categories like Process, Argument, State, “Adverb” (modifier), etc., which are used as

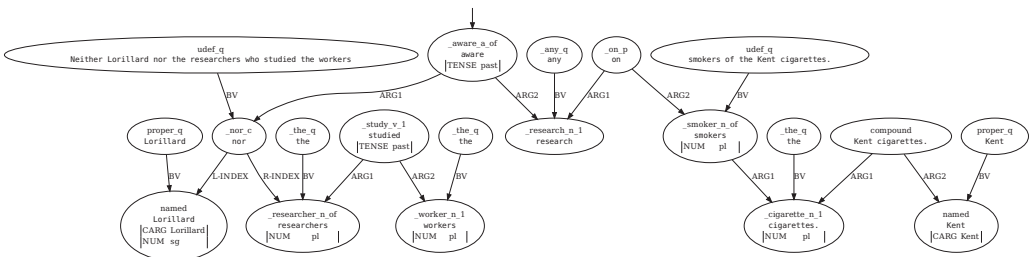


Figure 6. EDS representation of the English sentence *Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes*. The arcs denote either (untyped) argument relations between the nodes (ARG1, ARG2), bound variables (BV), or mark the items conjoined by conjunctions (L-INDEX, R-INDEX)



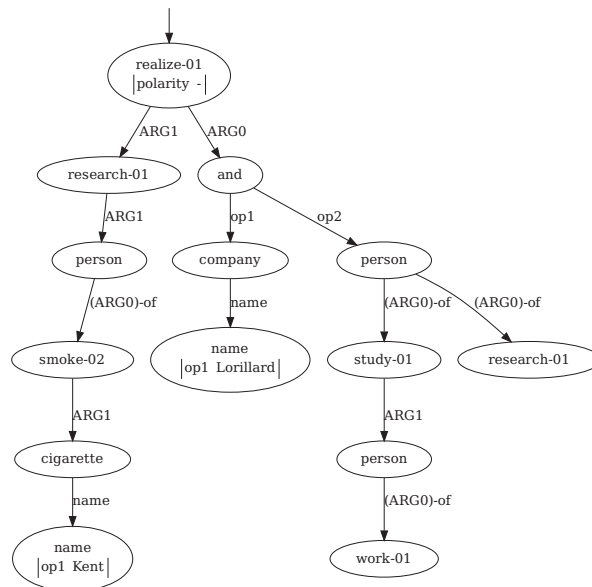


Figure 7. AMR representation of the English sentence *Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes*

labels on edges linking unlabeled nodes representing semantic units and surface word forms. See Figure 8 for the UCCA representation of our sample English sentence.

Prague Tectogrammatical Graphs (PTG, Zeman & Hajic 2020), is another graph-based meaning representation formalism derived from the Prague Functional Generative Description (FGD, Sgall et al. 1986), a dependency-based formalism, retaining a subset of Prague Tectogrammatical annotation. As Figure 9 suggests, PTG is a more syntax-oriented formalism with typed dependency links.

Finally, Discourse Representation Graphs (DRG, Abzianidze et al. 2020) is a graph encoding of Discourse Representation Structures (DRS), the meaning representations at the core of Discourse Representation Theory (DRT, Kamp & Reyle 1993). This model handles many challenging semantic phenomena from quantifiers to presupposition accommodation, and discourse structure. Figure 10 presents the DRG representation of a (different) English sentence.

With the exception of EDS, the other formalisms mentioned cover other languages besides English: there exists annotated data in Czech for PTG, in German and French for UCCA, in German for DRG, and in Chinese for AMR.

There is a parser, called PERIN, which is capable of turning raw text into any of the flavors of meaning representation graphs mentioned above (Samuel & Straka 2020). It was developed at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague. The parser is available at the ÚFAL GitHub repo¹² including the

¹²<https://github.com/ufal/perin>.



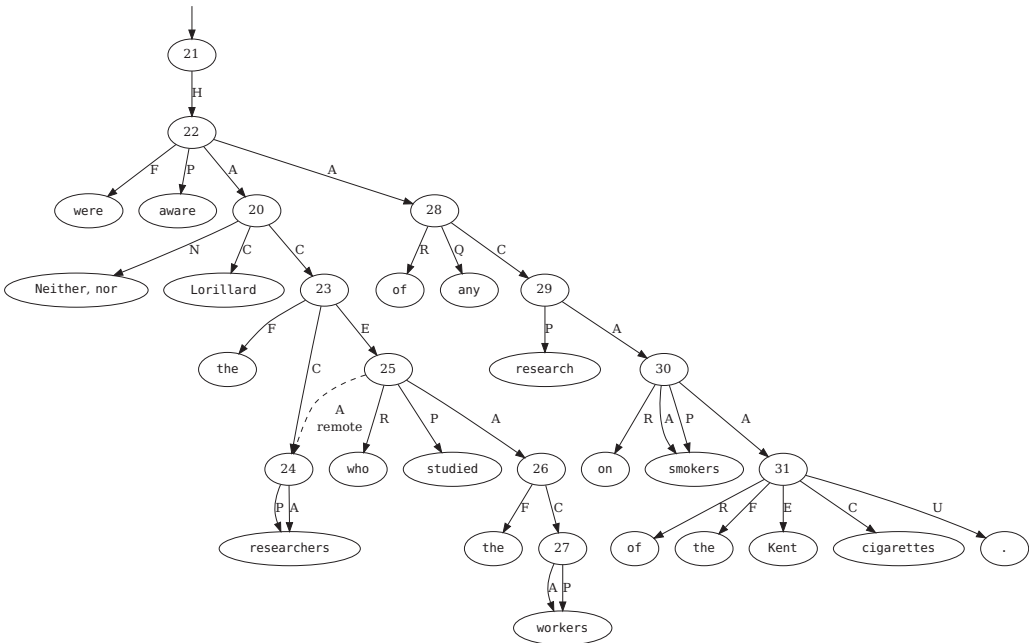


Figure 8. UCCA representation of the English sentence *Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes*. A: argument, P: process, S: state, C: center (main element of a non-scene unit), E: elaborator (of a center), N: connector (of centers), Q: quantifier, R: relator (usu. preposition), F: function (e.g. the copula), D: adverb (modifier), H: parallel scene, L: linker (of parallel scenes)

trained parser models. There is also a link to an Interactive demo on Google Colab, which makes testing the models on various inputs easy. Positive subjective impressions of the performance of the English and especially the Czech PTG model of the parser on Hungarian input prompted us to perform the experiment described here evaluating the zero-shot cross-lingual performance of the model. Of the models available, we selected PTG, because

- the categories/concepts it operates with was immediately familiar,
- the annotation it generated seemed reasonable and detailed,
- the non-English model covers Czech, a language sharing many typological features with Hungarian (rich morphology, relatively free word order, pro drop, etc.),
- the model was trained on a sizable 740k-token corpus,
- a rather detailed annotation manual (Mikulová et al. 2006) of the underlying Prague tectogrammatical annotation is available in English, and
- performance of PERIN on the Czech PTG data is relatively high as reported in Samuel & Straka (2020).

Concerning the other formalisms mentioned above, we had the following impressions, further motivating our model selection:



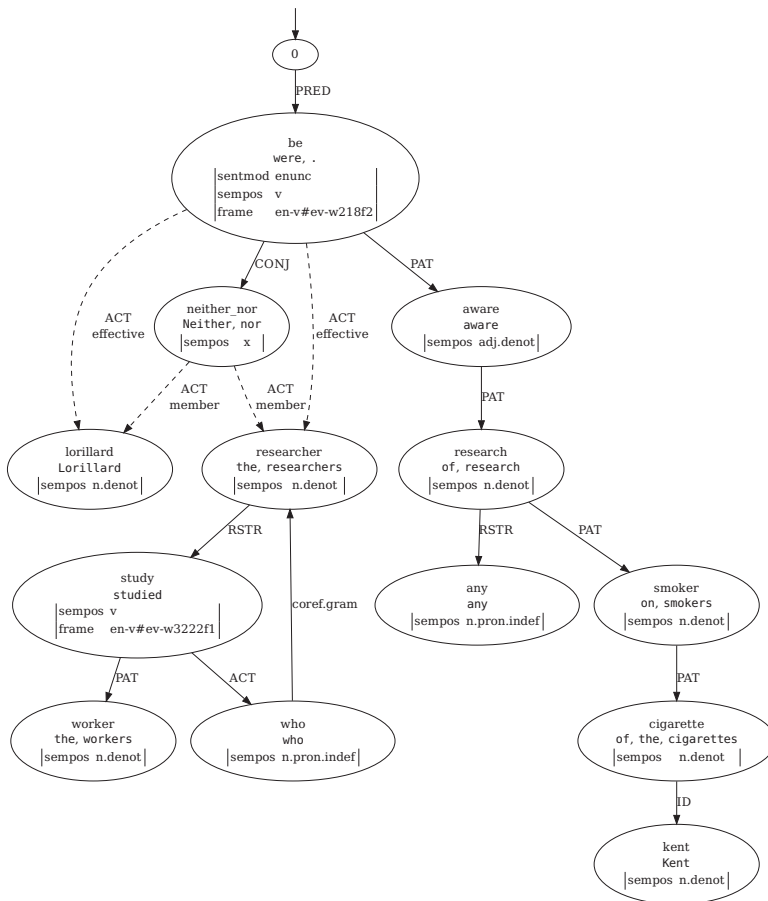


Figure 9. PTG representation of the English sentence *Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes*

- Annotation in the UCCA foundational layer is rather coarse-grained compared to PTG (a handful of edge label types, no annotation on nodes).¹³ In spite of this, parsers perform relatively poorly ($F_1 < 0.5$ on edge labels) on UCCA. This might indicate consistency problems with the UCCA annotation.
- While reported performance of the parser is generally good on DRG, DRG output generated for our Hungarian test corpus seemed to make relatively little sense.
- Performance of the parser is also good on EDS. However, an EDS-style model is trained only for English. The model struggles on Hungarian input often completely misinterpreting important constructions. This seems to be partly due to typological differences between

¹³On the positive side, some distinctions present in UCCA, such as state vs. process are orthogonal to those in other annotation schemes, and these would be worth porting.



