Научная статья / Original article
УДК 81-114.2
doi:10.37632/PI.2024.69.51.003

О ВЛИЯНИИ ГЕНЕТИЧЕСКОГО РОДСТВА ЯЗЫКОВ НА КАЧЕСТВО МЕЖЪЯЗЫКОВОГО ПЕРЕНОСА АВТОМАТИЧЕСКОЙ СЕМАНТИЧЕСКОЙ РАЗМЕТКИ*

А.М. Ивойлова
Российский государственный гуманитарный университет

Аннотация: Данная статья посвящена анализу результатов zero-shot межъязыкового переноса автоматической лингвистической разметки в стандарте CoBaLD с русского языка на близкородственные и неродственные языки. Исследование показывает, что качество переноса зависит как от генетического родства языков, так и от количества обучающих данных для языков в используемой языковой модели.

Ключевые слова: автоматическая семантическая разметка, лингвистическая разметка, межъязыковой перенос

Для цитирования: Ивойлова А.М. О влиянии генетического родства языков на качество межъязыкового переноса автоматической семантической разметки // Типология морфосинтаксических параметров. 2024. Том 7, вып. 2. С. 43–59. (На английском.). doi:10.37632/PI.2024.69.51.003

ON THE INFLUENCE OF GENETIC RELATEDNESS OF LANGUAGES ON THE QUALITY OF CROSS-LINGUAL TRANSFER OF AUTOMATIC SEMANTIC ANNOTATION**

* Автор выражает искреннюю благодарность анонимным рецензентам и А.В. Циммерлингу за их ценные замечания и вопросы по теме данного исследования.

** The author expresses sincere gratitude to the anonymous reviewers and to A. Zimmerling who gave their notes and opinions on the topic.

Alexandra Ivoylova
Russian State University for the Humanities

Abstract: This paper is dedicated to the analysis of the results of zero-shot cross-lingual transfer of linguistic annotation in the CoBaLD standard from Russian to genetically related and unrelated languages. It appears that the quality of zero-shot cross-lingual transfer depends on genetic relatedness of donor and recipient languages as well as on the amount of training data for a language in the language model used as a backbone for the parser.

Keywords: automatic semantic annotation, linguistic annotation, cross-lingual transfer

For citation: Ivoylova A. On the influence of genetic relatedness of languages on the quality of cross-lingual transfer of automatic semantic annotation. *Typology of Morphosyntactic Parameters*. 2024. Vol. 7, iss. 2. Pp. 43–59. doi:10.37632/PI.2024.69.51.003

1. Introduction

Currently, modern linguistics increasingly incorporates methods involving computer technologies for research purposes. These methods include creating text corpora and developing machine learning algorithms to address both theoretical and practical tasks. On the one hand, no corpus can exist without at least the most basic linguistic annotation (typically part-of-speech tagging). On the other hand, information about the grammar and semantics of natural language can serve as additional features when solving Natural Language Processing (NLP) tasks.

According to the definition provided in [Ide 2017], “linguistic annotation involves associating descriptive or analytical tags with linguistic data.” Initially, linguistic annotation was used exclusively for theoretical research, primarily to confirm or refute linguistic theories, and was therefore closely tied to corpus creation. Evidently, to make linguistic annotation applicable in practice, it is necessary to establish unified standards for annotation. Any standard must be grounded in some theoretical framework, such as the “Meaning ⇔ Text” theory by I. Mel’čuk [Мельчук 1974] or the Head-Driven Phrase Structure Grammar by C. Pollard and I. Sag [Pollard, Sag 1994]. Linguistic annotation can describe one or more levels of language; almost always, the morphological level is included. However, particularly in recent years, syntactic and semantic annotation have become increasingly important for both applied NLP tasks and computational linguistics research.

Two significant issues currently attracting the attention of researchers in computational linguistics are, first, the question of how modern language models (Large Language Models, LLMs) “understand” natural language semantics, given their well-known “black-box” nature, and second, the challenges related to processing data for low-resource languages. For most applied tasks (and often theoretical ones as well), annotated data are essential, as supervised learning algorithms still demonstrate the highest performance. However, creating such data is a labor-intensive and expensive process. Consequently, annotated datasets are available for far fewer natural languages than needed, and even when they exist, they are often insufficient in volume or quality.

Thus, active research today focuses on developing methods that reduce resource requirements or leverage resources created for other languages. Beyond their practical significance, studies in automatic processing of low-resource languages, particularly those involving the technique of cross-lingual transfer, can provide additional insights into the typological characteristics of various languages and be of interest to linguists.

