Discourse-Related Language Contrasts in English-Croatian Human and Machine Translation

Margita Šoštarić[†*] **Christian Hardmeier**[*] **Sara Stymne**[*]

[†]University of Zagreb
Dept. of English
margita.sostaric1@ffzg.hr

[*]Uppsala University
Dept. of Linguistics and Philology
first.last@lingfil.uu.se

Abstract

We present an analysis of a number of coreference phenomena in English-Croatian human and machine translations. The aim is to shed light on the differences in the way these structurally different languages make use of discourse information and provide insights for discourse-aware machine translation system development. The phenomena are automatically identified in parallel data using annotation produced by parsers and word alignment tools, enabling us to pinpoint patterns of interest in both languages. We make the analysis more fine-grained by including three corpora pertaining to three different registers. In a second step, we create a test set with the challenging linguistic constructions and use it to evaluate the performance of three MT systems. We show that both SMT and NMT systems struggle with handling these discourse phenomena, even though NMT tends to perform somewhat better than SMT. By providing an overview of patterns frequently occurring in actual language use, as well as by pointing out the weaknesses of current MT systems that commonly mistranslate them, we hope to contribute to the effort of resolving the issue of discourse phenomena in MT applications.

1 Introduction

Every natural language has means of marking elements belonging to the same coreference chain in order to achieve cohesion and coherence in running text. These discourse phenomena are crucial for understanding texts and their misrepresentation harms text intelligibility. Despite their significance, machine translation (MT) systems have been known to struggle with adequately transferring coreference relations from the source to the target language, which is partly due to the great differences in the way languages express these relations. In our approach, we extend the framework introduced by Lapshinova-Koltunski and Hardmeier (2017), who identify discourse discrepancies in parallel data for English and German by predefining and automatically extracting discourse patterns of interest, and then utilize word alignment information to determine which of the extracted phenomena lack the equivalent counterpart in the other language. We use the same procedure to automatically extract phenomena, but extend the methodology to include cases where the phenomenon does have an equivalent construction in the other language, despite the alignment data suggesting that it is more frequently left unaligned.

In this research, we perform an in-depth study of the way in which diverse discourse phenomena are handled in translation from English to Croatian. We investigate both human translation and the output of different types of MT systems. In the first step, we use the extended methodology of Lapshinova-Koltunski and Hardmeier (2017) to extract interesting diverging discourse patterns that commonly occur in the parallel data. While reflections on the relevant linguistic intuitions are given as a reference, the selection of the phenomena chosen for further examination is primarily based on the data obtained from corpora. This makes our approach strongly usage-based and provides ample space for making observations unconstrained by a particular theoretical framework.

In the second step, we construct a dataset with sentences containing challenging discourse phenomena identified in the analysis of human translations. The constructed dataset can be used for further research in the field of corpus linguistics and translation studies, but it is also useful for gaining insight about language contrasts that is of relevance to MT researchers. We therefore use it to test and evaluate the performance of several types of MT systems and to that end devise a weighted error classification, tailored to accommodate the complexity of the problem at hand.

The paper is structured as follows: in Section 2 we explain the motivation for the research and

Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 36–48
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64005

in Section 3 we give an overview of related work. In Section 4 we describe the used parallel corpora and present the approach to and findings of their analysis. In Section 5 we describe the MT experiment and our approach to error classification. Section 6 contains a discussion of the obtained results and the paper ends with concluding remarks and ideas for future research in Section 7.

2 Motivation

As a South Slavic language, Croatian is a morphologically rich language with relatively free word order. Its pronouns and determiners are grammatically marked for seven cases, three genders and two numbers. Additionally, the forms of determiners and some pronouns show the distinction between animate and inanimate, whereas personal pronouns have long and short variants that reflect emphasis, the choice between them affecting the word order and information flow in the sentence. Croatian is a pro-drop language, meaning that pronouns in the subject position tend to be omitted if the agent can be inferred from other features, such as verbal inflection. In comparison, English is a morphologically less diverse and syntactically stricter language, which suggests that the two languages potentially employ quite different mechanisms to express coreference links.

Apart from the inevitable structural differences, there are several general points of divergence that quickly come to light when handling parallel data for the two languages. First of all, although Croatian has means of expressing the notion of definiteness, it does not have articles, which have a prominent role in the English language. Instead, demonstratives and possessives are sometimes used, as well as definite forms of adjectives and, to a certain extent, restrictive relative clauses. There is also a general tendency to avoid passive constructions and inanimate subjects in Croatian, with these structures commonly rephrased using impersonal verb forms with the reflexive pronoun *se*. Moreover, there is no need for cleft constructions, as information flow can be manipulated through word order, which makes pleonastic pronouns largely redundant in Croatian. Finally, it does not easily create participial constructions, preferring to elaborate the concise English participial expressions into full, finite relative clauses using the relative pronoun *koji*.

