Testing the Syntactic Competence of Large Language Models with a Translation Task

Dative Ambiguity in Russian

Edyta Jurkiewicz-Rohrbacher[1,2]

[1]Universität Hamburg, Mittelweg 177, 22222, Hamburg, Germany
[2]Universität Regensburg, Universitätsstr. 34, 93333, Regensburg, Germany

Abstract

The paper explores opportunities for using a translation task to obtain knowledge about the syntactic competence of large language models. It reports the accuracy achieved in a Russian–English translation task on Russian sentences containing highly ambiguous structures with two dative personal pronouns. Seven tools (systems and agents) based on pre-trained generative models were tested in their function as machine translators on a data set obtained from several web corpora. The study shows that the principles of reference assignment relevant to the syntax of human language users (referential prominence and linear order of pronouns and predicates) are also statistically relevant for pre-trained generative models.

Keywords

syntax, ambiguity, translation task, linguistics, linguistic competence of large language models,

1. Introduction

The rapid developments in generative pre-trained language models have resulted in agents that deliver relatively well-formed texts in various natural languages. The economic result of this process is a large number of cheaply produced but relatively well-written machine-generated texts (MGT) freely circulating and spreading online. For linguists this means that language users are being exposed to automatically generated content on an equal footing with human-generated content. The language varieties emerging from MGTs are thus quite naturally becoming an object of linguistic research next to human varieties such as slangs, dialects, idiolects, etc. Consequently, linguistics as a discipline is facing new challenges pertaining to the methods through which knowledge about artificially emerging lects can be obtained. The new question before the linguistic community is: Can the established corpus-, psycho- and neurolinguistic methods be applied in research on rapidly emerging LLMs? In general, tasks intuitively formulated as instructions, where the input has a similar structure to the output are better processed in zero-shot prompts than tasks presented in other ways, for example, as finishing an incomplete sentence [1].This study aims to explore to what extent using translation, a method well-known from typological questionnaires, can be applied to explore the syntactic competence of LLMs. In the subsections that follow, translation as a task and a selected phenomenon of dative case ambiguity in Russian are described. Section 2 presents the study design. The central quantitative results are provided in Section 3, while minor results, which might feed into future studies, are presented in Section 4.

1.1. Translation task

Translation has been used as a data elicitation task in typological and psycholinguistic research in various ways. In linguistic fieldwork, translational questionnaires are frequently constructed to examine how a particular area of grammar with a known representation in language A is represented by its

4th Workshop on Humanities-Centred Artificial Intelligence 2024 (CHAI 2024)
edyta.jurkiewicz-rohrbacher@uni-hamburg.de (E. Jurkiewicz-Rohrbacher)
https://www.slm.uni-hamburg.de/slavistik/personen/jurkiewicz-rohrbacher.html (E. Jurkiewicz-Rohrbacher)
0000-0001-6737-7847 (E. Jurkiewicz-Rohrbacher)

native speakers in language B [cf. 2]. In psycholinguistics, translation is in itself an object of study as a cognitive process [3], but it is also used as a method for accessing the linguistic performance and competence of multilingual speakers, e.g., for exploring their multilingual lexicons [4].

Previous studies [5, 6] suggest that pre-trained generative models do capture syntactic information. However, accessing this information seems computationally demanding, and due to various practical reasons, impossible in the case of very large, commercially developed models. To address this, the present study employs a translation task to access knowledge of the principles governing syntactic parsing by having various types of pre-trained systems or agents perform a translation task.