This study combines the two outlined areas and is aimed at conducting experiments on transferring automatic linguistic annotation in a recently proposed multi-level annotation format CoBaLD from Russian to genetically related and unrelated languages, specifically Bulgarian, Hungarian, Serbian, and Turkish, and to analyze the results.

The selection of languages for this study is guided by two main criteria. First, this is the genetic proximity of the target languages to Russian, which serves as the donor language. For instance, Bulgarian and Serbian belong to the Slavic language group, while Hungarian and Turkish, on the other hand, are non-Indo-European languages. Second, for experiments involving machine learning, the availability of language support in existing multilingual language models is crucial. For example, in the training data of XLM-RoBERTa (XLM-R) [Conneau 2019], the model used for experiments in this study, Hungarian and Bulgarian are represented in approximately equal volumes, whereas Serbian and Turkish have lower representation. Russian, meanwhile, ranks second in training data volume in XLM-R, following English.

The hypothesis is that the quality of transfer will be higher for (a) genetically related languages and (b) languages with larger training data volumes in the language model. The results of this study can also be of interest for theoretical studies in the field of linguistic typology as they provide an additional method of quantitatively determining the differences in languages.

2. Related work

Due to the multilayered nature and large number of parameters in neural networks (modern neural models can have billions of parameters), their training requires a sufficient quantity of data. Ideally, an algorithm is trained on a large amount of labeled data with a distribution similar to the test data [Zhuang et al. 2020]. However, in practice, such a scenario is rarely achievable. As data requirements grow, training and fine-tuning techniques that reduce these demands become increasingly relevant. For instance, semi-supervised learning can partially address this issue by using a mix of labeled and unlabeled data, though acquiring even unlabeled data may sometimes be challenging.

Currently, the most well-known and widely used technique for training models with minimal labeled data is transfer learning [Weiss et al. 2016]. This approach enables a model to leverage knowledge acquired during training on a different task, domain, or language.

Cross-lingual transfer (CLT) technique, thus, is a means widely used nowadays to develop NLP models for low-resourced languages.

There are two concepts related to knowledge transfer depending on the presence or absence of labeled data for the target task or domain. When labeled data is completely unavailable, it is referred to as *zero-shot transfer learning*, which involves applying a model trained on other data or tasks to new data or tasks. Conversely, when at least a small amount of labeled data is available, this is termed *few-shot transfer learning*, where fine-tuning is used. In this study the former technique was applied.

In the case of CLT, the primary distinction lies in the language of the data. The language of the source model's data is typically referred to as the *donor language*, while the language for transfer is called the *recipient language*.

Since the appearance of multilingual language models such as mBERT [Devlin et al. 2018] and XLM-RoBERTa the use of CLT became available, although there have been attempts to align monolingual embeddings for the task [Duong et al. 2016; Zhang et al. 2017; Artetxe et al. 2017]. One of the first attempts to thoroughly analyze the ability of these models for cross-lingual generalization were conducted by [Wu et al, 2019] and [Hu et al, 2020]; in the latter study it was revealed that the quality of zero-shot CLT normally doesn't drop below 25% for English as a donor language.

There have been many investigations in the area of CLT since then, including the question of language choice for a donor language [Lin et al. 2019; Er-

onen et al. 2023], analyzing the impact of linguistic features [Dolicki, Spanakis 2021] or using several languages as donors [Lim et al. 2024], which revealed that diverse donor languages result in more robust results.

One of commonly used downstream tasks for the evaluation of CLT in recent research is dependency parsing [Ahmad et al. 2021; Choenni et al. 2023]; the reason for this may be the availability of multilingual datasets annotated in the Universal Dependencies (UD) standard [de Marneffe et al. 2021]. There have also been attempts to transfer semantic annotation such as [Fei et al. 2020; Sherborne, Lapana 2021], but none of them took typological aspects into account.

There have also been a lot of investigations devoted to the impact of linguistic similarity on the quality of CLT [Philippy et al. 2023]; although mostly it is shown that genetic relatedness and geographical distance between languages does affect CLT, there is also research where it is shown that other factors influence it, e.g. the amount of pre-training data in the models [ibid.]. Other research shows that shared vocabulary of a model can also be important [Patil et al. 2022].

