Test Suite Evaluation: Morphological Challenges and Pronoun Translation

Marion Weller-Di Marco and Alexander Fraser
Center for Information and Language Processing
LMU Munich
{dimarco,fraser}@cis.uni-muenchen.de

Abstract

This paper summarizes the results of our test suite evaluation with a main focus on morphology for the language pairs English to/from German. We look at the translation of morphologically complex words (DE–EN), and evaluate whether English noun phrases are translated as compounds vs. phrases into German. Furthermore, we investigate the preservation of morphological features (gender in EN–DE pronoun translation and number in morpho-syntactically complex structures for DE–EN). Our results indicate that systems are able to interpret linguistic structures to obtain relevant information, but also that translation becomes more challenging with increasing complexity, as seen, for example, when translating words with negation or non-concatenative properties, and for the more complex cases of the pronoun translation task.

1 Introduction

Evaluating MT output is challenging. Document-levels metrics give a rather coarse-grained estimation of the overall translation quality, but cannot determine how well a system operates for particular challenges. Translations do not have a deterministic solution, but there are always several possibilities for a valid translation, making a focused evaluation of particular phenomena difficult.

The annual WMT Shared Task provides the possibility to submit custom test suites to be translated in addition to the regular test sets, which allows the investigation of the translation performance of state-of-the-art systems when presented with particular translation tasks. In this test suite, we focus on morphological challenges for English to/from German translation: For German–English translation, we look at the translation of morphologically complex words, in addition to a small set of sentences where a subtle difference (singular vs. plural) needs to be detected. For English–German, we study how complex noun phrases are translated – as compounds or rather as multi-word phrases. Furthermore, we add a pronoun translation task and evaluate the translation of the English pronoun *it* into its German equivalents *er/sie/es*, depending on the gender of the noun it refers to.

The test suite does not aim at measuring a system's general translation performance – this is already assessed by means of a manual evaluation and various other metrics in the main shared task – but rather at evaluating the translational behaviour for carefully selected words or phrases. As the sentences in the test suite are not parallel, we opt for a semi-automatic approach where translation options for the words in question are manually collected and then matched with the translation output. Thus, only the translation of the relevant word is considered, whereas the rest of the sentence is ignored.

2 Data Creation and Evaluation

In the following, we outline the process of composing and evaluating the test suite.

Selection of words The sets for the analysis of translating morphologically complex words, compound variants and compounds for re-translating into English are mostly based on a word-frequency list from DeWac[1] (Baroni et al., 2009). The words were morphologically analyzed with SMOR (Schmid et al., 2004). Based on this analysis, words for the aforementioned categories were selected:

- Morphologically complex words: Words with a high degree of complexity and properties such as different forms (e.g. with/without *Umlaut*) in stem and derivations; with negation prefixes or particles or verbal components.
- Compound variants: compounds for which both variants NN1 NN2 and NN2 NN1 exist.
- Compounds for re-translation: adjectives and nouns with up to four components.

[1] https://wacky.sslmit.unibo.it/doku.php?id=frequency_lists

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 458 - 468

Sentence selection We manually retrieved sentences containing the selected words, using Google and the search function provided by the corpus platform DWDS (Geyken et al., 2017) with the corpus *Webmonitor*[2] which is daily updated. The search was (mostly) restricted to newspaper entries from this year, in order to obtain "new" data that was not previously seen in the (monolingual) training data.

Evaluation We identified the translation hypotheses of the relevant words using word alignment (Eflomal (Östling and Tiedemann, 2016)), which were then matched with a manually composed lexicon containing translations options. This step is semi-automatic in the sense that yet unseen translation options need to be verified and added to the lexicon. For the verification, we took into account the sentence context. Being mainly interested in adequacy (i.e. reproducing the meaning of the source word) we allowed for some leeway at the level of fluency, which is difficult to determine anyway in sentences that are not always fully grammatical.

3 DE–EN Translation

This section summarized the design and the outcome of the four categories in DE–EN translation.

3.1 Morphologically Complex Words

We are interested in the translation of morphologically complex words that contain interesting morphological properties such as negation, particles, verbal elements or non-concatenative derivation, which often pose a challenge for translation.

Consider, for example, the word *abrissunwillig: abreißen + un + willig* (*tear down + un + willing*: *unwilling to tear down*), which consists of a nominalization ($abreißen_{\mathrm{V}} \rightarrow Abriss_{\mathrm{N}}$), a negation prefix (*un-*) and an adjective (*willig*: *willing*). In addition to being complex, there is also a non-concatenative operation in the derivation, namely the stem change in *abriss-* vs. *abreißen*. This makes it difficult for linguistically uninformed splitting approaches to find a segmentation into meaningful splits that match with, and thus benefit from, other instances of related words with the same stem.