1.2. Test Case: Dative Ambiguity in Russian

Recent reports show that neural machine translation (NMT) systems still have shortcomings in the area of co-reference resolution and lexical cohesiveness, which results in inaccurate translation of pronouns [7]. Syntactically ambiguous structures pose another type of problem [8], which I assume is challenging for the correlates of syntactic constituency parsing that might be found in generative pre-trained language models. A typical example of such structure is the prepositional phrase attachment, as in the often-cited sentence *A man saw a woman with a telescope*, where the phrase *with a telescope* can be parsed as an attribute of *a woman* or as an adjunct to the predicate *saw*. Nevertheless, ambiguity is an inherent feature of natural languages. Some scholars propose that it is a desirable quality because it facilitates efficient, that is, short and simple communication [9]. To explore the feasibility of using translation as a task in research on the syntactic competence of large language models, I examined ambiguous Russian structures containing two personal pronouns in the dative case placed adjacently in a complex sentence.

Although Slavic languages have extremely flexible word order, ambiguity in syntactic role assignment is rather rare because of their rich morphology. However, when two arguments have identical lexico-grammatical properties and the same morphological marking on the sentence surface, ambiguity is possible even in the case of two full NPs, as shown in example (1).

(1) *Miškata vižda kotkata.*
mouse.F.SG.DEF see.IPFV.PRS.3SG cat.F.SG.DEF
'The mouse sees the cat / The cat sees the mouse' (Bg, [10])

Studies on the Russian dative case with infinitive structures [11, 12, 13, 14, 15] mention in passing that the co-occurrence of two dative arguments in one sentence is possible, being predominantly observed in sentences with a *free infinitive*.[1] Such sentences are ambiguous because in several types of Russian clauses the dative case is not assigned solely to the syntactic role of indirect object (third argument), but also to the so-called 'logical subject',[2] as shown in (2):

(2) *Mne zvonit' nekomu.*
me.DAT call.INF nobody.DAT
Reading 1: 'Nobody should call me.'
Reading 2: 'I have nobody to call.'

It is suggested [13] that such structures might generally be avoided in language use. However, where they occurred, semantic-syntactic role assignment would follow the linear principle, correlating with the syntactic hierarchy of arguments (Agent over Recipient or other Participant). Others [14] claim that a word order of dative arguments which is at variance with the syntactic hierarchy is marked only

[1] A sentence where the main predicate is expressed as an infinitive.

[2] There is no general agreement as to which syntactic role should be assigned to such datives; for a recent review of the topic see [16]. I refrain from generalization on this matter here, as the present study also involves object control structures where the syntactically highest dative argument carries the syntactic role of indirect object in the matrix predicate, but at the same time also assigns the semantic role of the non-overt subject in the complement clause.

prosodically. Therefore, such structures could pose a challenge for large language models that are not trained on acoustic data.

Another work [15] argues that the context clarifies role assignment. For example, in (3), it is the negative personal pronoun *nekomu* 'to/for nobody', and not the linearly first dative *mne* 'to/for me' which is higher in the syntactic structure, and therefore more subject-like.

(3) *Mne zvonit' nekomu - ja i ne slušaju.*
me.DAT call.INF nobody.DAT I FOC NEG listen.1SG
'Nobody calls me so I'm not listening (for the telephone).' [I. Grekova. *Letom v gorode* (1962)] (after [15])

Finally, scholars have yet to provide an overview of structures in which two dative arguments interact in one sentence in Russian, and limit themselves to at least overtly one-predicate infinitive structures.[3] Hence, the order of predicates governing dative arguments is usually neglected as a factor.

A study on adjacent dative pronouns in Russian natural data originating from written text corpora [21] establishes that such structures do occur in language use, albeit mostly in *overtly bipredicative* structures, in combination with embedded infinitival complements, as shown in example (4).[4]

(4) *Istočnik takže upominaet nekotorye interesnye spekuljacii otnositel'no planov Intel*
source also mention.3.SG some interesting speculations regardint plan.GEN.PL I.
i NVIDIA, no im_2 nam_1 by $chotelos'_1$ $posvjatit'_2$ otdel'nyj material.
and N. but them.DAT us.DAT COND wish.REFL.SG dedicate.INF material
The source also mentions some interesting speculations regarding the plans of Intel and NVIDIA, but we would like to dedicate a separate article to them.'