3 Related Work

The study by Lapshinova-Koltunski and Hardmeier (2017) examines discrepancies in discourse structures for the language pair English-German. The structures are defined as linguistic patterns using part-of-speech and dependency annotation and the discrepancies are identified using alignment information by finding elements with no corresponding structure in the parallel sentence. This approach allows for a corpus-based contrastive analysis, since the discrepancies might be an indication of systematic linguistic differences or examples of explicitation and implicitation phenomena in the translation process. The mentioned study is mostly focused around the former and the authors manually investigate definite articles and pronouns in subject position as the most frequent unaligned patterns in both languages. Through the analysis, they were able to obtain quantitative proof of tendencies regarding, for example, article use in generics and differences in the use of relative clauses.

Although our approach largely follows the above described methodology, Lapshinova-Koltunski and Hardmeier (2017) were hardly the first to recognize the need to address discourse phenomena in translation. Given the immense variety of linguistic phenomena that fall within the scope of the term, research on discourse phenomena in translation has often focused on a limited group of phenomena (e.g. Furkó, 2014; Zinsmeister et al., 2012; Bührig and House, 2004), which frequently have to be studied in reference to particular registers (Kunz and Lapshinova-Koltunski, 2015). Moreover, the pronouncedly language-specific character of their form has led to examinations of explicitation and implicitation of these phenomena in translation (Blum-Kulka, 1986). On a similar note, Meyer and Webber (2013) compare implicitation tendencies in human and machine translation and find that the latter displays more cases where the phenomena are kept in translation. Scarton and Specia (2015) assess the impact of discourse structures on MT quality through quantitative analysis, while Lapshinova-Koltunski (2017) compares human and machine translations to identify and describe variation in the distribution of different cohesive devices.

On the other hand, a variety of approaches have also been proposed to incorporate discursive information in the workflow of MT systems. The approaches of Le Nagard and Koehn (2010), Hardmeier and Federico (2010) and Guillou (2012) are based on the projection of the source side annotation of coreferring pronouns. A number of discourse-oriented pronoun prediction systems, statistical and rule-based, have also been developed for the submission for the DiscoMT shared task (Hardmeier et al., 2015). The systems experimented with different coreference resolution techniques to improve the translation of pronouns. In recent approaches, Voita et al. (2018), Jean et al. (2017), Wang et al. (2017), Tiedemann and Scherrer (2017) and Bawden et al. (2018) all attempt to improve the translation of discourse phenomena using context-aware NMT systems. Although the degree of their success varies, all papers notably report improvement over the baseline systems.

However, the evaluation of these systems remains problematic, as MT evaluation research has typically been focused on providing an overall score for documents, either through automatic metrics like BLEU (Papineni et al., 2002), or through human evaluation, such as the ranking of systems in the WMT evaluations (Bojar et al., 2017). There have been attempts at error analysis where specific errors are identified and classified into typologies (Vilar et al., 2006; Stymne and Ahrenberg, 2012; Comelles et al., 2016), but these classifications usually do not target discourse-related phenomena. Taking a more specific approach to MT evaluation, Burlot and Yvon (2017) describe how test suites can be created and used automatically for the evaluation of MT systems on morphological phenomena, while the test suite PROTEST, developed by Guillou and Hardmeier (2016), enables relative comparisons between MT systems in terms of pronoun translation. Bawden et al. (2018) construct a contrastive test set to evaluate anaphoric pronouns, cohesion and coherence by having NMT systems rank a correct and an incorrect translation of an input sentence, whereas Sennrich (2017) describes a ranking approach for evaluating NMT systems on grammaticality.

Some of the above work has specifically targeted the differences in performance between NMT and SMT (Burlot and Yvon, 2017; Sennrich, 2017). There are also other types of error analysis targeting this difference, e.g. based on post-edits (Bentivogli et al., 2016). For Croatian in particular, Klubička et al. (2017) conducted an error analysis of SMT and NMT systems, finding that the translation of function words in general is considerably improved in NMT. However, they do not present separate results for pronouns or other elements with coreference functions.

4 Human Translation Analysis

In this section we give an overview of the used datasets and their preprocessing. We also describe the extraction process and the selected phenomena, along with the observations based on the manual data analysis.

4.1 Parallel Corpora

As the use of coreference phenomena varies across different registers and text types, we decided to perform the analysis on corpora from three different domains:

- DGT-TM (Steinberger et al., 2012): EU legal texts, 950K sentences
- SETIMES2 (Tiedemann, 2009): newspaper articles, 200K sentences
- TedTalks (Tiedemann, 2012): prepared speeches, 86K sentences

The three datasets cover an interesting range from very formal, strictly standardized and highly repetitive texts (DGT) to fairly loose and informal translation of speeches (TedTalks). For the purposes of the analyses, English is taken as the source and Croatian as the target language. The corpora were tokenized, tagged for parts of speech and parsed using the pre-trained models for the respective languages developed for the annotation pipeline UDPipe (2017). The parallel data were then aligned at word-level with efmaral (Östling and Tiedemann, 2016), using the grow-diag-final-and heuristic (Koehn et al., 2003).

