NRC Publications Archive
Archives des publications du CNRC

A challenge set approach to evaluating machine translation
Isabelle, Pierre; Cherry, Colin; Foster, George

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. /
La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version
acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

Computer Science, 2017-08-27

National Research
Council Canada

Conseil national de
recherches Canada

Canada

A Challenge Set Approach to Evaluating Machine Translation

Pierre Isabelle and **Colin Cherry** and **George Foster**
National Research Council Canada
`first.last@nrc.gc.ca`

Abstract

Neural machine translation represents an exciting leap forward in translation quality. But what longstanding weaknesses does it resolve, and which remain? We address these questions with a challenge set approach to translation evaluation and error analysis. A challenge set consists of a small set of sentences, each hand-designed to probe a system's capacity to bridge a particular structural divergence between languages. To exemplify this approach, we present an English-French challenge set, and use it to analyze phrase-based and neural systems. The resulting analysis provides not only a more fine-grained picture of the strengths of neural systems, but also insight into which linguistic phenomena remain out of reach.

1 Introduction

The advent of neural techniques in machine translation (MT) (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014) has led to profound improvements in MT quality. For "easy" language pairs such as English/French or English/Spanish in particular, neural (NMT) systems are much closer to human performance than previous statistical techniques (Wu et al., 2016). This puts pressure on automatic evaluation metrics such as BLEU (Papineni et al., 2002), which exploit surface-matching heuristics that are relatively insensitive to subtle differences. As NMT continues to improve, these metrics will inevitably lose their effectiveness. Another challenge posed by NMT systems is their opacity: while it was usually clear which phenomena were ill-handled by previous statistical systems—and why—these questions are more difficult to answer for NMT.

Src	The repeated calls from his mother **should** have alerted us.
Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
Sys	Les appels répétés de sa mère devraient nous avoir alertés.
Is the subject-verb agreement correct (y/n)? **Yes**	

Figure 1: Example challenge set question.

We propose a new evaluation methodology centered around a *challenge set* of difficult examples that are designed using expert linguistic knowledge to probe an MT system's capabilities. This methodology is complementary to the standard practice of randomly selecting a test set from "real text," which remains necessary in order to predict performance on new text. By concentrating on difficult examples, a challenge set is intended to provide a stronger signal to developers. Although we believe that the general approach is compatible with automatic metrics, we used manual evaluation for the work presented here. Our challenge set consists of short sentences that each focus on one particular phenomenon, which makes it easy to collect reliable manual assessments of MT output by asking direct yes-no questions. An example is shown in Figure 1.

We generated a challenge set for English to French translation by canvassing areas of linguistic divergence between the two language pairs, especially those where errors would be made visible by French morphology. Example choice was also partly motivated by extensive knowledge of the weaknesses of phrase-based MT (PBMT). Neither of these characteristics is essential to our method, however, which we envisage evolving as NMT progresses. We used our challenge set to evaluate in-house PBMT and NMT systems as well as Google's GNMT system.

In addition to proposing the novel idea of a challenge set evaluation, our contribution includes our annotated English-French challenge set, which we provide in an appendix and will make available in a separate machine-readable file. We also supply further evidence that NMT is systematically better than PBMT, even when BLEU score differences are small. Finally, we give an analysis of the challenges that remain to be solved in NMT, an area that has received little attention thus far.

2 Related Work

A number of recent papers have evaluated NMT using broad performance metrics. The WMT 2016 News Translation Task (Bojar et al., 2016) evaluated submitted systems according to both BLEU and human judgments. NMT systems were submitted to 9 of the 12 translation directions, winning 4 of these and tying for first or second in the other 5, according to the official human ranking. Since then, controlled comparisons have used BLEU to show that NMT outperforms strong PBMT systems on 30 translation directions from the United Nations Parallel Corpus (Junczys-Dowmunt et al., 2016a), and on the IWSLT English-Arabic tasks (Durrani et al., 2016). These evaluations indicate that NMT performs better on average than previous technologies, but they do not help us understand what aspects of the translation have improved.

Some groups have conducted more detailed error analyses. Bentivogli et al. (2016) carried out a number of experiments on IWSLT 2015 English-German evaluation data, where they compare machine outputs to professional post-edits in order to automatically detect a number of error categories. Compared to PBMT, NMT required less post-editing effort over-all, with substantial improvements in lexical, morphological and word order errors. NMT consistently out-performed PBMT, but its performance degraded faster as sentence length increased. Later, Toral and Sánchez-Cartagena (2017) conducted a similar study, examining the outputs of competition-grade systems for the 9 WMT 2016 directions that included NMT competitors. They reached similar conclusions regarding morphological inflection and word order, but found an even greater degradation in NMT performance as sentence length increased, perhaps due to these systems' use of subword units.

Most recently, Sennrich (2016) proposed an approach to perform targeted evaluations of NMT through the use of contrastive translation pairs. This method introduces a particular type of error automatically in reference sentences, and then checks whether the NMT system's conditional probability model prefers the original reference or the corrupted version. Using this technique, they are able to determine that a recently-proposed character-based model improves generalization on unseen words, but at the cost of introducing new grammatical errors.

Our approach differs from these studies in a number of ways. First, whereas others have analyzed sentences drawn from an existing bitext, we conduct our study on sentences that are manually constructed to exhibit canonical examples of specific linguistic phenomena. This challenge set methodology allows us to emphasize the difficult cases in an otherwise "easy" language pair. These sentences are designed to allow us to dive deep into phenomena of interest, and do a much finer-grained analysis of the strengths of NMT than has come before. However, this strategy also necessitates that we work on many fewer sentences. We leverage the small size of our challenge set to manually evaluate whether the system's actual output correctly handles our phenomena of interest. Manual evaluation side-steps some of the pitfalls that can come with Sennrich (2016)'s contrastive pairs, as a ranking of two contrastive sentences may not necessarily reflect whether the error in question will occur in the system's actual output.