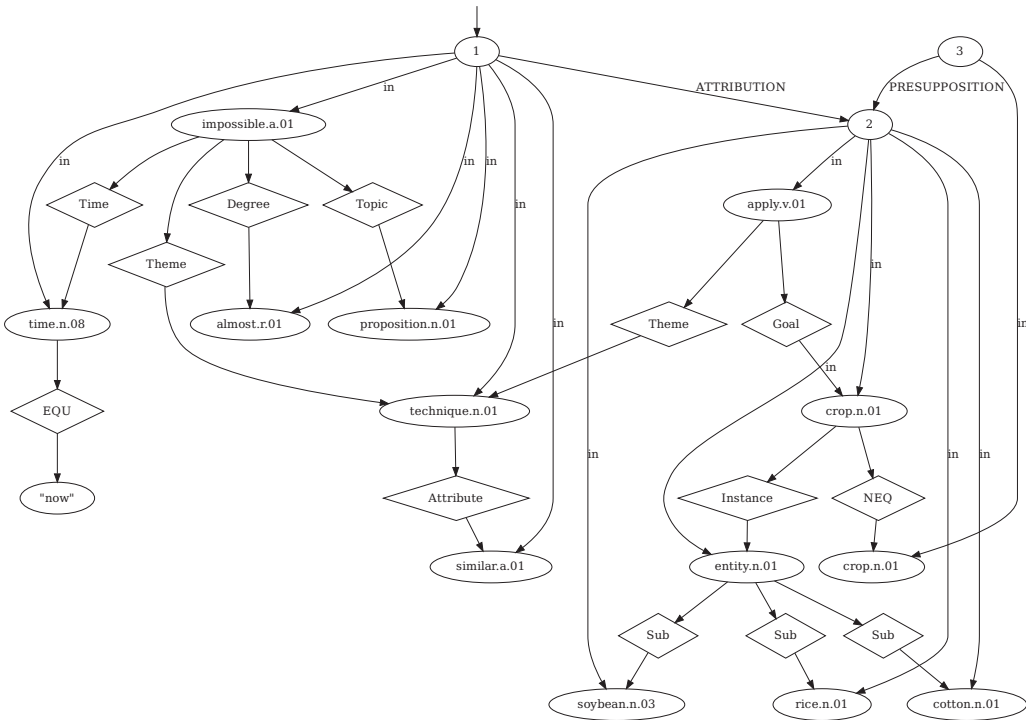


Figure 10. DRG representation of the English sentence *A similar technique is almost impossible to apply to other crops, such as cotton, soybeans and rice*

Hungarian and English. E.g. grammatical relations expressed by prepositions and word order are mainly expressed by suffixes in Hungarian, the latter being an agglutinative language. The EDS model often fails to properly recognize these relations (locations, times, possessive constructions, constituents not in canonical positions for English, etc.), because suffixes are not independent tokens in Hungarian while the EDS annotation scheme assumes that they should be.¹⁴ There is also pro drop in Hungarian, and this phenomenon affects a high proportion of clauses (see section 3.3.2). The EDS model fails to recover all such covert pronouns. Figure 11 shows the output of the EDS parser trained on English data for Hungarian input. It makes little sense. PDT output for the same sentence can be seen in Figure 12.

The PTG annotation the Czech PERIN model was trained on is derived from the Prague Tectogrammatical Annotation, an elaborate system of deep linguistic analysis based on a many-decade-long tradition of dependency-grammar-based linguistic research. The Prague

¹⁴Note, however, that the English PTG model, which utilizes a rich set of edge label categories to encode grammatical relations, seems to be much less affected by these typological differences, as that formalism does not depend on these relations to be expressed as independent tokens.



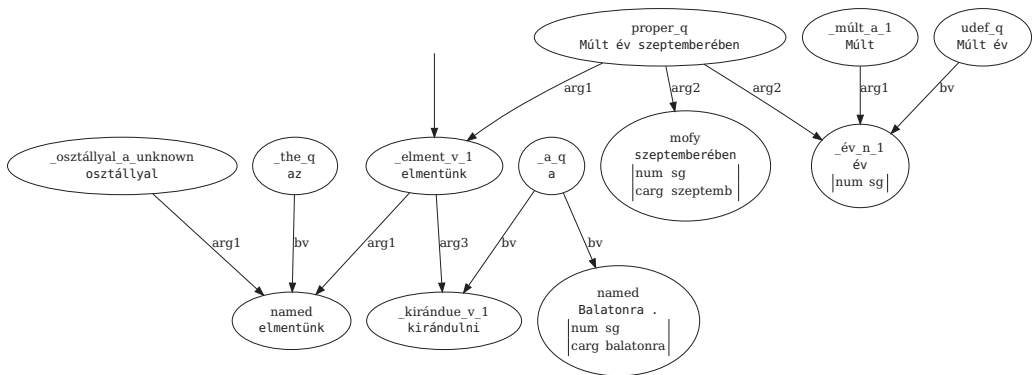


Figure 11. Output of the EDS parser for the sentence *Múlt év szeptemberében az osztályal elmentünk kirándulni a Balatonra*. ‘In September last year, the class and I went on a trip to Lake Balaton.’ The ‘analysis’ makes little sense

Dependency Treebank (PDT) and the Prague Czech–English Dependency Treebank (PCEDT, Hajič et al. 2012), from which PTG data was derived, embody an immense amount of annotation work. In addition to the deep syntactic annotation we review here, P(CE)DT annotation includes morphological annotation and a dependency-based shallow ‘analytical’ syntactic annotation of the underlying text. The tectogrammatical analysis was generated based on these surface-level representations, and then manually checked and corrected.

The Prague dependency annotation scheme was ported to languages other than Czech or English, examples including the Slovak Dependency Treebank (Gajdošová et al. 2016), the PAWS Treebank (including Polish and Russian in addition to English and Czech, Nedoluzhko et al. 2018), and the Prague Arabic Dependency Treebank (Hajič & Zemánek 2004). However, all these syntactic annotations were created manually. Here we examine automatic cross-lingual annotation transfer.

This approach can save much manual annotation work, but evaluation still requires manual effort and getting acquainted with the annotation guidelines.

3.2. Our approach

We had a 150-sentence fragment of the Hungarian Szeged Corpus (Csendes et al. 2004) annotated by the Czech PTG model. We manually corrected the output of the parser following the annotation manual of the tectogrammatical level of the Prague Dependency Treebank (Mikulová et al. 2006), turning it into gold standard annotation. Members of the annotation team were trained in theoretical and computational linguistics encompassing dependency syntax and formal semantics. Nevertheless, we had to understand and learn details of the annotation scheme during the process, which required substantial effort. Examples in the Annotation Guidelines have English translations, but unfortunately only a few examples have a full tree representation. We converged on an annotation that we considered consistent with what is described in the PDT annotation guidelines reiterating and rediscussing our solutions several times, as our understanding of the annotation scheme evolved during the process. Access to



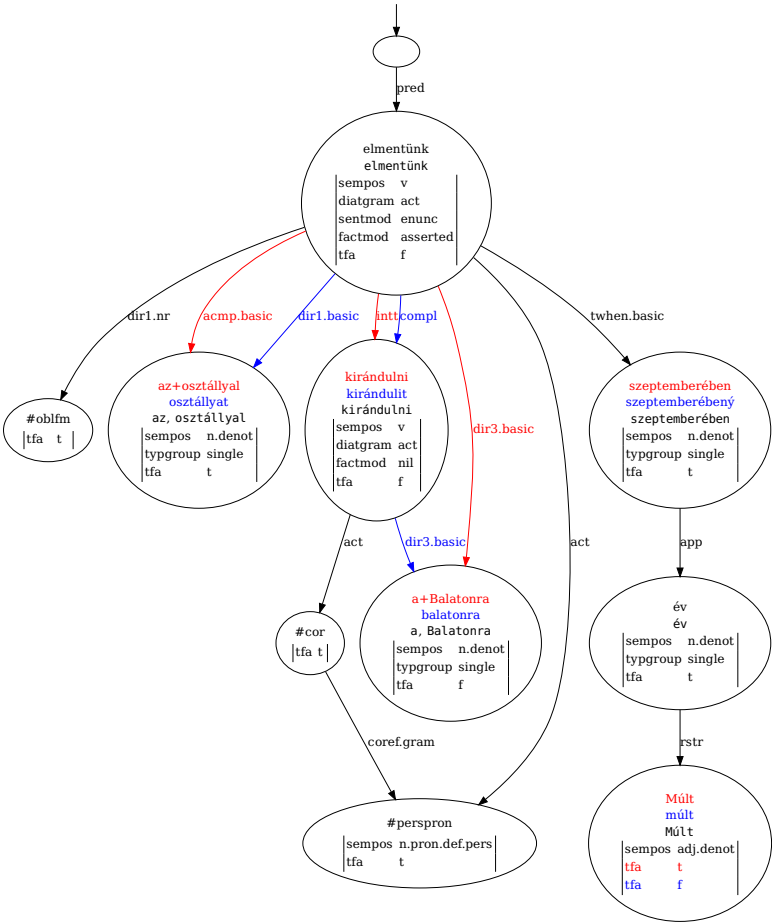


Figure 12. Analysis of the sentence *Múlt év szeptemberében az osztállyal elmentünk kirándulni a Balatonra*. ‘In September last year, the class and I went on a trip to Lake Balaton.’: output of the Czech-trained parser compared with gold annotation. Gold edges, labels and features are in red, those in the parser output in blue when there is a mismatch. Applying Czech lemmatization patterns results in erroneous lemmata (here simply replaced by surface forms instead of real Hungarian lemmata)

PCEDT would have been very helpful, however, only the Czech part of treebank is available online,¹⁵ so we could not take a look at the English equivalents or efficiently search for specific constructions not being speakers of Czech.

In order to have a fair evaluation of model transfer, we had to refrain from making modifications to the annotation scheme during manual correction. We tried to refrain from interpreting dubious situations ‘the way we would have made it’, we tried to figure out instead, how

¹⁵https://lindat.mff.cuni.cz/services/pmltq/#!/treebank/pcedt20_cz/query/.



ÚFAL experts would do it. We assumed that if the parser more-or-less consistently generates some sort of sufficiently sensible annotation for a specific construction, it reflects a deliberate annotation pattern in the training data.

3.2.1. The PERIN parser. The parser embodies a permutation-invariant model that predicts all nodes at once in parallel and is trained using a permutation-invariant loss function not sensitive to the ordering of nodes (Samuel & Straka 2020).

The language model the parser uses as a neural representation of the input when inferring the graph annotation is XLM-RoBERTa (base). XLM-R (Conneau, Khandelwal et al. 2020) is the encoder part of a transformer model pretrained originally on 2.5TB of filtered CommonCrawl data in 100 languages including Czech and Hungarian to predict masked word forms. This underlying multilingual neural language model makes an essential contribution to the decent cross-lingual performance we encountered, enabling the parser to output sensible annotation for input in a language the parser itself was not trained to handle originally.

The PERIN model uses relative string encodings to predict node labels that map token strings onto label strings. Specifically, in the PTG model, lemmata ('t-lemmata') are used as node labels. This mechanism works well when parsing text in the language the model was trained on. However, it is not surprising that applying Czech lemmatization patterns to Hungarian word forms results in strange lemmata like attaching the Czech infinitive ending *-it* to the Hungarian stem *kirándul* 'make an excursion' and the mostly adjectival ending *-ý* to the inflected form *szeptemberében* 'in September of' as shown in Figure 12. But since nodes are anchored to spans in the input (practically to tokens), external lemmatization can be used to fix the node labels. Our current solution was just replacing lemmata of non-zero elements with the corresponding surface forms. Since tokens can be linked to nodes using the anchors in the annotation, it is possible to evaluate the annotation ignoring the ill-formed lemmata.