3. Linguistic annotation

The annotation standard used in this study is CoBaLD [Petrova et al. 2023], which describes all three language levels (morphology, syntax and semantics). It is a new annotation format based on prior work; for morphosyntactic annotation, it is based on UD, and its semantic part is adopted from the ABBYY Compreno model [Manicheva 2012]. In this research, the semantic level was the one investigated, but its connection to other language levels is interesting as well.

The aim of CoBaLD development was to make this format fully compatible with the most popular annotation standard, UD, and easy to use. For annotation representation, it uses the CONLL-U (.conllu) file standard, which contains ten columns in its basic version (CoBaLD uses the extended version of CONLL-U which contains twelve columns; the first ten of them are the same as in UD). The annotation is word-centric: CoBaLD treats “words” as the fundamental elements of a text, assuming they possess morphological properties and participate in syntactic relations. The developers of the UD standard claim to adhere to the principle of **lexical integrity** [Chomsky 1970], and suppose that words are linguistic units constructed from other structural elements and based on principles distinct from those of syntactic constructions. CoBaLD inherited these principles.

Each token represents a row in a CONLL-U table (see fig. 1). Basic token information and morphological data are stored in the first six columns.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS
1	It	it	PRON	Pronoun	Case=Nom|Gender=N	2	nsubj	2:nsubj
2	trivialises	trivialize	VERB	Verb	Mood=Ind|Number=S	0	root	0:root
3	them	they	PRON	Pronoun	Case=Acc|Number=Pl	2	obj	2:obj
4	.	.	PUNCT	PUNCT	_	2	punct	2:punct

Figure 1. An example of morphosyntactic annotation in CoBaLD

The first column of the format contains the token index (its sequential number in the sentence), the second column specifies the word form, and the next four columns represent the core morphological annotations.

Within the format, three categories associated with the morphological level are annotated: lemma, part of speech, and morphological features. The part of speech can be represented in two ways: the UPOS column contains one of the 17 universal part-of-speech tags standardized in UD, while the XPOS column can include designations from other tagsets if the annotation was converted from a different standard, or it can be filled with placeholders if not used. The FEATS column lists the morphological features relevant to the part of speech (universal features). These features are recorded in alphabetical order, with grammatical categories and their specific values (e.g., 'Aspect = Imp' indicates that the verb is in the imperfective aspect). The number of features may vary depending on the token.

The syntactic level of annotation in CoBaLD, as in UD, is represented by three columns following the morphological information: HEAD, DEPREL, and DEPS. The final column, MISC (not shown in fig. 1), typically contains technical details, such as the absence of a space following the token, spelling errors, or other auxiliary information.

The HEAD column specifies the index of the head (governing token) for the current token in the sentence. The DEPREL column indicates the dependency relation between the current token and its head. As for the DEPS column, it is used for additional annotation in the Enhanced UD standard [Schuster, Manning 2016]. Enhanced UD (E-UD) is an extension of the basic UD format, proposed by the same authors. In E-UD, some of the restrictions of the basic UD are lifted: for instance, the prohibition against having more than one root in a dependency tree and the prohibition of cyclic dependencies.

Additionally, E-UD allows for the annotation of certain types of ellipsis and includes the marking of phenomena such as referential relations (in a limited form) and raising/control structures.

Just like the original UD standard, CoBaLD is also presented in two versions, Base and Enhanced.

One of the main purposes for the development of CoBaLD format is the enrichment of UD with semantics. This standard is grounded in the formalism of the ABBYY Compreno model.

Historically, semantics has been the least explored level of natural language, although in contemporary computational linguistics, the primary focus is on semantics and how language models “understand” it. Given the variety of existing theories, the formalized representation of semantics can also vary significantly. This variability arises in part from the inevitable question of which specific characteristics of natural language we aim to represent.

In computational linguistics, two tasks directly related to semantic analysis are typically formulated: Semantic Role Labeling (SRL) and Semantic Parsing (SP). Among these, SRL is a more narrowly defined task, as it involves labeling only semantic roles.

The concept of semantic roles was largely shaped by the ideas of Charles Fillmore [Fillmore 1968; Fillmore 2006], Robert Van Valin [Van Valin 1993], and others. Semantic roles generally refer to the participants in predicate structures (i.e., verb arguments). Semantic role labeling attempts to recover the deep predicate structure, unlike syntactic dependency parsing, which annotates surface structure. Technically, these tasks are similar, as they both involve establishing categorized dependencies between words and are thus often solved using similar methods.