The linearly first dative pronoun *im* 'them' is governed by the infinitival complement *posvjatit'* 'dedicate', while the linearly second and adjacent pronoun *nam* 'us' is governed by the complement-taking matrix predicate *chotelos'* 'wish.REFL'. Note that this sentence does not contain an explicit subject in the nominative.

According to the analysis in question [21], two factors significantly impact the probability of obtaining different word orders of arguments: the order of the main and embedded predicate, and the type of referential prominence that the pronouns represent, locuphoric pronouns (first/second person) being more likely to be assigned the agentive role than aliophoric ones (third person) [22].

For better comprehension the model is shown in Figure 1. Considerable variation is observed in sentences with an infinitival complement preceding the matrix verb in the linear order of the sentence (CM type marked on the abscissa). In such environments, two locuphoric pronouns are more likely to comply with the deep syntactic order of the predicates rather than with the shallow order suggested by the surface.

In combinations with at least one aliophoric pronoun, the picture is more complex (marked with red circle). Sentences where the referential prominence hierarchy is retained show the highest variation in pronoun order, as the probabilities of the two word orders occurring are nearly equal. When the referential prominence hierarchy is violated (an aliphoric pronoun takes a high position in syntax), the pronoun order corresponding to the order of the predicates on the surface is preferred, and so is significantly more likely to occur. Nonetheless, without context it is impossible to distinguish between these two conditions, which is a source of ambiguity.

It may be predicted that in a translation task, adjacent dative pronouns in Russian structures with embedded infinitives would be a source of error for LLMs. A particularly high error rate is to be expected for sentences with an infinitival complement preceding the matrix predicate in the surface linear order of a sentence, and sentences with a combination of a locuphoric and an aliophoric pronoun.

[3] It is unclear whether sentences with a free infinitive are monoclausal [17, 18] or biclausal [19, 20]. In the latter case, scholars assume the existence of a copula, which is not marked overtly in present- tense sentences.

[4] https://overclockers.ru/hardnews/show/93720/kazhdyj-kvartal-sledujuschego-goda-budet-prinosit-novye-graficheskie-resheniya-amd

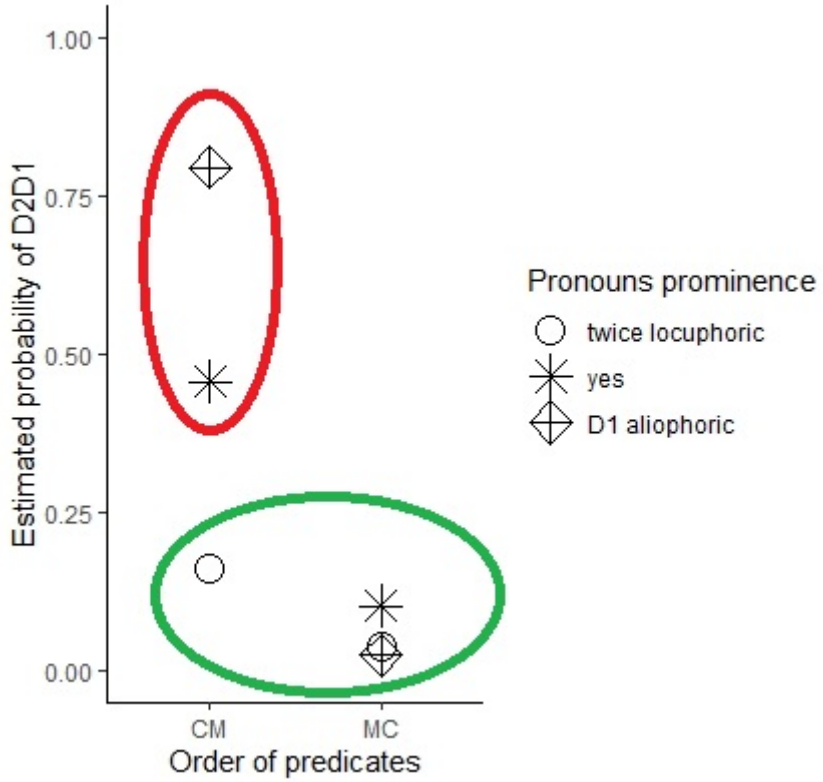

Figure 1: Probability of obtaining a reversed order of pronominal arguments in bipredicative structures with two dative arguments [21]. Red marked items are predicted to be source for error in interpretation for LLMs.

