Linguistic evaluation of German-English Machine Translation using a Test Suite

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel and Hans Uszkoreit
German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
firstname.lastname@dfki.de

Abstract

We present the results of the application of a grammatical test suite for German→English MT on the systems submitted at WMT19, with a detailed analysis for 107 phenomena organized in 14 categories. The systems still translate wrong one out of four test items in average. Low performance is indicated for idioms, modals, pseudo-clefts, multi-word expressions and verb valency. When compared to last year, there has been a improvement of function words, non verbal agreement and punctuation. More detailed conclusions about particular systems and phenomena are also presented.

1 Introduction

For decades, the development of Machine Translation (MT) has been based on either automatic metrics or human evaluation campaigns with the main focus on producing scores or comparisons (rankings) expressing a generic notion of quality. Through the years there have been few examples of more detailed analyses of the translation quality, both automatic (HTER (Snover et al., 2009), Hjerson (Popović, 2011)) and human (MQM Lommel et al., 2014). Nevertheless, these efforts have not been systematic and they have only focused on few shallow error categories (e.g. morphology, lexical choice, reordering), whereas the human evaluation campaigns have been limited by the requirement for manual human effort. Additionally, previous work on MT evaluation focused mostly on the ability of the systems to translate test sets sampled from generic text sources, based on the assumption that this text is representative of a common translation task (Callison-Burch et al., 2007).

In order to provide more systematic methods to evaluate MT in a more fine-grained level, recent research has relied to the idea of test suites (Guillou and Hardmeier, 2016; Isabelle et al., 2017). The test suites are assembled in a way that allows testing particular issues which are the focus of the evaluation. The evaluation of the systems is not based on generic text samples, but from the perspective of fulfilling a priori quality requirements.

In this paper we use the DFKI test suite for German→English MT (Burchardt et al., 2017) in order to analyze the performance of the 16 MT Systems that took part at the translation task of the Fourth Conference of Machine Translation. The evaluation focuses on 107 mostly grammatical phenomena organized in 14 categories. In order to apply the test suite, we follow a semi-automatic methodology that benefits from regular expressions, followed by minimal human refinement (Section 3). The application of the suite allows us to form conclusions on the particular grammatical performance of the systems and perform several comparisons (Section 4).

2 Related Work

Several test suites have been presented as part of the Test Suite track of the Third Conference of Machine Translation (Bojar et al., 2018a). Each test suite focused on a particular phenomenon, such as discourse (Bojar et al., 2018b), morphology (Burlot et al., 2018), grammatical contrasts (Cinkova and Bojar, 2018), pronouns (Guillou et al., 2018) and word sense disambiguation (Rios et al., 2018). In contrast to the above test suites, our test suite is the only one that does such a systematic evaluation of more than one hundred phenomena. A direct comparison can be done with the latter related paper, since it focuses at the same language direction. Its authors use automated methods to extract text items, whereas in our test suite the test items are created manually.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 644–653
Florence, Italy, August 1-2, 2019.

3 Method

The test suite is a manually devised test set whose contents are chosen with the purpose to test the performance of the MT system on specific phenomena or requirements related to quality. For each phenomenon a subset of relevant test sentences is chosen manually. Then, each MT system is requested to translate the given subset and the performance of the system on the particular phenomenon is calculated based on the percentage of the phenomenon instances that have been properly translated.

For this paper we use the latest version of the DFKI Test Suite for MT on German to English. The test suite has been presented in (Burchardt et al., 2017) and applied extensively in last year's shared task (Macketanz et al., 2018b). The current version contains 5560 test sentences in order to control 107 phenomena organised in 14 categories. It is similar to the method used last year, with few minor corrections. The number of the test instances per phenomenon varies, ranging between a 20 and 180 sentences. A full list of the phenomena and their categories can be seen as part of the results in the Appendix. An example list of test sentences with correct and incorrect translations is available on GitHub[1].

3.1 Construction and application of the test suite

The construction and the application of the test suite follows the steps below, also indicated in Figure 1:

(a) Produce paradigms: A person with good knowledge of German and English grammar devises or selects a set of source language sentences that may trigger translation errors related to particular phenomena. These sentences may be written from scratch, inspired from previous observations on common MT errors or drawn from existing resources (Lehmann et al., 1996).

(b) Fetch sample translations: The source sentences are given as an input to easily accessible MT systems and their outputs are fetched.

(c) Write regular expressions: By inspecting the MT output for every given sentence, the annotator writes rules that control whether the output contains a correct translation regarding the respective phenomenon. The rules are written as positive or negative regular expressions, that signify a correct or an incorrect translation respectively.

Lexical Ambiguity	
Das Gericht gestern Abend war lecker.	
The court last night was delicious.	fail
The dish last night was delicious.	pass
Conditional	
Er würde einkaufen gehen, wenn die Geschäfte nicht geschlossen hätten.	
He would go shopping if the stores didn't close.	fail
He would go shopping if the shops hadn't closed.	pass
Passive voice	
Es wurde viel gefeiert und getanzt.	
A lot was celebrated and danced.	fail
There was a lot of celebration and dancing.	pass

Table 1: Examples of passing and failing MT outputs

(d) Fetch more translations: When the test suite contains a sufficient number of test items with the respective control rules, the test suite is ready for its broad application. The test items are consequently given to a large number of MT systems. This is done in contact with their developers or through the submission process of a shared task, as is the case described in this paper.

(e) Apply regular expressions: The control rules are applied on the MT outputs in order to check whether the relevant phenomena have been translated properly. When the MT output matches a positive regular expression, the translation is considered correct (*pass*) whereas when the MT output matches a negative regular expression, the translation is considered incorrect (*fail*). Examples can be seen in Table 1. In case an MT output does not match either a positive or a negative regular expression, or in case these contradict to each other, the automatic evaluation results in a uncertain decision (*warning*).

(f) Resolve warnings and refine regular expressions: The *warnings* are given to the annotator, so that they manually resolve them and if possible refine the rules to address similar cases in the future. Through the iterative execution of steps (e) and (f) (which are an extension of steps (c) and (d) respectively) the rules get more robust and attain a better coverage. If needed, the annotator can add full sentences as rules, instead of regular expressions.