Many of the selected words emerged from creative use of language and are rather low-frequency. This is to challenge the systems to analyze the words rather than having them already memorized. The words can be loosely grouped as follows:

[2]https://www.dwds.de/d/korpora/webmonitor

		JDExplore Academy	Lan-Bridge	LT22	Online-A	Online-B	Online-G	Online-W	Online-Y	PROMT
NEGATION	correct	49	51	15	50	55	47	54	51	49
	incorrect	8	6	42	7	2	10	3	6	8
	→ polarity	–	4	9	1	2	4	–	3	1
	→ lex.	8	2	30	6	–	6	3	3	7
	→ untransl.	–	–	3	–	–	–	–	–	–
VERB	correct	12	13	2	12	15	13	14	11	11
	incorrect	4	3	14	4	1	3	2	5	5
UML	correct	50	49	18	48	49	43	50	43	40
	incorrect	2	3	34	4	3	9	2	9	12
NON CONC	correct	43	40	5	28	44	35	36	34	24
	incorrect	8	11	46	23	7	16	15	17	27
PART INF	correct	62	64	24	63	62	64	64	60	58
	incorrect	14	12	52	13	14	12	12	16	18

Table 1: Morphologically complex words.

Negation: words containing the negation morphemes *un-* (*unschmelzbar*: *unmeltable*) or *-los* (*knopflos*: *without buttons*). We are in particular interested how the negation is realized, i.e. as an isomorphic, word-internal negation vs. word-external negation. This group comprises 57 sentences.

Verbal elements: the form of verbal elements in derivations often differs from that of the verb stem (*aufbruchsbereit*: *ready to go*; *aufbrechen*: *to leave*). This group comprises 16 sentences.

Stem change (Umlaut): words containing an *Umlaut* in the derivation but not in the stem: *blümchenbedruckt/Blume* (*printed with little flowers/flower*). This group comprises 52 sentences.

Non-concatenative words: adjectives derived from nouns with non-concatenative properties, e.g. *langwimprig/Wimper*: *long-lashed/lash*. This group comprises 51 sentences.

Complex words: words containing particles, such as *mitzittern* (lit: *tremble-with*; *to sympathize, share somebody's emotions*) and words containing *-zu-* infixes (e.g. *aufzutürmen*: *to stack up*). Words of this group are often difficult to translate directly. This group comprises 76 sentences.

Table 1 gives an overview for all five categories. For words containing **negation**, we find that most systems made between 6 and 10 errors (out of 57), with three systems being much better or worse.

For the errors, we distinguish between *lexically incorrectly* translated and *wrong polarity*[3]. For the lexically bad translations, we found that a majority still contained a negation morpheme (such as

[3]The negation is incorrectly reproduced in the translation, either through omission or by a word of the opposite meaning.

nuancen**los**	nuanced (3), nuances (1)
manövrier**un**fähige	maneuverable (3),manoeuvrable (1), manoeuvring (1)
familien**un**freundliche	family-friendly (1)
datenschutz**un**freundliche	data protection-friendly (2), data-protection-friendly (1)
keim**un**fähig	viable (1), germinate (2)
kunden**un**freundlichen	client-friendly (1)
klima**un**freundliches	climate-friendly (1)
fahrrad**un**freundlichste	most bicycle-friendly (1)
unverblasst	still faded (1)
unaufgetaut	unfrozen (1)
unadeligen	aristocratic (1)
kalorien**lose**	calories (1)

Table 2: Translations with wrong polarity (all systems).

knopflos (*buttonless*) → *headless*). For translations with wrong polarity, we observed that in particular words with an infix negation morpheme are error-prone, especially when considering that the test set contains only 13 sentences with such words. Table 2 lists the (otherwise lexically correct) translations with wrong polarity.

Among the correct translations, we observe the entire range between no translation variation (e.g. *unwählbar* ↔ *unelectable* and *vorwarnungslos* → *without warning*) and lexical and local structural variation, as shown in table 3.

For words containing **verbal elements** that differ from the lemma of the verb, the systems' performances range from nearly all correct to nearly all incorrect. For this subset, there was no clear trend of error, certainly also due to its small size. One thing that we observed was that for *abrissbedroht* (*threatened by demolition*), *abrissbereit, abrissreife* (*ready to be demolished*), *abrissgeweihten* (*marked for demolition*), *abrisswilligen* (*willing to demolish*) a common mistranslation was just *demolished*, even though the state of being actually demolished is not described by any of these words.

The words with a **stem change (Umlaut)** lead to mixed results; for the words with **non-concatenative properties**, we observe even more errors. Among the incorrectly translated words, there is a tendency that the part with the non-concatenative properties is mistranslated, whereas the other, more easy part, is correct (cf. table 4). Finally, the words containing **particles or infixes** were challenging to translate, even though some words were considerably more difficult. In particular, the set included some verbs that cannot be translated isomorphically. One example is the combination of *kaputt* (*broken*) + *verb*, in analogy to *kaputtmachen* (*to break*, lit. *kaputt-make*): *kaputt-sparen* (*to destroy through excessive money saving*) or *kaputtsanieren* (*to destroy through excessive renovating*). With the exception of *kaputtschlägt* and *kaputtzukriegen* (*to break*), they were nearly always translated incorrectly. In particular *kaputtsparen* was often translated as *saved from damage* or similar, the opposite of the intended meaning. In contrast, for *schönreden* (*to gloss over, to sugar coat*, lit: *beautiful* + *talk*), a generally similar construction, about half of the translations were correct.

quittungslos	without receipt (9), without receipts (2), without a receipt (2), receiptless (1), receipt-free (1)
nuancenlos	without nuances (4), nuanceless (3), nuance-free (3),nuance-less (2), unnuanced (1), lacking in nuance (1)
unaufgetaut	unthawed (3), without thawing (1), without defrosting (1), undefrosted (1), before it is thawed (1)
unverblasst	unfaded (2), still vivid (1), not yet faded (1), not faded (1), still fresh (1)

Table 3: Translation variants for words with negation morphemes (only correct translations shown).

	correct	incorrect
langwimprigen	long-lashed (4)	long-drawn (1), long tail (1), long-winded (1), long-eyed (1) long-wimprigen (1)
sonnenbebrillt	in sun glasses (2), with sunglasses (1), wearing sunglasses (1)	bespectacled by the sun (1), in the sun (1), sunglassed (1)
löwenmähnige	lion-maned (15)	lion-eyed (1) lion-like (1), duel-like (1)

Table 4: Translating non-concatenative words.