3 Challenge Set Evaluation

Our challenge set is meant to measure the ability of MT systems to deal with some of the more difficult problems that arise in translating English into French. This particular language pair happened to be most convenient for us, but similar sets can be built for any language pair.

One aspect of MT performance that we aimed to exclude from our evaluation is robustness to sparse data. To control for this, when crafting source and reference sentences, we chose words that occurred at least 100 times in the training corpus described in section 4.1.[1]

---

[1]With three principled exceptions: *boeuf* (87 occurrences) and *spilt* (58 occurrences)—both part of idiomatic phrases—and *guitared* (0 occurrences).

3.1 Building the Challenge Set

The challenging aspect of the test set we are presenting stems from the fact that the source English sentences have been chosen so that their closest French equivalent will be *structurally divergent* from the source in some crucial way. Translational divergences have been extensively studied in the past – see for example (Vinay and Darbelnet, 1958; Dorr, 1994). We expect the level of difficulty of an MT test set to correlate well with its density in divergence phenomena. We classify divergence phenomena into three main types: morpho-syntactic, lexico-syntactic and purely syntactic divergences.

Morpho-syntactic divergences

In some languages, word morphology (e.g. inflections) carries more grammatical information than in others. When translating a word towards the richer language, there is a need to recover additional grammatically-relevant information from the context of the target language word. Note that we only include in our set cases where the relevant information is available in the *linguistic* context.[2]

We lack the space to describe all the subtypes of morpho-syntactic divergences that appear in our challenge set, but illustrate through representative examples. One particularly important case is that of *subject-verb agreement*. French verbs typically have more than 30 different inflected forms, while English verbs typically have 4 or 5. As a result, English verb forms strongly underspecify their French counterparts. Much of the missing information must be filled in through forced agreement in person, number and gender with the grammatical subject of the verb. But extracting these parameters can prove difficult. For example, the agreement features of a coordinated noun phrase are a complex function of the coordinated elements: a) the gender is feminine if all conjuncts are feminine, otherwise masculine wins; b) the conjunct with the smallest person (p1<p2<p3) wins; and c) the number is always plural when the coordination is "et" but the case is more complex with "ou".

A second example of morpho-syntactic divergence between English and French is the more explicit marking of the *subjunctive mood* in French subordinate clauses. In the following example, the verb "partiez", unlike its English counterpart, is marked as subjunctive:

> He demanded that you leave immediately. → Il a exigé que vous *partiez* immédiatement.

When translating an English verb within a subordinate clause, the context must be examined for possible subjunctive triggers. Typically these are specific lexical items found in a governing position with respect to the subordinate clause: verbs such as "exiger que", adjectives such as "regrettable que" or subordinate conjunctions such as "à condition que".

Lexico-syntactic divergences

Syntactically governing words such as verbs tend to impose specific requirements on their complements: they *subcategorize* for complements of a certain syntactic type. But a source language governor and its target language counterpart can diverge on their respective requirements. The translation of such words must then trigger adjustments in the target language complement pattern. We can only examine here a few of the subtypes instantiated in our challenge set.

A good example is *argument switching*. This refers to the situation where the translation of a source verb V_s as V_t is correct but only provided the arguments (usually the subject and the object) are flipped around. The translation of "to miss" as "manquer à" is such a case:

> John misses Mary → Mary *manque à* John.

Failing to perform the switch results in a severe case of mistranslation.

A second example of lexico-syntactic divergence is that of "crossing movement" verbs. Consider the following example:

> Terry swam across the river → Terry *a traversé* la rivière *à la nage*.

The French translation could be glossed as, "Terry crossed the river by swimming." A literal translation such as "Terry a nagé à travers la rivière," is ruled out.

---

[2] The so-called Winograd Schema Challenges (https://en.wikipedia.org/wiki/Winograd_Schema_Challenge) often involve cases where common-sense reasoning is required to correctly choose between two potential antecedent phrases for a pronoun. Such cases become En → Fr translation challenges when the pronoun in the source sentence is *they* and its alternative antecedents happen to have different grammatical genders in French: *they → ils/elles*.

Syntactic divergences

Some syntactic divergences are not relative to the presence of a particular lexical item but rather stem from differences in the basic set of available syntactic patterns. Source-language instances of structures that do not exist in the target language must be mapped onto equivalent structures. Here are some of the subtypes appearing in our challenge set.

The position of French pronouns is a major case of divergence from English. French is basically an SVO language just like English but it departs from that canonical order when post-verbal complements are pronominalized: the pronouns must then be *cliticized*, that is phonetically attached to the verb, in this case to the left side of the verb.

> He gave Mary a book. $\rightarrow$ Il a donné un livre à Marie.
>
> He gave$_i$ it$_j$ to her$_k$. $\rightarrow$ Il *le$_j$ lui$_k$* a donné$_i$.

Another example of syntactic divergence between English and French is that of *stranded prepositions*. When forming a relative clause or a question on a prepositional phrase, English can leave the preposition stranded, fronting only the pronominalized object of that preposition. In French, the preposition needs to be fronted alongside its object:

> The girl whom$_i$ he was dancing with$_j$ is rich. $\rightarrow$ La fille *avec$_j$ qui$_i$* il dansait est riche.

A final example of syntactic divergence is the use of the so-called *middle voice*. While English uses the passive voice in subjectless generic statements, French tends to prefer the use of a special pronominal construction where the pronoun "se" has no real referent:

> Caviar is eaten with bread. $\rightarrow$ Le caviar *se mange* avec du pain.

This completes our exemplification of morpho-syntactic, lexico-syntactic and purely syntactic divergences. Our actual test set includes several more subcategories of each type. The ability of MT systems to deal with each such subcategory is then tested using at least three different test sentences. We use short test sentences so as to keep the targeted divergence in focus. The 108 sentences that constitute our current challenge set can be found in Appendix B.