For every system we calculate the phenomenon-specific translation accuracy as the the number of the test sentences for the phenomenon which were translated properly, divided by the number of all test sentences for this phenomenon:

[1] https://github.com/DFKI-NLP/TQ_AutoTest

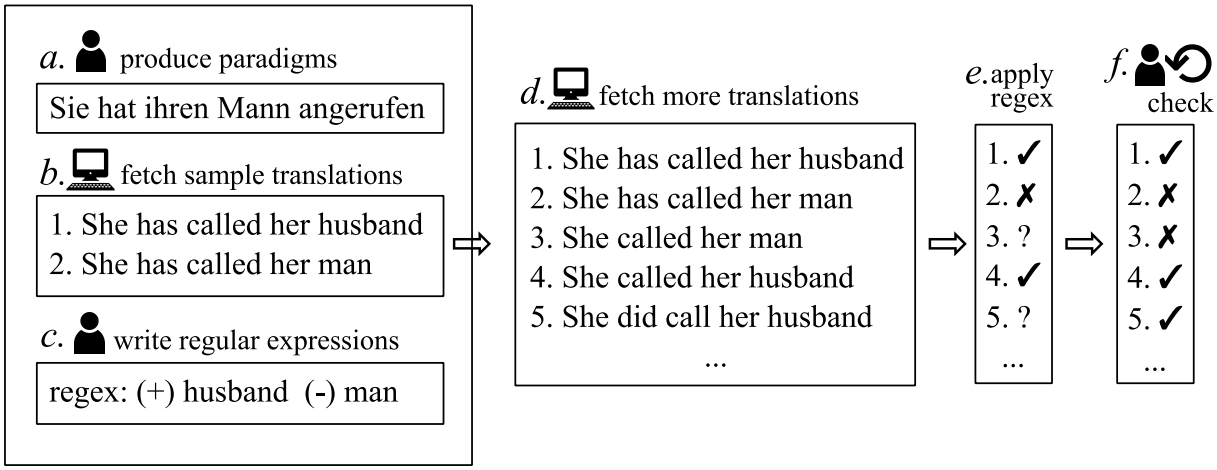

Figure 1: Example of the preparation and application of the test suite for one test sentence

$$\text{accuracy} = \frac{\text{correct translations}}{\text{sum of test sentences}}$$

When doing comparisons, the significance of every comparison is confirmed with a one-tailed Z-test with $\alpha = 0.95$.

3.2 Experiment Setup

In the evaluation presented in the paper, MT outputs are obtained from the 16 systems that are part of the *news translation task* of the Fourth Conference on Machine Translation (WMT19). According to the details that the developers have published by the time this paper is written, 10 of the systems are declared to be Neural Machine Translation (NMT) systems and 9 of them confirm that they follow the Transformer paradigm, whereas for the rest 6 systems no details were given. For the evaluation of the MT outputs the software TQ-AutoTest (Macketanz et al., 2018a) was used.

After processing the MT output for the 5560 items of the test suite, the automatic application of the regular expressions resulted to about 10% warnings. Consequently, one human annotator (student of linguistics) committed about 70 hours of work in order to reduce the warnings to 3%. The final results were calculated using 5393 test items, which, after the manual inspection, did not have any warning for any of the respective MT-outputs.

Since we applied the same test suite as last year, this year's automatic evaluation is profiting from the manual refinement of the regular expressions that took place then. The first application of the test suite in 2018 resulted in about 10-45% of warnings depending on the system, whereas after this year's application, we only had 8-28%. This year's results are therefore based on 16% more valid test items, as compared to last year.

4 Results

The results of the test suite evaluation can be seen in Tables 3 and 4, where the significantly best systems for every category or phenomenon are boldfaced. The average accuracy per system is calculated either based on all test items (with the assumption that all items have equal importance) or based on the categories (with the assumption that all categories have equal importance). In any case, since the averages are calculated on an artificial test suite and not on a sample test set, one must be careful with their interpretation.

4.1 Linguistic categories

Despite the significant progress of NMT and the recent claims for human parity, the results in terms of the test suite are somewhat mediocre. The MT systems achieve 75.6% accuracy in average for all given test items, which indicates that one out of four test items is not translated properly. If one considers the categories separately, only five categories have an accuracy of more than 80%: **negation**, where there are hardly any mistakes, followed by **composition**, **function word**, **subordination** and **non-verbal agreement**. The lowest-performing categories are the **multi-word expressions** (MWE) and the **verb valency** with about 66% accuracy.

4.2 Linguistic phenomena

Most MT systems seem to struggle with **idioms**, since they could only translate properly only 11.6% of the ones in our test set, whereas a similar

situation can be observed with resultative predicates (17.8%). **Negated modal pluperfect** and **modal pluperfect** have an accuracy of only 23-28%. Some of the phenomena have an accuracy of about 50%, in particular the domain-specific terms, the pseudo-cleft sentences and the modal of pluperfect subjunctive II (negated or not). We may assume that these phenomena are not correctly translated because they do not occur often enough in the training and development corpora.

On the other side, for quite a few phenomena an accuracy of more than 90% has been achieved. This includes several cases of verbs declination concerning the transitive, intransitive and ditransitive verbs mostly on perfect and future tenses, the passive voice, the polar question, the infinitive clause, the conditional, the focus particles, the location and the phrasal verbs.

4.3 Comparison between systems

As seen in Table 3, the system that significantly wins most categories is Facebook with 11 categories and an average of 87.5% (if all categories counted equally), followed by DFKI and RWTH which are in the best cluster for 10 categories. When it comes to averaging all test items, the best systems are RWTH and Online-A. On specific categories, the most clear results come in **punctuation** where NEU has the best performance with 100% accuracy, whereas Online-X has the worst with 31.7%. Concerning **ambiguity**, Facebook has the highest performance with 92.6% accuracy. In **verb tense/aspect/mood**, RWTH Aachen and Online-A have the highest performance with 84% accuracy, whereas in this category, MSRA.MADL has the lowest performance with 60.4%. For the rest of the categories there are small differences between the systems, since more than five systems fall into the same significance cluster of the best performance.

When looking into particular phenomena (Table 4), Facebook has the higher accuracy concerning **lexical ambiguity** with an accuracy of 93.7%. NEU and MSRA.MADL do best with more than 95% on **quotation marks**. The best system for translating **modal pluferect** is online-A with 75.6%, whereas at the same category, Online-Y and Online-G perform worse, with less than 2.2%. On **modal negated - preterite**, the best systems are RWTH and UCAM with more than 95%. On the contrary, MSRA.MADL achieves the worst accuracy, as compared to other systems, in phenomena related to modals (perfect, present, preterite, negated modal Future I), where it mistranslates half of the test items. One system, Online-X, was the worst on quotation marks, as it did not convey properly any of them, compared to other systems that did relatively well. Online-Y also performs significantly worse than the other systems on domain-specific terms.