While applying Czech lemmatization patterns to Hungarian obviously makes little sense, our initial probing of the model indicated that dependency relations among content words (edge labels in the graph, 'functors' and 'subfunctors' in PDT terminology) seem to carry over relatively well to Hungarian (of course, with some errors, as is evident in Figure 12). It was this aspect of the annotation that we wanted to concentrate on.

3.3. Results

We used the mtool¹⁶ evaluation, conversion and visualization tool to evaluate the zero-shot output of both the Czech and the English PTG models against the gold standard version of the test corpus. English PTG has less node features than the Czech model, and also the edge labels generally lack subfunctor annotation. The English model also uses different patterns to generate node labels (lemmata), so the performance of the models is only comparable after applying some normalization to the annotations. The normalization included a) replacement of node labels (lemmata) by the sequence of tokens anchored to the node (except for unanchored tokens, which retained their labels, see Figure 12), and b) removal of subfunctor annotation from edge

¹⁶<https://github.com/cfmpr/mtool>.



Table 1. Zero-shot performance of the English and Czech PTG models (PTG-en vs. PTG-cz) on the Hungarian test set

PTG-en	P	R	F
tops	1	1	1
labels	0.843	0.763	0.801
anchors	0.864	0.839	0.852
edges	0.682	0.558	0.614
attributes	0.714	0.546	0.619
PTG-cz	P	R	F
tops	1	1	1
labels	0.842	0.857	0.849
anchors	0.844	0.854	0.849
edges	0.704	0.690	0.697
attributes	0.745	0.665	0.703

Bold values indicate best performance.

labels (except for subtypes of *coref* and *bridging* relations, as these also have subfunctors in the English annotation).

Performance of the normalized output of the models as returned by mtool is compared in Table 1. The Czech model definitely performs better at identifying grammatical relations (edges, attributes) and the difference in node label recall reflects mainly its advantage at identifying zero nodes (due to handling of pro drop and richer annotation of argument coreference relations in light verb constructions).

Below we discuss specific details of the inspiringly good performance of the Czech model on Hungarian input. Some shortcomings of the cross-lingual annotation seem to be related to properties of the original annotation, or the the lack of a similar construction in the source language.

3.3.1. Node properties. The PTG graph representations were created by automatic conversion from tectogrammatical trees in P(CE)DT. English data comes from the Prague Czech–English Dependency Treebank 2.0 (Hajič et al. 2012) while the source of Czech data was the Prague Dependency Treebank 3.5 (Hajič et al. 2020).

Quite unfortunately, important grammatical features (number, person, tense, modality, degree, etc.) were omitted during conversion. This results in that almost all relevant information is lost in the annotation of e.g. covert pronouns (see Section 3.3.2) or modal auxiliaries (corresponding to *can*, *must*, *will*, etc.). The latter are not represented in PDT annotation as independent nodes: they only contribute a feature to the node of verb they combine with. This feature, however, was lost in conversion: e.g. the node pertaining to *el kellett menni* ‘we had to go’ does not have any feature corresponding to the deontic auxiliary in Figure 13, neither does the node of *ha tudnék futni* ‘if I



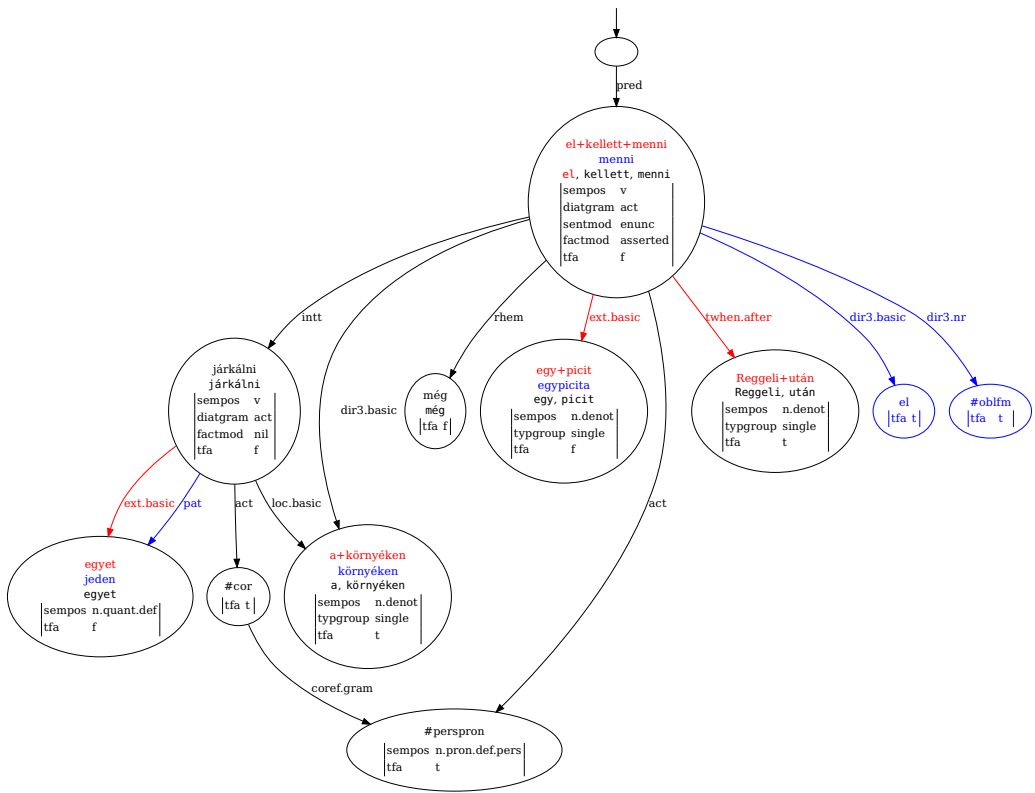


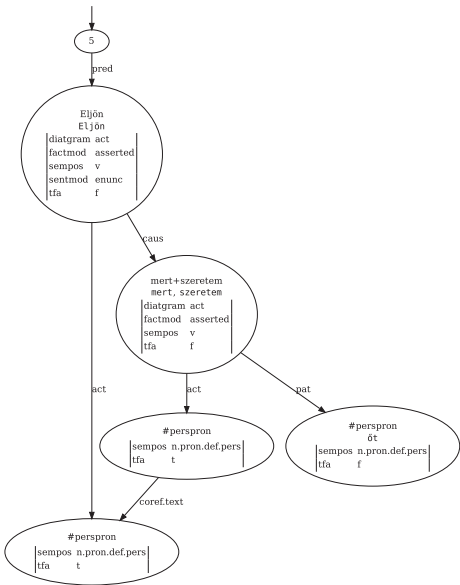
Figure 13. Analysis of the sentence *Reggeli után még egy picit el kellett menni a környéken járkalni egyet.* ‘After breakfast (we) had to go around the area for some time.’ The node corresponding to *el kellett menni* ‘we had to go’ does not have any feature corresponding to the deontic auxiliary due to loss of data during conversion from PDT to PTG. Annotation errors in the parser output are marked in red (missing) and blue (surplus)

could run’ in Figure 14b have one corresponding to the dispositional auxiliary. The lack of crucial grammatical features in the representation may play an important role in the parser making errors like establishing coreference relations between pronouns and noun phrases of different person/number (e.g. between ‘she’ and ‘I’ in the parse of *Eljön, mert szeretem őt.* ‘She will come because I love her.’ instead of linking ‘she’ and ‘her’: see Figure 14a).

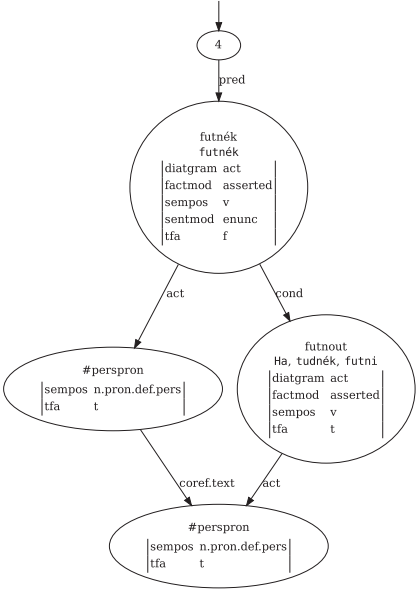
Since much of what we would like to see there is not there, and some of what we do have is irrelevant, we have not performed an exhaustive quantitative evaluation of the mapping of node features. Nevertheless, we make some qualitative observations concerning the performance of the parser with regard to specific node features present in the annotation in the following sections.

Part of speech Lexical nodes have at least a part-of-speech property, which is termed ‘semantic’ in PDT terminology, but it is less semantic than what one would expect. E.g. nominalized verbs are ‘semantic’ nouns. There are just a few deviations from syntactic part of speech:





(a) Analysis of the sentence *Eljön, mert szeretem őt.* ‘She will come because I love her.’ The subject of the matrix verb is erroneously linked with the subject of the subordinate clause instead of the object possibly due to the lack of key grammatical features in the representation.



(b) Analysis of the sentence *Ha tudnék futni, futnék.* ‘If I could run, I would run.’ The modality is erroneously identified as an assertion instead of potential.

Figure 14. Some types of erroneous analyses in the automatic zero-shot PTG annotation

deadjectival adverbs corresponding to English *-ly* adverbs are tagged ‘semantic’ adjectives, and numerals as adjectival or nominal quantifiers. Morphological negation is a Czech-specific feature reflected in the part of speech category set that does not apply to Hungarian. Non-ly adverbs are sometimes tagged as adjectives by the model, but otherwise part of speech is accurately identified.

Topic-focus articulation The Czech model also contains a feature related to topic-focus articulation (tfa). This is an advanced feature rarely found in computational meaning representations. We would have, however, expected four possible values instead of the actual three: *t* = contextually bound expression (topic), *f* = contextually non-bound expression (new information), *c* = (contextually bound) contrastive expression. We think that it would be relevant to distinguish contextually non-bound contrastive elements (focus proper) from contextually bound contrastive elements (contrastive topic). We could not determine from the limited description in the annotation manual how specific constructions (e.g. contrasting predicates) should fit into the annotation scheme used in PDT. The parser often assigns values to this feature that seem reasonable, but there are also cases where the annotation is obviously wrong



(e.g. assignment of the *f* value to topicalized definite expressions). The source of these problems could be among others that word order constraints concerning contrastive elements (focus/contrastive topic) are quite different in Czech and Hungarian (Czech: clause final, Hungarian: preverbal) and that there is no definite article in Czech.