The SP task, however, is broader and may include SRL as a subtask. Semantic annotation is generally described as the construction of formalized, machine-readable representations of natural language semantics (typically at the sentence level, but sometimes at the text level). However, the concept of a formalized, machine-readable representation can be interpreted in various ways, leading to a wide range of approaches to semantic annotation.

The importance of semantic annotation for applied NLP tasks is clear: such annotation could enhance the quality of solutions to well-known tasks such as named entity recognition, sentiment analysis, text classification, dialogue system construction, information extraction, and more. However, the question of what exactly semantic annotation should include to be effective for practical

applications remains unresolved. As a result, there is no dominant semantic annotation format analogous to UD for morphosyntax, although some approaches are more popular than others — often because their creators continue to actively develop them.

The ABBYY Compreno model was a proprietary linguistic model developed by ABBYY over two decades for rule-based automatic machine translation. Today, development of the model has ceased due to the advent of new machine translation algorithms that eliminate the need for rule-based systems like Compreno.

Nevertheless, Compreno is not merely a rule-based translation system but also a sophisticated formalism suitable for linguistic annotation for various purposes, including theoretical research. Compreno is integrative, encompassing descriptions of morphological, syntactic, and semantic levels of language.

The semantic level of the model is the one adopted for CoBaLD. Despite its high level of detail in the original formalism, it is conceptually straightforward. Compreno represents semantics as a hierarchy of semantic classes (word meanings) connected by hyperonymy and hyponymy relations. Each class is associated with a specific set of deep slots (semantic roles) that it can assign to dependents. For example, the verb class **TO_GIVE** (*дать* and similar verbs) can assign dependents in roles such as **Agent**, **Possessor**, or **Object**.

The number of semantic classes in Compreno exceeds 200,000, and the number of deep roles (over 200) is also substantial, covering not only actants but also circumstantial roles.

Due to the proprietary nature of the model, no annotated datasets or automated parsers for the Compreno standard are publicly available. The only parser capable of producing analyses in the Compreno format is the rule-based parser owned by ABBYY. However, in 2022, the company released a simplified version of the semantic hierarchy for free academic use[1].

The primary feature of CoBaLD is the inclusion of a semantic annotation layer compatible with the UD formalism. Semantic information was integrated from the Compreno model but was reworked and simplified to enhance practical usability and compatibility. For instance, while the Compreno model contains additional information, the CoBaLD format retains only data on the semantic roles of tokens and their meanings (semantic classes) and the amount of those is also greatly reduced. In the Enhanced version of the format, there is

[1] https://github.com/compreno-semantics

also partial information about referential relations (in accordance with Enhanced UD principles). However, the CoBaLD format can potentially be extended to include additional information from the original Compreno model.

Thus, in the CONLL-U Plus format, which is also used for UD, two additional columns have been introduced. One column contains the semantic role category (referred to as the "deep slot" in Compreno formalism), and the other indicates the semantic class category. It is assumed that the heads of semantic relations align with the syntactic heads.

4. Data

The dataset used in this study for the source language is CoBaLD Rus[2] [Ivoylova et al. 2023], which consists of Russian news texts taken from NewsRu.Com. Its size is around 400,000 tokens and it was manually annotated in Compreno standard and then converted to CoBaLD. Additionally, an English dataset CoBaLD Eng[3] [Petrova et al. 2024] was used which contains approx. 180,000 tokens of BBC news texts.

The data used for CLT are the following:

- *Bulgarian*: UD Bulgarian BTB[4], which contains literary texts;
- *Hungarian*: UD Hungarian Szeged[5], which also contains news texts from different sources;
- *Serbian*: UD Serbian SET[6], comprised of news texts from SETimes;
- *Turkish*: Turkish news dataset taken from Kemik Group (Yıldız Technical University)[7].

5. Automatic linguistic annotation

In 2024, a neural network based parser capable to produce automatic annotation in both versions of CoBaLD standard was created [Баюк 2024]; its code[8] and trained models[9] are publicly available. This is a multi-task learning model