3.2 Evaluation Methodology

Given the very small size of our challenge set, it is easy to perform a human evaluation of the respective outputs of a handful of different systems. The obvious advantage is that the assessment is then absolute instead of relative to one or a few reference translations.

The intent of each challenge sentence is to test one and only one system capability, namely that of coping correctly with the particular associated divergence subtype. As illustrated in Figure 1, we provide annotators with a question that specifies the divergence phenomenon currently being tested, along with a reference translation with the areas of divergence highlighted. As a result, judgments become straightforward: was the targeted divergence correctly bridged, yes or no?[3] There is no need to mentally average over a number of different aspects of the test sentence as one does when rating the global translation quality of a sentence, e.g. on a 5-point scale. However, we acknowledge that measuring translation performance on complex sentences exhibiting many different phenomena remains crucial. We see our approach as being complementary to evaluations of overall translation quality.

One consequence of our divergence-focused approach is that faulty translations will be judged as successes when the faults lie outside of the targeted divergence zone. However, this problem is mitigated by our use of short test sentences.

4 Machine Translation Systems

We trained state-of-the-art neural and phrase-based systems for English-French translation on data from the WMT 2014 evaluation.

4.1 Data

We used the LIUM shared-task subset of the WMT 2014 corpora,[4] retaining the provided tokenization and corpus organization, but mapping characters to lowercase. Table 1 gives corpus statistics.

---

[3] Sometimes the system produces a translation that circumvents the divergence issue. For example, it may dodge a divergence involving adverbs by reformulating the translation to use an adjective instead. In these rare cases, we instruct our annotators to abstain from making a judgment, regardless of whether the translation is correct or not.

[4] http://www.statmt.org/wmt14/translation-task.html http://www-lium.univ-lemans.fr/~schwenk/nnmt-shared-task

corpus	lines	en words	fr words
train	12.1M	304M	348M
mono	15.9M	—-	406M
dev	6003	138K	155K
test	3003	71K	81K

Table 1: Corpus statistics. The WMT12/13 eval sets are used for dev, and the WMT14 eval set is used for test.

4.2 Phrase-based systems

To ensure a competitive PBMT baseline, we performed phrase extraction using both IBM4 and HMM alignments with a phrase-length limit of 7; after frequency pruning, the resulting phrase table contained 516M entries. For each extracted phrase pair, we collected statistics for the hierarchical reordering model of Galley and Manning (2008).

We trained an NNJM model (Devlin et al., 2014) on the HMM-aligned training corpus, with input and output vocabulary sizes of 64K and 32K. Words not in the vocabulary were mapped to one of 100 mkcls classes. We trained for 60 epochs of 20K x 128 minibatches, yielding a final dev-set perplexity of 6.88.

Our set of log-linear features consisted of forward and backward Kneser-Ney smoothed phrase probabilities and HMM lexical probabilities (4 features); hierarchical reordering probabilities (6); the NNJM probability (1); a set of sparse features as described by Cherry (2013) (10,386); word-count and distortion penalties (2); and 5-gram language models trained on the French half of the training corpus and the French monolingual corpus (2). Tuning was carried out using batch lattice MIRA (Cherry and Foster, 2012). Decoding used the cube-pruning algorithm of Huang and Chiang (2007), with a distortion limit of 7.

We include two phrase-based systems in our comparison: PBMT-1 has data conditions that exactly match those of the NMT system, in that it does not use the language model trained on the French monolingual corpus, while PBMT-2 uses both language models.

4.3 Neural systems

To build our NMT system, we used the Nematus toolkit,[5] which implements a single-layer neural sequence-to-sequence architecture with attention (Bahdanau et al., 2015) and gated recurrent

units (Cho et al., 2014). We used 512-dimensional word embeddings with source and target vocabulary sizes of 90K, and 1024-dimensional state vectors. The model contains 172M parameters.

We preprocessed the data using a BPE model learned from source and target corpora (Sennrich et al., 2016). Sentences longer than 50 words were discarded. Training used the Adadelta algorithm (Zeiler, 2012), with a minibatch size of 100 and gradients clipped to 1.0. It ran for 5 epochs, writing a checkpoint model every 30K minibatches. Following Junczys-Dowmunt et al. (2016b), we averaged the parameters from the last 8 checkpoints. To decode, we used the AmuNMT decoder (Junczys-Dowmunt et al., 2016a) with a beam size of 4.

While our primary results will focus on the above PBMT and NMT systems, where we can describe replicable configurations, we have also evaluated Google's production system,[6] which has recently moved to NMT (Wu et al., 2016). Notably, the "GNMT" system uses (at least) 8 encoder and 8 decoder layers, compared to our 1 layer for each, and it is trained on corpora that are "two to three decimal orders of magnitudes bigger than the WMT." The evaluated outputs were downloaded in December 2016.

5 Experiments

The 108-sentence English-French challenge set presented in Appendix B was submitted to the four MT systems described in section 4: PBMT-1, PBMT-2, NMT, and GNMT. We employed three bilingual native speakers of French who had no prior knowledge of the challenge set. They rated each translated sentence as either a success or a failure according to the protocol described in section 3.2. For example, the 26 sentences of the sub-categories S1-S5 of Appendix B are all about different cases of subject-verb agreement. The corresponding translations were judged successful if and only if the translated verb correctly agrees with the translated subject.

The different system outputs for each source sentence were grouped together to reduce the burden on the annotators. That is, in figure 1, annotators were asked to answer the question for each of four outputs, rather than just one as shown. The outputs were listed in random order, without identification. Questions were also presented in ran-

---

[5]https://github.com/rsennrich/nematus

[6]https://translate.google.com

dom order to each annotator. Appendix A contains the instructions shown to the annotators.

5.1 Quantitative comparison

Table 2 summarizes our results in terms of percentage of successful translations, globally and over each main type of divergence. For comparison with traditional metrics, we also include BLEU scores measured on the WMT 2014 test set.