3.2 Compound Variations

Compounds are commonly occurring in German and their translational behaviour has been studied extensively. An important aspect in compound translation is to correctly reproduce the relation between the head and modifier in noun-noun compounds, which we aim to investigate in this category by looking at compound pairs that consist of the variants NN1 NN2 and NN2 NN1, such as *Oliven*|*öl* and *Öl*|*olive* (*olive oil* vs. *oil olive*) or *Leder*|*stiefel* and *Stiefel*|*leder* (*leather boot* vs. *boot leather*).

The compound variants NN1 NN2 and NN2 NN1 have different heads and thus a different meaning (as opposed to variation in hyponymy/hypernymy) and are not generally interchangeable[4]. We thus

[4]We found, however, that in some cases, there is an acceptable one-word translation for both variants, e.g. *Absatzschuh*

	JDExplore Academy	Lan-Bridge	LT22	Online-A	Online-B	Online-G	Online-W	Online-Y	PROMT
correct	57	58	29	59	59	60	59	59	57
wrong order	–	–	2	–	1	–	–	–	–
head missing	1	–	7	–	–	–	–	–	1
mod missing	–	–	5	1	–	–	1	–	–
bad transl.	2	2	17	–	–	–	–	1	2

Table 5: Translation results of compound pairs. *wrong order*: wrong order of head and modifier; *head/mod missing*: only translated head or modifier; *bad transl*: translation was either missing or wrong.

retrieved different sentences for each variant: this means that the compound variants are not analyzed in a minimal pair setting, but each variant is presented in an appropriate and natural context.

This category is somewhat inspired by one of the error types introduced by Sennrich (2017), where translation probabilities for contrastive sentences containing compound variants (a correct vs. a wrong translation consisting of a compound with switched components) are compared.

Table 5 shows the results for 15 compound pairs, with 2 examples per variant in most cases, resulting in 60 sentences total. All systems, with the exception of one, translated most compounds correctly. Furthermore, there is no dominant error type for cases with incorrect translation. This indicates that through most systems, there is a generally good understanding of compound structure and subsequent translation, even in cases such as the high-frequency *Olivenöl* (21M google hits[5]) vs. the low-frequency *Ölolive* (507 google hits).

3.3 Compound Translation

In this section, we look at the translation of compounds consisting of two to four components. This set of compounds[6] also serves as a basis for the experiment in section 4.1 which studies how English noun phrases are translated into German.

This word set contains some "newish" words, i.e. words that are not new per-se, but became considerably more frequent recently, such as *Gasengpass* (*gas bottleneck*) and some Covid-related terms such as *Impfbereitschaft* (*willingness to be vaccinated*).

→ *heel, heeled shoe* and *Schuhabsatz* → *heel, shoe heel*.

[5]Search of the citation form in double quotes. Numbers reported by Google give only a rough idea of the true frequency on the web, but are sufficient to estimate the order of magnitude.

[6]This set of words is not exactly the same as in section 4.1.

	JDExplore Academy	Lan-Bridge	LT22	Online-A	Online-B	Online-G	Online-W	Online-Y	PROMT
correct	68	70	34	70	70	71	69	68	71
wrong	5	3	35	3	3	2	3	5	2
missing/copy	–	–	4	–	–	–	1	–	–

Table 6: Results for translating compounds into English.

Most words in this set are compositional, and very few are non-compositional compounds, such as *Dornröschendasein* (*Sleeping Beauty existence*).

Table 6 shows the results for translating compounds into English; most systems did quite well. Among the compounds with the most consistent translations are *Sonnenblumenkernöl* (*sunflower seed oil*) and *Haifischflossensuppe* (*shark fin soup*), i.e. compounds with a straightforward literal translation. Similarly, the somewhat new *Testmüdigkeit: test fatigue (17), testing fatigue(1)* (occurrences in two sentences) leads to consistent translations. One of the more difficult words was *distanzlernende (distance learning)*, which 5 of the 9 systems translated correctly. The incorrect translations did not quite capture the meaning and translated into *learning distance*, *learning about distance* and *to dance* (probably due to an incorrect splitting that contained the German "tanz").

3.4 Preserving Morphological Information in Syncretic Forms

Understanding the precise meaning of a word and its function in the sentence is crucial to obtain a good translation. This includes the comprehension of relevant morphological features.

While German is rich in different inflected forms, there is also a certain degree of syncretism (forms with different morphological features sharing the same surface form). For example, *Hund* (*dog*) can be dative, accusative and nominative, *Unternehmen* (*company*) can be singular and plural. Usually, this can be resolved by the context, often by means of the determiner: dem_{DAT}/den_{ACC}/der_{NOM} *Hund* and das_{SG}/die_{PL} *Unternehmen*.