2. Study Design

In this section I describe the translation task conducted in the study. The primary sources of data were the Russian Timestamped JSI web corpus 2014-2021 [23] and the ruTenTen17 corpus [24], from which I extracted 74 stimuli excerpts of 200–1100 characters each.[5] Every excerpt contained a sentence with a two-predicate structure,[6] in which the embedded predicate (complement) was placed earlier in the linear structure of the sentence than the embedding predicate (matrix), and which contained two adjacent personal pronouns in the dative case, one locuphoric and the other aliophoric (see example 5

Further in the paper I use the following notation: M stands for matrix predicate (syntactically higher predicate), C for complement predicate (syntactically lower predicate, embedded by M), D1 for dative pronoun governed by M, D2 for dative pronoun governed by C.

(5) *Vodički prosili prostoj, a nam*$_{D1}$ *im*$_{D2}$ *dat'*$_{C}$ *bylo nečego*$_{M}$*.*
water.GEN ask.3PL simple.GEN CONJ us.DAT them.DAT give.INF AUX.3SG nothing.GEN
'They asked for plain water, but we had nothing to give them.'

The obtained data set was used to test the performance of three specialized translation tools based on neural network architectures, DeepL[7], Google Translate[8], Yandex[9], and four chatbots with similar

[5]I applied a CQL query for two adjacent pronominal lowercase word forms using the Sketch Engine corpus manager. The obtained data were manually controlled by a native speaker annotator and controlled for error by a second native speaker annotator. Because the context was always available, the task was usually straightforward. Still, for the current task only such sentences were chosen that did not raise any doubts in either of the annotators.
[6]For instance subject or object control constructions, predicatives, or modal-existential wh-predicates.
[7]https://www.deepl.com, licenced account
[8]https://translate.google.com, free user account
[9]https://translate.yandex.com, free user account

architectures: Google Gemini[10], Perplexity AI[11], ChatGPT Turbo and ChatGPT Omni[12].

The choice of commercial tools has clear drawbacks. First, the exact details of commercial models' architectures and the structure of the training data are not disclosed. Second, the computations performed by the models cannot be controlled, nor can the models be fine-tuned to improve accuracy of performance. However, training methods are beyond the scope of this study. My objective was to verify to what extent errors in MGTs can be predicted on the basis of statistically significant regularities detected in the behavior of language users. The selected pre-trained generative agents are all based on encoder-decoder architectures. This paper treats them similarly to human agents in usage-based theories of language acquisition [25]. In these theories, language acquisition is possible not thanks to universal grammar but is based on cognitive skills, in particular intention-reading and pattern-finding. Since the latter is clearly relevant to pre-trained generative models, I assume that their linguistic competence is emergent [26].

The training data represents the performance of multiple competent language users and is fed during the training process, from which the linguistic competence emerges, with the difference that each model has been exposed to a much larger amount of linguistic (written) data than any human agent can ever be. Knowledge about the linguistic competence of LLMs is accessed indirectly in this study, by evaluating performance in a specially developed translation task, just as it is done when the linguistic competence of human beings is studied. Therefore, for the selection of tools high competence had a greater priority than control over a language model.

Another important argument in favor of commercial models was that typologically interesting Slavic languages, characterized by rich morphology and very flexible word order, are still rarely available in open-source multilingual models such as the Llama family.[13] Although ambiguity per se is not rare in language, keeping as many factors as possible constant, and thus focusing on only one type of ambiguity, leads to considerable data reduction. The construction examined in this study is rather complex and relatively rare. Therefore, correct performance requires big computational capacities and state-of-art technologies.