Factual and sentence modality The model is able to differentiate appeals, requests and questions from assertions, however, quite surprisingly, it often fails to identify potential ('would') and counterfactual ('would have')¹⁷ modalities: see e.g. Figure 14b.

Identification of grammatical/semantic relations Although an F_1 score of around 0.7 for edges/attributes (see Table 1) might not seem very high at first sight, this is not bad, especially considering the rich variety of possible labels and the fact that the model was trained to parse Czech, not Hungarian. It is not very much worse than the performance of the same parser model for Czech input (edges: $F_1 = 0.84$, edge labels (attributes): $F_1 = 0.78$, Samuel & Straka 2020). The model is especially good at identifying adjuncts (time, place, directional and manner adverbials).

In contrast, the annotation of predicate argument relations in PDT, which in most cases is limited to two relations called *act* and *pat*, was a source of disappointment. These have nothing to do with real thematic roles like *agent* or *patient*, but are mostly simply placeholders for the first two arguments of any predicate. E.g. the subject of *the window broke* is *act* (Figure 15a), while the predicate argument in the valency frame of the copula (i.e. *blue* in *the car was blue*) is marked as *pat* (Figure 15b). But PTG is not alone having uninformative argument labels among meaning representation schemes: others mentioned in Section 3 have ARG0, ARG1, etc.

3.3.2. Empty elements. The model also relatively successfully predicts empty elements, such as dropped pronouns and ellipsis as long as similar patterns apply to both the source and the target language.

Pro drop For example, Czech, similarly to Hungarian, features pro drop: i.e. subject pronouns may optionally be omitted in neutral sentences, as shown in the two side-by-side one-word sentences in (1). In the case of Hungarian, there is a lack of overt subject pronouns in most cases when the pronoun is not emphasized. It is a nice feature of the model that it includes existentially bound optional arguments in the analyses it generates (e.g. *Olvasok.* is interpreted as 'I am reading something.', see Figure 16a)

- (1) *Olvasok.* *Čtu.*
 read.PRS.1SG read.PRS.1SG
 'I am reading. = I am reading (something).'

In Hungarian, the same sentence with an overt pronoun has different interpretations depending on the intonation pattern (2).

- (2) *Én olvasok.*
 'I am reading.' (neutral, rare)
 'It is me who is reading.' (focus)
 'As for me, I do read.' (contrastive topic)

¹⁷In PDT the value 'irreal' is used for counterfactual.



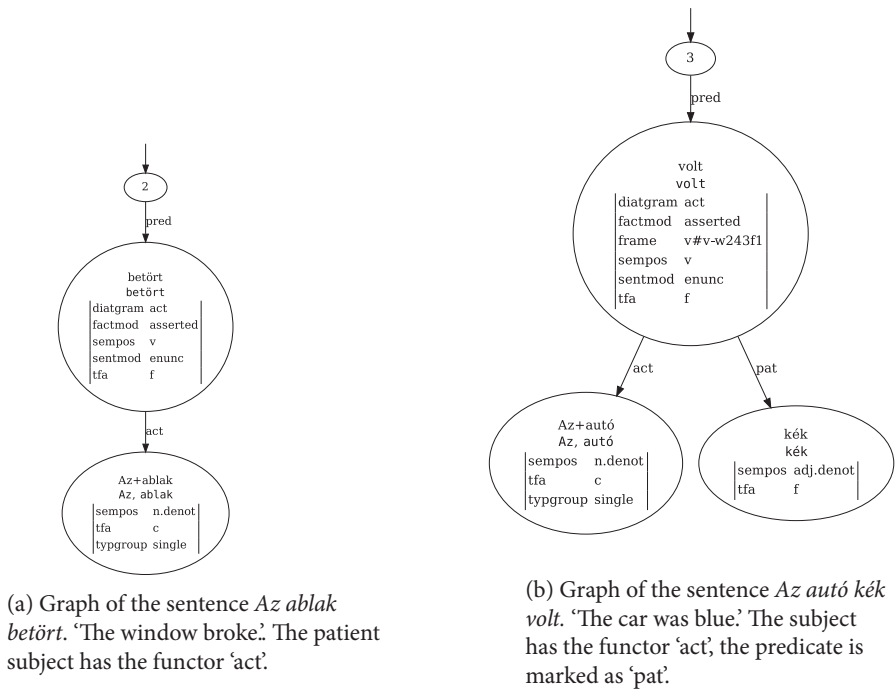


Figure 15. Some examples of the uninformative argument relations ‘act’ and ‘pat’ in the PTG annotation

However, in cases where the same phenomenon (i.e. pro drop) does not apply to certain pronouns in the source language, the model always fails to predict such covert pronouns in the target language. In Hungarian, for example, object pronouns also undergo pro drop. What makes this possible is that verbal morphology encodes not only subject agreement but also the definiteness of the object, as illustrated in (3). If the morphology of verb form implies the presence of a definite object, then the lack of an overt object implies the presence of an object pronoun (4a). In contrast, there is no object pro drop in Czech (4b), thus the model fails to predict covert object pronouns for Hungarian. Instead, we get the same interpretation with an existentially bound object that we get for *Olvasok* (1). The output of the model for these constructs is shown in Figure 16.

- (3) a. *Olvasok egy könyvet.*
read.PRS.1SG a book.SG.ACC
‘I am reading a book.’
- b. *Olvasom a könyvet.*
read.PRS.DEF.1SG the book.SG.ACC
‘I am reading the book.’
- (4) a. *Olvasom.*
read.PRS.DEF.1SG



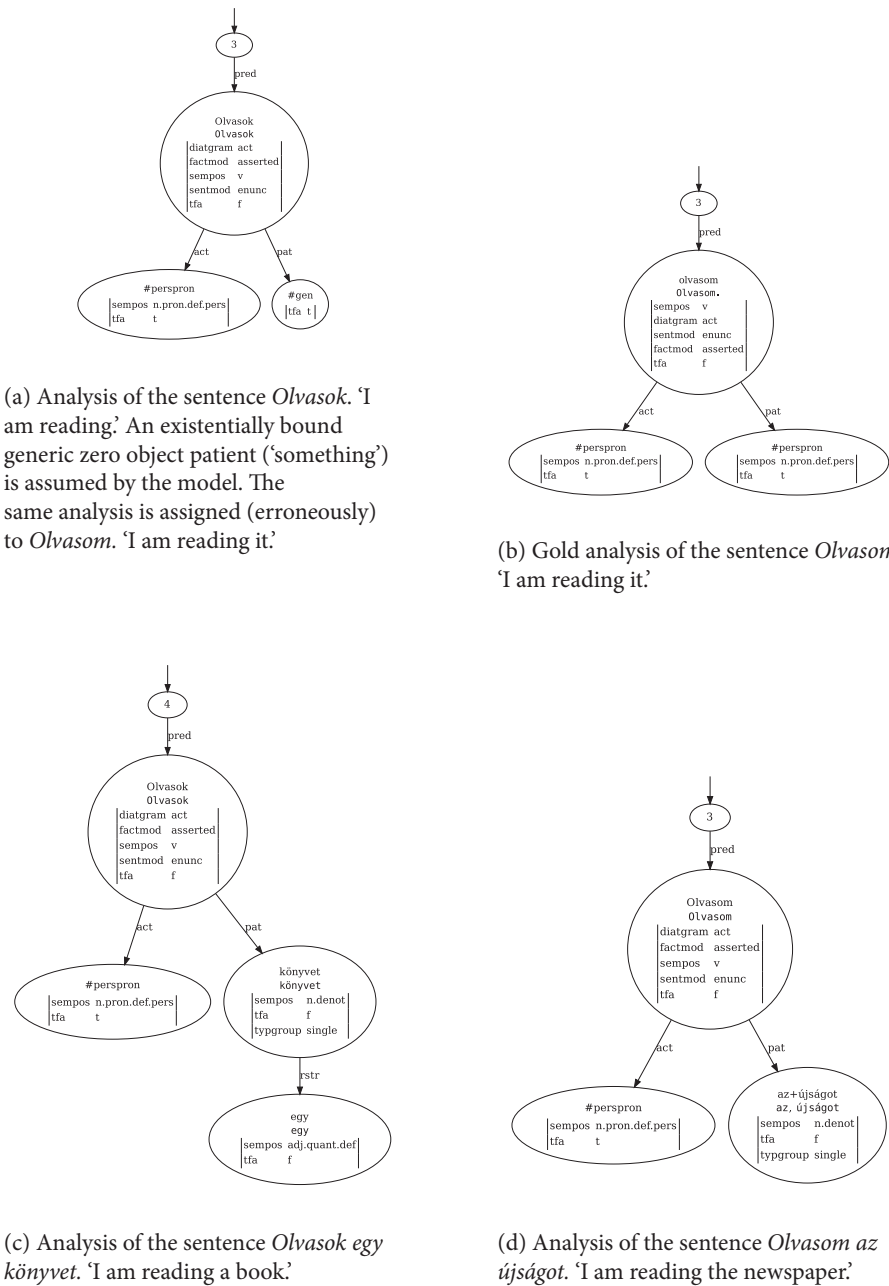


Figure 16. Some examples of how the Czech PTG model handles pro drop and existentially bound arguments



- b. *Čtu to.*
 read.PRS.1SG that.ACC
 ‘I am reading it.’

Possessive constructions involving pronouns The same applies to possessive constructions involving personal pronouns. The Czech (or English) version of these constructs involves a possessive pronoun determiner followed by a noun, optionally modified by adjuncts (5b). In Hungarian (and many similar agglutinating languages), the construct involves possessive suffixes attached to the noun as inflection, and the presence of an overt pronoun is optional, and, again, is mostly limited to cases of emphasis on the pronoun (5a). Since the possessive pronoun is obligatory in Czech (it is the key element of the construction), the parser trained on Czech data always fails to predict empty personal pronouns involved in possessive constructions in Hungarian, too.

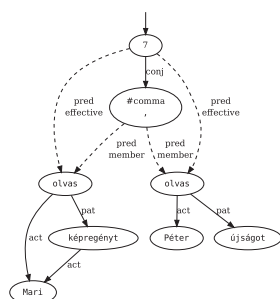
- (5) a. *az (én) anyám*
 the I mother.POSS.1SG
 b. *moje matka*
 my.FEM.SG.NOM mother.FEM.SG.NOM
 ‘my mother’

Ellipsis The model performs reasonably well predicting and reconstructing elliptical structures as long as a similar elliptical construction is present in the language the model was trained on. Both Czech and Hungarian feature gapping in the second clause in coordinated clauses. However, in Hungarian (similarly to e.g. Turkish), gapping in the first clause is also a frequently used construction. As shown in Figure 17, the parser fails to properly recognize the elliptical structure if the gap is in the first clause (not an option in Czech or English). For the given examples, we get a perfect parse only if the gap is in the second clause, and word order in the first clause is SVO (Figure 17c).