[2] https://github.com/CobaldAnnotation/CobaldRus
[3] https://github.com/CobaldAnnotation/CobaldEng
[4] https://github.com/UniversalDependencies/UD_Bulgarian-BTB
[5] https://github.com/UniversalDependencies/UD_Hungarian-Szeged
[6] https://github.com/UniversalDependencies/UD_Serbian-SET
[7] https://www.kaggle.com/datasets/furkanozbay/turkish-news-dataset
[8] https://github.com/CobaldAnnotation/CobaldParser
[9] https://huggingface.co/CoBaLD

which has a separate classifier for ellipsis restoration and several jointly trained heads for morphology, syntax and semantics prediction. Its overall F1-score is 93%; however, in order to assess the quality of CLT performed with it, one should analyze its errors on the source language data. For analysis, both English and Russian versions were used. The results show that there are errors related to deep slots, which may hypothetically arise from incorrectly determined semantic classes of heads. In the CoBaLD standard, the semantic head determines the permissible deep slots for its dependents. Nonetheless, in approximately 56% of cases, the parser assigns incorrect deep slots despite correct semantic classes. Notably, errors are most frequent in core arguments (e.g., **Experiencer** *vs* **Agent**) rather than adjuncts. This is likely because core arguments are typically distinguished only by the semantics of their head, whereas adjuncts are easier to identify based on context and structure.

During CLT qualitative analysis these errors weren't taken into account as linguistically non-specific.

6. Results

In order to get the results, a trained parser model for Russian (based on XLM-R) was taken and applied directly to the samples of the datasets for the chosen languages; no additional training was performed. Three of the UD datasets were already tokenized according to the UD principles, and the dataset for Turkish had been processed by UDPipe first, so that the UD principles should also apply to its tokenization. Afterwards, the labels were corrected manually by trained annotators (two annotators reviewed each dataset). The results of the annotation were discussed by the annotators and a specialist in Compreno semantics. Finally, the amount of hand-made corrections was automatically calculated and is presented in Table 1.

Table 1. The amount of hand-made corrections for the CLT results

	Bulgarian	**Serbian**	**Hungarian**	**Turkish**
Deep Slots	5.62%	8.49%	11.6%	14.8%
Semantic Classes	14.59%	13.34%	14.6%	14.9%

6.1. Bulgarian and Serbian

Both Bulgarian and Serbian belong to the South Slavic subgroup of languages. Bulgarian is a typical representative of the Balkan language area and exhibits features such as the presence of articles and the almost absent case system.

Serbian, historically situated on the periphery of the Balkan sprachbund, shares some similar characteristics, such as the lack of infinitive verb forms, but is considered non-Balkanized.

It is expected that CLT from Russian to these two languages would yield high-quality results. Indeed, manual validation of automatic annotation revealed that the correction rate for deep slots was only 8.49%, and for semantic classes, it was 13.34%; these figures are comparable for Bulgarian. Note that the parser's annotation quality for Russian reaches 90.8% for deep slots and 93.6% for semantic classes, indicating that the neural model performs better with semantic classes.

This difference in transfer quality can be explained quite straightforwardly. The CoBaLD framework includes certain semantic classes that are not entirely universal; they may either be absent in the data of a specific language or expressed in fundamentally different ways, such as articles, prepositions, and particles. CoBaLD assumes that a semantic class is assigned to every token except punctuation, which necessitates assigning a class even to these functional words. This situation does not arise with deep slots, which are more universal and may not apply to functional parts of speech.

In the case of Serbian, one reason for the relatively high number of manual corrections in semantic classes is the particle/conjunction *da*, which lacks a direct counterpart in Russian. *Da* can function similarly to Russian conjunctions such as *что* 'what' or *чтобы 'in order to'*, e.g., in sentences like *Iskreno se nadam da se to neće desiti* ('I sincerely hope that this won't happen'). However, it is also used as an untranslatable particle in syntactic structures equivalent to Russian constructions involving the infinitive, e.g., *zatvorska kazna mogla bi da zaplaši medije* ('a prison sentence could frighten the media') or *Da li očekujete i promenu ekonomske situacije u zemlji?* ('Do you also expect a change in the economic situation in the country?'); for more information on *da* refer to [Иванова 2022].

In the UD standard for Serbian, *da* is always tagged as SCONJ (subordinating conjunction) and typically linked with the syntactic relation *mark*, which usually indicates the relationship between a subordinating conjunction and the head of a subordinate clause. Notably, Bulgarian also has an analogous *da*, which, depending on the grammar, can be defined as either a particle or a conjunction (e.g., see [Маслов 1981]). For a comparison of *da*-constructions across South Slavic languages, including Bulgarian and Serbian, refer to [Иванова 2018].

When transferring annotations from Russian to Serbian or Bulgarian, the parser often annotates *da* as CONJUNCTIONS (i.e., a conjunction). However, in many cases, the semantic class is simply omitted (a placeholder is left) or sometimes duplicates the semantic class of a neighboring verb. Thus, a significant portion of manual corrections involves fixing issues related to *da*.

