Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2019"

Moscow, May 29—June 1, 2019

MULTILINGUAL PARALLEL CORPORA AS A SOURCE FOR QUANTITATIVE CROSS-LINGUISTIC GRAMMAR RESEARCH (THE CASE OF VOICE CONSTRUCTIONS)

Bonch-Osmolovskaya A. A. (abonch@gmail.com),
Nesterenko L. V. (lnesterenko@hse.ru)

National Research University "Higher School of Economics",
Moscow, Russia

Multilingual parallel corpora make possible the application of quantitative methods in cross-linguistic research. Due to the lack of appropriate resources, this has not become a widespread technique among linguists, but the studies based on this idea tend to emerge. In our work, we focus on the application of logistic regression for the research of passive voice constructions with an overtly expressed agent. The study is conducted on the data extracted from a multilingual parallel corpus that was created for this purpose. The issue we find noteworthy about voice alternation is the motivation for choosing active instead of passive, i.e. when a person would say 'This essay was written by Mary' instead of 'Mary wrote this essay'. Relying on theoretical studies, we selected a bunch of features claimed to be important for this kind of choice and used them for training logistic regression models. As a result, based on the model coefficients we can detect which features appear to be passive triggers.

Key words: multilingual parallel corpora, quantitative methods, logistic regression, feature selection, passive voice, semantic roles

МУЛЬТИЯЗЫЧНЫЕ ПАРАЛЛЕЛЬНЫЕ КОРПУСА КАК РЕСУРС ДЛЯ КВАНТИТАТИВНЫХ МЕЖЪЯЗЫКОВЫХ ИССЛЕДОВАНИЙ В ГРАММАТИКЕ (НА ПРИМЕРЕ ЗАЛОГОВЫХ КОНСТРУКЦИЙ)

Бонч-Осмоловская А. А. (abonch@gmail.com),
Нестеренко Л. В. (lnesterenko@hse.ru)

Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Ключевые слова: мультиязычные параллельные корпуса, логистическая регрессия, анализ признаков, страдательный залог, семантические роли

1. Introduction

Parallel corpora are known to be a powerful resource for research and NLP engineering. Mostly, they have been used in machine translation and development of multilingual applications. However, since linguists pointed out the benefits of parallel corpora usage in typological studies [Dahl 2007], [Wälchli 2007], researchers tend to set aside traditional techniques of collecting language data and perform their experiments on parallel texts. There are few typological studies, e.g. [Mayer, Cysouw 2012], [Cysouw 2014], [Östling 2015], [Östling 2016], [Asgari, Schütze 2017], [Bonch-Osmolovskaya, Nesterenko 2018], presenting application of quantitative methods to multilingual parallel data. Even though the idea of using parallel data and quantitative analysis gains popularity, the lack of suitable data, such as deeply annotated multilingual parallel corpora containing texts of different genres, slows down the progress in this field [Nesterenko 2019].

In this paper, we show how a multilingual parallel corpus and machine learning techniques can be used for the cross-linguistic analysis of a grammatical phenomenon, namely, passive voice constructions with overtly expressed agent. Despite the fact that there is vast variety of literature on voice constructions [Melchuk, Kholodovich 1970], [Keenan, Dryer 1981], [Frajzyngier 1982], [Shibatani 1985], [Comrie 1988], [Shibatani 1988], [Fox, Hopper 1994], [Givón 1994], [Kallulli 2006], [Tsunoda, Kageyama 2006], only a few papers include quantitative [Berez, Gries 2010] or corpus-based analysis [Jisa et. al. 2002], [Plecháčková 2007], [Xiao 2007].

The sentences like *Mary wrote this essay* and *This essay was written by Mary* represent the same situation that has different language encoding. The question is what linguistic factors motivate the passive occurence instead of active. We assume that different combinations of factors that might influence the choice between active and passive are relevant for different languages. Modelling of this problem demands the use of multilingual parallel corpora and the samples from the same context environment. That is important because it allows to evaluate how a fixed set of factors works for different languages. Using a corpus that contains parallel texts in 9 European languages and feature annotation tranfer prodedure, for each language we train a logistic regression model that predicts the probability of passive occurrence in particular context relying on a set of linguistic features. The resulting parameters of the models determine which features have stronger impact on the passive occurrence, and help to reveal differences between language strategies for choosing passive.