In this experiment, we look at number in non-subject words as (i) number is the only feature of nominal inflection that is shared between German and English, and (ii) there are no further ramifications to the rest of the sentence. We designed a setting in which the disambiguating context, a definite article, is not directly adjacent to the word in question, but separated by an inserted phrase.

Die Verzögerungen sind auf Engpässe bei **den** mit der Umsetzung beauftragten **Unternehmen** und auf ... zurückzuführen. The delays are to bottlenecks at **the** with the implementation charged **companies** and to ... due
Die Verzögerungen sind auf Engpässe bei **dem** mit der Umsetzung beauftragten **Unternehmen** und auf ... zurückzuführen. The delays are to bottlenecks at **the** with the implementation charged **company** and to ... due
The delays are due to bottlenecks at **the companies/company** charged with the implementation and to ... .

Table 7: Example for minimal sentence pairs.

	JDExplore Academy	Lan-Bridge	LT22	Online-A	Online-B	Online-G	Online-W	Online-Y	PROMT
Correct	15	16	8	15	16	17	17	18	17
Incorrect	3	2	7	3	2	1	1	–	1
NA	–	–	3	–	–	–	–	–	–

Table 8: Preserving number information: *Correct*: the noun in singular and plural was translated correctly. *Incorrect*: for at least one noun, the number was incorrect. *NA*: not translated or otherwise impossible to judge.

We created 18 minimal sentence pairs with the only difference being a singular vs. a plural article, in order to test whether the noun (with identical forms in both sentences) is correctly translated. The sentences contain "nested prepositional phrases" where an inserted prepositional phrase separates the article and the noun, cf. table 7.

Table 8 shows the results for the task of preserving number information: most systems can handle this problem reasonably well, indicating that the systems have the ability to interpret the sentence structure and to identify the relevant context.

4 EN–DE Translation

In this section, we look at re-translating compounds and present a pronoun translation task.

4.1 Compound Creation

To prepare the test set, we translated the German compounds from section 3.3 into English, including structural or lexical variations if possible (cf. table 9 for some examples) and retrieved English sentences with these translations, resulting in a set of 102 sentences. We distinguish between "phrase" (PHR), containing a preposition (such as *interpreter for sign language*) and "compound" (COMP) where the order of the words corresponds to a compound (as in *sign language interpreter*).

The results in table 10 show a tendency to keep the structure, i.e. translating a compound-like structure into a compound, and a phrase into a phrase rather than a compound, even though there are differences depending on the word.

Gebärdensprach-dolmetscher	sign language interpreter interpreter for sign language
Obstbaumschnittkurs	fruit tree pruning workshop workshop on fruit tree pruning
Kleinkläranlagen-betreiber	small wastewater treatment plant operators; operators of small wastewater treatment plants
Kreuzworträtselfrage	crossword question crossword puzzle question

Table 9: Structural and lexical variants in the compound translation task.

	JDExplore Academy	Lan-Bridge	Online-A	Online-B	Online-G	Online-W	Online-Y	OpenNMT	PROMT
Comp → Comp	58	56	60	57	58	55	47	49	55
Comp → Phr	13	15	10	11	11	17	11	13	6
Phr → Comp	5	9	2	9	5	7	2	3	4
Phr → Phr	23	18	24	18	23	21	25	18	19
Wrong transl.	2	3	6	6	5	2	16	19	18
Copied EN	1	1	–	1	–	–	1	–	–

Table 10: Translating English complex phrases.

For example, the variants *wearers of headscarves* and *headscarf wearers* were mostly translated by the compound *Kopftuchträger(innen)*, with only two instances of *Träger von Kopftüchern*. In contrast, both *pacemaker wearer* and *pacemaker carrier* have a more equal distribution of *Träger von Herzschrittmachern* and *(Herz)Schrittmacherträger*. A more complex example, *willingness to get vaccinated*, was translated to the corresponding compound *Impfbereitschaft* (6 times), as *Bereitschaft zur Impfung* (3 times) and *Bereitschaft, sich impfen zu lassen* (9 times). The variant *unwillingness to get vaccinated* proved more problematic: only three systems obtained correct translations: *Impfunwilligkeit* (2) and *Unwilligkeit, sich impfen zu lassen* (1). The translation *Impfunbereitschaft*, while transporting the correct message, is questionable. In the remaining 5 cases, the negation was ignored.

4.2 ContraCat: Translating Pronouns

The translation of pronouns is often more difficult than it seems at a first glance: a translation system requires diverse linguistic information to produce a

1	The mouse ate the cookie and the **bear** drank the milk. **It** drank the milk quickly.
2	The **tiger** ate the ice cream. **It** was happy.
3	The giraffe ate the **steak**. **It** was cooked.

Table 11: Examples for the ContraCat template set.

target-language pronoun with the correct morphological features such as gender, number or case.