4.4 Comparison with last year's systems

One can attempt to do a vague comparison of the statistics between two consequent years (Table 2). Here, the last column indicates the percentage of improvement from the average accuracy of all systems from last year's shared task[2] to the average accuracy of all systems of this year. Although this is not entirely accurate, since different systems participate, we assume that the large amount of the test items allows some generalisations to this direction. When one compares the overall accuracy, there has been an improvement of about 6%. When focusing on particular categories, the biggest improvements are seen at function words (+12.5%), non-verbal agreement (+9.7%) and punctuation (+8%). The smallest improvement is seen at named entity and terminology (+0.3%).

We also attempt to perform comparisons of the systems which were submitted with the same name both years. Again, the comparison should be done under the consideration that the MT systems are different in many aspects, which are not possible to consider at the time this paper is written. The highest improvement is shown by the system Online-G, which has an average accuracy improvement of 18.7%, with most remarkable the one concerning negation, function words and non-verbal agreement. Online-A has also improved at composition, verb issues and non-verbal agreement and RWTH and UEDIN at punctuation. On the contrary, we can notice that UCAM deteriorated its accuracy for several categories, mostly for coordination and ellipsis (-13.1%), verb issues (-7.6%) and composition (-4.7%). JHU and Online-G and RWTH show some deterioration for three categories each, whereas Online-A seems to have worsened considerably regarding punctuation (-21.6%) and UEDIN regarding negation (-10.5%).

[2] unsupervised systems excluded

category	#	JHU	MLLP	onlA	onlB	onlG	onlY	RWTH	UCAM	UEDIN	avg
Ambiguity	74	-2.7	21.6	4.1	0.0	4.1	10.8	-1.3	2.7	12.1	6.9
Composition	42	4.8	0.0	14.3	0.0	9.5	2.4	-2.4	-4.7	7.1	5.2
Coordination and ellipsis	23	8.7	-4.4	0.0	0.0	13.1	0.0	0.0	-13.1	0.0	7.3
False friends	34	-3.0	5.8	0.0	3.0	-5.9	23.6	5.9	-5.8	14.7	6.8
Function word	41	-2.5	7.3	4.9	0.0	41.4	0.0	-7.4	-2.4	9.7	12.5
LDD & interrogatives	38	10.6	10.6	-2.7	0.0	5.3	0.0	0.0	5.3	7.9	5.6
MWE	53	5.6	7.5	5.7	0.0	1.9	1.9	3.8	-1.8	3.8	4.7
Named entity and terminology	34	5.9	3.0	5.9	0.0	-3.0	-5.9	8.9	0.0	5.9	0.3
Negation	19	0.0	0.0	0.0	0.0	42.1	0.0	0.0	0.0	-10.5	6.6
Non-verbal agreement	48	12.5	10.4	12.5	0.0	22.9	2.1	-2.1	0.0	12.5	9.7
Punctuation	51	5.9	2.0	-21.6	0.0	-7.9	1.9	27.5	0.0	23.5	8.0
Subordination	31	3.3	6.5	-6.5	3.2	19.4	3.2	6.5	0.0	0.0	5.0
Verb tense/aspect/mood	3995	-4.0	-5.9	12.9	0.2	19.8	1.6	5.6	-7.6	5.1	6.0
Verb valency	30	10.0	0.0	0.0	0.0	13.4	6.6	0.0	0.0	3.4	5.8
average (items)	4513	-3.1	-4.3	11.6	0.2	18.7	2.0	5.3	-6.8	5.4	6.1
average (categories)		3.9	4.6	2.1	0.5	12.6	3.4	3.2	-2.0	6.8	6.5

Table 2: Percentage (%) of accuracy improvement or deterioration between WMT18 and WMT19 for all the systems submitted (averaged in last column) and the systems submitted with the same name

5 Conclusion and Further Work

The application of the test suite results in a multitude of findings of minor or major importance. Despite the recent advances, state-of-the-art German→English MT still translates erroneously one out of four test items of our test suite, indicating that there is still room for improvement. For instance, one can note the low performance on MWE and verb valency, whereas there are issues with idioms, resultative predicates and modals. Function words, non verbal agreement and punctuation on the other side have significantly improved.

One potential benefit of the test suite would be to investigate the implication of particular development settings and design decisions on particular phenomena. For some superficial issues, such as punctuation, this would be relatively easy, as pre- and post-processing steps may be responsible. But for more complex phenomena, further comparative analysis of settings is needed. Unfortunately, this was hard to achieve for this shared task due to the heterogeneity of the systems, but also due to the fact that at the time this paper was written, no exact details about the systems were known. We aim at looking further on such an analysis in future steps.

Acknowledgments

This research was supported by the German Federal Ministry of Education and Research through the projects DEEPLEE (01IW17001) and BBDC2 (01IS18025E).

References

Ondej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors. 2018a. *Proceedings of the Third Conference on Machine Translation*. Association for Computational Linguistics, Belgium, Brussels.

Ondej Bojar, Jií Mírovský, Kateina Rysová, and Magdaléna Rysová. 2018b. EvalD Reference-Less Discourse Evaluation for WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 545–549, Belgium, Brussels. Association for Computational Linguistics.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108:159–170.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 Morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation*, pages 550–564, Belgium, Brussels. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Silvie Cinkova and Ondej Bojar. 2018. Testsuite on Czech–English Grammatical Contrasts. In *Proceedings of the Third Conference on Machine Translation*, pages 565–575, Belgium, Brussels. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. *Tenth International Conference on Lan- guage Resources and Evaluation (LREC 2016)*.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, pages 576–583, Belgium, Brussels. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. In *EMNLP 2017: Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Herve Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSNLP - Test Suites for Natural Language Processing. *Proceedings of the 16th . . .* , page 7.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172. Croatian Language Technologies Society, European Association for Machine Translation.

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. 2018a. TQ-AutoTest An Automated Test Suite for (Machine) Translation Quality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-2018), 11th, May 7-12, Miyazaki, Japan*. European Language Resources Association (ELRA).

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018b. Fine-grained evaluation of German-English Machine Translation based on a Test Suite. In *Proceedings of the Third Conference on Machine Translation (WMT18)*, Brussels, Belgium. Association for Computational Linguistics.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96(-1):59–68.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 594–602, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, number March in StatMT '09, pages 259–268, Stroudsburg, PA, USA. Association for Computational Linguistics.