Logistic regression is widely applied for modelling of grammatical phenomena on monolingual data and annotation transfer is a common procedure in studies on parallel data, in our paper we combine these two methods, which, as far as we know, were previously used only in separation.

The paper is organized as follows. First, in **Section 2**, we discuss the factors that might influence the passive occurrence and determine a set of features used for training logistic regression models. Then, in **Section 3** we describe our multilingual corpus and emphasize some essential points in the process of its development. In **Sections 4** and **5** we describe the dataset and the models for the passive occurrence prediction. Finally, in **Section 6** we discuss the obtained results.

2. Factors determining the choice of voice construction

According to the framework of St. Petersburg typology group the crucial concept for voice description is diathesis. It is defined as “a pattern of mapping of semantic arguments onto syntactic functions (grammatical relations)” [Kulikov 2011]. The basic transitive diathesis has the following structure: the first (macro)role Actor is mapped onto the grammatical relation of Subject, while the semantic (macro)role Undergoer is mapped onto the grammatical relation of Direct Object (for the macroroles framework see [Foley & Van Valin 1984]).

Table 1. Active and passive diathesis

	Active			Passive	
Semantic argument level	X (Actor)	Y (Undergoer)	⇒	X (Actor)	Y (Undergoer)
Syntactic functional level	Subject	Direct Object		Direct Object/—	Subject

In this paper, we will consider only the passives with overtly expressed agent and leave agentless passives for future research.

The semantic features of the participants play an important role in voice alternation. There are examples which demonstrate that if there is no prototypical agent in a clause, then the initial active construction cannot be reformulated as passive with overtly expressed agent.

(1) a. *$250000 won't buy this kind of house any more.*
b. **This kind of house won't be bought by $250000 any more.*
[Shibatani 1985]

(2) a. *The refugees have {seen/witnessed} some traumatic events.*
b. *Some traumatic events have been {seen/witnessed} by the refugees.*
c. *{This country/The last decade} has {seen/witnessed} some traumatic events.*
d. **Some traumatic events have been {seen/witnessed} by {this country/the last decade}.*
[Langacker 2006]

In [Keenan, Dryer 1981] authors formulate a similar constraint based on the notion of a patient. They emphasize that the verbs without a patient are not easily passivizable. These assumptions indicate that besides other features we should focus on characteristics of the participants.

A detailed description of semantic roles and parameters that characterize participant relations and might be relevant for the explanation of voice alternation can be found in [Lehmann 2006]. Lehmann discribes the roles of participants regarding their empathy, involvement and control. The features involvement and control form a continuous scale where the roles can be placed. Involvement indicates if the situation is inconceivable without a particular participant, and control parameter presupposes that the participant has control over the situation or the participant is being controlled. The empathy parameter is based on the empathy hierarchy and adds another dimension to Lehmann’s system.

One of the generalizations in [Keenan, Dryer 1981] tells us that some languages, mostly Asian, like Chinese or Vietnamese, distinguish two types of passive constructions depending on negative vs. positive nature of subject affectedness. There are no indications if this might influence the choice of passive occurrence, but we include this factor in our set of features in order to check whether it has any importance for passives in European languages.

In Jespersen's list of factors motivating passive occurrence [Jespersen 1924], among others, we find "The passive turn may facilitate the connection of one sentence with another". Shibatani [Shibatani 1985], referring to [Dixon 1979], indicates that passives can create a syntactic pivot so that syntactic transformations can take place. For our study, we engage the idea of connection between sentences in two ways. First, we take into consideration the presence of contextual refereces to the verb arguments. Second, we pay attention to the fact whether the sentences are in contrast/opposition relation.

All in all, our set of features include semantic roles and their characteristics (empathy, involvement), characteristic of the verb meaning (whether it denotes a negative action), contextual fatures such as contrast/opposition relation between sentences and mention of the arguments in the previous context. The elaboration of features we describe in more detail in **Section 4**.