To translate the English *it* into German, a translation system needs to identify the noun *it* refers to and to have knowledge about that noun's gender[7] in German, as illustrated below:

... a dog ... it ... → *... ein Hund*$_{\text{MASC}}$ *... er ...*
... a cat ... it ... → *... eine Katze*$_{\text{FEM}}$ *... sie ...*
... a zebra ... it ... → *... ein Zebra*$_{\text{NEUT}}$ *... es ...*

To analyze the translation of pronouns, we make use of the template set *ContraCat* (Stojanovski et al., 2020) which consists of sentence pairs where several nouns are introduced in the first sentence, and a pronoun *it* in the second sentence either refers to one of these nouns, or is generic as in *it is raining*. The sentences are constructed in a way that the relevant noun/context can be derived through either world knowledge or through analyzing the structure of the sentence. Furthermore, the sentences are designed such that the nouns of e.g. the two subjects (mouse and bear in sentence 1 in table 11) have translations into German with different genders (*Maus*$_{\text{FEM}}$ and *Bär*$_{\text{MASC}}$) in order to allow for an unambiguous evaluation.[8] Table 11 shows three examples; an overview of all template categories can be found in table A.

Technically, each sentence consists of two short sentences. As this might be disadvantageous in some system settings, we generated a second version where we joined the two short sentences with "and" into one sentence. In the evaluation, these variants will be referred to as 2S and AND.

In its original form, the template set provides three translation hypotheses, each with a different translation option (male/female/neutrum) for *it*, for which the system's likelihood to produce the correct translation is then measured.

To be used in an actual translation scenario, we adapt the evaluation process: given the template structure, we first identify the antecedent (the noun that is referenced by the pronoun *it*), and then its translation and the translation of the pronoun *it* in the target sentence using word alignment (Eflomal, Östling and Tiedemann (2016)). The translation options of the nouns observed in the different systems' outputs are listed in a manually compiled dictionary[9], alongside their German grammatical gender. With this, the translated pronoun can be automatically matched with the noun's gender.

4.2.1 Test Set Creation

From the original test suite[10], we randomly selected 100 sentences for each of the 20 categories (cf. table A for an overview), with the exception of the category *world knowledge*, for which 200 sentences were selected as this category comprises the scenario of addressing an animate noun (animal) vs. inanimate noun (food). Doubling the sentences for the AND variant results in 4200 sentences total.

4.2.2 Evaluation and Results

Table 12 shows the results of translating pronouns. For the categories *event_** and *pleo_**, where the translation *it* → *es* is always expected, nearly all systems have a perfect score. The other categories where the antecedent needed to be derived from the context are more challenging, however without a clear pattern between the systems. We can observe two tendencies, even though not consistent through all systems: first, the variant AND often leads to better results, probably due to the fact that sentences are often the "standard unit" for translation, whereas the two sentences in variant 2S might be considered separately, depending on the systems' architecture. Second, sentences where the antecedent is the second NP of the first sentence, i.e. closer to the *it*, tend to get better results.

Looking further into the errors, we find that *es*$_{\text{NEUT}}$ is often preferred over a a feminine or masculine form. This might simply be the case because *it* → *es* is the default translation, and also because the generic *es* can oftentimes be considered grammatical, even though a translation into the gender-specific pronoun would be possible.

A general problem with this template approach is the degree of freedom in the translation process: sometimes the pronoun is just not translated (cf. table 13), and in some cases, it is possible to formulate the sentence such that the pronoun *es* leads to

[7] Further features leading to variations at the level of grammatical case and number will be ignored here.

[8] This was guaranteed for the pre-defined translations in the original setting. In actual translations, there can be more variation, for example *deer* → *Hirsch*$_{\text{MASC}}$, *Reh*$_{\text{FEM}}$, *Wild*$_{\text{NEUT}}$.

[9] The dictionary comprises entries for 141 English nouns, with one to four translation options.

[10] https://github.com/BennoKrojer/ContraCAT

	JDExplore Academy		Lan-Bridge		Online-A		Online-B		Online-G		Online-W		Online-Y		OpenNMT		PROMPT	
	2S	AND	2S	AND	2S	AND	2S	AND	2S	AND	2S	AND	2S	AND	2S	AND	2S	AND
event_chaos	100	100	100	100	100	100	100	100	100	100	100	99	100	100	100	100	100	100
event_happened	100	100	100	100	100	100	100	100	100	100	100	99	100	100	99	100	100	100
event_situation	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
event_surprise	100	100	100	100	100	100	100	93	100	100	100	99	100	100	100	100	100	100
gender_step	95	92	22	95	21	84	22	92	21	91	88	97	21	92	80	66	22	89
obj_drink	77	96	22	95	18	57	22	76	35	50	81	81	18	30	26	38	29	27
obj_eat	1	27	37	6	23	26	37	26	32	38	42	10	23	24	14	23	20	28
obj_verb_drink	100	100	20	98	16	53	20	95	38	53	84	92	22	57	24	41	18	41
obj_verb_eat	1	1	36	2	25	30	33	11	29	36	43	8	26	32	2	25	28	41
pleo_believe	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
pleo_rain	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
pleo_seem	100	100	100	100	100	100	100	100	100	100	99	99	100	100	100	100	100	100
pleo_shame	5	100	100	100	100	100	100	100	100	100	100	100	100	100	97	100	100	100
subj_drink	36	45	34	34	34	34	34	37	34	34	45	84	34	34	85	68	34	34
subj_eat	45	44	45	45	45	45	45	45	45	45	44	24	45	45	18	35	45	45
subj_verb_drink	99	99	34	100	34	94	34	100	34	98	100	100	34	75	100	93	34	64
subj_verb_eat	22	20	34	20	34	11	34	13	34	34	54	2	34	22	0	14	34	25
verb_drink	100	100	24	98	22	72	25	90	26	68	84	96	22	72	71	19	22	43
verb_eat	2	2	27	2	18	18	26	11	25	33	43	20	18	21	28	8	21	45
world_knowl.	180	194	77	200	71	124	73	195	67	133	185	181	76	126	97	77	77	97

Table 12: Results for pronoun translation using the ContraCat template, the number indicates the amount of correct pronoun translations (out of 100 for all except for world_knowledge, which has 200 test sentences). (Note: in JDExploreAcademy–pleo_shame, *it is a shame* is nearly always translated as *"Schade."*, i.e. without a pronoun.)