Both Bulgarian and Serbian also contain idiomatic expressions that lack direct parallels in Russian. For example, in Serbian, expressions like *bilo koja druga* ('any other'), *nakon što* ('after'), and *zato što* ('because') do not always have clear semantic class mappings (although deep slots tend to be more straightforward).

In conclusion, the quality of cross-linguistic transfer for deep slots to South Slavic languages is quite high, nearly comparable to the quality of automatic annotations for Russian. Errors made during transfer are consistent with errors observed in the source language. Regarding semantic classes, issues primarily arise for non-semantic phenomena (such as *da* and other functional words) or idiomatic expressions without direct analogues in the donor language.

6.2. Hungarian and Turkish

The Hungarian language belongs to the Uralic family (Ugric group). Among its key grammatical features distinguishing it from Russian are the presence of articles, a greater number of cases, and the use of postpositions. Much the same can be said about Turkish, which belongs to the Turkic family. In addition to these characteristics, both languages have a word order different from Russian: SOV in Turkish and OV constructions in Hungarian.

For Hungarian, the percentage of manual corrections for automatic annotations transferred from Russian was 11.6% for deep slots and 14.6% for semantic classes. For Turkish, the percentage of corrections for deep slots was slightly higher at 14.8%, while the percentage for semantic classes was comparable. These results are fairly predictable, given the size of the training data for these languages in the language model and their genetic distance from Russian.

As with the Slavic languages, the quality of semantic class transfer inevitably declines due to the presence of functional words, for which no semantic class can be assigned because they are absent from the parser's training data. For example, in Hungarian, the definite article *a* or *az*, especially in locative constructions, is often erroneously assigned the semantic class PREPOSITION — a

noteworthy error. In Turkish, the indefinite article *bir* is usually annotated with the semantic class CH_REFERENCE_AND_QUANTIFICATION and the deep slot **Quantity**, as the parser "confuses" it with the numeral *one*. It is worth noting that when annotations are transferred from English to Turkish, this issue with articles disappears, and the parser correctly assigns the class ARTICLES.

However, both Hungarian and Turkish face challenges related to the presence of postpositions and adjectives derived from them (in Hungarian). Postpositions are often misinterpreted as nouns and annotated accordingly. For instance, in the Turkish sentence *Dışişleri Bakanı Davutoğlu, Yunanistan ile Türkiye arasındaki farlılıkların ortak vizyon ile çözülebileceğini söyledi* ('Foreign Minister Davutoğlu stated that the differences between Greece and Turkey can be resolved with a shared vision'), the postposition *arasındaki* ('between') is annotated with the deep slot **Locative** and the semantic class CH_OF_CONNECTIONS. However, the correct annotation would omit the deep slot and assign the semantic class PREPOSITION which is used for functional words that express some relation between two content words such as nouns etc. (this label was adopted during manual validation for these languages and is used to denote postpositions, despite its somewhat misleading name).

Finally, in Turkish, the decline in annotation quality for deep slots can also be attributed to the presence of structures like the izafat. Errors frequently occur in izafat constructions, e.g., in *Lübnan'ın başkenti Beyrut'ta* ('In Beirut, the capital of Lebanon'), the token *Lübnan'ın* (Lebanon) should receive the deep slot **Whole**, but the parser assigns the deep slot **Locative** to all three tokens. Errors in deep slots within izafat constructions can vary but are clearly caused by insufficient embedding alignment.

7. Conclusion

As the number of hand-made corrections implies, not only genetic relatedness, but the amount of pre-training data in the language model influences the quality of CLT, besides, the annotation standard itself may have some flaws concerning the semantic classes (e.g., it is disputable if one should have semantic classes for functional words). The results of our study indicate that, firstly, certain features within the developed annotation standard itself may hinder transfer, and secondly, despite this, the quality of transfer is remarkably high, particularly for deep slots. This finding not only underscores the universality of

semantics as a linguistic level but also reinforces the adequacy of the CoBaLD standard for describing natural language semantics, although, naturally, the standard should be improved[10]. One of suggested improvements (by reviewers as well) is the introduction of multi-word semantic units; the developers have been discussing the means of labeling such units without the need to change the UD-compatible format. The other question is the connection between syntax and semantics in the standard, which may also influence the quality of automatic semantic annotation not only for CLT, but for source language parsing as well. A reviewer has reasonably argued that some of the principles of CoBaLD may be incorrect and semantics should not be strictly tied to syntax, but it is indiscrepancies in the current version of the format that most probably causes troubles (e.g., the unresolved difference in copula treatment in semantics and syntax which is inherited from Compreno and UD, respectively). Anyway, it is a disputable issue.