Relying on the approach of Lapshinova-Koltunski and Hardmeier (2017), we used POS-tags and dependency information to extract a high-recall list of pronouns and determiners in both languages, in order to identify potentially interesting coreference patterns. The main criterion for their extraction was the *pron* or *det* tag, as the original research has found this approach to permit reliable identification of phenomena, even with multi-word units. Similarly to Lapshinova-Koltunski and Hardmeier (2017), we couple the

POS-tags with syntactic information to create linguistic patterns in the format *lemma, POS-tag, dependency label* (e.g. *it, pron, nsubj:pass*) and use word-alignment information to identify the equivalent structure in the other language, if it exists. This gave us a dataset of sentence pairs with indicated coreference phenomena, grouped according to the described linguistic patterns.

Although our approach was largely open and we looked into a variety of phenomena, an initial overview analysis of the data revealed noise both in the output of the tools and in the corpora themselves. As a result, the phenomena chosen for closer examination were selected based on the combination of several factors: the interesting tendencies in their translation identified in the brief overall examination of the data, the tentative interpretation of the frequency of their occurrence across the corpora and the purely practical criterion of having a pattern that enables reliable extraction, meaning that we opted for phenomena which were in most cases correctly handled by the parsing and alignment tools.

4.2 Analysis of Discourse Phenomena

This subsection contains the description of the studied phenomena[1] and the observations made in relation to the specific datasets. The number of phenomena occurrences per corpus is shown in Table 1.

KOJI, det, unaligned. The high frequency of cases where the relative pronoun *koji* is present in Croatian with no corresponding phenomenon on the English side (*who, whom, whose, which, that*) led us to further examine its use. We separate the phenomenon into two groups, depending on whether the relative pronoun has the function of the subject (nominative case) or object (oblique cases) in the relative clause. A major source of unaligned instances with object function seems to be the omission of *that* in English. In both syntactic functions, *koji* is often introduced as a result of elaboration of participial clauses into finite relative clauses. Interestingly enough, introducing relative clauses seems to be a way of dealing with lexical gaps, as illustrated by the example:

EN: *a resealable bag*

vrećica	***koja***	*se*	*može*	*ponovno*	*zatvoriti*
bag	that	REFL	can	again	to seal

'a bag that can be sealed again'

Moreover, it is a way of making the concise English phrases more natural and understandable in Croatian:

EN: *women-run entreprises*

poduzeća	***koja***	*vode*	*žene*
enterprises	that	run	women

'enterprises that are run by women'

Essentially, clauses with *koji* seem to constitute a fairly neutral means of paraphrasing, but their overuse might yield unnecessarily elaborate and clumsy constructions. In SETIMES2 we notice a tendency to resort to such paraphrases in order to maintain a more neutral style:

EN: *the beheaded mother*

majka	***koja***	*je*	*ostala*	*bez*	*glave*
mother	who	is	left	without	head

'the mother who has lost her head'

Here the entire relative clause could be substituted with the Croatian adjective *obezglavljen*, whose meaning is equivalent to that of 'beheaded', but whose use is slightly stylistically marked.

ARTICLES, det, aligned. We have already mentioned that Croatian has alternative ways of representing definiteness, the most straightforward example of this being through the use of demonstratives and possessives[2]. We were interested to see whether specific contextual features could be distinguished that make the explicitation of these coreference links necessary. In that respect, the function of articles seems to vary among the corpora: while the DGT deploys a strict coreferencing system to ensure precision, cohesion and consistency, in TedTalks articles are more pronouncedly used for emphasis and achieving immediacy and closeness in delivering a speech in front of an audience. SETIMES2 generally retains definiteness for the purposes of cohesion and boosting the effect of reader engagement by making news appear as more relevant:

EN: *to address* ***the*** *problem, he says...*

kako	*bi*	*se*	*uspješno*	*nosilo*	*s*
in order to	would	REFL	successfully	deal	with
ovim	*problemom,*	*kaže*	*Simitis*		
this	problem	says	Simitis		

'to successfully deal with this problem, says Simitis'

[1] The patterns used to refer to phenomena have the following format: *phenomenon, pos-tag, alignment information*. The last feature specifies whether or not the equivalent structure exists in the other language. At a more specific level, phenomena are defined in reference to the Universal Dependency Treebank labels (Nivre et al., 2015).

[2] The automatic annotation of adjective definiteness was not reliable enough to be used for automatic extraction.