The mouse ate the cookie and the **sheep**$_{\text{SG/PL}}$ drank$_{\text{SG/PL}}$ the tea.
It$_{\text{SG}}$ liked the tea.
Die Maus aß den Keks und die **Schafe**$_{\text{PL/NT}}$ tranken$_{\text{PL}}$ den Tee.
Er$_{\text{SG/MASC}}$ mochte den Tee.

Table 13: Example for incorrectly passed-on number.

The cow ate and the dog drank. It drank a lot.
Die Kuh aß und der Hund trank viel.
The cow ate and the dog drank a lot.
The frog ate the fruit and **it** had a sour taste.
Der Frosch aß die Nuss und ∅ hatte einen sauren Geschmack.
The frog ate the fruit and had a sour taste.

Table 14: Examples for pronoun omission.

a grammatical sentence. For example, *the -animal- liked it* (*it* →*food item*) can be translated as *Dem -Tier- gefiel/schmeckte es*. This is a valid translation, even though not strictly in the sense of the intended meaning as in *dem -Tier- schmeckte er (→ Apfel*$_{\text{MASC}}$*)* vs. *dem -Tier- schmeckte sie (→ Banane*$_{\text{FEM}}$*)* . For the sake of evaluation, we count a translation only as correct if the pronoun exists and matches in gender with the noun it refers to.

While this experiment only focused on *gender*, we also observed some cases that extended to *number*, namely in a few cases where the English singular and plural forms are the same. In the example in table 13, the number of *sheep* is not directly visible in the first part of the sentence, but can be disambiguated through the singular form *it*. The translation contains *Schafe* in plural, but *er* as translation of *it* is singular/masculine (*Schaf* is neutrum).

5 Related Work

The linguistic and morphological compentence of translation systems is a topic of previous and ongoing research. Isabelle et al. (2017) present a challenge set for English to French translation targeting linguistic divergence between the two language pairs. Their hand-crafted set has a focus on morpho-syntactic, lexico-syntactic and syntactic divergences. Burlot and Yvon (2017) present an analysis of minimal pairs representing a contrast that is expressed syntactically in EN and morphologically in a morphologically rich language (DE, CZ and LV). For a source test sentence (the base), variant(s) containing exactly one difference with the base (e.g. person/number/tense of a verb or number/case of a noun/adjective or polarity) are generated and automatically evaluated, counting a translation as correct if the targeted feature is produced correctly in the target language. The work of Burchardt et al. (2017) and Avramidis et al. (2019) comprises the DFKI test suite for German to English MT. Their test set consists of over 5k sentences to analyze over 100 categories, including negation, composition, function words, subordination, non-verbal agreement, multi-word expressions, verb tense/aspect/mood, lexical ambiguity and punctuation. LingEval97 (Sennrich, 2017) is a large-scale data set of 97000 contrastive English–German translation pairs where errors (on the level of agreement, auxiliaries, verb particles, polarity and swapped compound components) haven been automatically created. It is then measured whether

a reference translation is more probable than the corresponding contrastive translation containing an inserted error. An obvious question with this approach is whether forced translation mimics the MT system's "natural behaviour", i.e. whether the presented sentence `e` is the system's best choice given the source sentence `f`. This question is addressed in Vamvas and Sennrich (2021) where it is argued that test data should be chosen such that there is minimal discrepancy between the training data and the data to be evaluated. They recommend that test sentences be created from machine generated text rather than using human-written references. The paper proposes an updated version of LingEval97.

6 Conclusion and Future Work

This paper summarizes the results of our WMT22 Test Suite, looking at the translation of morphologically complex words, compounds and a set of minimal pairs to assess the preservation of number. Our evaluation shows that on one side, the translation of morphologically complex words is not without challenges, in particular for low-frequency words and when containing negation. On the other hand, the handling of the (structurally much simpler) compounds NN1 NN2 vs NN2 NN1 and the preservation of the number feature worked quite well. The results for the pronoun translation experiment were mixed.

Our results indicate that the systems have a generally good understanding of linguistic structures, but also that at a certain degree of (morphological) complexity, problems start to arise. For research in MT, this means that modeling morphology, particularly negation and non-concatenative processes, might be worthwhile.

The test suite, with the exception of the pronoun translation task, is based on a manually created set of sentences alongside matching dictionaries. While this has the advantage of presenting the selected words/phrases in a natural context, it comes with a comparatively high amount of manual effort, making it difficult to upscale. In contrast, the artificial data used in the pronoun translation task allows for a comparatively straightforward evaluation, but sounds unnatural and likely differs considerably from the MT training data, which might even bias the results to a certain extent.