3. The Corpus

In our study, we use a corpus that consists of the collection of J. K. Rowling Harry Potter books series from 1 to 7 in 9 languages: English, German, Swedish, French, Italian, Spanish, Russian, Czech and Bulgarian. The size of the text data per language is about 1 million tokens. In order to be used productively in research, a multilingual parallel corpus should satisfy some requirements. First, an essential attribute of a parallel corpus is alignment, because it allows to compare the same parts of the texts in different languages and, if necessary, transfer their characteristics, e.g., from one pivot language to the others. In our corpus, texts are aligned pairwise at sentence and word level. For sentence alignment, we processed the texts with Gale & Church algorithm [Gale, Church 1991] and for the word alignment, we used the Efmaral toolkit [Östling, Tiedeman 2016]. Another thing, crucial for effective data extraction from a multilingual parallel corpus is morphological and syntax annotation in a unified format. The unified annotation format simplifies data extraction and the search for words or phrases with the same grammatical properties across all languages in the corpus. For these purposes, we used a UDPipe parser [Straka, Straková 2017], that runs on models trained for different languages within the Universal Dependencies tags set [Nivre et al. 2016].

For our experiment, we extracted corresponding active and passive sentences in 9 languages relying on the alignment and UDPipe annotation. After that, the English samples were manually annotated with the features we previously defined as relevant for active and passive distinction and this annotation was transferred to the data from all other languages. In the next section we describe the dataset and the feature annotation more precisely.

4. The Dataset

For our experiments (the methods itself we describe in **Section 5**), we designed the dataset according to the basic principles of machine learning. Sentences are considered as objects of classification, linguistic factors build up feature vectors for learning, and there are two class labels — active and passive.

Most of the sentences from the corpus have alignments in all languages of the sample, i.e., each translational unit consists of translational equivalents in 9 languages, and all of them are marked for the voice construction type. There are 236 fully aligned samples and there are partially aligned additional samples that we had to use in order to avoid high class imbalance in our the data. The distribution of active and passive constructions in the datasets and the total number of samples is presented in **Table 2**.

Table 2. The proportions of active and passive sentences by language

	EN	DE	SE	IT	ES	FR	RU	CZ	BG
Active (full alignment)	169	195	196	183	212	186	216	217	187
Passive (full alignment)	67	41	40	53	24	50	20	19	49
Additional passive (partial alignment)	103	23	52	53	50	51	40	37	52
Total	339	259	288	289	286	287	276	273	288

EN — English, DE — German, SE — Swedish, IT — Italian, ES — Spanish,
FR — French, RU — Russian, CZ — Czech, BG — Bulgarian

Linguistic features of the samples were annotated manually. The existence of fully aligned sets of the sentences allowed us to focus only on the English data and transfer the values of features to the data of the other languages. Alignment is not the only reason why we can easily project the features from one language to the other ones. The characteristics we use are applicable to the situation and its participants, i.e., those that are not affected by the translation and hold the same for all languages. Since our goal is not to distinguish between active and passive constructions, like in the task of morphological or syntax parsing, but to predict the probability of passive occurrence, we do not use features that directly indicate the presence of passive or active (e.g., oblique case, verb form).

The features we selected for our experiment refer to the semantic level of language representation. Formulating the features, we adopted the scheme of the basic transitive diathesis from [Kulikov 2011] and most of the features are assigned to X and Y.

Table 3. The basic transitive diathesis

Semantic argument level	X (Actor)	Y (Undergoer)
Syntactic functional level	Subject	Direct Object

There are two binary features that represent the ***semantic role status*** of X and Y. One feature encodes if X is a prototypical **actor** and the other one encodes if Y is a prototypical **undergoer**[1]**. For example, in a sentence** *John hit the dog*, 'John' is a prototypical agent and 'the dog' is a prototypical undergoer, in this case both features get 1 as their value. In the sentence *John saw the dog*, 'John' is an experiencer and 'the dog' is a theme, both features are of value 0.

For encoding of ***involvement*** feature we used idea of continuum proposed in [Lehmann 2006], we decided to assign the value of 1 to the most central participants and 0 value was assigned to the most peripheral participants. The participants located in the middle of continuum got values between 1 and 0, depending to which end of the continuum they were closer. Involvement is assigned for both X and Y.

For the ***empathy*** encoding we used the hierarchy proposed in [Lehmann 2006], see **Figure 1**.