In the period 05-18.06.24, the sentences in their authentic contexts were fed into the translation systems as chunks of 200–800 characters. The size of each chunk depended on the place in the text where the context necessary for disambiguation was located. It was mainly found either before or after the tested sentence. In rare cases, both the pre- and post-context were necessary for an unambiguous interpretation of pronouns. The chatbots were zero-shot prompted with the command "Translate the following passage from Russian to English: [passage]".[14] For each stimulus, the chat was restarted and no feedback regarding performance was given. In this way, 518 observations were obtained.

In the study, I focused only on the correct assignment of syntactic roles to the dative pronouns, as rendered in the process of translation. Other types of translation errors were disregarded.

3. Results

Table 1 demonstrates that the dative pronoun disambiguation task is not straightforward and that error rates vary primarily between the specialized translation systems and the agents. The best-performing ChatGPT Omni achieved an accuracy of nearly 0.95, while all the other chatbots had a strikingly similar accuracy of 0.89. The best translation system performed under this rate, reaching an accuracy of 0.85. The other two systems showed far poorer accuracy: 0.74 in the case of Google Translate and 0.67, of Yandex.

Interestingly, single instances of misclassification were observed only for translation systems, but not for generative agents. In other words, if an agent misclassified, there existed a system that misclassified too.

[10] https://gemini.google.com/app , free user account
[11] https://www.perplexity.ai/, free user account
[12] Access to both models via https://uhhgpt.uni-hamburg.de provided via the Universität Hamburg's license.
[13] The newest Llama 3.2 supports officially only English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai.
[14] Some prompts had to be modified for Google Gemini, which sometimes failed to perform the task for various reasons.

Table 1
Accuracy in the translation task.

Translation Systems	**DL**	**Google**	**Yandex**	
Accuracy	0.85	0.74	0.66	
Chatbots	**Omni**	**Turbo**	**Gemini**	**Perplexity**
Accuracy	0.95	0.89	0.89	0.89

Table 2
Distribution of the studied factors.

Incorrectly classified			Correctly classified		
Prominence hierarchy	**no**	**yes**	Prominence hierarchy	**no**	**yes**
Word order			word order		
D1D2	24 (0.49)	43(0.17)	D1D2	25	216
D2D1	14 (0.15)	2(0.02)	D2D1	77	117

In order to find out whether the word order and the prominence hierarchy principle were considerable factors, the data were annotated for these two features. The distributions of errors across them are presented in Table 2.

I observe that D2D1 word order with kept prominence hierarchy is clearly the easiest to classify causing barely any errors.

It may be observed that the D2D1 word order which preserved the prominence hierarchy was clearly the easiest to classify, causing barely any errors. Violation of one of the principles led to a considerable deterioration in the models' performance. I observed a decrease in accuracy when either the surface word order of the pronouns did not replicate the surface order of the predicates or the pronouns did not comply with the prominence hierarchy, i.e., when an aliophoric pronoun was higher in the syntactic hierarchy than a locuphoric one. However, if both of these principles were violated, the language models seemed to act quite randomly: performance dropped to 0.51.

A logistic-regression model with mixed-effects[15] [27] performed in R Environment [28], where stimuli and translator were treated as random effects, confirmed the intuition formulated above. The model shows that both factors (word order and prominence hierarchy) play a significant role in modeling the performance of the studied LLMs (cf. Table 3), and they have positive impact on performing the reference assignment in the translation task. and have a positive impact on correct reference assignment in the translation task. According to the studied model, the probability that sentences violating both principles will be accurately classified is 0.55. Sentences complying with both principles have a probability of 0.99 of being classified correctly. For sentences where only the hierarchy prominence is violated, the model predicts correct reference assignment with a probability of 0.93, while for sentences violating only the surface word order correspondence between governors and pronouns the respective number is 0.94.