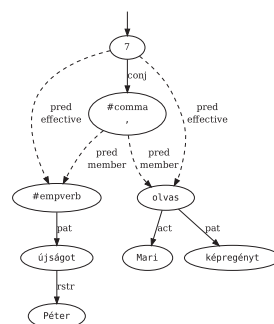
Zero copula PDT much predates the Universal Dependencies (UD) project, and in contrast to the lexical content head solution to copula constructions there, the copula is the head in PDT/PTG. In Hungarian, there is a zero copula in the default 3rd person singular present indicative case, so we needed to introduce a new zero copula (#zerocop) item to accommodate the annotation of zero copula constructions to the scheme applied in PDT and PTG. In the case of the clause *Az autó kék.* ‘The car is blue.’, #zerocop would assume the same position assumed by the copula *volt* ‘was’ in the past tense version of the sentence *Az autó kék volt.* ‘The car was blue.’, see Figure 15b. As it is, the model fails to parse zero copula constructions due to the analysis above and the fact that there is always an overt copula in Czech (see e.g. the completely failed analysis in Figure 19b). A UD-style copula annotation where the predicative noun/adjective is considered the head of construction and the copula is assumed an (optional) function word would naturally fit Hungarian, and such a model would probably have fewer problems parsing zero copula constructions.

3.3.3. Coordination/parataxis. In contrast to subordination, coordination (in PDT terminology: parataxis) is problematic for dependency-based annotation schemes because it is not an

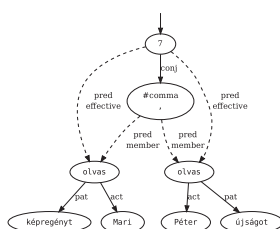




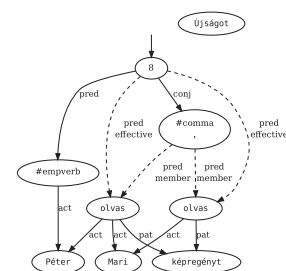
(a) Péter újságot olvas, Mari képregényt. 'Peter is reading a newspaper, Mary a comic.' – Gap in second clause, SOV word order in first clause. Minor error in the analysis.



(b) Péter újságot, Mari képregényt olvas. – Gap in first clause, SOV word order in second clause. Wrong analysis of first clause.



(c) Péter olvas újságot, Mari képregényt. – Gap in second clause, SVO word order in first clause. Perfect analysis.



(d) Újságot Péter, képregényt Mari olvas. – Gap in first clause, OSV word order in second clause. Wrong analysis.

Figure 17. Output of the Czech PTG grammar for various Hungarian gapping constructions. Only the analysis in 17c is completely correct

endocentric construction. The solution applied in the PTG implementation of PDT structures makes coordinating conjunctions or, in the absence of these, punctuation (commas) the head of coordinate structures, as shown in Figure 17c. The coordinated predicates (the two *olvas* 'read' nodes) are attached as *pred-members* to the *#comma* node (technical head of the coordinate structure), and they are also linked directly by *pred-effective* edges to the node dominating the whole structure, here the root of the sentence graph. The technical head (*#comma*) is attached using a relation characterizing the paratactic structure (e.g. conjunction, disjunction, apposition, etc., here: *conj*) to the node dominating the coordinate structure. This solution is again different from the one applied in UD (where the left conjunct is promoted as the head of the construction



and all other conjuncts are linked to it by *conj* edges): it is analogous to the way coordination is represented, e.g. in EDS.

Two aspects of this solution are problematic, however. Certain types of coordination, like cause, consequence or confrontation, express an asymmetric relation. These types of relations were doubled in the annotation scheme only because they also have a subordinating variant (coordinating *confr*, *reas* vs. subordinating *contrd*, *caus*). The distinction is purely syntactic, and most speakers would have an extremely hard time making a distinction between the subordinating and the coordinating variant. The analyses, however, are totally different. What is more, the representation of the paratactic variant of these constructions completely fails to distinguish which conjunct plays which role, e.g. what is the cause and what is the consequence. These unnecessary syntactic distinctions gave us a hard time during correction of the gold standard data (and can be considered a rather unmotivated artifact in the PDT annotation).

Coordinated predicates were analyzed by the model as verb phrase coordination sharing a single subject rather than assuming coreferring covert pronouns, as illustrated in Figures 18. We accepted this solution assuming that similar constructions must have been analyzed analogously in the Czech PDT treebank.

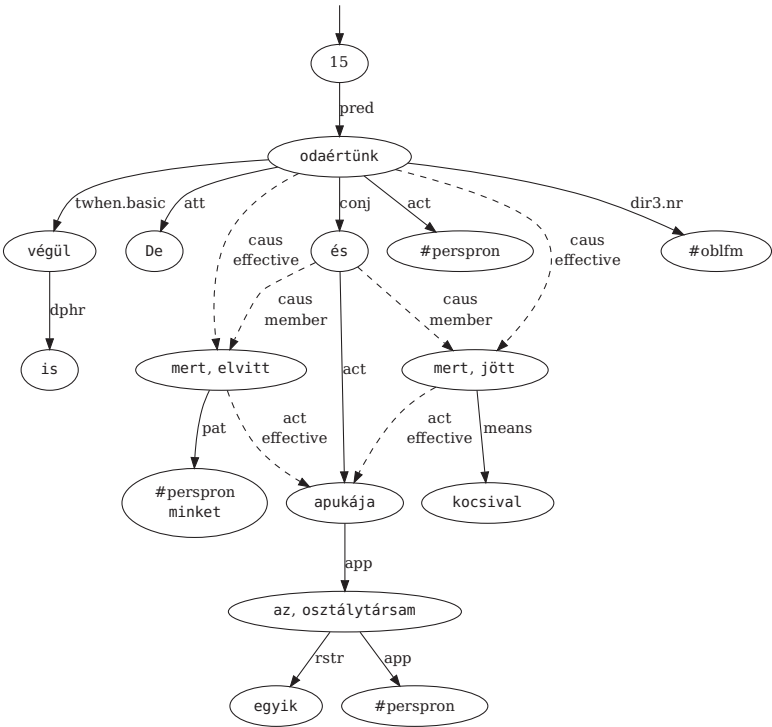


Figure 18. Analysis of the sentence *De végül is odaértünk, mert jött az egyik osztálytársam apukája kocsival és elvitt minket*. ‘But we ended up there because one of my classmates’ dad came in a car and took us there.’ *Apukája* ‘dad’ is a shared argument of both *jött* ‘came’ and *elvitt* ‘took’



3.3.4. Further problems. The model sometimes fails to integrate parts of the analysis into the whole structure, or, in some cases, completely ignores some part of the input. This often seems to be related to covert elements not attested in the source language like a zero copula or gapping in the first conjunct (see e.g. Figure 17d).

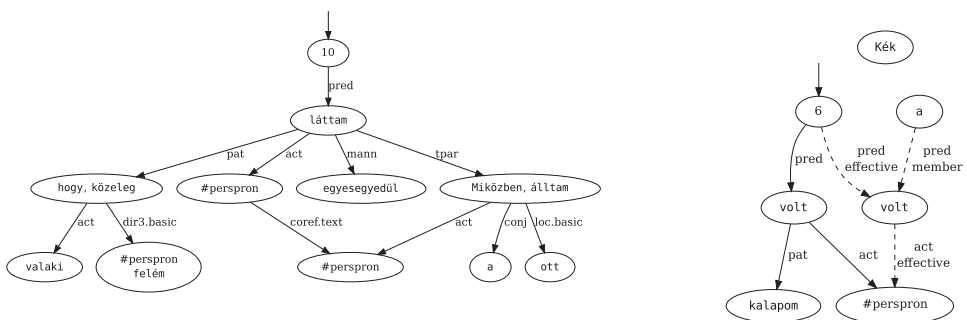
Short function words are sometimes confused with short frequent function words in the source language, and this may result in wrong and/or incomplete analysis. E.g. Hungarian *s* ‘and’ and *a* ‘the’ are sometimes confused with Czech *s* ‘with’ and *a* ‘and’, respectively, see Figure 19. In the case of the sentence shown in Figure 19b, nothing is like it would be in Czech: inverted word order, the definite article and the pronominal possessive construction unattested in the source language all contribute to the failure of the analysis.

Function words (articles, postpositions, subordinating conjunctions, auxiliaries) are normally merged with content words (the node is anchored on several tokens), but in some cases a partial merge is performed (the function word is anchored both to an independent node of its own and to the node of a content word) as shown in Figure 20. This is an error.

The model tokenizes the input at hyphens, and the hyphen remains unanchored. This is problematic for Hungarian, because suffixes (e.g. case endings) are often attached with a hyphen to the stem, and such case endings become independent tokens in the analysis. This is mostly a technical issue, because the model usually ends up anchoring the tokens around the hyphen to the same node, similarly to the way it handles conjunctions, auxiliaries, postpositions and definite articles.

3.4. Concluding remarks on zero-shot meaning representation parsing

As we have seen, a good multilingual contextual language model can make zero-shot cross-lingual transfer of quite elaborate syntactic annotation possible. The model yields reasonable cross-lingual performance, and it can be feasibly applied in a semi-automatic annotation



(a) Analysis of the sentence *Miközben ott álltam a pusztán, egyesegyedül, láttam, hogy valaki közeleg felém.* ‘As I stood there in the plain, alone, I saw someone approaching me.’ *pusztán* ‘in the plain’ is dropped due to failed analysis of *a*.

(b) Completely failed analysis of the sentence *Kék volt a kalapom.* ‘My hat was blue.’

Figure 19. Examples where a ‘the’ is misinterpreted as the Czech conjunction *a* ‘and’



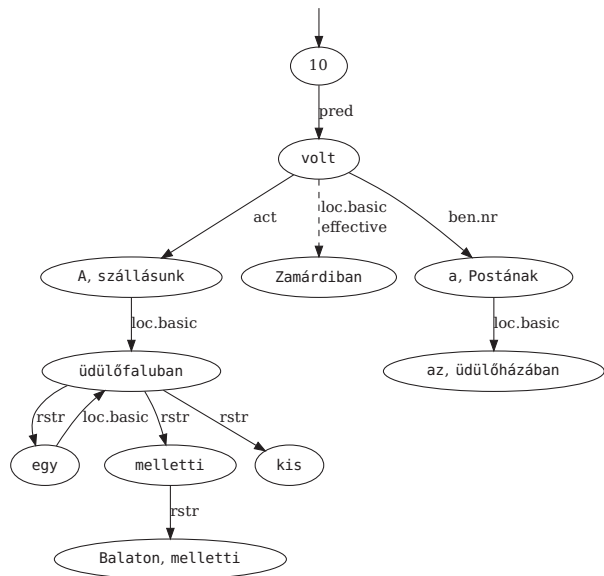


Figure 20. Graph output for the input sentence *A szállásunk egy Balaton melletti kis üdülőfaluban, Zamárdiban volt, a Postának az üdülőházában.* ‘Our accommodation was in a small holiday village near Lake Balaton, in Zamárdi, in the Posta’s holiday house.’ The postposition *melletti* ‘near’ is both merged with the noun it modifies and has a node of its own

scenario. Of the language pairs Czech–Hungarian vs. English–Hungarian, the former performs better because the source and the target language share more typological characteristics like rich morphology, free word order, pro drop, etc. even though they belong to different language families. Moreover, the PDT/PTG annotation scheme utilizing a rich set of dependency relations as edge labels seems to perform much better than, e.g. EDS where edge labels are completely abstract, and annotation crucially relies on the content of nodes, which are assumed to be independent tokens, an annotation scheme fitting isolating languages well, but transferring relatively poorly to agglutinating languages (where e.g. morphemes corresponding to prepositions are not independent tokens).