As for the possibility of using the proposed method as a means to measure the genetic relatedness of languages, it may yield stable results only if the compared languages have a similar amount of pre-training data for the language model used as a backbone. Besides, there are other factors which may influence the CLT quality. Still, the qualitative analysis of the CLT results can probably give some linguistic insights.

It is clear that the availability of even a small manually annotated dataset in the target language would enable few-shot CLT with minimal quality loss, provided the dataset includes the features described above. Additionally, a multi-source transfer strategy — where the model is initially trained on annotated corpora from multiple languages, such as English and Russian — appears highly promising.

References

Ahmad et al. 2021 — Ahmad W.U., Li H., Chang K.W., Mehdad Y. Syntax-augmented multilingual BERT for cross-lingual transfer. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, August 1–6, 2021. Pp. 4538–4554.

[10] One of the reviewers has suggested using lexical functions for enhancement of the standard, but in our opinion, the restrictions imposed on semantic classes with regard to which semantic slots they can fill are equivalent to these.

Artetxe et al. 2017 — Artetxe M., Labaka G., Agirre E. Learning bilingual word embeddings with (almost) no bilingual data. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017. Pp. 451–462.

Choenni et al. 2023 — Choenni R., Garrette D., Shutova E. Cross-Lingual Transfer with Language-Specific Subnetworks for Low-Resource Dependency Parsing. Computational Linguistics. 2023. Vol. 49. No. 3. Pp. 613–641.

Chomsky 1970 — Chomsky N. Remarks on nominalization. Jacobs R.A., Rosenbaum P.S. (eds.). Readings in English transformational grammar. Waltham, MA: Ginn, 1970. Pp. 184–221.

Conneau 2019 — Conneau A. Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451. July 5–10, 2020.

De Marneffe et al. 2021 — De Marneffe M.C., Manning C.D., Nivre J., Zeman D. Universal dependencies. Computational linguistics. 2021. 47(2). Pp. 255–308.

Dolicki, Spanakis 2021 — Dolicki B., Spanakis G. Analysing the impact of linguistic features on cross-lingual transfer. arXiv:2105.05975.

Duong et al. 2016 — Duong L., Kanayama H., Ma T., Bird S., Cohn T. Learning crosslingual word embeddings without bilingual corpora. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016. Pp. 1285–1295.

Eronen et al. 2023 — Eronen J., Ptaszynski M., Masui F. Zero-shot cross-lingual transfer language selection using linguistic similarity. Information Processing & Management. 2023. Vol. 60. No. 3. Article 103250.

Fei et al. 2020 — Fei H., Zhang M., Li F., Ji D. Cross-lingual semantic role labeling with model transfer. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020. Vol. 28. Pp. 2427–2437.

Fillmore 1968 — Fillmore C.J. The Case for Case. Bach E., Harms R. T. (eds.). Universals in Linguistic Theory. London: Holt, Rinehart and Winston, 1968. Pp. 1–25.

Fillmore 2006 — Fillmore C.J. Frame semantics. Cognitive linguistics: Basic readings. 2006. Vol. 34. Pp. 373–400.

Hu et al. 2020 — Hu J., Ruder S., Siddhant A., Neubig G., Firat O., Johnson M. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. International Conference on Machine Learning. PMLR, 2020. Pp. 4411–4421.

Ide 2017 — Ide N. Introduction: The handbook of linguistic annotation. Netherlands: Springer Netherlands, 2017.

Ivoylova et al. 2023 — Ivoylova A., Dyachkova D., Petrova M., Michurina M. The problem of linguistic markup conversion: the transformation of the Compreno markup into the UD format. International Conference on Computational Linguistics and Intellectual Technologies “Dialog”. 2023.

Lin et al. 2019 — Lin Y.H., Chen C.Y., Lee J., Li Z., Zhang Y., Xia M., Neubig G. Choosing transfer languages for cross-lingual learning. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. Pp. 3125–3135.

Manicheva et al. 2012 — Manicheva E., Petrova M., Kozlova E., Popova T. The Compreno Semantic Model as Integral Framework for Multilingual Lexical Database. Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon. 2012. Pp. 215–230.

Patil et al. 2022 — Patil V., Talukdar P., Sarawagi S. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland. 2022. Pp. 219–233.