	KOJIsub	KOJIobj	ARTICLES	ITnsubj	ITexpl	ITpass	ITobj	ITobl/nmod	POSSESSIVES
DGT	19747	6606	10558	8019	1576	3981	3113	2395	9645
SETIMES2	2844	1532	8304	3801	400	448	1648	401	6842
TedTalks	618	300	1758	4411	185	134	4919	1758	3043

Table 1: Overall number of occurrences of each phenomenon in the respective language per corpus.

IT, pron, both. The semantically vague English pronoun *it* can be used in a variety of functions and roles. Given that our approach is based on the patterns produced by the dependency parser, we generally split the phenomenon according to its syntactic function (subject or object), and then break down the two groups into more fine-grained categories. *It* as the subject is hence analysed according to three different patterns: *it* as the subject of an active clause (nsubj), as the subject of a passive clause (nsubj:pass) and as an expletive (expl). In the first case, the behaviour of *it* generally follows that of other Croatian pronouns, i.e. it is frequently omitted. The two latter cases, by contrast, frequently require paraphrasing of varying extent and level of conventionality in Croatian. These typically entail changing the word order and using impersonal constructions:

EN: ***It** is necessary to make them from scratch.*

Potrebno	*ih*	*je*	*stvoriti*	*od*	*početka.*
Necessary	them	is	to create	from	beginning

‘It is necessary to make them from scratch.’

In the example, the expletive *it* is dropped and the adjective in singular neuter form is shifted to the initial position in the sentence.

Unfortunately, the diversity of forms of *it* in Croatian makes it a tricky task for word alignment tools, which especially comes to light when *it* is in object position and varies both lexically and grammatically depending on the antecedent. Due to the inability to reliably separate aligned from unaligned instances, we abstracted away from this information in analysing how this phenomenon is handled in translation. For *it* as an object we looked at two phenomena, depending on whether the object is preceded by a preposition (obl/nmod) or not (obj). *It* in object position is more frequently retained in Croatian, which is understandable as it is often required by verb valency.

POSSESSIVES, det, unaligned. Finally, we noticed that possessives, especially reflexive possessives, are frequently left out on the Croatian side when their introduction is clumsy or redundant. Notably, possessives are sometimes omitted when there are other clear markers of possession in the sentence, encoded for example by verb inflection or indirect objects:

EN: *it did not deny **my** right to vote*

nije	*mi*	*uskratila*	*pravo*	*da*	*glasujem*
did not	to me	deny	right	to	vote

‘it did not deny me the right to vote’

The specification of possession in the example above is made redundant by the use of the personal pronoun in dative case *mi*. Similar situations frequently occur in the more informal TedTalks corpus, where personal pronouns in dative case are often introduced to denote a degree of familiarity with the audience. Given the nature of the corpus, there is also a relatively large proportion of cases where the possessives are dropped in phrases that are closely tied to the agent (referring to e.g. one’s body parts or family members). On the other hand, SETIMES2 and DGT are somewhat stricter in style and often omit possessives, an interesting tendency being the omission of reflexive possessives in cataphoric reference:

EN: *Shortly after **their** arrival, the royal couple...*

Nedugo	*nakon*	*dolaska,*	*kraljevski*	*par*
Shortly	after	arrival	royal	couple

‘Shortly upon arrival, the royal couple’

In the example, the reflexive possessive *svoj* referring to the subject is omitted from the adverbial phrase that precedes it. In the DGT data we also notice the tendency to substitute possessives with explicit noun phrases:

EN: *the value of the procurement over **its** entire duration*

vrijednost	*nabave*	*tijekom*	*cijelog*	*razdoblja*
value	procurement	during	entire	period
trajanja	*nabave*			
duration	procurement			

‘the value of the procurement during the entire duration of the procurement’

As can be seen, the noun *nabava* is repeated in the translation instead of using the possessive *njezin*.

5 MT Experiment

After analysing the parallel data and identifying interesting tendencies regarding the coreference

	TRAIN	DEV	PREPROCESSING	CONFIGURATION	BLEU
SMT	1.23M	4500	Standard preprocessing: data tokenized and truecased, max. sentence length 80.	Training and tuning using the Moses default settings, order of the n-gram model: 3.	33.54
NMT1	1.18M	4500	Tokenization, max. sentence length 60, min. word frequency 5.	Encoder: 3-layer bidirectional LSTM, hidden size 500. Decoder: 3-layer LSTM, hidden size 500.	38.14
NMT2	1.18M	4500	Tokenization, max. sentence length 60, individual BPE, min. frequency 5.	Encoder: 3-layer bidirectional LSTM, hidden size 500. Decoder: 3-layer LSTM, hidden size 500.	36.56

Table 2: MT systems – training configurations.

phenomena, we wanted to see how they were handled by different types of MT systems. Using our linguistic patterns, we extracted a subset of sentences, targeting the constructions that are handled differently by the two languages. The number of sentences per phenomenon corresponds to the overall frequency of their occurrence, while the proportion of sentences taken from each corpus roughly reflects the differences in corpus sizes. We added a couple of manually selected examples (cases of lexical gaps and unaligned reflexive pronoun *se* in Croatian) to construct a test set comprising a total of 1899 sentence pairs with indicated phenomena of interest[3]. We have made the dataset publicly available[4].