As we can see, the two PBMT systems fare very poorly on our challenge set, especially in the morpho-syntactic and purely syntactic types. Their relatively better handling of lexico-syntactic cases probably reflects the fact that PBMT systems are naturally more attuned to lexical cues than to morphology or syntax. The two NMT systems are clear winners in all three categories. The GNMT system is best overall with a success rate of 68%, likely due to the data and architectural factors mentioned in section 4.3.[7]

WMT BLEU scores correlate poorly with challenge-set performance. The large gap of 2.3 BLEU points between PBMT-1 and PBMT-2 corresponds to only a 1% gain on the challenge set, while the small gap of 0.4 BLEU between PBMT-2 and NMT corresponds to a 21% gain.

Inter-annotator agreement (final column in table 2) is excellent overall, with all three annotators agreeing on almost 90% of system outputs. Syntactic divergences appear to be somewhat harder to judge than other categories.

5.2 Qualitative assessment of NMT

We now turn to an analysis of the strengths and weaknesses of neural MT through the microscope of our divergence categorization system, hoping that this may help focus future research on key issues. In this discussion we ignore the results obtained by PBMT-2 and compare: a) the results obtained by PBMT-1 to those of NMT, both systems having been trained on the same dataset; and b) the results of these two systems with those of Google NMT which was trained on a much larger dataset.

In the remainder of the present section we will reference the sentences of our challenge set using the subcategory-based numbering scheme S1-S26 as assigned in Appendix B.

---

[7]We cannot offer a full comparison with the pre-NMT Google system. However, in October 2016 we ran a smaller 35-sentence version of our challenge set on both the Google system and our PBMT-1 system. The Google system only got 4 of those examples right (11.4%) while our PBMT-1 got 6 (17.1%).

Strengths of neural MT

Overall, both neural MT systems do much better than PBMT-1 at bridging divergences. Their dramatic advantage on morpho-syntactic divergences (a jump from 16% to 72% in the case of our two local systems) results from achievements such as the following:

- The subject's head noun agreement features get correctly passed to the verb phrase across intervening noun phrase complements (sentences S1a-c).

- Subject agreement marks appear to be correctly distributed to each element of a coordinated verb phrase (S3a-c).

- Much of the calculus that is at stake in determining the agreement features of a subject noun phrase (cf. our relevant description in section 3.1) appears to be correctly captured in the 12 translations of S4.

- Most instances of the difficult case of past participle agreement after the "avoir" auxiliary are correctly handled (S5b-e).

The NMT systems are also better at handling lexico-syntactic divergences. For example:

- They can perform the required restructuring of English double object constructions (sentences S8a-S8c).

- They can discriminate between an NP complement and a sentential complement starting with an NP: cf. *to know NP* versus *to know NP is VP* (S11b-e)

- They often correctly restructure English NP-to-VP complements (S12a-c).

Finally, NMT systems also turn out to better handle purely syntactic divergences. For example:

- The differences in yes-no question syntax is correctly bridged (S17a-c).

- English pronouns in verb complement position are often correctly cliticized, that is, moved before the main verb and case-inflected correctly (S23a-e).

- The Google NMT system manages to correctly translate tag questions (S18a-c), most cases of the "inalienable possession" construction (S25a-e), zero relative pronouns (S26a-c) and constructions with stranded prepositions (S19a-f).

Divergence type	PBMT-1	PBMT-2	NMT	Google NMT	Agreement
Morpho-syntactic	16%	16%	72%	65%	94%
Lexico-syntactic	42%	46%	52%	62%	94%
Syntactic	33%	33%	40%	75%	81%
Overall	31%	32%	53%	68%	89%
WMT BLEU	34.2	36.5	36.9	—	—

Table 2: Summary performance statistics for each system under study, including challenge set success rate grouped by linguistic category, as well as BLEU scores on the WMT 2014 test set. The final column gives the proportion of system outputs on which all three annotators agreed.

The large gap observed between the results of the in-house and Google NMT systems indicates that current neural MT systems are extremely data hungry. But given enough data, they can successfully tackle some challenges that are often thought of as extremely difficult. A case in point is that of stranded prepositions, in which English and French happen to diverge in their handling of the celebrated "WH-movement" long-distance dependencies. Specifically, in the French translation, the preposition must be repatriated with its fronted WH object no matter how far on the left it happens to be.

Weaknesses of neural MT

In spite of its clear edge over PBMT, NMT is not without some serious shortcomings. Some of them have been mentioned already, such as the tendency of system output to degrade with sentence length. By design this particular problem could not be observed with our challenge set. But many others get highlighted by an analysis of our results. Globally, we note that even using a staggering quantity of data and a highly sophisticated NMT model, the Google system fails to reach the 70% mark on our challenge set. Thus, there is ample room for improvement. The fine-grained error categorization associated with the challenge set makes it possible to single out precise areas where more research is needed. A first analysis of our results yields the following observations.

Incomplete generalizations. In several cases, while partial results might suggest that NMT has correctly captured a basic generalization about linguistic data, further instances reveal that this is not fully the case. Here are some examples:

- The calculus governing the agreement features of coordinated noun phrases (see section 3.1) appears to be handled correctly most of the time. However unlike our NMT system, the Google NMT system gets into difficulty with mixed-person subjects (sentences S4d1-3).

- While some subjunctive mood triggers are correctly captured (e.g. "demander que" and "malheureux que"), others such as the very common subordinate conjunction *provided that → à condition que* are getting missed (sentence S6a).

- The NMT systems often appear to have successfully captured the semantic relation that ties together the two nouns of an English noun compound, thereby giving rise to the correct preposition in the French translation $N_1 N_2 → N_2 Prep N_1$. However, some cases that one might think of as easy are being missed. For example, the Google translation of "steak knife" (sentence S14c) fails to convey that this is a knife intended to cut steak; similarly, the Google translation of "paper filter" (sentence S14i) suggests the filter is intended to filter paper rather being made of it.