A Appendices

	#	DFKI	FB	JHU	MMLP	MSRA	NEU	onlA	onlB	onlG	onlX	onlY	PROMT	RWTH	Tartu	UCAM	UEDIN	avg
Ambiguity	81	70.4	**92.6**	64.2	76.5	80.2	75.3	69.1	76.5	72.8	50.6	76.5	48.1	77.8	60.5	75.3	59.3	70.4
Composition	48	**93.8**	**97.9**	87.5	85.4	83.3	87.5	**93.8**	**95.8**	83.3	58.3	**93.8**	81.2	85.4	81.2	**89.6**	87.5	86.6
Coordination & ellipsis	74	**85.1**	**89.2**	**78.4**	**85.1**	75.7	**81.1**	**85.1**	**85.1**	60.8	**79.7**	**78.4**	74.3	**86.5**	68.9	**78.4**	**81.1**	79.6
False friends	36	72.2	75.0	55.6	63.9	63.9	55.6	72.2	**77.8**	72.2	72.2	**91.7**	72.2	72.2	55.6	58.3	66.7	68.6
Function word	60	**88.3**	**91.7**	78.3	**91.7**	**83.3**	**90.0**	**88.3**	80.0	**90.0**	65.0	**88.3**	**85.0**	**83.3**	76.7	**88.3**	**88.3**	84.8
LDD & interrogatives	160	**82.5**	**85.0**	**79.4**	**82.5**	**81.2**	**81.2**	73.1	**78.8**	66.2	63.1	75.6	71.2	**83.8**	76.2	**85.0**	69.4	77.1
MWE	77	**68.8**	**77.9**	**64.9**	**66.2**	**66.2**	**67.5**	**67.5**	**70.1**	**68.8**	48.1	**71.4**	55.8	**70.1**	61.0	**63.6**	62.3	65.7
Named entity & terminology	87	**80.5**	**82.8**	**83.9**	**81.6**	**82.8**	**79.3**	**81.6**	**85.1**	66.7	48.3	**82.8**	64.4	**85.1**	**79.3**	**80.5**	**83.9**	78.0
Negation	20	100.0	100.0	100.0	100.0	95.0	100.0	100.0	95.0	100.0	100.0	100.0	90.0	100.0	100.0	100.0	90.0	98.1
Non-verbal agreement	61	**85.2**	**91.8**	**83.6**	**88.5**	**86.9**	78.7	**83.6**	**86.9**	80.3	65.6	80.3	70.5	**82.0**	80.3	78.7	**82.0**	81.6
Punctuation	60	85.0	93.3	70.0	68.3	95.0	**100.0**	76.7	76.7	58.3	31.7	80.0	83.3	88.3	91.7	58.3	90.0	77.9
Subordination	168	**89.3**	**89.9**	**88.7**	**89.9**	**88.1**	**85.7**	75.6	**85.7**	**83.3**	70.8	**86.3**	79.2	**88.7**	**83.9**	**89.9**	76.2	84.4
Verb tense/aspect/mood	4375	77.1	79.4	70.3	78.8	60.4	77.1	**84.1**	74.3	66.2	70.2	72.7	75.4	**83.9**	71.7	79.2	81.4	75.1
Verb valency	86	**72.1**	**79.1**	**68.6**	**67.4**	**70.9**	**66.3**	**67.4**	**68.6**	**67.4**	55.8	**66.3**	54.7	**72.1**	62.8	**68.6**	60.5	66.8
average (items)	5393	78.0	80.9	71.6	79.2	64.3	77.7	**82.8**	75.5	67.5	68.4	74.1	74.4	**83.6**	72.3	79.2	80.2	75.6
average (categories)	5393	82.2	**87.5**	76.7	80.4	79.5	80.4	79.9	81.2	74.0	62.8	81.7	71.8	82.8	75.0	78.1	77.0	78.2

Table 3: Accuracies of successful translations for 16 systems and 14 categories. Boldface indicates significantly best systems in each row

	#	DFKI	FB	JHU	MLLP	MSRA	NEU	onlA	onlB	onlG	onlX	onlY	PROMT	RWTH	Tartu	UCAM	UEDIN	avg
Ambiguity	81	70.4	**92.6**	64.2	76.5	80.2	75.3	69.1	76.5	72.8	50.6	76.5	48.1	77.8	60.5	75.3	59.3	70.4
Lexical ambiguity	63	73.0	**93.7**	65.1	77.8	81.0	74.6	73.0	82.5	79.4	55.6	82.5	50.8	81.0	58.7	76.2	66.7	73.2
Structural ambiguity	18	**61.1**	**88.9**	**61.1**	**72.2**	**77.8**	**77.8**	55.6	55.6	50.0	33.3	55.6	38.9	**66.7**	**66.7**	**72.2**	33.3	60.4
Composition	48	**93.8**	**97.9**	87.5	85.4	83.3	87.5	**93.8**	**95.8**	83.3	58.3	**93.8**	81.2	85.4	81.2	**89.6**	87.5	86.6
Compound	28	**92.9**	**96.4**	**82.1**	78.6	78.6	**82.1**	**92.9**	**96.4**	**82.1**	50.0	**89.3**	67.9	**82.1**	**85.7**	**92.9**	78.6	83.0
Phrasal verb	20	**95.0**	**100.0**	**95.0**	**95.0**	**90.0**	**95.0**	**95.0**	**95.0**	85.0	70.0	**100.0**	**100.0**	**90.0**	75.0	85.0	**100.0**	91.6
Coordination & ellipsis	74	**85.1**	**89.2**	**78.4**	**85.1**	75.7	**81.1**	**85.1**	**85.1**	60.8	**79.7**	**78.4**	74.3	**86.5**	68.9	**78.4**	**81.1**	79.6
Gapping	19	**94.7**	**100.0**	**94.7**	**100.0**	**100.0**	**89.5**	**89.5**	**89.5**	57.9	**89.5**	73.7	73.7	**94.7**	78.9	**94.7**	**89.5**	88.2
Right node raising	20	**80.0**	**85.0**	**80.0**	**75.0**	55.0	**85.0**	**85.0**	**85.0**	50.0	**70.0**	**75.0**	**70.0**	**80.0**	**60.0**	**60.0**	**60.0**	72.2
Sluicing	18	88.9	88.9	83.3	88.9	88.9	88.9	88.9	88.9	83.3	77.8	88.9	83.3	88.9	88.9	88.9	88.9	87.2
Stripping	17	**76.5**	**82.4**	52.9	**76.5**	**58.8**	**58.8**	**76.5**	**76.5**	52.9	**82.4**	**76.5**	**70.6**	**82.4**	47.1	**70.6**	**88.2**	70.6
False friends	36	72.2	75.0	55.6	63.9	63.9	55.6	72.2	**77.8**	72.2	72.2	**91.7**	72.2	72.2	55.6	58.3	66.7	68.6
Function word	60	**88.3**	**91.7**	78.3	**91.7**	**83.3**	**90.0**	**88.3**	80.0	**90.0**	65.0	**88.3**	**85.0**	**83.3**	76.7	**88.3**	**88.3**	84.8
Focus particle	20	**95.0**	**100.0**	**95.0**	**90.0**	**100.0**	**95.0**	85.0	**95.0**	**90.0**	85.0	**95.0**	85.0	**95.0**	**95.0**	**95.0**	**100.0**	93.4
Modal particle	22	**81.8**	**81.8**	**81.8**	**86.4**	**72.7**	**81.8**	**81.8**	**77.3**	**90.9**	63.6	**81.8**	**86.4**	68.2	**77.3**	**77.3**	68.2	78.7
Question tag	18	**88.9**	**94.4**	55.6	**100.0**	77.8	**94.4**	**100.0**	66.7	**88.9**	44.4	**88.9**	83.3	**88.9**	55.6	**94.4**	**100.0**	82.6
LDD & interrogatives	160	**82.5**	**85.0**	**79.4**	**82.5**	**81.2**	**81.2**	73.1	**78.8**	66.2	63.1	75.6	71.2	**83.8**	76.2	**85.0**	69.4	77.1