5.1 MT Systems

For the experiment we trained a baseline SMT system and several baseline NMT systems. We used open-source toolkits, the phrase-based SMT package Moses (Koehn et al., 2007) and the OpenNMT toolkit (Klein et al., 2017) respectively, and followed the standard training procedures. The NMT systems were based on a sequence-to-sequence architecture with general attention (Luong et al., 2015) and were trained for 13 epochs. We also experimented with sub-word segmentation with byte pair encoding (Sennrich et al., 2016), trained both individually and jointly, for which 10,000 operations were performed. However, only the two models with the highest BLEU scores were retained for the manual analysis. An overview of the chosen MT systems is given in Table 2[5].

The BLEU scores seem to be in line with what could generally be expected from standard MT systems used on relatively repetitive data, except that the performance of NMT systems slightly drops with the application of byte-pair encoding. This calls for further investigation in the future, with some adaptations possibly needed to be made in the training process. However, the BLEU scores are given only as a reference, as it remains questionable whether this evaluation metric can capture the quality of performance on such specific instances as those that are examined in this study. We hence turn to the manual error analysis.

5.2 Error Analysis

For the purposes of the manual analysis, the original human translations are taken as a reference and the order of the machine translations is randomized to reduce bias. The MT output is evaluated only with regard to the relevant antecedent and the syntactic structure containing the specific phenomenon, with the rest of the sentence not being taken into account. Based on our initial data analysis, we devised a classification of errors in terms of translation variation acceptability. The categories used in classification are listed in Table 3. The evaluation was performed by one of the authors, who is a native speaker of Croatian.

To reflect the scalar nature of error severity, we assign a penalty to each error category. This also enables us to produce a provisional score for relative comparison and evaluation of the systems. Some clarification might be needed for categories 4 to 6. Agreement error means that the phenomenon does not agree with the grammati-

[3] Due to the nature of the extraction process, the study is largely focused on intra-sentential phenomena. Although the segmented nature of the artificially constructed test set might be considered a constraint, it is difficult to find an alternative way of testing such a variety of phenomena, while retaining as much data as possible for training.

[4] `http://hdl.handle.net/11234/1-2855`

[5] The test and development sets are kept constant, but the training data used for the two NMT systems had to be further filtered due to technical issues.

error description	category	penalty
mistranslation	1	1
phen. misrepresented, unacceptable translation	2	1
different formulation, unacceptable translation	3	1
phen. represented, agreement error	4	0.75
phen. represented, lexical error	5	0.5
phen. represented, grammatical error	6	0.5
phen. misrepresented, unacceptable due to style/register	7	0.25
phen. misrepresented, acceptable in the style/register	8	0
different formulation, acceptable translation	9	0
identical translation	10	0

Table 3: MT error classification.

cal categories of its antecedent, whereas the lexical and grammatical errors refer to cases such as antecedent mistranslation or errors in the grammatical form of the surrounding elements contained within the same phrase as the phenomenon. As we are primarily interested in the coreference element, we assign a higher penalty for cases where the coreference link gets lost due to agreement mismatch between the antecedent and the observed phenomenon. On a different note, the choice between errors 2, 7 and 8 sometimes seems particularly subjective, but as we conducted the analysis for three different corpora, it was interesting to retain this fine-grained distinction. As an illustration of the complexity of the categorization task, as well as of our approach to it, we provide an example from the created dataset:

SOURCE: *... modifications for the feed currently legally placed on the market...*

REFERENCE:

... izmjene	*hrane za životinje*	***koja***
f.pl.nom	f.sg.gen prep f.pl.acc	f.sg.nom
modifications	feed	which

se	*trenutačno*	*zakonito*	*stavlja*	*na*
refl.pron.	adv	adv	3.sg	prep
REFL	currently	legally	place	on

tržište...
n.sg.acc
market

‘modifications for the feed which is currently legally placed on the market’

SMT:

... izmjene	*hrane za životinje*	*trenutačno*
f.pl.nom	f.sg.gen prep f.pl.acc	adv
modifications	feed	currently

zakonito	*stavljeni*	*na*	*tržište...*
adv	m.pl.nom	prep	n.sg.acc
legally	placed	on	market

‘modifications for the feed currently legally placed on the market’

NMT1:

... izmjene	*hrane za životinje*	*već*
f.pl.nom	f.sg.gen prep f.pl.acc	adv
modifications	feed	already

pravno	*stavljene*	*na*	*tržište...*
adv	f.sg.gen	prep	n.sg.acc
juridically	placed	on	market

‘modifications for the feed already juridically placed on the market’

NMT2:

... izmjene	*za*	*hranu za životinje*	***koje***
f.pl.nom	prep	f.sg.acc prep f.pl.acc	f.pl.nom
modifications	for	feed	which

su	*trenutačno*	*zakonito*	*stavljene*	*na*	*tržište...*
3.pl	adv	adv	f.pl.nom	prep	n.sg.acc
are	currently	legally	placed	on	market

‘modifications relating to the feed which are currently legally placed on market’

The observed phenomenon here is the unaligned relative pronoun *koji* in subject position, which means we evaluate the translation of the noun phrase whose head noun is *feed*, or *hrana*. The reference translation uses the relative pronoun and an impersonal verb form (*se stavlja*) instead of the participial post-modification. SMT keeps the participial form, which would arguably be an acceptable translation in the DGT corpus (error category 8). However, there is an agreement mismatch between the head noun *hrane* (feminine, singular, genitive case) and the participle *stavljeni* (masculine, plural, nominative case). As the phenomenon present in the reference translation is not represented and there are additional errors which make the translation unacceptable, this is an example of error category 2.

The translation produced by NMT1 uses the correct participial form *stavljene*, but makes inadequate lexical choices in the translation of other elements contained in the phrase, translating *currently* and *legally* by *već* and *pravno* instead of *trenutačno* and *zakonito*, respectively. Regardless of the correct participial form, using the relative clause is generally more acceptable in the translation of this particular sentence, so we treat it as a case of misrepresented phenomenon and opt for a more severe punishment by marking it as error category 2, and not 5. Finally, NMT2 uses the relative pronoun *koji*, but the post-modification does not agree with the head noun in number. It is therefore categorized as error 4. As a side note, all three machine translations also lack the durative aspect, which is one of the morphological properties of

	total	SMT	NMT1	NMT2
DGT	931	41.78	43.29	38.78
SETIMES2	628	17.36	37.9	38.85
TedTalks	340	11.76	30	27.65

Table 4: Percentage of acceptable translations out of the total number of sentences for each corpus.

	acceptable	unacceptable	score
SMT	538	1361	1219.5
NMT1	743	1156	980
NMT2	699	1200	1036

Table 5: Overall number of acceptable and unacceptable translations and the score based on summed-up penalties.

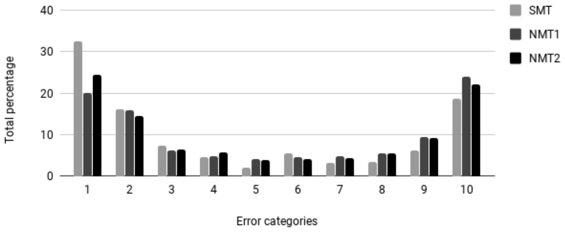

Figure 1: Percentages of error categories for each system.

verbs in Croatian (e.g. *stavljane* instead of *stavljene*), which means that they all belong to error category 6 as well. However, if multiple categories are applicable, we give precedence to the one with the severest penalty so that the overall error scores do not get distorted by single examples.

5.3 Results

As already mentioned, the different properties of individual corpora were taken into account in the analysis, but for brevity's sake we focus more on the overall results in the discussion. However, we should point out that all systems generally perform better on the DGT dataset, which is hardly surprising given that it is the largest and most repetitive corpus. As can be seen in Table 4, the variance in performance across corpora is most pronounced in SMT, which produces 42% of acceptable translations for the DGT and only 12% for the TedTalks data.

While for individual phenomena SMT invariably performs best on DGT, there is some variation in the NMT systems, with NMT2 notably performing best on SETIMES2 for all three cases of *it* in subject position and for *koji* as object. Interestingly enough, when it comes to the retention of articles and the omission of possessives, both NMT systems perform best on TedTalks. However, a closer look at the data reveals that the good performance on articles is largely due to NMT producing differently phrased translations (category 9), whereas their performance on possessives is explained by the fact that the informal style and overall proliferation of determiners and pronouns frequently make the retention of possessives seem acceptable (category 8). Finally, we take note of the poor performance of all systems on *it* in obl/nmod function in the TedTalks corpus, with the majority of errors belonging to one of the first three categories and the NMT systems producing the lowest percentage of acceptable translations.

Looking at the overall results, it should be pointed out that the systems generally perform relatively badly on the examined phenomena. As can be seen in Table 5, the systems in total produce more unacceptable than acceptable translations, although the penalty score does seem to loosely reflect the difference in overall translation quality measured by BLEU. For individual phenomena, shown in Table 6, NMT1 consistently performs best, except on possessives and the miscellaneous examples where NMT2 achieves a better score. All systems are most successful in translating *it* as an expletive and passive subject. On the other end of the scale, SMT performs worst on possessives, NMT1 on articles and NMT2 on *it* as object.