4. FURTHER EXPERIMENTS

Our further experiments involved cross-lingual transfer in another structured prediction task: named entity recognition (NER). Although named entity recognizers have been there since the 1990s, their creation required extensive manual effort for each target language. The cross-lingual transfer paradigm we discuss in this paper seems to have changed that situation by providing reasonable cross-lingual performance and thus substantially reducing the human effort needed to create such models: as we will see, even what we get with quasi-zero effort is quite reasonable.



4.1. Named entity recognition transfer for Hungarian

We performed a number of experiments concerning zero-shot transfer on named entity recognition. In one experiment, we performed research concerning the automatic enhancement of the annotation of the large NYTK-NerKor named entity corpus (Simon & Vadász 2021) containing Hungarian texts of various genres. We have almost doubled the number of annotated elements in the corpus and made the number of distinguished classes 7 times bigger. For this, we applied cross-lingual transfer, which proved to be efficient according to our evaluation, but the performance of various transfer models showed significant differences. We relied on zero-shot application of transformer-based named entity recognition models trained on resources in other languages: English and Czech, algorithmic merging with the original annotation, and semi-automatic and manual correction.

4.1.1. Enriching the annotation using cross-lingual transfer and semi-automatic correction. NYTK-NerKor is a typical 4-class (persons, organizations, places, ‘miscellaneous’) named entity corpus, which we wanted to upgrade to something that includes useful entity classes instead of the kitchen sink ‘miscellaneous’ category, and also annotation for numerical and time expressions.

First we applied two models trained on the English OntoNotes 5 corpus to the Hungarian corpus. The first model was created by the DeepPavlov team (Burtsev et al. 2018) fine-tuning multilingual BERT (Devlin et al. 2019). The other model, flair/ner-english-ontonotes-large (Schweter & Akbik 2020), is based on XLM-RoBERTa (Conneau, Khandelwal et al. 2020), a multilingual contextual language model trained on a significantly bigger multilingual corpus than multi-BERT.

We merged the annotation from the models with the original annotation. The merging algorithm considered the spans in the input annotations gold standard in the case of overlapping entity spans, and if the generated annotation contained a compatible entity subtype, the entity type was updated accordingly.

While cross-lingual mapping resulted in some anomalies like inclusion of definite articles in names,¹⁸ it had other side-effects that we found useful. E.g. since English prepositional phrases of names (which are obviously annotated as named entities) often correspond to adjectives derived from the given name in Hungarian, the output of the models also included entity annotation for these adjectives, definitely another step forward from the annotation scheme used in all legacy Hungarian NER corpora that systematically left adjectives derived from names unannotated.

We also applied a third model to the corpus. We used the Czech model of the NameTag 2 neural named entity tagger (Straková et al. 2019) trained on the Czech Named Entity Corpus CNEC 2 (Ševčíková et al. 2007). This model is based on a fine-grained hierarchy of entity classes having many subclasses within the broader categories like a distinction of companies vs. governmental/political institutions vs. academic/educational/cultural/sports institutions and

¹⁸For some named entity types, like names of organizations, journals, titles of works of art, etc., a definite article is present in Hungarian when the name is incorporated in the sentence structure – but not in parentheses –, while there is no article in English: *Peter works at IBM.* – *Peter az IBM-nél dolgozik.* Originally, we thought that this correspondence is the source of the tendency of definite articles being included in named entities by the OntoNotes-based model. Later having acquired access to the OntoNotes corpus, we found that the real source of the problem seems rather to be that, in the OntoNotes corpus, determiners are generally included in named entity annotations.



conferences/contests (the latter are also considered a subclass of organizations). NameTag 2 is capable of returning nested annotations (with a maximal depth of two overlapping entities). Since there is no definite article in Czech, we expected this model to have a problem with definite articles similarly to the English models, but it turned out to have this problem only in the case of sentence-initial capitalized definite articles (probably due to a constraint on capitalization that might be included in the algorithm). A more prevalent problem with this model was that it often assigned different classes to different occurrences of the same entity (and usually this was an error rather than real ambiguity) and often left the same entity unannotated. Identification of the span of the entities was also less accurate than what the English-based models generated.

The article problem and certain ill-formed quantifier expressions were easy to fix using regular expression-based substitution patterns. We also created a gazetteer-like lemmatized named entity inventory based on the output of the models that could be used to mark entries for correction. Marked entries were then mass-corrected in their inflected forms as well in the corpus.

Further details of the creation of the updated version of the NerKor corpus and a detailed description of the distribution of entity types can be found in [Novák & Novák \(2022a\)](#) (in Hungarian).

4.1.2. Evaluation of the models. We evaluated the zero-shot performance of the transfer-based models: the OntoNotes 5-based Flair and DeepPavlov models and the Czech NameTag 2 tagger on the test set of the corpus. We also performed the evaluation with the tagset normalized to the tags present in the original model (e.g. in the case of the OntoNotes 5-based models, labels not present in the original model (typeset in bold in [Table 3](#)) were normalized (e.g. CAR → PRODUCT, PROJ → EVENT) or ignored (e.g. MISC or AWARD) during evaluation). We also trained a neural tagger model based on the Hungarian huBERT contextual language model ([Nemeskey 2021](#)) on the training set of the corpus using the HuggingFace Transformers library ([Wolf et al. 2020](#)) with an improved Viterbi-like decoding that eliminates invalid tag sequences from the output ([Nemeskey 2020](#)). The performance of these models is shown in [Table 2](#). We report P, R and F₁ scores as percentage.

Models using language transfer performed quite well, but among the English models trained on the same corpus, the XLM-RoBERTa-based Flair model performed significantly (about 10% F-measure) better. The Flair model using a “stronger” language model obtained higher precision and recall values across the board for all named entity types, than the weaker model. The performance of these models increased (by 5–6% F-measure) when the automatic regular-expression-based correction of definite articles was applied to their output. The zero-shot performance of the Flair model on entity types in common with those in the final version (i.e. normalizing/ignoring the MEDIA, SMEDIA, PROJ, etc. tags not present in the original model) is quite convincing. This performance made our re-annotation effort feasible.

The apparently quite weak performance of the Czech model is partly explained by the fact that it works with a much more fine-grained tagset, thus in order to measure its performance, the normalization of tags was unavoidable. Its performance, however, lags far behind the other models after normalization, too. The Czech training set is much smaller than that of the other models, and the more complex algorithm allowing nested entities might also play a role in its weaker performance.



Table 2. Performance of models on the test set, CZ: Czech model NameTag 2, DP: DeepPavlov OntoNotes/m-BERT, FL: Flair-OntoNotes-Large/XLM-RoBERTa, NKC: NerKor + Cars/huBERT, test: precision of the test set before manual correction

version Model	final tagset			Det fixed			normalized labels			norm. labels, Det fixed		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
CZ	15.82	11.39	13.25	15.89	11.44	13.30	64.57	52.92	58.16	64.63	52.97	58.22
DP	66.32	60.41	63.23	71.66	65.27	68.31	68.79	63.42	65.99	74.63	68.81	71.60
FL	74.81	70.73	72.71	80.59	76.19	78.33	77.68	74.34	75.97	83.90	80.29	82.06
NKC	91.07	88.12	89.57				91.64	89.18	90.39			
test	91.92	87.65	89.73									

Best results are in bold.



Table 3. Performance of the best model trained on NerKor + Cars on each entity type compared to the performance a similar model trained on the original NerKor annotation. Tags ordered within the categories with descending frequency in the test set top to bottom. Tags in bold are not present in the OntoNotes 5 tag set

NerKor	F ₁	NerKor + Cars	F ₁
		DATE	88.85
		CARDINAL	83.78
		NORP	87.12
		ORDINAL	94.67
		LAW	82.12
		QUANTITY	91.11
		DUR	74.67
		PERCENT	84.21
		TIME	66.67
		LANGUAGE	83.33
		AGE	100.00
		MONEY	87.50
ORG	88.45	ORG	93.33
PER	95.32	PER	97.11
		GPE	91.98
LOC	92.28	LOC	76.60
		FAC	80.00
		WORK_OF_ART	90.27
		PROD	79.37
		MEDIA	91.53
		CAR	92.86
MISC	81.85	SMEDIA	73.33
		EVENT	72.73
		MISC-ORG	47.06
		PROJ	66.67
		AWARD	100.00
		MISC	66.67
	91.02		89.57/92.05



4.1.3. Comparison of models trained on the original and the enhanced corpus. We also compared the performance of the best tagger model with that of the same algorithm trained on the original four-class NerKor annotation to see how the division of some entity classes (especially MISC) into several subclasses impacts performance. For the sake of comparability, we partitioned entity types into non-entities, which are not part of the original annotation (numerical and time expressions, language names, NORP adjectives and law references), and named entities. The results are shown in Table 3: non-entities at the top half (ordered by decreasing frequency in the test set), entities at the bottom half, with aggregate scores at the bottom row (for the NerKor + Cars model: $F_1 = 89.57$ on all entities, $F_1 = 92.05$ on named entities).

The huBERT-based model with the Viterbi-based decoding performed similarly on named entities to a similar model (emBERT) trained on the original version of the corpus (Simon et al. 2022).

The F-score on locations is lower than in the case of the model trained on the original corpus partly due to missed adjectival GPE entities (not present in the original annotation), massive ambiguity of *Europe* as a continent or a reference to the EU (which was not consistently marked in the not yet checked portion of the training corpus), ORG vs. FAC ambiguity of institutions (universities) and obscure place names (GPE vs. LOC ambiguity). Most confusion is within subtypes of locations.