	#	DFKI	FB	JHU	MLLP	MSRA	NEU	onlA	onlB	onlG	onlX	onlY	PROMT	RWTH	Tartu	UCAM	UEDIN	avg
Extended adjective construction	18	**83.3**	**83.3**	**66.7**	**83.3**	**61.1**	**77.8**	**66.7**	**66.7**	44.4	38.9	**66.7**	**61.1**	**83.3**	**61.1**	**72.2**	**66.7**	67.7
Extraposition	18	44.4	61.1	55.6	55.6	72.2	66.7	55.6	61.1	50.0	50.0	61.1	66.7	61.1	55.6	61.1	55.6	58.3
Multiple connectors	20	**90.0**	**85.0**	**80.0**	**80.0**	**80.0**	**85.0**	**75.0**	**75.0**	**65.0**	**80.0**	55.0	**85.0**	**85.0**	**85.0**	**85.0**	**70.0**	78.8
Pied-piping	19	**84.2**	**84.2**	**89.5**	**78.9**	**89.5**	**84.2**	**78.9**	**73.7**	**73.7**	52.6	**84.2**	**73.7**	**84.2**	68.4	**94.7**	**73.7**	79.3
Polar question	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	84.2	**100.0**	84.2	**100.0**	**94.7**	**94.7**	**100.0**	**100.0**	**100.0**	84.2	95.7
Scrambling	17	**76.5**	**88.2**	**70.6**	**76.5**	**76.5**	**70.6**	52.9	**70.6**	**64.7**	29.4	**70.6**	29.4	**64.7**	**70.6**	**88.2**	35.3	64.7
Topicalization	18	**83.3**	**83.3**	**77.8**	**83.3**	**77.8**	**72.2**	**61.1**	**77.8**	**66.7**	**66.7**	**66.7**	**61.1**	**88.9**	**66.7**	**83.3**	55.6	73.3
Wh-movement	31	**90.3**	**90.3**	**87.1**	**93.5**	**87.1**	**93.5**	**93.5**	**93.5**	74.2	74.2	**93.5**	**83.9**	**93.5**	**90.3**	**90.3**	**93.5**	88.9
MWE	77	**68.8**	**77.9**	**64.9**	**66.2**	**66.2**	**67.5**	**67.5**	**70.1**	**68.8**	48.1	**71.4**	55.8	**70.1**	61.0	**63.6**	62.3	65.7
Collocation	19	68.4	**94.7**	57.9	68.4	**78.9**	**73.7**	**78.9**	**84.2**	**78.9**	52.6	**89.5**	57.9	**73.7**	63.2	57.9	63.2	71.4
Idiom	20	15.0	20.0	15.0	5.0	15.0	5.0	15.0	15.0	10.0	10.0	10.0	5.0	20.0	10.0	5.0	10.0	11.6
Prepositional MWE	19	**100.0**	**100.0**	**100.0**	**94.7**	**89.5**	**100.0**	84.2	**94.7**	**94.7**	57.9	**89.5**	73.7	**100.0**	78.9	**100.0**	**89.5**	90.5
Verbal MWE	19	**94.7**	**100.0**	**89.5**	**100.0**	84.2	**94.7**	**94.7**	**89.5**	**94.7**	73.7	**100.0**	**89.5**	**89.5**	**94.7**	**94.7**	**89.5**	92.1
Named entity & terminology	87	**80.5**	**82.8**	**83.9**	**81.6**	**82.8**	**79.3**	**81.6**	**85.1**	66.7	48.3	**82.8**	64.4	**85.1**	**79.3**	**80.5**	**83.9**	78.0
Date	20	**85.0**	**90.0**	**90.0**	**95.0**	**95.0**	**90.0**	**95.0**	**95.0**	50.0	55.0	**95.0**	50.0	**90.0**	**95.0**	**95.0**	**95.0**	85.0
Domainspecific term	19	**57.9**	**68.4**	**63.2**	**52.6**	**57.9**	**52.6**	**52.6**	**68.4**	**42.1**	21.1	**47.4**	**36.8**	**68.4**	**52.6**	**57.9**	**57.9**	53.6
Location	20	**95.0**	**95.0**	**100.0**	**95.0**	**90.0**	**90.0**	**100.0**	**90.0**	80.0	65.0	**90.0**	**90.0**	**95.0**	**90.0**	**95.0**	**100.0**	91.2
Measuring unit	19	84.2	84.2	**94.7**	**89.5**	**89.5**	**89.5**	**89.5**	**89.5**	**89.5**	63.2	**100.0**	**89.5**	**94.7**	78.9	78.9	**100.0**	87.8
Proper name	9	77.8	66.7	55.6	66.7	77.8	66.7	55.6	77.8	77.8	22.2	77.8	44.4	66.7	77.8	66.7	44.4	63.9
Negation	20	100.0	100.0	100.0	100.0	95.0	100.0	100.0	95.0	100.0	100.0	100.0	90.0	100.0	100.0	100.0	90.0	98.1
Non-verbal agreement	61	**85.2**	**91.8**	**83.6**	**88.5**	**86.9**	78.7	**83.6**	**86.9**	80.3	65.6	80.3	70.5	**82.0**	80.3	78.7	**82.0**	81.6
Coreference	20	**75.0**	**85.0**	**75.0**	**80.0**	**80.0**	**65.0**	**75.0**	**70.0**	**65.0**	45.0	**65.0**	**65.0**	**65.0**	**70.0**	**70.0**	**80.0**	70.6
External possessor	21	**85.7**	**95.2**	**76.2**	**90.5**	**81.0**	**76.2**	**81.0**	**90.5**	**81.0**	61.9	**81.0**	57.1	**85.7**	**81.0**	71.4	71.4	79.2
Internal possessor	20	95.0	95.0	100.0	95.0	100.0	95.0	95.0	100.0	95.0	90.0	95.0	90.0	95.0	90.0	95.0	95.0	95.0
Punctuation	60	85.0	93.3	70.0	68.3	95.0	**100.0**	76.7	76.7	58.3	31.7	80.0	83.3	88.3	91.7	58.3	90.0	77.9
Comma	20	100.0	100.0	100.0	95.0	95.0	100.0	100.0	95.0	95.0	95.0	100.0	100.0	100.0	95.0	100.0	95.0	97.8
Quotation marks	40	77.5	90.0	55.0	55.0	**95.0**	**100.0**	65.0	67.5	40.0	0.0	70.0	75.0	82.5	90.0	37.5	87.5	68.0
Subordination	168	**89.3**	**89.9**	**88.7**	**89.9**	**88.1**	**85.7**	75.6	**85.7**	**83.3**	70.8	**86.3**	79.2	**88.7**	**83.9**	**89.9**	76.2	84.4
Adverbial clause	20	90.0	90.0	95.0	90.0	95.0	90.0	85.0	90.0	90.0	75.0	95.0	90.0	95.0	80.0	90.0	85.0	89.1
Cleft sentence	19	**94.7**	**94.7**	**94.7**	**94.7**	**100.0**	**94.7**	84.2	**89.5**	84.2	84.2	84.2	78.9	**100.0**	**94.7**	**100.0**	**89.5**	91.4
Free relative clause	18	94.4	83.3	83.3	94.4	94.4	94.4	94.4	88.9	88.9	94.4	88.9	94.4	94.4	88.9	88.9	94.4	91.3
Indirect speech	19	**73.7**	**84.2**	**89.5**	**89.5**	**73.7**	68.4	42.1	**94.7**	**84.2**	57.9	**73.7**	63.2	**78.9**	42.1	**84.2**	36.8	71.1
Infinitive clause	20	**100.0**	**100.0**	**95.0**	**90.0**	**100.0**	**90.0**	85.0	**95.0**	**95.0**	**90.0**	**100.0**	**90.0**	**100.0**	**100.0**	**100.0**	85.0	94.7
Object clause	20	**95.0**	**100.0**	**90.0**	**95.0**	**95.0**	**95.0**	85.0	**95.0**	**95.0**	85.0	**90.0**	85.0	**95.0**	85.0	**95.0**	**90.0**	91.9
Pseudo-cleft sentence	18	**66.7**	**72.2**	**66.7**	**72.2**	**61.1**	**55.6**	22.2	**50.0**	**55.6**	22.2	**61.1**	**44.4**	**55.6**	**77.8**	**61.1**	22.2	54.2
Relative clause	18	94.4	83.3	83.3	94.4	77.8	83.3	83.3	77.8	77.8	83.3	94.4	83.3	83.3	88.9	88.9	88.9	85.4
Subject clause	16	**93.8**	**100.0**	**100.0**	**87.5**	**93.8**	**100.0**	**100.0**	**87.5**	75.0	37.5	**87.5**	81.2	**93.8**	**100.0**	**100.0**	**93.8**	89.5
Verb tense/aspect/mood	4375	77.1	79.4	70.3	78.8	60.4	77.1	**84.1**	74.3	66.2	70.2	72.7	75.4	**83.9**	71.7	79.2	81.4	75.1
Conditional	19	**100.0**	**89.5**	84.2	**100.0**	**89.5**	**100.0**	**94.7**	**100.0**	**89.5**	68.4	**100.0**	**94.7**	**100.0**	84.2	**100.0**	**100.0**	93.4
Ditransitive - future I	36	**100.0**	**100.0**	**100.0**	**100.0**	**97.2**	83.3	**100.0**	**100.0**	91.7	**100.0**	**100.0**	**100.0**	**100.0**	**94.4**	83.3	**100.0**	96.9
Ditransitive - future I subjunctive II	36	**100.0**	**100.0**	91.7	**100.0**	**100.0**	80.6	**100.0**	**97.2**	**97.2**	**100.0**	**100.0**	**97.2**	**100.0**	88.9	83.3	**97.2**	95.8
Ditransitive - future II	36	83.3	**100.0**	58.3	**100.0**	86.1	83.3	86.1	**100.0**	72.2	63.9	50.0	77.8	**100.0**	**97.2**	83.3	80.6	82.6
Ditransitive - future II subjunctive II	36	83.3	**100.0**	**100.0**	**100.0**	**94.4**	83.3	83.3	**100.0**	80.6	**100.0**	**100.0**	**97.2**	**100.0**	91.7	83.3	77.8	92.2