4. Discussion and Future Prospects

The results obtained in the study are in line with prior predictions that "sentences with reversed order of predicates (CM), where two dative pronouns represent different levels of the prominence hierarchy can pose interpretation problems for NMT systems and other tools for NLP" [21]. It appears that in such structures, contextually available information might not be sufficient for correct disambiguation by a machine; for example, the key features might not be identified, as in sentence (6)[16] which was misclassified by six out of seven translators:

[15]Formula: Correct ∼ Order + hierarchy + Order*hierarchy + (1 | SentID) + (1|Translator)
[16]https://viktorkotl.livejournal.com/167122.html

Table 3
Results of the performed regression model

Random effects: Groups	Name	Variance	Std.Dev.	
SentID (Intercept)	3.5459	1.8831		
Translator (Intercept)	0.9743	0.9871		
Fixed effects:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.2053	0.9090	0.226	0.82133
D2D1	2.3771	1.0624	2.237	0.02525*
D1 locuphoric	2.5861	0.9511	2.719	0.00655*
D2D1:D1 locuphoric	0.9296	1.5092	0.616	0.53791

(6) *Ja ne kriču i ne zovu na pomošč' – ėto bessmyslenno: rebjata naverchu, dostatočno*
I NEG shout.1SG and NEG call on help it pointless kids top enough
daleko ot menja, kinut'$_C$ im$_{D1}$ mne$_{D2}$ nečego$_M$ (a esli by i ***kinuli***
far from me throw.INF them.DAT me.DAT nothing CONJ if COND FOC throw.PST.3PL
verëvku, to čem za neë uchvatit'sja?).
rope then INS for she.ACC catch.INF.REFL
'I'm not shouting or calling for help – it's pointless: the guys upstairs are far enough away from me, they have nothing to throw to me (and even if **they threw** a rope, how would I grab it?).'

Both hierarchical prominence and concordance of the word order of the governors and pronouns turned out to be relevant factors that might facilitate or hinder the task of disambiguation for the purpose of role assignment, for instance if there are no contextual cues. It should be pointed out that typically, sentences with two dative pronouns do not contain a nominative phrase, which in traditional syntax would be interpreted as the canonical subject. Consequently, such sentences most likely place a greater burden on processing, assuming that a correlate of syntactic parsing emerges in LLMs [5, 6]. In other words, phenomena studied in theoretical linguistics and typology seem to be relevant and retrievable from the linguistic behavior of large language models, also through established linguistic methods. Although the way pre-trained language models process language is still comparable to a black box, I argue that methods used to study the linguistic behavior of the human species can be adjusted to studying the linguistic behavior of machines. If not human natural language users can be considered a black box too, since linguistic knowledge is never directly accessible. Another interesting finding of this study is that models pre-trained for performing various tasks communicated within a conversation performed better than specially trained machine translators. This result should by no means suggest that agents are in general better than translation systems at performing translation tasks, as only one particular aspect was evaluated. Nonetheless, in the future it should be examined whether this observation holds for other phenomena and what the reason might be. I cannot rule out that it is related to the number of parameters, the size of the context window, or the type of training data. Nevertheless, it is important to repeat that in this study, agents misclassified only structures which were misclassified by systems, that is, a subset of stimuli misclassified by systems and not a disjoint set of stimuli. This suggests the systematicity of errors made by the agents. Note that for all stimuli, disambiguation was always potentially possible due to the available context. Presumably, agents use context better than translation systems do. This could be due to the fact that agents are trained to be multifunctional. Contextually given information is necessary to perform other types of tasks and therefore better used, also in translation. Multilingualism is in a sense a byproduct. Verifying this claim would certainly require further studies on context processing in reference resolution tasks.