For future work, we intend to look into the generation of meaningful sentences with particular properties that allow for a systematic evaluation of MT.

Limitations

There are several limitations to this work: first, the work is obviously limited in terms of data-set size and the small number of language pairs considered. As there is a certain amount of manual selection and annotation required, this is generally a tricky problem to address. As mentioned previously, we plan to work on more sophisticated test data generation as a basis for a more focused evaluation. Another limitation is a lack of generalizability: the presented analyses offer only partial insights and provide but a first glimpse into understanding to what extent morphological information is captured and passed on in machine translation.

Ethics Statement

We have no ethical concerns about the research presented in this paper. The data selection and annotation work was carried out by the first author.

Acknowledgements

This work was supported by the DFG (grant FR 2829/4-1).

References

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

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.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Alexander Geyken, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand, and Lothar Lemnitzer. 2017. Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). *Zeitschrift für Germanistische Linguistik*, 45(2):327–344.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

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

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1263–1266, Lisbon, Portugal.

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, Valencia, Spain. Association for Computational Linguistics.

Dario Stojanovski, Benno Krojer, Denis Peskov, and Alexander Fraser. 2020. ContraCAT: Contrastive coreference analytical templates for machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4732–4749, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

Words with negation: *abgaslose, abgaslosen, akzentlosen, datenschutzunfreundliche fahrradunfreundlichste, fahrunfähig, familienunfreundlich, familienunfreundliche, flugunfähigen, handlungsunfähig, kalorienlose, keimunfähig, klimaunfreundliches, knopflosen, kundenunfreundlichen, lückenlose, manövrierunfähige, nuancenlos, quittungslos, Rücksichtlose, tageslichtlose, unabgefüllten, unabgeschirmt, unablösliche, unadeligen, unanfechtbar, unanfechtbare, unangemeldete, unaufgetaut, unausgereift, unauswechselbar, unbegehbar, unfußballerisch, ungepflügten, ungeschliffen ,ungeschliffene, unkontrollierbare, unliebenswert, unregierbaren, unreparierbar, unreparierbarer, unsanierten, unschmelzbaren, unverblasst, unverderblich, unverhangene, unverwundbaren, unwählbar, unzerknittert, unzerschnittene, vorwarnungslos*

Words with verbal element: *abrissbedroht, abrissbereiten, abrissgeweihten, abrissreife, abrisswilligen, abwieglerisch, aufbruchsbereiten, aufbruchsicher, aufwieglerisch, aufwieglerischen, ausbruchsartigen, ausbruchsicher, ausstiegswillige, bestbesprochenen, weitergesponnen*

Words with Umlaut: *ananasförmigen, Anemonenblütige, barhändig, blümchenbedruckte, blümchenbedruckten, blümchentapetigen, doppelbödig, doppelköpfig, doppelköpfigen, einblättrig, einblättrigen, einsträngig, einsträngige, einsträngigen, engräumig, fädenziehende, fältchenmindernden, gehirnwäscherische, gehirnwäscherischen, großäugig, großäugigen, großräumig, höhergeschossigen, höherrangiger, höherwüchsiger, hundertäugigen, hütchenförmige, kaltblütig, kannenförmige, kannenförmiges, kleinräumig, kurzfädige, kurzfädigen, pünktchenförmig, rot-schnäblige, Rundbäuchig, rundbäuchige, sanftäugige, sanftäugigen, schnellfüßige, schnellfüßigen, spitztürmige, städtebauliche, städtebaulichen, städteübergreifend, täschchenlosen, viersträngig*

Words with non-concatenative properties: *angsthasig, aprilwettrig, dreistreifig, dunkelschalige, dünnschalig, dünnschalige, eigenpfotig, einhöckrigen, einstreifig, engmaschig, erdbeerartigen, flinkfingrige, flinkfingriger, grobbrockige, grobmaschig, grobmaschigen, grobmaschiger, großfenstrigen, großmaschig, großnasigen, hellschalig, hochgiebligen, hornbrilligen, langwimprigen, leichtpfotig, löwenmähnige, rotschalig, rotwangige, samtpfotigen, schmalhüftige, schnarchnasig, sonnenbebrillt, spitzgiebligen, Unbebrillt, zartschalig, zweihöckrige, zweistreifig*

Words with particle/*zu*-infix : *anföhnen, angeföhnt, aufdimensioniert, aufeinandergestapelt, aufgetürmt, aufgetürmten, auftürmen, aufzutürmen, beschuhten, coronabedingter, dichtgedrängten, eingerahmte, eingeschnürt, einrahmende, einschnürende, erdzugewandten, Fehlbefüllte, Fehlbefüllung, fehlbesetzt, fehlgeleitete, Fehlübersetzung, feindosiert, feindosierte, feingekleidete, fernsteuerbar, fernsteuerbarer, fertiggepackten, festbetoniert, festgerostet, festgeschraubt, geheimgehaltene, geheimzuhalten, geheimzuhaltenden, gutriechende, heißbegehrter,hitzebedingt, hochaufgetürmte,*

hochbeschuhten, kaputtanalysiert, kaputtgespart, kaputtgestanden, kaputtsanieren, kaputtschlägt, kaputtzukriegen, kaputtzusparen, kontinentübergreifende, kostümbedingt, krankheitsbedingt, mitgezittert, mitzittern, mitzitternden, nachzubauen, notbedingt, pandemiebedingt, plattgebügelten, plattgetrampelt, redimensioniert, sanktionsbedingten, schiefgelaufenen, schiefgelaufener, schiefstehende, schöngeredet, schönreden, schönzureden, sonnenzugewandten, straßenzugewandten, überdimensioniert, unterdimensioniert, vollgesprüht, vollgestapelt, vorbeiflanieren, weltzugewandter, zukunftszugewandte