- The so-called French "inalienable possession" construction arises when an agent performs an action on one of her body parts, e.g. *I brushed my teeth*. In such cases the French translation normally follows a pattern that can be glossed as *He brushed the teeth to himself*. In our dataset, the Google system gets this right for examples in the first and third persons (sentences S25a,b) but fails to do the same with the example in the second person (sentence S25c).

Then there are also phenomena that current NMT systems, even with massive amounts of data, appear to be completely missing:

- *Idioms*. While PBMT-1 produces an acceptable translation for half of the idiomatic ex-

pressions of S15 and S16, the local NMT system misses them all and the Google system does just slightly better. It looks as if NMT systems lack sufficient capacity for raw memorization.

- *Control verbs*. Two different classes of verbs can govern a subject NP, an object NP plus an infinitival complement. With verbs of the "object-control" class (e.g. "persuade"), the object of the verb is understood as the semantic subject of the infinitive. But with those of the subject class (e.g. "promise"), it is rather the subject of the verb which plays that semantic role. None of the systems tested here appear to get a grip on subject control cases, as evidenced by the lack of correct feminine agreement on the French adjectives in sentences S2b-d.

- *Argument switching verbs*. All systems tested here mistranslate sentences S7a-c by failing to perform the required argument switch: *NP₁ misses NP₂ → NP₂ manque à NP₁*.

- *Crossing movement verbs*. None of the systems managed to correctly restructure the regular manner-of-movement verbs e.g. *swim across X → traverser X à la nage* in sentences S10a-c, let alone the even harder example S10d, in which the word "guitar" is spontaneously recast as a manner-of-movement verb.

- Middle voice. None of the systems tested here were able to recast the English "generic passive" of S21a-c into the expected French "middle voice" pronominal construction.

6 Conclusions

We have presented a radically different kind of evaluation for machine translation systems: the use of challenge sets designed to stress-test MT systems on "hard" linguistic material, while providing a fine-grained linguistic classification of their successes and failures. This approach is not meant to replace our community's traditional evaluation tools but to supplement them.

Our proposed error categorization scheme makes it possible to bring to light different strengths and weaknesses of PBMT and neural MT. With the exception of idiom processing, in all cases where a clear difference was observed it turned out to be in favor of neural MT. A key

factor in NMT's superiority appears to be its ability to overcome many limitations of n-gram language modeling. This is clearly at play in dealing with subject-verb agreement, double-object verbs, overlapping subcategorization frames and last but not least, the pinnacle of Chomskyan linguistics, WH-movement (in this case, stranded prepositions).

But our challenge set also brings to light some important shortcomings of current neural MT, regardless of the massive amounts of training data it may have been fed. As may have been already known or suspected, NMT systems struggle with the translation of idiomatic phrases. Perhaps more interestingly, we notice that neural MT's impressive generalizations still seem somewhat brittle. For example, the NMT system can appear to have mastered the rules governing subject-verb agreement or inalienable possession in French, only to trip over a rather obvious instantiation of those rules. Probing where these boundaries are, and how they relate to the neural system's training data and architecture is an obvious next step.

7 Future Work

It is our hope that the insights derived from our challenge set evaluation will help inspire future MT research, and call attention to the fact that even "easy" language pairs like English-French still have many linguistic issues left to be resolved. But there are also several ways to improve and expand upon our challenge set approach itself.

First, though our human judgments of output sentences allowed us to precisely assess the phenomena of interest, this approach is not scalable to large sets, and requires access to native speakers in order to replicate the evaluation. It would be interesting to see whether similar scores could be achieved through automatic means. The existence of human judgments for this set provides a gold-standard by which proposed automatic judgments may be meta-evaluated.

Second, the construction of such a challenge set is as much an art as a science, and requires in-depth knowledge of the structural divergences between the two languages of interest. A method to automatically create such a challenge set for a new language pair would be extremely useful. One could imagine approaches that search for divergences, indicated by atypical output configurations, or perhaps by a system's inability to repro-

duce a reference from its own training data. Localizing a divergence within a difficult sentence pair would be another useful subtask.

Finally, and perhaps most interestingly, we would like to explore how to train an MT system to improve its performance on these divergence phenomena. This could take the form of designing a curriculum to demonstrate a particular divergence to the machine, or altering the network structure to more easily capture such generalizations.

Acknowledgments

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the Third International Conference on Learning Representations (ICLR)*. San Diego, USA. http://arxiv.org/abs/1409.0473.

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*. Association for Computational Linguistics, Austin, Texas, pages 257–267. https://aclweb.org/anthology/D16-1025.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W16-2301.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 22–31. http://www.aclweb.org/anthology/N13-1003.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, pages 427–436. http://www.aclweb.org/anthology/N12-1047.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1724–1734. http://www.aclweb.org/anthology/D14-1179.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 1370–1380. http://www.aclweb.org/anthology/P14-1129.

Bonnie J. Dorr. 1994. Machine translation divergences: a formal description and proposed solution. *Computational Linguistics* 20:4.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2016. QCRI machine translation systems for IWSLT 16. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, Washington.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 848–856. http://www.aclweb.org/anthology/D08-1089.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 144–151. http://www.aclweb.org/anthology/P07-1019.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016a. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, Washington.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016b. The amu-uedin submission to

the wmt16 news translation task: Attention-based nmt models as feature functions in phrase-based smt. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 319–325. http://www.aclweb.org/anthology/W16-2316.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1700–1709. http://www.aclweb.org/anthology/D13-1176.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Rico Sennrich. 2016. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. *CoRR* abs/1612.04629. http://arxiv.org/abs/1612.04629.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., pages 3104–3112.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus statistical machine translation for 9 language directions. In *Proceedings of the The 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Valencia, Spain.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais*, volume 1. Didier, Paris.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144. http://arxiv.org/abs/1609.08144.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR* abs/1212.5701. http://arxiv.org/abs/1212.5701.