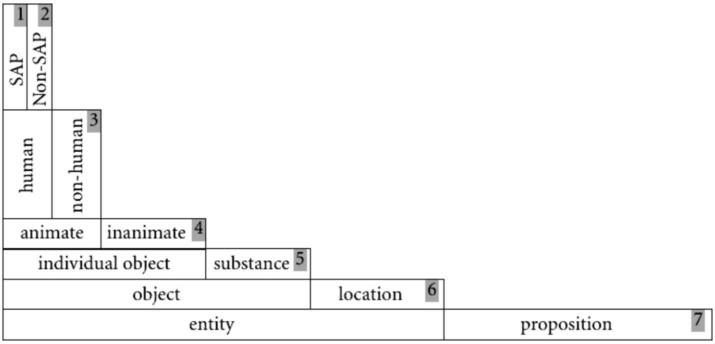

Fig. 1. The empathy hierarchy [Lehmann 2006]

According to this hierarchy there are seven degrees of empathy, where 1 is the highest value that is attested to speech act participants and 7 is the lowest value attested to propositions.

Besides the empathy itself, we include two features that express relation of the X empathy and the Y empathy**.** The first one is the ***absolute difference of X and Y empathy values***, the other one is the ***ratio of the absolute difference and the sum of X and Y empathy values***.

We also include a feature that represents if a verb has a ***negative*** component in its meaning, i.e. verbs like *attack* or *interrupt* get 1 as a feature value.

The next two features are a contextual, they encode, whether the arguments of the verb ***were mentioned*** in 3 preceding sentences, again, separetely for X and Y.

(3) When my sister was six years old, she was attacked, set upon, by three Muggle boys.

The ***contrast*** feature unlike the other ones is related to the situation as a whole. Let us consider some examples, which represent the cases of a positive value for this feature.

(4) a. If you continue behaving like that, *you will be punished by the headmaster.*
b. She tried to reach him, but *he was swallowed by the darkness*.

[1] This coincides with the scheme of control feature described in [Lehmann 2006].

If the target situation is somehow opposed to the other situation or two situations are in relation of contrast then this feature gets a value of 1 and 0 otherwise.

As a result, we end up with a list of twelve features:

1) X is actor
2) X's empathy
3) X was mentioned
4) X's involvement
5) Y's empathy
6) Y was mentioned
7) Y' involvement
8) Y's undergoer
9) Contrast
10) Empathy distance
11) Empathy distance 2
12) Negative action

Further we describe the models trained on these features and show, how they influence the choice of passive occurrence in the languages from our sample.

5. The Models

The goal of this experiment is to determine which of the factors appear to be the most important triggers for the passive occurrence, and to what extent the languages from our sample differ in their mechanism for choosing between active and passive constructions. To meet that goal, for each language in the sample we train a logistic regression model that predicts the probability of passive construction occurrence.

To avoid overfitting, for evaluation we used stratified cross-validation with 10 folds and checked these results by the permutation test [Ojala, Garriga 2010], which shows if the classification result is significant compared to a model trained on data with randomly assigned class labels. Since the datasets for most of the languages are unbalanced we decided to use the "class_weight" parameter of logistic regression function from the sklearn library in Python and assign weights to classes. In table 4 we present the evaluation results.

Table 4. Cross-validation results with permutation test scores

Metric	EN	DE	SE	IT	ES	FR	RU	CZ	BG
F1(macro average)/ Accuracy	0.88/ 0.88	0.82/ 0.85	0.83/ 0.85	0.85/ 0.86	0.79/ 0.82	0.83/ 0.86	0.7/ 0.77	0.72/ 0.77	0.84/ 0.85
PTS p-value	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009

PTS — Permutation test score, EN — English, DE — German, SE — Swedish, IT — Italian, ES — Spanish, FR — French, RU — Russian, CZ — Czech, BG — Bulgarian

All the results have low p-values of the permutation test, it indicates that our results are significant and not random. The F1 and accuracy values for Russian and Czech seem to be rather low, what questions the reliability of the results we get for these models. That might be explained by the fact of the dataset unbalance. However, the German data have similar class ratio but the results are much better, which makes us think that Russian and Czech data probably need different set of features for modelling this problem.

The Figure 2 represents the feature coefficients distribution in the languages, we included only those features that appeared to be statistically significant in the models, i.e., they had p-values < 0.05[2]. The insignificant features are assigned to a zero coefficient value.