Words from section 3.2 *Bekleidungsberuf – Berufsbekleidung, Stiefelleder – Lederstiefel, Fettbauch – Bauchfett, Zugluft – Luftzug, Stallkühe – Kuhstall, Drahtmaschen – Maschendraht, Teppichwolle – Wollteppich, Dauerprojekte – Projektdauer, Öloliven – Olivenöl, Schalenobst – Obstschale, Schachtelpappe – Pappschachtel, Tütenpapier – Papiertüten, Absatzschuhe – Schuhabsatz, Stoffschichten – Schichtstoffe, Druckkunst – Kunstdrucken*

Words from section 3.3: *Energieentlastungspakets, Energieentlastungspakete, Energieentlastungspaketen, Entlastungspaket, Gasengpässen, Gasengpasses, Gasengpass, Gasengpässe, Mindestfüllstände, Mindestfüllstand, Mindestfüllständen, Mindestfüllstands, Halbleiterengpässe, Halbleiterengpass, Halbleiterengpasses, Halbleiterengpässen, Testmüdigkeit, Testmüdigkeit, Endlos-Lockdown, Distanzlernens, Distanzlernen, distanzlernende, Distanzlernenden, ansteckungsfrei, ansteckungsfreien, ansteckungsfreiem, ansteckungsfreies, bemaskt, Impfbereitschaft, impfbereit, Impffrust, Herzschrittmacherträger, Parkraumbewirtschaftungskonzept, Fluggastdatensätze, Herzschrittmachertypen, Musiktauschbörse, Kochbuchautorinnen, Kochbuchautor, Kochbuchautoren, Atomkraftgegner, Kopfsteinpflasterpassage, massenvernichtungswaffenfreien, Hausstaubmilbenallergikern, Gebärdensprachdolmetscher:innen, Gebärdensprachdolmetscher, Kopftuchträgerin, Schilddrüsenhormontabletten, Sonnenblumenkernöl, Haifischflossensuppe, Abwasserbeseitigungspflicht, Knochenmarkspenderregister, Muttermilchersatzprodukten, Kinderbuchautorin, Maiglöckchenduft, Herrenarmbanduhr, Kunstrasenspielfeld, Kunstrasenspielfeldes, Dornröschendasein, Gabelstaplerführerschein, Festnetztelefonnummer, Massentierhaltungsanlagen, Mauerblümchendasein, Obstbaumschnittkurs, Kreuzworträtselfrage, Medizinjournalismus, Blutzuckerteststreifen, Kuhmilcheiweißallergie, Kleinkläranlage, Kläranlagenbetreiber, Kleinkläranlagenbetreiber*

event_chaos	The tiger ate the fruit . **It** resulted in chaos .
event_happened	The wolf ate the apple . **It** actually happened .
event_situation	The lion ate the carrot . **It** was a funny situation .
event_surprise	The owl ate the cake . **It** came as a surprise .
gender_step	I saw a **pineapple** . **It** was big .
object_overlap_eatdrinkdrink	The zebra ate the fruit and the **monkey** drank the tea . **It** liked the tea .
object_overlap_eatdrinkeat	The **lion** ate the fruit and the zebra drank the milk . **It** liked fruit .
object_verb_overlap_eatdrinkdrink	The mouse ate the cookie and the **bear** drank the milk . **It** drank the milk quickly .
object_verb_overlap_eatdrinkeat	The **zebra** ate the fruit and the lion drank the water . **It** ate the fruit quickly .
pleo_believe	The lion ate the ice cream . **It** is hard to believe this is true .
pleo_rain	The lion ate the pizza . **It** was raining .
pleo_seem	The frog ate the cookie . **It** seemed this was unnecessary .
pleo_shame	The giraffe ate the cheese . **It** is a shame .
subject_overlap_eatdrinkdrink	The turtle ate the bread and the dog drank the **tea** . The dog liked **it** .
subject_overlap_eatdrinkeat	The dove ate the **fruit** and the zebra drank the tea . The dove liked **it** .
subject_verb_overlap_eatdrinkdrink	The dove ate the apple and the frog drank the **water** . The frog drank **it** quickly .
subject_verb_overlap_eatdrinkeat	The mouse ate the **fruit** and the lion drank the tea . The mouse ate **it** quickly .
verb_overlap_eatdrinkdrink	The zebra ate and the **bear** drank . **It** drank quickly .
verb_overlap_eatdrinkeat	The **zebra** ate and the lion drank . **It** ate a lot .
world_knowledge	The **tiger** ate the ice cream . **It** was happy .
world_knowledge	The giraffe ate the **steak** . **It** was cooked .

Table 15: Overview of the different categories of reference in ContraCat. The noun that is referred to by the *it* in question, as well as the *it* itself, are marked in bold face. For the categories *event_** and *pleo_**, the *it* does not refer to a noun.