	#	DFKI	FB	JHU	MLLP	MSRA	NEU	onlA	onlB	onlG	onlX	onlY	PROMT	RWTH	Tartu	UCAM	UEDIN	avg
Ditransitive - perfect	36	86.1	**97.2**	**100.0**	**100.0**	**100.0**	83.3	**100.0**	**100.0**	88.9	**94.4**	**97.2**	**100.0**	**100.0**	**100.0**	83.3	**100.0**	95.7
Ditransitive - pluperfect	35	48.6	**100.0**	51.4	60.0	82.9	57.1	91.4	45.7	25.7	25.7	20.0	51.4	91.4	80.0	77.1	74.3	61.4
Ditransitive - pluperfect subjunctive II	36	83.3	**100.0**	**97.2**	**94.4**	**100.0**	86.1	**100.0**	77.8	88.9	**100.0**	**94.4**	**94.4**	**100.0**	86.1	83.3	**94.4**	92.5
Ditransitive - present	36	**94.4**	**100.0**	**97.2**	86.1	**97.2**	86.1	**100.0**	**100.0**	61.1	88.9	72.2	63.9	**97.2**	66.7	77.8	**100.0**	86.8
Ditransitive - preterite	35	80.0	**94.3**	71.4	68.6	65.7	71.4	68.6	**85.7**	77.1	60.0	**85.7**	62.9	**85.7**	74.3	65.7	62.9	73.8
Ditransitive - preterite subjunctive II	36	69.4	75.0	63.9	58.3	63.9	58.3	58.3	69.4	69.4	52.8	72.2	58.3	72.2	72.2	55.6	52.8	63.9
Imperative	20	**70.0**	**85.0**	**70.0**	**65.0**	**65.0**	**70.0**	**70.0**	**85.0**	**85.0**	**60.0**	**85.0**	50.0	**80.0**	**60.0**	**65.0**	**60.0**	70.3
Intransitive - future I	36	**97.2**	**97.2**	**97.2**	**97.2**	88.9	**97.2**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.2**	**97.2**	**97.2**	**94.4**	97.6
Intransitive - future I subjunctive II	36	**100.0**	**100.0**	80.6	**100.0**	77.8	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	91.7	**100.0**	**100.0**	96.9
Intransitive - future II	42	92.9	**100.0**	57.1	90.5	85.7	92.9	**97.6**	81.0	88.1	78.6	**100.0**	**97.6**	**95.2**	85.7	85.7	85.7	88.4
Intransitive - future II subjunctive II	36	**97.2**	**100.0**	83.3	75.0	88.9	91.7	**100.0**	**100.0**	77.8	**100.0**	**100.0**	**100.0**	88.9	83.3	75.0	83.3	90.3
Intransitive - perfect	80	**97.5**	92.5	90.0	**97.5**	88.8	93.8	**98.8**	**98.8**	**97.5**	**100.0**	**100.0**	**97.5**	**98.8**	85.0	**98.8**	90.0	95.3
Intransitive - pluperfect	36	**83.3**	77.8	30.6	75.0	58.3	66.7	**94.4**	47.2	19.4	44.4	16.7	55.6	**83.3**	63.9	80.6	80.6	61.1
Intransitive - pluperfect subjunctive II	36	**97.2**	**100.0**	77.8	**100.0**	91.7	**94.4**	**100.0**	**100.0**	88.9	**100.0**	**97.2**	86.1	**94.4**	72.2	91.7	80.6	92.0
Intransitive - present	36	**100.0**	**100.0**	**97.2**	**100.0**	52.8	**100.0**	**97.2**	**97.2**	**97.2**	**100.0**	**97.2**	**94.4**	**100.0**	91.7	**100.0**	**94.4**	95.0
Intransitive - preterite	65	80.0	**96.9**	70.8	73.8	67.7	76.9	81.5	86.2	80.0	69.2	**96.9**	75.4	**93.8**	69.2	89.2	72.3	80.0
Intransitive - preterite subjunctive II	35	**65.7**	**80.0**	**60.0**	**62.9**	51.4	**68.6**	**62.9**	**65.7**	**71.4**	42.9	**71.4**	54.3	**62.9**	51.4	**71.4**	**65.7**	63.0
Modal - future I	180	76.1	**77.2**	75.6	71.1	61.1	**84.4**	**80.0**	74.4	65.0	**79.4**	**78.3**	**80.0**	**78.3**	73.3	**79.4**	**80.0**	75.9
Modal - future I subjunctive II	173	74.6	71.7	76.9	71.7	38.7	**81.5**	**82.7**	71.1	65.3	79.2	72.3	72.8	**82.7**	61.3	77.5	**87.9**	73.0
Modal - perfect	168	88.7	73.2	72.6	83.3	34.5	83.9	62.5	69.0	73.2	42.3	91.7	85.7	**98.8**	78.0	79.2	66.1	73.9
Modal - pluperfect	179	20.1	29.1	11.2	40.2	7.3	22.9	**76.5**	30.7	1.7	7.3	2.2	34.1	49.7	17.9	46.4	58.1	28.5