Performance on frequent and easier-to-distinguish subtypes of MISC (WORK OF ART, MEDIA and CAR) is better than on the generic MISC category, while for rare and difficult-to-categorize entities (as well as for products) we got worse-than-average performance. Nevertheless, the division of the MISC class to several subclasses (even with mainly automatic methods) did not result in a substantial drop in the performance of the system, the aggregate F_1 score for named entities turned out to be even better than for the same type of model trained on the original four-class annotation (92.05 vs. 91.02).

We illustrate the difference between the output of the model trained on the original four-class corpus and that of the one trained on the new 28-class version in Figure 21.

4.2. Zero-shot transfer for Azerbaijani NER

Another experiment we performed involved zero-shot NER transfer from English to Azerbaijani (Ibiyev & Novák 2021). In that experiment we used the same OntoNotes 5-based English models like in our experiment with Hungarian, and we used the Azerbaijani part of the WikiAnn dataset (Pan et al. 2017) as the base NER corpus (the only Azerbaijani NER corpus available).

WikiAnn is a massively multilingual ‘silver standard’ named entity corpus covering 282 languages automatically generated from Wikipedia extracting fragments containing internal links and annotating the link spans according to the coarse type¹⁹ of the Wikipedia entry the link points to. Due to its volume, WikiAnn is often used as a benchmark, it is even part of the multilingual XTREME benchmark (Hu et al. 2020), although it is full of errors: e.g. we found that almost half of the entities marked as organizations in the Azerbaijani part are in error. As shown in Table 4, the overall quality of the original Azerbaijani WikiNer annotation is worse

¹⁹WikiAnn contains only person, organization and location entities.



A Porto **ORG** történetének legnagyobb hazai vereségét a Liverpool **ORG** ellen szenvedte el Európában **LOC**, 2018 februárjában **DATE** - többek között Sadio Mané **PER** mesterhármásával - 5-0-ra nyertek az angolok Portugáliában **LOC**. Az első bécsi döntést november 2.-án hirdették ki, amelynek értelmében Magyarország **LOC** visszakapott 11 927 km²-nyi területet, 869 299 lakossal, amelynek megközelítőleg 84%-a magyar volt. A szépfü **MISC**, eredeti címén Bel Ami **MISC**, Guy de Maupassant **PER** francia író második regénye, mely 1885-ben jelent meg és az első négy hónapban 37 kiadást ért meg. Kolumbia **ORG** a múlt héten értesült a 35 éves Rosa Elvira Cely **PER** esetéről, akire május 24-ére virradóra találtak rá Bogotában **LOC**, a Nemzeti Park **LOC** egy elhagyott részén. A spanyol **ORG** sajtó szerint a londoni **LOC** klub 100 millió fontot **MONEY** kért a 28 éves **AGE** támadóért a Realtól **ORG**, ám Sarri **PER** közölte, hogy a vezetőség nem szívesen adná el őt. 2012. július 10-én **DATE** került fel a YouTube **MISC** -ra az a videó, melyen közel húsz férfi moleztált egy 16 éves indiai lányt egy Guwahati-i kocsmában. @ 2011feb17 **PER**: Libia **LOC** első május elsejei (Munkások Napja **MISC**) ünnepe 42 év után! Összehasonlításképp, tavaly az első helyezett Volvo XC40 **MISC** 325, a második Seat Ibiza **ORG** pedig 242 pontot zsebelt be, tehát 2018-ban messze nem volt olyan szoros a verseny, mint idén **DATE**. Ez a törvény volt érvényben Norvégiában **LOC** és Feröeren **LOC** egészen 1604-ig, amikor IV. Keresztély **PER** dán király dán nyelvre fordította, felülvizsgáltatta „Norske Lov **MISC**” (norvég törvény) néven. Ghazzawi **PER** az Egyesült Államokban **LOC** született szíriai származású blogger és aktív Twitter **MISC**-felhasználó, az angol nyelvű Global Voices Online **MISC** és Global Voices Advocacy **MISC** oldalak szerzője. A szezon végén ő kapta meg az „Ajax **ORG** év tehetsége **MISC**” díjat mivel 40 mérkőzést játszott le a szezonban és sikerült debütálnia a holland válogatottban is.

A Porto **ORG** történetének legnagyobb hazai vereségét a Liverpool **ORG** ellen szenvedte el Európában **LOC**, 2018 februárjában **DATE** - többek között Sadio Mané **PER** mesterhármásával - 5-0-ra nyertek az angolok Portugáliában **LOC**. Az első bécsi döntést november 2.-án **DATE** hirdették ki, amelynek értelmében Magyarország **LOC** visszakapott 11 927 km² **QUANTITY**-nyi területet, 869 299 **CARDINAL** lakossal, amelynek megközelítőleg 84%-**PERCENT** a magyar **NORP** volt. A szépfü **WORK_OF_ART**, eredeti címén Bel Ami **WORK_OF_ART**, Guy de Maupassant **PER** francia **NORP** író második **CARDINAL** regénye, mely 1885-ben **DATE** jelent meg és az első négy hónapban **DATE** 37 **CARDINAL** kiadást ért meg. Kolumbia **LOC** a múlt héten **DATE** értesült a 35 éves **AGE** Rosa Elvira Cely **PER** esetéről, akire május 24-ére **DATE** virradóra **TIME** találtak rá Bogotában **LOC**, a Nemzeti Park **LOC** egy elhagyott részén. A spanyol **NORP** sajtó szerint a londoni **LOC** klub 100 millió **fontot MONEY** kért a 28 éves **AGE** támadóért a Realtól **ORG**, ám Sarri **PER** közölte, hogy a vezetőség nem szívesen adná el őt. 2012. július 10-én **DATE** került fel a YouTube **SMEDIA** -ra az a videó, melyen közel húsz **CARDINAL** férfi moleztált egy 16 éves **AGE** indiai **NORP** lányt egy Guwahati-**LOC** i kocsmában. @ 2011feb17 **PER**: Libia **LOC** első **CARDINAL** május elsejei **DATE** (Munkások Napja **EVENT**) ünnepe 42 év **DATE** után! Összehasonlításképp, tavaly **DATE** az első **CARDINAL** helyezett Volvo XC40 **CAR** 325 **CARDINAL**, a második **CARDINAL** Seat Ibiza **CAR** pedig 242 **point QUANTITY** zsebelt be, tehát 2018-ban **DATE** messze nem volt olyan szoros a verseny, mint idén **DATE**. Ez a törvény volt érvényben Norvégiában **LOC** és Feröeren **LOC** egészen 1604-**DATE** ig, amikor IV. Keresztély **PER** dán **NORP** király dán **NORP** nyelvre fordította, felülvizsgáltatta „Norske Lov **LAW**” (norvég törvény) néven. Ghazzawi **PER** az Egyesült Államokban **LOC** született szíriai **NORP** származású blogger és aktív Twitter **SMEDIA**-felhasználó, az angol **LANGUAGE** nyelvű Global Voices Online **MEDIA** és Global Voices Advocacy **MEDIA** oldalak szerzője. A szezon **DATE** végén ő kapta meg az „Ajax év tehetsége **AWARD**” díjat mivel 40 **CARDINAL** mérkőzést játszott le a szezonban **DATE** és sikerült debütálnia a holland **NORP** válogatottban is.

Figure 21. Output of the model trained on the original four-class corpus (left) and the one trained on the enhanced 28-class version (right)

Table 4. Comparison of the accuracy of the original Azerbaijani WikiAnn annotation to the zero-shot performance of the Flair OntoNotes model on LOC/PER/ORG entities in the Hungarian NerKor + Cars-OntoNotes++ test set (best scores in bold). With an overall $F_1 = 0.82$, using WikiAnn as a benchmark is quite questionable

AzWikiAnn	P	R	F_1	occ.
LOC	89.00	79.56	84.02	15,362
ORG	54.58	78.70	64.46	3,492
PER	94.25	81.70	87.53	4,888
all	85.02	79.96	82.41	23,743
FLO on NKC	P	R	F_1	occ.
LOC	93.69	76.94	84.50	317
ORG	73.16	89.21	80.39	339
PER	91.04	87.24	89.10	413
all	86.16	84.11	85.12	1,069



Table 5. Comparison of zero-shot performance of the OntoNotes-based DeepPavlov (DPO) and Flair (FLO) models on corrected Azerbaijani WikiAnn annotation (only on LOC/ORG/PER entities)

entities	DPO on AzWikiAnn			FLO on AzWikiAnn		
	P	R	F ₁	P	R	F ₁
LOC	76.10	65.40	70.35	78.06	75.44	76.73
ORG	61.09	51.53	55.90	65.84	59.54	62.53
PER	72.12	77.67	74.80	74.26	80.10	77.07
all	73.70	66.81	70.09	76.06	74.96	75.50

Best results are in bold.

than the zero-shot performance of the Flair English OntoNotes model on LOC/PER/ORG entities in the Hungarian NerKor + Cars-OntoNotes++ test set, which makes using WikiAnn as a benchmark to evaluate multilingual models quite questionable. Note also that the distribution of entities is much more balanced in the NerKor + Cars test set than in WikiAnn.

We have also found the zero-shot models, especially the Flair model, perform worse on Azerbaijani WikiAnn than on NerKor, and the performance gap between the mBERT and XLM-RoBERTa-based model was smaller than in the case of the Hungarian corpus, although the Flair model also performs better here across the board; see Table 5. This may be due to the fact that languages with smaller monolingual corpora available for exploitation in multilingual transformer pretraining (Azerbaijani in this case) perform worse in zero-shot transfer in general. Another factor may be peculiarities of WikiAnn: a significant proportion (17.13%) of Azerbaijani WikiAnn segments are short fragments consisting of a single entity or a list of entities without any context except some punctuation, thus the models can only rely on lexical knowledge in the absence of contextual clues. These ‘no context’ segments cover 14% of all entity occurrences.

5. CONCLUSIONS

In this paper, we shortly reviewed how recent artificial deep neural models acquire linguistic capacities that were not attainable by earlier generative-grammar-based models by being only exposed to raw linguistic data applying a learning paradigm based on an objective of predicting missing or corrupted parts of the data. We also argued that these results can be regarded as further arguments against Chomsky’s theory about a poverty of stimulus during language acquisition and the assumed existence of a specific human language acquisition device, as these models are based on generic neural architectures and are not exposed either to any multimodal contextual clues, negative examples, explanation or annotation of any sort. We have demonstrated that these models contain fairly accurate syntactic knowledge without any training to perform parsing. The success of these models also proves that, in contrast to Chomsky, distributionalists were on the right track.

Models pre-trained in this fashion can subsequently efficiently be fine-tuned to perform specific linguistic tasks including not only end-to-end skills like question answering or natural



language inference but also ‘structured prediction’ tasks such as deep linguistic annotation or identification of names and the type of their referents. Moreover, models fine-tuned to perform specific tasks from multilingual pre-trained models can efficiently transfer their specific skills acquired in one language to other languages covered by the underlying multilingual model due to massive parameter sharing in the internal representations formed during training in the higher layers of these models.