A Instructions to Annotators

The following instructions were provided to annotators:

You will be presented with 108 short English sentences and the French translations produced for them by each of four different machine translation systems. You will not be asked to provide an overall rating for the machine-translated sentences. Rather, you will be asked to determine whether or not a highly specific aspect of the English sentence is correctly rendered in each of the different translations. Each English sentence will be accompanied with a yes-no question which precisely specifies the targeted element for the associated translations. For example, you may be asked to determine whether or not the main verb phrase of the translation is in correct grammatical agreement with its subject.

In order to facilitate this process, each English sentence will also be provided with a French reference (human) translation in which the particular elements that support a yes answer (in our example, the correctly agreeing verb phrase) will be highlighted. Your answer should be "yes" if the question can be answered positively and "no" otherwise. Note that this means that any translation error which is unrelated to the question at hand should be disregarded. Using the same example: as long as the verb phrase agrees correctly with its subject, it does not matter whether or not the verb is correctly chosen, is in the right tense, etc. And of course, it does not matter if unrelated parts of the translation are wrong.

In most cases you should be able to quickly determine a positive or negative answer. However, there may be cases in which the system has come up with a translation that just does not contain the phenomenon targeted by the associated question. In such cases, and only in such cases, you should choose "not applicable" regardless of whether or not the translation is correct.

B Challenge Set

We include a rendering of our challenge set in the pages that follow, along with system output for the PBMT-1, NMT and Google systems. Sentences are grouped by linguistic category and subcategory. For convenience, we also include a reference translation, which is a manually-crafted translation that is designed to be the most straightforward solution to the divergence problem at hand. Need-

less to say, this reference translation is seldom the only acceptable solution to the targeted divergence problem. Our judges were provided these references, but were instructed to use their knowledge of French to judge whether the divergence was correctly bridged, regardless of the translation's similarity to the reference.

In all translations, the locus of the targeted divergence is highlighted in boldface and it is specifically on that portion that our annotators were asked to provide a judgment. For each system output, we provide a summary of our annotator's judgments on its handling of the phenomenon of interest. We label the translation with a ✓ if two or more annotators judged the divergence to be correctly bridged, and with an ✗ otherwise.

We also release a machine-readable version of this same data, including all of the individual judgments, in the hope that others will find interesting new uses for it.

Morpho-Syntactic

S-V agreement, across distractors

Is subject-verb agrement correct? (Possible interference from distractors between the subject's head and the verb).

S1a	Source	The repeated calls from his mother **should** have alerted us.
	Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
	PBMT-1	Les appels répétés de sa mère aurait dû nous a alertés. ✗
	NMT	Les appels répétés de sa mère devraient nous avoir alertés. ✓
	Google	Les appels répétés de sa mère auraient dû nous alerter. ✓
S1b	Source	The sudden noise in the upper rooms **should** have alerted us.
	Ref	Le bruit soudain dans les chambres supérieures **aurait** dû nous alerter.
	PBMT-1	Le bruit soudain dans les chambres supérieures auraient dû nous a alertés. ✗
	NMT	Le bruit soudain dans les chambres supérieures devrait nous avoir alerté. ✓
	Google	Le bruit soudain dans les chambres supérieures devrait nous avoir alerté. ✓
S1c	Source	Their repeated failures to report the problem **should** have alerted us.
	Ref	Leurs échecs répétés à signaler le problème **auraient** dû nous alerter.
	PBMT-1	Leurs échecs répétés de signaler le problème aurait dû nous a alertés. ✗
	NMT	Leurs échecs répétés pour signaler le problème devraient nous avoir alertés. ✓
	Google	Leur échec répété à signaler le problème aurait dû nous alerter. ✓

S-V agreement, through control verbs

Does the flagged adjective agree correctly with its subject? (Subject-control versus object-control verbs).

S2a	Source	She asked her brother not to be **arrogant**.
	Ref	Elle a demandé à son frère de ne pas se montrer **arrogant**.
	PBMT-1	Elle a demandé à son frère de ne pas être arrogant. ✓
	NMT	Elle a demandé à son frère de ne pas être arrogant. ✓
	Google	Elle a demandé à son frère de ne pas être arrogant. ✓
S2b	Source	She promised her brother not to be **arrogant**.
	Ref	Elle a promis à son frère de ne pas être **arrogante**.
	PBMT-1	Elle a promis son frère à ne pas être arrogant. ✗
	NMT	Elle a promis à son frère de ne pas être arrogant. ✗
	Google	Elle a promis à son frère de ne pas être arrogant. ✗
S2c	Source	She promised her doctor to remain **active** after retiring.
	Ref	Elle a promis à son médecin de demeurer **active** après s'être retirée.
	PBMT-1	Elle a promis son médecin pour demeurer actif après sa retraite. ✗
	NMT	Elle a promis à son médecin de rester actif après sa retraite. ✗
	Google	Elle a promis à son médecin de rester actif après sa retraite. ✗
S2d	Source	My mother promised my father to be more **prudent** on the road.
	Ref	Ma mère a promis à mon père d'être plus **prudente** sur la route.
	PBMT-1	Ma mère, mon père a promis d'être plus prudent sur la route. ✗
	NMT	Ma mère a promis à mon père d'être plus prudent sur la route. ✗
	Google	Ma mère a promis à mon père d'être plus prudent sur la route. ✗

S-V agreement, coordinated targets

Do the marked verbs/adjective agree correctly with their subject? (Agreement distribution over coordinated predicates)