In terms of total error counts, SMT produces significantly more complete mistranslations, while NMT2 makes more agreement errors than the other two systems. Both NMT systems also produce more translations that are generally acceptable, but do not fit the given register/style. Overall percentages of individual error categories in the output of each system are shown in Figure 1. We also notice that most cases fall into the extreme ends of the spectrum, i.e. identical translations and mistranslations.

6 Discussion

It is often pointed out that NMT systems generally produce more fluent, albeit sometimes inaccurate output compared to SMT. We can therefore hypothesize that the two NMT systems will per-

		SMT		NMT1		NMT2	
phenomenon	**instances**	**acceptable**	**score**	**acceptable**	**score**	**acceptable**	**score**
KOJI, det, subject, unaligned	237	30.8	148.5	40.51	120.5	40.93	126.25
KOJI, det, object, unaligned	247	33.2	143.25	46.56	102.75	43.72	110.25
ARTICLES, det, aligned	327	23.85	231.25	27.83	211	26.61	212.5
IT, pron, nsubj, both	109	33.03	64.75	44.04	50.25	39.45	57.5
IT, pron, expl, both	138	40.58	67.25	57.97	44.75	54.35	49
IT, pron, nsubj:pass, both	137	37.23	78.5	53.28	56.25	51.82	60.5
IT, pron, obj, both	263	22.81	180.75	32.7	148.75	25.48	165.5
IT, pron, obl/nmod, both	132	27.27	86.5	36.36	74	28.03	86.25
POSSESSIVES, pron, unaligned	297	21.89	209	33.33	166.75	35.69	164.25
MISCELLANEOUS	12	8.33	9.75	58.33	5	66.67	4

Table 6: Total scores and percentages of acceptable translations for each system per phenomenon.

form better on unaligned phenomena, especially when the omission or insertion of elements on the target side is more a matter of degree of expression idiomaticity than a strict rule. This is confirmed by our analysis, as NMT systems outperform SMT on all three unaligned phenomena. Moreover, SMT performs worst on possessives, which are generally indeed frequently retained in Croatian, and NMT seems to do a better job at identifying contexts in which they should be left out. As for the relative pronoun *koji* in object position, NMT2 does the best job at recognizing when it is necessary to introduce it on the target side, producing 31.98% of translations identical to the original.

The fluency of NMT could also result in better translations of *it* as an expletive or passive subject, as these instances typically require rephrasing in Croatian. This is confirmed in our analysis to some extent as well, with both NMT systems producing the highest percentage of acceptable translations for these phenomena. However, this is also the case for the SMT system, even if its percentages are much lower, which suggests that the patterns used to paraphrase these two phenomena are fairly standardized in Croatian, and hence frequently occur in the corpora. On the other hand, all systems tend to make mistakes when the rephrasing entails moving a noun into the subject position:

***it** is not possible for the controls*

kontrole	*ne*	*mogu*
controls	not	can

‘the controls cannot’

When it comes to restructuring participial clauses into finite relative clauses using *koji*, the situation is similar. The systems rarely produce the less natural literal translations of participial structures, despite the existence of grammatically equivalent forms in the Croatian language. However, when the translation requires more imaginative paraphrasing, the MT systems in most cases fail to deliver, which highlights their incapability to deal with creative language use and satisfactorily handle lexical gaps. Most cases of such mistranslations, manifested as either omission or retention of the source side element, are noticed for the phenomena of unaligned *koji* and in the small group of miscellaneous examples, which comprises a number of cases chosen specifically to see what the systems will do in situations where the translation and use of coreference phenomena are less straightforward.

For instance, let us consider the innovative phrase *non-carbon-based life*, which in the reference is translated as

život	*koji*	*se*	*ne*	*bazira*	*na*	*ugljiku*
life	which	REFL	not	base	on	carbon

‘life which is not based on carbon’

and is entirely mistranslated by all three systems. The SMT system leaves the unknown word in source language, misinterprets the dependency relations and substitutes the relative clause with an impersonal verb construction with *se*:

non-carbon	*se*	*temelje*	*na*	*životu*
non-carbon	REFL	based	on	life

‘non-carbon are based on life’

Both NMT systems leave out the entire unknown part and translate the phrase only as *život* (‘life’).

The systems also fail to cope with idiomatic expressions, frequently omitting or producing word-for-word translations for idiomatic uses of *it* in object position (e.g. *make it, get it*). The translation of multi-word units is another well-known stumbling block of MT systems, but this particular discourse phenomenon seems to be problematic for another reason, and that is the already mentioned diversity of grammatical forms this pronoun can take in the object position in Croatian. Incidentally, *it* in object position is the phenomenon for which all three systems produce the largest percentage of agreement errors: well above 20% of errors made by the systems on this phenomenon belong to category 4, compared to the usual average of around 3% of agreement errors produced in the translation of other phenomena. Finally, the relative performance of all three systems lies closest in the case of aligned articles, but that is because all systems perform poorly, probably due to the very strong tendency not to translate these elements that permeate the English side of the corpus.