To demonstrate this, we applied models trained in other languages (English and Czech) to Hungarian data evaluating their performance. In one application, we generated dependency-based meaning representations in the style of the Prague Tectogrammatical Dependency Annotation for Hungarian, having found that typological similarity, i.e. the existence of specific shared constructions between the source and target language affects the efficiency of transfer in this case.

The other application we presented was transforming a sizable Hungarian named entity resource into one that distinguishes 7-times as many different entity classes in a quite moderately labor-intensive manner, and found that a model trained on the new richly annotated corpus version is as accurate as one trained on the original limited and coarse-grained annotation.

ACKNOWLEDGMENTS

The research presented in this paper was 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. We would like to thank Csilla Novák for the annotation work she has done correcting the Hungarian PTG annotations presented in Section 3.1 on zero-shot meaning representation transfer.

REFERENCES

- Abend, Omri and Ari Rappoport. 2013. UCCA: A semantics-based grammatical annotation scheme. Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers. 1–12.
- Abzianidze, Lasha, Johan Bos and Stephan Oepen. 2020. DRS at MRP 2020: Dressing up discourse representation structures as graphs. Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing. 23–32.
- Bahdanau, Dzmitry, Kyunghyun Cho and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint. arXiv:1409.0473.
- Bai, Jiangang, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu and Yunhai Tong. 2021. Syntax-BERT: Improving pre-trained transformers with syntax trees. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 3011–3020.
- Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer and Nathan Schneider. 2013. Abstract meaning representation for sembanking. Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. 178–186.



- Bouma, Gosse, Gertjan van Noord and Rob Malouf. 2000. Alpino: Wide-coverage computational analysis of Dutch. *Computational Linguistics in the Netherlands 2000: Selected Papers from the Eleventh CLIN Meeting (Language and Computers: Studies in Practical Linguistics 37)*. 45–59.
- Burtsev, Mikhail, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhrev and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. *Proceedings of ACL 2018, System Demonstrations*. 122–127.
- Chi, Ethan A., John Hewitt and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5564–5577.
- Clark, Kevin, Minh-Thang Luong, Quoc V. Le and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *International Conference on Learning Representations*. arXiv. arXiv:2003.10555.
- Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.
- Conneau, Alexis, Shijie Wu, Haoran Li, Luke Zettlemoyer and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6022–6034.
- Copestake, Ann, Dan Flickinger, Carl Pollard and Ivan Sag. 2005. Minimal recursion semantics: An introduction. *Research On Language And Computation* 3. 281–332.
- Csendes, Dóra, János Csirik and Tibor Gyimóthy. 2004. The Szeged corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*. 19–22.
- Dalrymple, Mary. 2001. *Syntax and semantics*, Vol. 34: Lexical functional grammar. New York, NY: Academic Press.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)*. 4171–4186.
- Dufter, Philipp and Hinrich Schütze. 2020. Identifying elements essential for BERT’s multilinguality. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4423–4437.
- Firat, Orhan, Kyunghyun Cho and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 866–875.
- Flickinger, Dan, Emily M. Bender and Stephan Oepen. 2014. Towards an encyclopedia of compositional semantics: Documenting the interface of the English Resource Grammar. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. 875–881.
- Flickinger, Dan, Stephan Oepen and Emily M. Bender. 2017. Sustainable development and refinement of complex linguistic annotations at scale. Dordrecht: Springer Netherlands. 353–377.



- Gajdošová, Katarína, Mária Šimková and et al. 2016. Slovak Dependency Treebank. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague.
- Glavas, Goran and Ivan Vulic. 2020. Is supervised syntactic parsing beneficial for language understanding? An empirical investigation. CoRR. abs/2008.06788.
- Hajič, Jan and Petr Zemánek. 2004. Prague Arabic dependency treebank: Development in data and tools. Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools. 110–117.
- Hajič, Jan, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek and Barbora Štěpánková. 2020. Prague Dependency Treebank – Consolidated 1.0. Proceedings of the 12th Language Resources and Evaluation Conference. 5208–5218.
- Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdenka Uřešová and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). 3153–3160.
- Harris, Zellig. 1954. Distributional structure. Word 10(2–3). 146–162.
- Haxby, James V., Elizabeth A. Hoffman and M. Ida Gobbini. 2000. The distributed human neural system for face perception. Trends in Cognitive Sciences 4(6). 223–233.
- Hebb, Donald Olding. 1949. The organization of behavior: A neuropsychological theory. New York, NY: John Wiley & Sons.
- Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 4129–4138.
- Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. CoRR. abs/2003.11080.
- Ibiyev, Kamran and Attila Novák. 2021. Using zero-shot transfer to initialize azWikiNER, a gold standard named entity corpus for the Azerbaijani language. In P. Sojka, I. Kopeček, K. Pala and A. Horák (eds.) Text, speech, and dialogue. Cham: Springer International Publishing. 305–317.
- Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics 5. 339–351.
- Kamp, Hans and Uwe Reyle. 1993. From discourse to logic: Introduction to model-theoretic semantics of natural language, formal logic and discourse representation theory (Studies in Linguistics and Philosophy 42). Dordrecht: Springer.
- Kaplan, Ronald M., John T. Maxwell, III, Tracy Holloway King and Richard Crouch. 2004. Integrating finite-state technology with deep LFG grammars. Proceedings of the ESSLLI'04 Workshop on Combining Shallow and Deep Processing for NLP.
- Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 423–430.
- Koehn, Philipp. 2009. Statistical machine translation, 1st edn. New York, NY: Cambridge University Press.
- Kondratyuk, Dan and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing



- and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2779–2795.
- Le, Quan Van, Lynne A. Isbell, Junpei Matsumoto, Minh Nguyen, Etsuro Hori, Rafael S. Maior, Carlos Tomaz, Anh Hai Tran, Taketoshi Ono and Hisao Nishijo. 2013. Pulvinar neurons reveal neurobiological evidence of past selection for rapid detection of snakes. *Proceedings of the National Academy of Sciences* 110(47). 19000–19005.
- Maudslay, Rowan Hall and Ryan Cotterell. 2021. Do syntactic probes probe syntax? Experiments with Jabberwocky probing. *CoRR*. abs/2106.02559.
- Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*. abs/1301.3781.
- Mikolov, Tomas, Quoc V. Le and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*. abs/1309.4168.
- Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Uřešová, Kateřina Veselá and Zdeněk Žabokrtský. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank: Annotation manual (Technical Report 30). ÚFAL MFF UK, Prague.
- Nedoluzhko, Anna, Michal Novák and Maciej Ogrodniczuk. 2018. PAWS: A multi-lingual parallel treebank with anaphoric relations. *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*. 68–76.
- Nemeskey, Dávid Márk. 2020. Egy emBERT próbáló feladat [A task testing emBERT]. XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020). 409–418.
- Nemeskey, Dávid Márk. 2021. Introducing huBERT. XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021). 3–14.
- Novák, Attila and Borbála Novák. 2018. Cross-lingual generation and evaluation of a wide-coverage lexical semantic resource. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 45–51.
- Novák, Attila and Borbála Novák. 2022a. Nerkor 1.41e. XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022). 389–402.
- Novák, Attila and Borbála Novák. 2022b. POS, ANA and LEM: Word embeddings built from annotated corpora perform better. *Computational linguistics and intelligent text processing*. Cham: Springer International Publishing.
- Pan, Xiaoman, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. *Proceedings of ACL 2017*. 1946–1958.
- Pollard, Carl and Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. Chicago, IL: The University of Chicago Press.
- Pullum, Geoffrey K. and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19(1–2). 9–50.
- Rakison, David H. and Jaime Derringer. 2008. Do infants possess an evolved spider-detection mechanism? *Cognition* 107(1). 381–393.
- Rosenblatt, Frank. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6). 386–408.
- Sachan, Devendra, Yuhao Zhang, Peng Qi and William L. Hamilton. 2021. Do syntax trees help pre-trained transformers extract information? *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2647–2661.



- Samuel, David and Milan Straka. 2020. ÚFAL at MRP 2020: Permutation-invariant semantic parsing in PERIN. *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, 53–64.
- Schweter, Stefan and Alan Akbik. 2020. FLERT: Document-level features for named entity recognition. *arXiv*. arXiv:2011.06993.
- Ševčíková, Magda, Zdeněk Žabokrtský and Oldřich Krůza. 2007. Named entities in Czech: Annotating data and developing NE tagger. In V. Matoušek and P. Mautner (eds.) *Text, Speech and Dialogue – 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007, Proceedings (Lecture Notes in Artificial Intelligence 4629 (Lecture Notes in Computer Science))*. Berlin & Heidelberg: Springer. 188–195.
- Sgall, Petr, Eva Hajičová and Jarmilla Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht: Reidel.
- Shatz, Marilyn. 2007. On the development of the field of language development. In E. Hoff and M. Shatz (eds.) *Blackwell handbook of language development*. John Wiley & Sons, Ltd. 1–15.
- Siklósi, Borbála. 2018. Using embedding models for lexical categorization in morphologically rich languages. *Computational linguistics and intelligent text processing*. Cham: Springer International Publishing. 115–126.
- Simon, Eszter and Noémi Vadász. 2021. Introducing NYTK-NerKor, a gold standard Hungarian named entity annotated corpus. In K. Ekštejn, F. Pártl and M. Konopík (eds.) *Text, Speech, and Dialogue – 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings (Lecture Notes in Artificial Intelligence (Lecture Notes in Computer Science 12848))*. Berlin & Heidelberg: Springer. 222–234.
- Simon, Eszter, Noémi Vadász, Dániel Lévai, Dávid Nemeskey, György Orosz and Zsolt Szántó. 2022. Az NYTK-NerKor több szempontú kiértékelése [A multi-faceted evaluation of NYTK-NerKor]. XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022). 403–416.
- Straková, Jana, Milan Straka and Jan Hajič. 2019. Neural architectures for nested NER through linearization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5326–5331.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. Attention is all you need. *CoRR*. abs/1706.03762.
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*. 353–355.
- Werbos, Paul John. 1994. *The roots of backpropagation: From ordered derivatives to neural networks and political forecasting*. New York, NY: Wiley-Interscience.
- Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.
- Wu, Qianhui, Zijia Lin, Börje F. Karlsson, Biqing Huang and Jian-Guang Lou. 2020. UniTrans: Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. 3926–3932.



- Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems* 32. 5753–5763.
- Zeman, Daniel and Jan Hajic. 2020. FGD at MRP 2020: Prague Tectogrammatical Graphs. *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*. 33–39.
- Zhen, Zonglei, Huizhen Fang and Jia Liu. 2013. The hierarchical brain network for face recognition. *Plos One* 8. e59886.

Open Access. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited, a link to the CC License is provided, and changes – if any – are indicated. (SID_1)