S3a	Source	The woman was very **tall** and extremely **strong**.
	Ref	La femme était très **grande** et extrêmement **forte**.
	PBMT-1	La femme était très gentil et extrêmement forte. ✗
	NMT	La femme était très haute et extrêmement forte. ✓
	Google	La femme était très grande et extrêmement forte. ✓
S3b	Source	Their politicians were more **ignorant** than **stupid**.
	Ref	Leurs politiciens étaient plus **ignorants** que **stupides**.
	PBMT-1	Les politiciens étaient plus ignorants que stupide. ✗
	NMT	Leurs politiciens étaient plus ignorants que stupides. ✓
	Google	Leurs politiciens étaient plus ignorants que stupides. ✓
S3c	Source	We **shouted** an insult and **left** abruptly.
	Ref	Nous **avons** lancé une insulte et nous **sommes** partis brusquement.
	PBMT-1	Nous avons crié une insulte et a quitté abruptement. ✗
	NMT	Nous avons crié une insulte et nous avons laissé brusquement. ✓
	Google	Nous avons crié une insulte et nous sommes partis brusquement. ✓

S-V agreement, feature calculus on coordinated source

Do the marked verbs/adjective agree correctly with their subject? (Masculine singular ET masculine singular yields masculine plural).

S4a1	Source	The cat and the dog **should** be **watched**.
	Ref	Le chat et le chien **devraient** être **surveillés**.
	PBMT-1	Le chat et le chien doit être regardée. ✗
	NMT	Le chat et le chien doivent être regardés. ✓
	Google	Le chat et le chien doivent être surveillés. ✓
S4a2	Source	My father and my brother **will** be **happy** tomorrow.
	Ref	Mon père et mon frère **seront heureux** demain.
	PBMT-1	Mon père et mon frère sera heureux de demain. ✗
	NMT	Mon père et mon frère seront heureux demain. ✓
	Google	Mon père et mon frère seront heureux demain. ✓
S4a3	Source	My book and my pencil **could** be **stolen**.
	Ref	Mon livre et mon crayon **pourraient** être **volés**.
	PBMT-1	Mon livre et mon crayon pourrait être volé. ✗
	NMT	Mon livre et mon crayon pourraient être volés. ✓
	Google	Mon livre et mon crayon pourraient être volés. ✓

Do the marked verbs/adjectives agree correctly with their subject? (Feminine singular ET feminine singular yields feminine plural).

S4b1	Source	The cow and the hen **must** be **fed**.
	Ref	La vache et la poule **doivent** être **nourries**.
	PBMT-1	La vache et de la poule doivent être nourris. ✗
	NMT	La vache et la poule doivent être alimentées. ✓
	Google	La vache et la poule doivent être nourries. ✓

S4b2	Source	My mother and my sister **will be happy** tomorrow.
	Ref	Ma mère et ma sœur **seront heureuses** demain.
	PBMT-1	Ma mère et ma sœur sera heureux de demain. ✗
	NMT	Ma mère et ma sœur seront heureuses demain. ✓
	Google	Ma mère et ma sœur seront heureuses demain. ✓
S4b3	Source	My shoes and my socks **will** be **found**.
	Ref	Mes chaussures et mes chaussettes **seront retrouvées**.
	PBMT-1	Mes chaussures et mes chaussettes sera trouvé. ✗
	NMT	Mes chaussures et mes chaussettes seront trouvées. ✓
	Google	Mes chaussures et mes chaussettes seront trouvées. ✓

Do the marked verbs/adjectives agree correctly with their subject? (Masculine singular ET feminine singular yields masculine plural.)

S4c1	Source	The dog and the cow **are nervous**.
	Ref	Le chien et la vache **sont nerveux**.
	PBMT-1	Le chien et la vache sont nerveux. ✓
	NMT	Le chien et la vache sont nerveux. ✓
	Google	Le chien et la vache sont nerveux. ✓
S4c2	Source	My father and my mother will be happy tomorrow.
	Ref	Mon père et ma mère **seront heureux** demain.
	PBMT-1	Mon père et ma mère se fera un plaisir de demain. ✗
	NMT	Mon père et ma mère seront heureux demain. ✓
	Google	Mon père et ma mère seront heureux demain. ✓
S4c3	Source	My refrigerator and my kitchen table **were stolen**.
	Ref	Mon réfrigérateur et ma table de cuisine **ont** été **volés**.
	PBMT-1	Mon réfrigérateur et ma table de cuisine ont été volés. ✓
	NMT	Mon réfrigérateur et ma table de cuisine ont été volés. ✓
	Google	Mon réfrigérateur et ma table de cuisine ont été volés. ✓

Do the marked verbs/adjectives agree correctly with their subject? (Smallest coordinated grammatical person wins.)

S4d1	Source	Paul and I **could** easily be **convinced** to join you.
	Ref	Paul et moi **pourrions** facilement être **convaincus** de se joindre à vous.
	PBMT-1	Paul et je pourrais facilement être persuadée de se joindre à vous. ✗
	NMT	Paul et moi avons facilement pu être convaincus de vous rejoindre. ✓
	Google	Paul et moi pourrait facilement être convaincu de vous rejoindre. ✗
S4d2	Source	You and he **could** be **surprised** by her findings.
	Ref	Vous et lui **pourriez** être **surpris** par ses découvertes.
	PBMT-1	Vous et qu'il pouvait être surpris par ses conclusions. ✗
	NMT	Vous et lui pourriez être surpris par ses conclusions. ✓
	Google	Vous et lui pourrait être surpris par ses découvertes. ✗

S4d3	Source	We and they **are** on different courses.
	Ref	Nous et eux **sommes** sur des trajectoires différentes.
	PBMT-1	Nous et ils sont en cours de différents. ✗
	NMT	Nous et nous sommes sur des parcours différents. ✗
	Google	Nous et ils sont sur des parcours différents. ✗

S-V agreement, past participles

Are the agreement marks of the flagged participles the correct ones? (Past participle placed after auxiliary AVOIR agrees with verb object iff object precedes auxiliary. Otherwise participle is in masculine singular form).