7 Conclusion and Future Work

In this paper, we apply the usage-based approach of Lapshinova-Koltunski and Hardmeier (2017) for automatic identification of unaligned patterns linked to discourse-related language discrepancies, and extend it to also include cases of interesting aligned phenomena. We focus on pronouns and determiners in two structurally different languages, English and Croatian, and study them in parallel corpora pertaining to three different registers. We were able to distinguish tendencies both at the general level (e.g. the omission of reflexive possessives in cataphoric position in Croatian) and at corpus-specific levels (e.g. the stricter regulation of representation of definiteness in the DGT corpus). We find that the data-driven nature of the approach makes it a useful framework for linguistic and translation studies, as it hardly makes any initial assumptions about the behaviour of phenomena.

The observations obtained from the parallel data analysis were used to pinpoint interesting linguistic patterns in the two languages, and we further study the way they are handled in MT. To that end, we trained several statistical and neural MT systems and constructed a test set targeting the challenging linguistic expressions. The test set has been made publicly available for further research. We devised a relatively fine-grained classification of errors to evaluate system performance and assigned a penalty to the different error categories in order to facilitate the comparison and ranking of systems in terms of translation acceptability. We provide insights for these diverse extracted phenomena both with regard to the different registers and to the general performance of several MT systems.

Overall, all systems seem to perform unsatisfactorily, especially so on the TedTalks corpus, which is smallest in size as well as linguistically informal and diverse. On the other hand, insofar as better handling of unaligned phenomena can be interpreted as a reflection of translation fluency, NMT systems seem to outperform SMT by producing a higher percentage of acceptable translations in cases which involve standard patterns of paraphrasing and the introduction/omission of coreference elements on the target side. However, all MT systems fall short when it comes to more creative language use, such as handling lexical gaps or idiomatic expressions. Our analysis highlights the complexity of the issue and offers an approach through which further insights can be obtained with a view to improve the translation of coreference phenomena. Lastly, we would like to point out that the research included Croatian, a language that is both under-resourced and under-researched in the field of MT. We also believe that many of the insights for English–Croatian could carry over to other closely related Slavic languages.

As part of future work it would be interesting to investigate other coreference phenomena, and experiment with basing the extraction patterns on some other linguistic features, such as pronoun function (cf. Guillou et al., 2014). As for MT system applications, our manual analysis suggests that the MT systems for this language pair are generally in need of some improvement to better support the study of such specific phenomena, despite obtaining reasonably high BLEU scores. Further inquiry into why the system performance dropped with the application of byte-pair encoding would certainly be advisable and experimenting with different architectures, notably the Transformer (Vaswani et al., 2017), would be desirable. Future work might also include attempts at integrating the output of coreference annotation systems in the workflow of MT systems, in order to make them more attuned to the translation of discourse phenomena.

Acknowledgements

Christian Hardmeier received funding from the Swedish Research Council under grant 2017-930. We acknowledge the computational resources provided by CSC in Helsinki through NeIC-NLPL (www.nlpl.eu).

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313. Association for Computational Linguistics.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*, pages 17–35. Tübingen: Günter Narr Verlag.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55. Association for Computational Linguistics.

Kristin Bührig and Juliane House. 2004. Connectivity in translation transitions from orality to literacy. *Multilingual communication*, 3:87–114.

Elisabet Comelles, Victoria Arranz, and Irene Castellón. 2016. Guiding automatic MT evaluation by means of linguistic features. *Digital Scholarship in the Humanities*, 32(4):761–778.

Bálint Péter Furkó. 2014. Perspectives on the translation of discourse markers: A case study of the translation of reformulation markers from English into Hungarian. *Acta Universitatis Sapientiae, Philologica*, 6(2):181–196.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 2014. ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation)*, pages 283–289, Paris, France.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 108:121–132.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Kerstin Kunz and Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies*, 14(1):258–288.

Ekaterina Lapshinova-Koltunski. 2017. Cohesion and translation variation: Corpus-based analysis of translation varieties. In Katrin Menzel andEkaterina Lapshinova-Koltunski and Kerstin Kunz, editors, *New perspectives on cohesion and coherence*, pages 105–130. Berlin: Language Science Press.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. 2017. Discovery of discourse-related language contrasts through alignment discrepancies in English-German translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26.

Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal dependencies 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Carolina Scarton and Lucia Specia. 2015. A quantitative analysis of discourse phenomena in machine translation. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, 16.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Jörg Tiedemann and Yves Scherrer. 2017. Machine translation with extended context. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of machine translation output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 697–702, Genoa, Italy.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831. Association for Computational Linguistics.

Heike Zinsmeister, Stefanie Dipper, and Melanie Seiss. 2012. Abstract pronominal anaphors and label nouns in German and English: Selected case studies and quantitative investigations. *Translation: Computation, Corpora, Cognition*, 2(1).