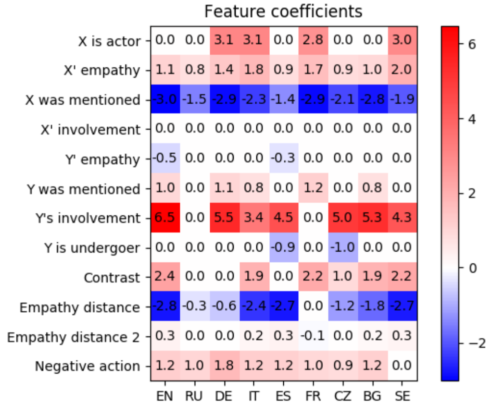

Figure 2. The model coefficients

The model coefficients reflect the impact of the features: the positive coefficient tells us that the feature contributes to the passive occurrence positively, negative coefficients indicate a positive impact on active occurrence. The features with values close to zero have very little impact on the classification.

In the final section we discuss the results and make conclusions.

6. Discussion and conclusion

As we can see from the Figure 2 the languages reveal similar coefficient distributions, but still they are not the same. Definitely there are features that have no impact in all or almost all languages, those are *X involvement*, *Y empathy*, *Y was mentioned*, *Y undergoer*, *Empathy distance 2*. The features relevant for majority of the languages are *Y involvement* (very strong), *X empathy* and *Negative action,* which are the triggers for passive occurrence, and also X was *mentioned*, which is a pro-active feature. The feature *X actor* is a strong passive trigger in German, Swedish, Italian, and French,

[2] The features Y involvement in Italian and Y empathy in Spanish were nearly significant but increased the evaluation results, so we also included they in the final models.

Contrast is a pro-passive feature for English, Italian, French, Czech, Bulgarian and Swedish, and *Empathy distance* is an pro-active feature for English, Italian, Spanish, Czech, Bulgarian and Swedish.

At this point, we do not make any claims about the nature of passives, and the results should be treated as preliminary. We demonstrated, how logistic regression can be applied to the multilingual parallel data for exploration of a language phenomenon. The determined set of language-independent features appeared to be relevant for prediction of passive occurrence in the majotiry of the considered languages. Except for Russian and Czech, the models have relatively good predictive power (the mean accuracy rate is about 82–85%). The robustness of current models should be tested under other conditions. The future research should include similar experiments based on 1) larger and more balanced datasets 2) multilingual parallel texts of other genres (e.g., movie subtitles, reports of Europarlament) 3) non-parallel corpus data. The current set of features may reveal different results for that data.

References

1. *Asgari E., Schütze H.* (2017), Past, Present, Future: A computational investigation of the typology of tense in 1000 languages, https://arxiv.org/abs/1704.08914
2. *Berez A. L. and Gries S. T.* (2010), Correlates to middle marking in Dena'ina iterative verbs. International Journal of American Linguistics, 76(1), pp.145–165.
3. *Bonch-Osmolovskaya A. A., Nesterenko L. V.* (2018), Networks as an instrument for "searches" and "findings" in multilingual parallel corpora [Seti kak instrument poiska i nakhodok v multijazychnykh parallelnykh korpusakh], EVRika! Papers on "searches" and "findings" to the anniversary of E. V. Rakhilina [EVRika! Sbornik statej o poiskakh i nakhodkakh k jubileju E. V. Rakhilinoj], Labirint, Moskva, pp. 305–320.
4. *Comrie B.* (1988). Passive and voice. Passive and voice, 16, pp. 9–24.
5. *Cysouw M.* (2014), Inducing semantic roles. Perspectives on semantic roles. Perspectives on semantic roles. Luraghi S., Narrog H. (eds.). Amsterdam: Benjamins, pp. 23–68.
6. *Dahl Ö.* (2007), From questionnaires to parallel corpora in typology. STUF-Sprachtypologie und Universalienforschung, 60(2), pp. 172–181.
7. *Foley W. A., Van Valin, R. D. Jr.* (1984), Functional Syntax and Universal Grammar. Cambridge: Cambridge University Press.
8. *Fox B. A., Hopper P. J.* (1994), Voice: Form and function (Vol. 27). John Benjamins Publishing.
9. *Frajzyngier Z.* (1982), Indefinite agent, passive and impersonal passive: a functional study. Lingua, 58(3–4), pp. 267–290.
10. *Gale W. A., Church K. W.* (1991), Identifying Word Correspondences in Parallel Texts. HLT, 91, pp. 152–157.
11. *Givón T.* (1994). Voice and inversion (Vol. 28). John Benjamins Publishing Company.
12. *Guyon I., Weston J., Barnhill S., Vapnik, V.* (2002), "Gene selection for cancer classification using support vector machines", Mach. Learn., 46(1–3), pp. 389–422.