Modal - pluperfect subjunctive II	178	**57.3**	**52.8**	**55.6**	**59.6**	41.0	**59.6**	**59.6**	**52.2**	42.7	**52.2**	**49.4**	**60.7**	**56.2**	**52.8**	**59.0**	**59.6**	54.4
Modal - present	179	90.5	**94.4**	92.2	**93.3**	48.6	86.6	**94.4**	**96.6**	59.8	**94.4**	77.7	88.8	**95.0**	85.5	**92.7**	**96.1**	86.7
Modal - preterite	179	95.5	**97.2**	86.6	96.6	52.0	89.4	93.9	95.0	89.4	95.0	**99.4**	89.4	**99.4**	86.0	**99.4**	81.6	90.4
Modal - preterite subjunctive II	173	**75.7**	**76.3**	**72.8**	**73.4**	48.6	**73.4**	**78.6**	**72.8**	**71.7**	**77.5**	**74.0**	**74.0**	**80.3**	64.7	**71.7**	**76.9**	72.7
Modal negated - future I	177	**76.3**	**78.0**	**75.7**	**75.1**	45.8	**81.4**	**80.2**	70.1	69.5	**80.2**	70.1	**80.2**	**78.5**	**75.7**	**79.7**	**81.4**	74.9
Modal negated - future I subjunctive II	175	78.3	71.4	76.6	77.7	60.6	83.4	84.0	69.7	67.4	81.7	72.0	78.9	81.7	69.7	81.7	**90.9**	76.6
Modal negated - perfect	175	93.1	73.1	80.6	92.6	65.1	83.4	91.4	69.1	68.0	77.1	70.9	86.9	**97.1**	76.6	79.4	89.7	80.9
Modal negated - pluperfect	173	10.4	13.9	0.0	34.7	8.7	6.4	**97.1**	16.8	0.0	20.8	0.0	15.6	46.2	9.2	16.2	80.3	23.5
Modal negated - pluperfect subjunctive II	170	51.2	60.0	32.9	**64.1**	51.2	**63.5**	**68.8**	33.5	38.2	43.5	50.0	**64.1**	**65.3**	58.8	**68.8**	**70.6**	55.3
Modal negated - present	177	**99.4**	96.0	90.4	**97.7**	71.2	96.6	**97.2**	68.9	72.3	77.4	67.2	92.7	96.6	83.6	**98.9**	96.0	87.6
Modal negated - preterite	178	93.8	96.6	83.7	98.3	79.8	89.3	98.3	96.1	81.5	93.8	94.4	88.2	**100.0**	91.0	**99.4**	83.1	91.7
Modal negated - preterite subjunctive II	171	66.7	**74.3**	64.9	**69.6**	67.3	**73.7**	**76.6**	**72.5**	66.1	**75.4**	**69.0**	**70.2**	**77.2**	**73.7**	**73.7**	**77.8**	71.8
Progressive	20	**65.0**	**85.0**	**60.0**	**70.0**	**80.0**	45.0	50.0	**60.0**	**60.0**	55.0	**80.0**	45.0	55.0	**65.0**	**70.0**	**60.0**	62.8
Reflexive - future I	32	**87.5**	**93.8**	**87.5**	**90.6**	**87.5**	**87.5**	**81.2**	**93.8**	**81.2**	65.6	**90.6**	**81.2**	**81.2**	75.0	**84.4**	**90.6**	85.0
Reflexive - future I subjunctive II	36	75.0	**88.9**	72.2	**80.6**	**80.6**	**83.3**	69.4	**91.7**	**83.3**	**80.6**	**91.7**	**77.8**	66.7	72.2	**80.6**	75.0	79.3
Reflexive - future II	33	75.8	**84.8**	33.3	78.8	66.7	**90.9**	69.7	**87.9**	48.5	33.3	81.8	57.6	**97.0**	72.7	81.8	**90.9**	72.0
Reflexive - future II subjunctive II	34	**82.4**	**94.1**	70.6	70.6	67.6	**85.3**	67.6	**88.2**	61.8	47.1	**88.2**	70.6	76.5	73.5	**82.4**	64.7	74.4
Reflexive - perfect	32	**96.9**	**90.6**	68.8	**90.6**	**84.4**	**93.8**	78.1	**93.8**	68.8	68.8	**87.5**	68.8	**87.5**	81.2	**84.4**	**96.9**	83.8
Reflexive - pluperfect	31	74.2	80.6	71.0	80.6	67.7	**96.8**	74.2	**93.5**	67.7	22.6	80.6	64.5	**83.9**	**93.5**	77.4	**90.3**	76.2
Reflexive - pluperfect subjunctive II	34	**76.5**	**79.4**	**79.4**	**76.5**	**67.6**	**79.4**	**61.8**	**76.5**	**79.4**	47.1	**79.4**	**61.8**	**70.6**	**67.6**	**79.4**	**82.4**	72.8
Reflexive - present	35	80.0	**82.9**	77.1	65.7	65.7	**91.4**	**82.9**	**94.3**	54.3	57.1	80.0	57.1	68.6	71.4	80.0	**82.9**	74.5
Reflexive - preterite	32	75.0	**96.9**	50.0	78.1	56.2	71.9	68.8	**84.4**	**84.4**	25.0	**87.5**	46.9	75.0	75.0	81.2	68.8	70.3