Limitations

Gicen that ambiguity is an intrinsic feature of natural language, this phenomenon is pertinent to the processing of any natural language by machines. This paper focuses on syntactic ambiguity which is associated with flexible word order and languages where a single morpheme can be used to encode various syntactic arguments. The Slavic branch is an illustrative example, where the scope of roles encoded by the dative is particularly extensive. This case study demonstrates that translation tasks can be employed to evaluate the capabilities of LLMs in a systemic manner and can serve as a foundation for future research.

The present study was limited to commercial products, which does not allow for evaluation of improvements on the training set. Moreover, the current tasks might permit improvement of the studied tools and thus the obtained results might not be replicable in the future.

To an extent, the study is limited by the small set of test sentences (which will be enlarged in the future) and neglecting of the contextual factors. However, the point of the study was to ascertain whether translation tasks can bring insights into the linguistic competences of LLMs. Furthermore, the same problem might be relevant for automatic dependency annotation.

Finally, the study was limited only to the automatic text processing tools only. Although it would be possible to perform a similar study with human language users would be possible, I do not expect that the results obtained in the same or similar task would surpass the best performing ChatGPT Omni. The set of stimuli is cognitively quite demanding. Therefore, I assume that in an artificial experimental setting, language users would base their choices on the two, linguistic principles discussed in this paper, rather than spending time on re-analyzing the full context because it is cognitively more demanding. However, it is precisely for this reason that I postulate that some notional linguistic rules, such as those observed in the theoretical and general linguistics, might be common to humans and machines,notwithstanding the fact, that one would expect the latter group to make fewer mistakes and to follow logic (provided by the context).

In addition, human users, unlike chatbots (agents), could make mistakes in different stimuli than the erroneous translation systems, which is an interesting result of this study worth further investigation.

Ethics Statement

This work complies with the ACL Ethics Policy. Prior to the current study, I had not taken any actions to pretrain the systems for the needs of the current task.

Acknowledgments

The research has been partly supported by the Representative for Equal Opportunities and Academic Research Sabbatical Fund of the University of Regensburg. I thank Roman Fisun, Konstanzia Lüke and Irina Maykova for help with the preparation of the data set.

References

[1] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, 2022. URL: https://arxiv.org/abs/2109.01652. arXiv:2109.01652.

[2] Ö. Dahl, From questionnaires to parallel corpora in typology, Sprachtypologie und Universalienforschung 60 (2007) 172–181.

[3] A. Ferreira, J. W. Schwieter (Eds.), Psycholinguistic and Cognitive Inquiries into Translation and Interpreting, John Benjamins, Amsterdam, 2015.

[4] T. M. Włosowicz, Some applications of translation to psycholinguistic research, Linguistica Silesiana 33 (2012) 127–145.