13. *Jespersen, O.* (1924), The Philosophy of Grammar. London: Allen & Unwin.
14. *Jisa H., Reilly J., Verhoeven L., Baruch E., Rosado E.* (2002), Passive voice constructions in written texts: A cross-linguistic developmental study. Written Language & Literacy, 5(2), pp. 163–81.
15. *Kallulli D.* (2006), A unified analysis of passives, anticausatives and reflexives. Empirical issues in formal syntax and semantics, 6, pp. 201–225.
16. *Keenan, E. L., Dryer, M. S.* (1981). Passive in the world's languages. Linguistic Agency, University of Trier.
17. *Kulikov L.* (2011), Voice typology. In The oxford handbook of linguistic typology. Oxford University Press, pp. 368–398.
18. *Lehmann, C.* (2006), Participant roles, thematic roles and syntactic relations. Voice and Grammatical Relations: In Honor of Masayoshi Shibatani, 65, pp. 153–174.
19. *Langacker, R. W.* (2006), Dimensions of defocusing. Voice and grammatical relations: In Honor of Masayoshi Shibatani, pp. 115–137.
20. *Manninen S., Nelson D.* (2004), What is a passive? The case of Finnish. Studia Linguistica, 58(3), pp. 212–251.
21. *Mayer T., Cysouw M.* (2012), Language comparison through sparse multilingual word alignment. Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH. Association for Computational Linguistics, 2012, pp. 54–62.
22. *Melchuk I. A. & Kholodovich A. A.* (1970), On the theory of voice [K teorii grammaticheskogo zaloga] Asian and African studies [Narody Azii i Afriki], 4, pp. 111–124.
23. *Nesterenko L. V.* (2019), Multilingual parallel corpora: Alternative resource of language data for typological studies, usage perspectives and problems [Multijazychnye parallelnye korpus: novyj istochnik dannykh dlya tipologicheskikh issledovanij, perspektivy ispolzovanija i problemy], Problems of Linguistics [Voprosy jazykoznanija], accepted
24. *Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Haji J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D.* (2016), Universal dependencies v1: A multilingual treebank collection. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pp. 1659–1666.
25. *Ojala M., Garriga, G. C.* (2010), Permutation tests for studying classifier performance. Journal of Machine Learning Research, 11(Jun), pp. 1833–1863.
26. *Östling R.* (2015), Word order typology through multilingual word alignment. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Vol. 2: Short papers, pp. 205–211.
27. *Östling R.* (2016), Studying colexification through massively parallel corpora. The lexical typology of semantic shifts, 58, pp. 157–176.
28. *Östling, R., Tiedemann, J.* (2016). Efficient word alignment with markov chain monte carlo. The Prague Bulletin of Mathematical Linguistics, 106(1), pp. 125–146.
29. *Plecháčková J.* (2007), Passive Voice in Translation: A Corpus-Based Study (MA dissertation), https://is.muni.cz/th/ew9h5/diplomka_final_version.pdf
30. *Shibatani M.* (1985), Passives and related constructions: A prototype analysis. Language, pp. 821–848.

31. *Shibatani M.* (1988), Passive and voice (Vol. 16). John Benjamins Publishing.
32. *Straka M., Straková J.* (2017), Tokenizing, POS-tagging, lemmatizing and parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 88–99.
33. *Tsunoda T., Kageyama T.* (2006). Voice and grammatical relations: in honor of Masayoshi Shibatani (Vol. 65). John Benjamins Publishing.
34. *Wälchli B.* (2007), Advantages and disadvantages of using parallel texts in typological investigations. STUF — Sprachtypologie und Universalienforschung, 60(2): 118–134.
35. *Xiao R. Z.* (2007), What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English. Indonesian JELT, 3(1), pp. 1–19.