Petrova et al. 2023 — Petrova M., Ivoylova A., Bayuk I., Dyachkova D., Michurina M. The CoBaLD Annotation Project: the Creation and Application of the full Morpho-Syntactic and Semantic Markup Standard. Proceedings of the International Conference “Dialogue”. 2023.

Petrova et al. 2024 — Petrova M.A., Ivoylova A.M., Tishchenkova A. CoBaLD Annotation: The Enrichment of the Enhanced Universal Dependencies with the Semantical Pattern. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024.

Philippy et al. 2023 — Philippy F., Guo S., Haddadan S. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada. 2023. Pp. 5877–5891.

Pollard, Sag 1994 — Pollard C., Sag I. A. Head-driven phrase structure grammar. University of Chicago Press, 1994.

Schuster, Manning 2016 — Schuster S., Manning C. D. Enhanced English universal dependencies: An improved representation for natural language understanding tasks. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. Pp. 2371–2378.

Van Valin 1993 — Van Valin Jr.R.D. A synopsis of Role and Reference Grammar. Advances in Role and Reference Grammar. Amsterdam: John Benjamins, 1993.

Weiss et al. 2016 — Weiss K., Khoshgoftaar T.M., Wang D. A survey of transfer learning. Journal of Big Data. 2016. Vol. 3. Article 1.

Wu et al. 2019 — Wu S., Conneau A., Li H., Zettlemoyer L., Stoyanov V. Emerging cross-lingual structure in pretrained language models. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2019. Pp. 6022–6034.

Zhang et al. 2017 — Zhang M., Liu Y., Luan H., Sun M. Adversarial training for unsupervised bilingual lexicon induction. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Volume 1: Long Papers. 2017. Pp. 1959–1970.

Zhuang et al. 2020 — Zhuang F., Qi Z., Duan K., Xi D., Zhu Y., Zhu H., et al. A comprehensive survey on transfer learning. Proceedings of the IEEE. 2020. Vol. 109. No. 1. Pp. 43–76.

Баюк 2024 — Баюк И.С. Создание трёхуровневого парсера для формата Enhanced CoBaLD. Магистерская диссертация, МФТИ, 2024. [Bayuk I.S. The Creation of a Three-Level Parser for the Enhanced CoBaLD Format. Master’s Thesis, MIPT, 2024.]

Иванова 2018 — Иванова Е.Ю. Да-конструкция как фактор синтаксической дифференциации славянских языков. // Славянское языкознание. XVI Международный съезд славистов. Белград, 20–27 августа 2018 г. Доклады российской делегации. М.: Институт славяноведения РАН, 2018. С. 171–205. [Ivanova E.Yu. The Da-Construction as a Factor of Syntactic Differentiation in Slavic Languages. Slavyanskoe yazykoznanie. XVI Mezhdunarodnyi sezd slavyanistov. Belgrad, August 20–27. M.: Institute of Slavic Studies RAS, 2018. Pp. 171–205.]

Иванова 2022 — Иванова Е.Ю. Балканославянская ирреальность в зеркале русского языка (южнославянские да-формы и их русские параллели). М.: Издательский Дом ЯСК, 2022. 288 с. [Ivanova E.Yu. Balkanoslavyanskaya irreal'nost' v zerkale russkogo yazyka (yuzhnoslavyanskie da-formy i ikh russkie paralleli). [Balkan-Slavic Irreality in the Mirror of the Russian Language (South Slavic da-Forms and Their Russian Parallels).] M.: Izdatel'skii Dom YASK, 2022. 288 p.]

Маслов 1981 — Маслов Ю.С. Грамматика болгарского языка. М.: Высшая школа, 1981. 407 с. [Maslov Yu.S. Bulgarian grammar. Moscow: Vysshaya Shkola, 1981. 407 p.]

Мельчук 1974 — Мельчук И.А. Опыт теории лингвистических моделей "Смысл – Текст". Семантика, синтаксис. 1974. 316 с. [Mel'čuk I.A. An Outline of the "Meaning – Text" Linguistic Models Theory. Semantika, sintaksis. 1974. 316 p.]

Статья поступила в редакцию 17.11.2024; одобрена после рецензирования 02.12.2024; принята к публикации 25.12.2024.
The article was received on 17.11.2024; approved after reviewing 02.12.2024; accepted for publication 25.12.2024.

Александра Михайловна Ивойлова
Российский государственный гуманитарный университет

Alexandra Ivoylova
Russian State University for the Humanities

a.m.ivoylova@gmail.com