	#	DFKI	FB	JHU	MLLP	MSRA	NEU	onlA	onlB	onlG	onlX	onlY	PROMT	RWTH	Tartu	UCAM	UEDIN	avg
Reflexive - preterite subjunctive II	34	**70.6**	**88.2**	47.1	**73.5**	44.1	61.8	58.8	**76.5**	**73.5**	20.6	**82.4**	47.1	**76.5**	**73.5**	**73.5**	58.8	64.2
Transitive - future I	41	100.0	100.0	100.0	100.0	97.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.8
Transitive - future I subjunctive II	36	100.0	100.0	97.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.2	100.0	100.0	99.7
Transitive - future II	36	**100.0**	**100.0**	86.1	**100.0**	**94.4**	**100.0**	**100.0**	**100.0**	**100.0**	91.7	**100.0**	**100.0**	**100.0**	**94.4**	**100.0**	**97.2**	97.7
Transitive - future II subjunctive II	36	**100.0**	**100.0**	**100.0**	83.3	**94.4**	**100.0**	**100.0**	**100.0**	**100.0**	**94.4**	**100.0**	**100.0**	**100.0**	**94.4**	**100.0**	83.3	96.9
Transitive - perfect	41	**95.1**	**100.0**	**100.0**	**100.0**	**100.0**	**97.6**	**100.0**	**100.0**	**100.0**	92.7	**100.0**	87.8	**100.0**	**100.0**	**100.0**	**100.0**	98.3
Transitive - pluperfect	36	**100.0**	**100.0**	72.2	69.4	**100.0**	80.6	**100.0**	72.2	44.4	91.7	41.7	83.3	**100.0**	91.7	**100.0**	88.9	83.5
Transitive - pluperfect subjunctive II	36	**94.4**	**100.0**	**97.2**	**97.2**	**100.0**	**100.0**	**100.0**	**94.4**	**97.2**	**97.2**	**100.0**	**97.2**	**100.0**	91.7	**100.0**	**97.2**	97.7
Transitive - present	48	**100.0**	**100.0**	**100.0**	**100.0**	**97.9**	**100.0**	**97.9**	**100.0**	93.8	**97.9**	**95.8**	**97.9**	**100.0**	85.4	**100.0**	**100.0**	97.9
Transitive - preterite	36	86.1	**97.2**	80.6	80.6	69.4	77.8	72.2	83.3	**97.2**	72.2	**97.2**	80.6	**100.0**	83.3	86.1	72.2	83.5
Transitive - preterite subjunctive II	36	47.2	**83.3**	58.3	61.1	47.2	**66.7**	55.6	58.3	**75.0**	30.6	**63.9**	52.8	**63.9**	44.4	**75.0**	58.3	58.9
Verb valency	86	**72.1**	**79.1**	**68.6**	**67.4**	**70.9**	**66.3**	**67.4**	**68.6**	**67.4**	55.8	**66.3**	54.7	**72.1**	62.8	**68.6**	60.5	66.8
Case government	27	77.8	**96.3**	**81.5**	74.1	**81.5**	74.1	70.4	74.1	77.8	63.0	70.4	55.6	**81.5**	70.4	70.4	63.0	73.8
Mediopassive voice	20	85.0	85.0	70.0	75.0	80.0	75.0	80.0	75.0	70.0	60.0	80.0	60.0	80.0	65.0	80.0	70.0	74.4
Passive voice	20	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**95.0**	80.0	**90.0**	**95.0**	**100.0**	**95.0**	**100.0**	**95.0**	96.2
Resultative predicates	19	21.1	26.3	21.1	15.8	15.8	15.8	15.8	21.1	21.1	15.8	21.1	5.3	21.1	15.8	21.1	10.5	17.8
average (items)	5393	78.0	80.9	71.6	79.2	64.3	77.7	**82.8**	75.5	67.5	68.4	74.1	74.4	**83.6**	72.3	79.2	80.2	75.6

Table 4: Accuracies (%) of successful translations for 16 systems and 107 phenomena organized in 14 categories. Boldface indicates the significantly best systems in each row.