[5] T. Limisiewicz, D. Mareček, Syntax representation in word embeddings and neural networks – a survey, in: M. Holeňa, T. Horváth, A. Kelemenová, F. Mráz, D. Pardubská, M. Plátek, P. Sosík (Eds.), Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020), CEUR-Workshop Proceedings, Košice, Slovakia, 2020, pp. 38–48.
[6] H. Zhao, A. Panigrahi, R. Ge, S. Arora, Do transformers parse while predicting the masked word?, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 16513–16542. URL: https://aclanthology.org/2023.emnlp-main.1029. doi:10.18653/v1/2023.emnlp-main.1029.
[7] A. Lopes, M. A. Farajian, R. Bawden, M. Zhang, A. F. T. Martins, Document-level neural MT: A systematic comparison, in: A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, M. L. Forcada (Eds.), Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 225–234.
[8] R. Bawden, Going beyond the sentence: Contextual Machine Translation of Dialogue, Ph.D. thesis, LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, France, 2018.
[9] S. T. Piantadosi, H. Tily, E. Gibson, The communicative function of ambiguity in language, Cognition 122 (2012) 280–291. doi:https://doi.org/10.1016/j.cognition.2011.10.004.
[10] M. Korytkowska, Gramatyka konfrontatywna bułgarsko-polska, volume V, Slavistyczny Ośrodek Wydawniczy, Warsaw, 1992.
[11] J. D. Apresjan, L. L. Iomdin, Konstrukcija tipa negde spat': sintaksis, semantika, leksikografija, Semiotika i informatika (1989) 34–92.
[12] D. Weiss, Infinitif et datif en polonais contemporain: un couple malheureux?, in: S. Karolak (Ed.), Complétude et incomplétude dans les langues romanes et slaves. Actes du VI Colloque international de linguistique romane et slave, Cracovie 29 sept.–3 oct. 1991, Cracow, 1993, pp. 443—-487.
[13] F. Maurice, Der modale Infinitiv in der modernen russischen Standardsprache, Peter Lang, Munich, 1996.
[14] A. Bonč-Osmolovskaja, Konstrukcii s dativnym subjektom v russkom jazyke, PhD thesis, Moscow: MGU, 2003.
[15] E. V. Padučeva, Otricatel'nye mestoimenija-predikativy (na ne-), in: Russkaja korpusnaja grammatika, accessed 26.02.2022, 2015. URL: http://rusgram.ru/new/chapter/pos/pronoun/#label_otritsateljnye_mestoimeniya-predikativy__na_ne-_.
[16] B. Hansen, Subject, in: M. L. Greenberg (Ed.), Encyclopedia of Slavic Languages and Linguistics Online, 2020. URL: http://dx.doi.org/10.1163/2589-6229_ESLO_COM_032471.
[17] A. Zimmerling, Dva tipa dativnych predloženij v russkom jazyke, in: Slovo – čistoe vesel'e: Sbornik statej v čest' A. B. Pen'kovskogo M., Jazyki slavjanskich kul'tur, Moscow, 2009, pp. 471–489.
[18] E. Tsedryk, Dative infinitive constructions in Russian: Are they really biclausal., in: B. Wayles, D. Miloje, N. Enzinna, S. Harmath de Lemos, R. Karlin, D. Zec (Eds.), Formal approaches to Slavic Linguistics 25 : The third Cornell meeting 2016, Michigan Slavic Publications, Ann Arbor, 2018, pp. 298–317.
[19] I. Livitz, Modal possessive constructions: Evidence from Russian, Lingua 122 (2012) 714–747.
[20] N. Kondrashova, R. Šimík, Quantificational properties of neg-wh items in Russian, in: Proceedings of North East Linguistics Society 40, Graduate Linguistic Student Association, Amherst: University of Massachusetts, 2013, pp. 15–28.
[21] E. Jurkiewicz-Rohrbacher, Dative ambiguity in Russian: A corpus induced study, Journal of Linguistics/Jazykovedný casopis 74 (2023) 70–80. doi:doi:10.2478/jazcas-2023-0025.
[22] M. Haspelmath, Role-reference associations and the explanation of argument coding splits, Linguistics 59 (2021) 123–174. URL: https://doi.org/10.1515/ling-2020-0252. doi:doi:10.1515/ling-2020-0252.
[23] M. Trampuš, B. Novak, The internals of an aggregated web news feed, in: Proceedings of

15th Multiconference on Information Society 2012 (IS-2012), 2012. URL: http://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf.
[24] M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, V. Suchomel, The TenTen corpus family, in: 7th International Corpus Linguistics Conference CL, 2013, pp. 125–127.
[25] M. Tomasello, Constructing a Language: A Usage-Based Theory of Language Acquisition, Harvard University Press, Harvard, 2003.
[26] P. J. Hopper, Emergent grammar, Berkeley Linguistics Society (1987) 139–157.
[27] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, Journal of Statistical Software 67 (2015) 1–48. doi:10.18637/jss.v067.i01.
[28] R. C. Team, R: A language and environment for statistical computing, 2017. URL: https://www.R-project.org/.

A. Online Resources

The full list of stimuli and their translations is available via GitHub.