S5a	Source	The woman who **saw** a mouse in the corridor is charming.
	Ref	La femme qui a **vu** une souris dans le couloir est charmante.
	PBMT-1	La femme qui a vu une souris dans le couloir est charmante. ✓
	NMT	La femme qui a vu une souris dans le couloir est charmante. ✓
	Google	La femme qui a vu une souris dans le couloir est charmante. ✓
S5b	Source	The woman that your brother **saw** in the corridor is charming.
	Ref	La femme que votre frère a **vue** dans le couloir est charmante.
	PBMT-1	La femme que ton frère a vu dans le couloir est charmante. ✗
	NMT	La femme que votre frère a vu dans le corridor est charmante. ✗
	Google	La femme que votre frère a vue dans le couloir est charmante. ✓
S5c	Source	The house that John has **visited** is crumbling.
	Ref	La maison que John a **visitée** tombe en ruines.
	PBMT-1	La maison que John a visité est en train de s'écrouler. ✗
	NMT	La maison que John a visitée est en train de s'effondrer. ✓
	Google	La maison que John a visité est en ruine. ✗
S5d	Source	John sold the car that he had **won** in a lottery.
	Ref	John a vendu la voiture qu'il avait **gagnée** dans une loterie.
	PBMT-1	John a vendu la voiture qu'il avait gagné à la loterie. ✗
	NMT	John a vendu la voiture qu'il avait gagnée dans une loterie. ✓
	Google	John a vendu la voiture qu'il avait gagnée dans une loterie. ✓

Subjunctive mood

Is the flagged verb in the correct mood? (Certain triggering verbs, adjectives or subordinate conjunctions, induce the subjunctive mood in the subordinate clause that they govern).

S6a	Source	He will come provided that you **come** too.
	Ref	Il viendra à condition que vous **veniez** aussi.
	PBMT-1	Il viendra à condition que vous venez aussi. ✗
	NMT	Il viendra lui aussi que vous le faites. ✗
	Google	Il viendra à condition que vous venez aussi. ✗
S6b	Source	It is unfortunate that he is not **coming** either.
	Ref	Il est malheureux qu'il ne **vienne** pas non plus.
	PBMT-1	Il est regrettable qu'il n'est pas non plus à venir. ✗
	NMT	Il est regrettable qu'il ne soit pas non plus. ✗
	Google	Il est malheureux qu'il ne vienne pas non plus. ✓

S6c	Source	I requested that families not **be** separated.
	Ref	J'ai demandé que les familles ne **soient** pas séparées.
	PBMT-1	J'ai demandé que les familles ne soient pas séparées. ✓
	NMT	J'ai demandé que les familles ne soient pas séparées. ✓
	Google	J'ai demandé que les familles ne soient pas séparées. ✓

Lexico-Syntactic

Argument switch

Are the experiencer and the object of the "missing" situation correctly preserved in the French translation? (Argument switch).

S7a	Source	**Mary** sorely misses **Jim**.
	Ref	**Jim** manque cruellement **à Mary**.
	PBMT-1	Marie manque cruellement de Jim. ✗
	NMT	Mary a lamentablement manqué de Jim. ✗
	Google	Mary manque cruellement à Jim. ✗
S7b	Source	**My sister** is really missing **New York.**
	Ref	**New York** manque beaucoup **à ma sœur**.
	PBMT-1	Ma sœur est vraiment absent de New York. ✗
	NMT	Ma sœur est vraiment manquante à New York. ✗
	Google	Ma sœur manque vraiment New York. ✗
S7c	Source	What **he** misses most is **his dog**.
	Ref	Ce qui **lui** manque le plus, c'est **son chien**.
	PBMT-1	Ce qu'il manque le plus, c'est son chien. ✗
	NMT	Ce qu'il manque le plus, c'est son chien. ✗
	Google	Ce qu'il manque le plus, c'est son chien. ✗

Double-object verbs

Are "gift" and "recipient" arguments correctly rendered in French? (English double-object constructions)

S8a	Source	John gave **his wonderful wife** a nice present.
	Ref	John a donné un beau présent **à sa merveilleuse épouse**.
	PBMT-1	John a donné sa merveilleuse femme un beau cadeau. ✗
	NMT	John a donné à sa merveilleuse femme un beau cadeau. ✓
	Google	John a donné à son épouse merveilleuse un présent gentil. ✓
S8b	Source	John told **the kids** a nice story.
	Ref	John a raconté une belle histoire **aux enfants**.
	PBMT-1	John a dit aux enfants une belle histoire. ✓
	NMT	John a dit aux enfants une belle histoire. ✓
	Google	John a raconté aux enfants une belle histoire. ✓
S8c	Source	John sent **his mother** a nice postcard.
	Ref	John a envoyé une belle carte postale **à sa mère**.
	PBMT-1	John a envoyé sa mère une carte postale de nice. ✗
	NMT	John a envoyé sa mère une carte postale de nice. ✗
	Google	John envoya à sa mère une belle carte postale. ✓

Fail to

Is the meaning of "fail to" correctly rendered in the French translation?		
S9a	Source	John **failed to** see the relevance of this point.
	Ref	John **n'a pas** vu la pertinence de ce point.
	PBMT-1	John a omis de voir la pertinence de ce point. ✗
	NMT	John n'a pas vu la pertinence de ce point. ✓
	Google	John a omis de voir la pertinence de ce point. ✗
S9b	Source	He failed to respond.
	Ref	Il **n'a pas répondu**.
	PBMT-1	Il n'a pas réussi à répondre. ✓
	NMT	Il n'a pas répondu. ✓
	Google	Il n'a pas répondu. ✓
S9c	Source	Those who fail to comply with this requirement will be penalized.
	Ref	Ceux qui **ne** se conforment **pas** à cette exigence seront pénalisés.
	PBMT-1	Ceux qui ne se conforment pas à cette obligation seront pénalisés. ✓
	NMT	Ceux qui ne se conforment pas à cette obligation seront pénalisés. ✓
	Google	Ceux qui ne respectent pas cette exigence seront pénalisés. ✓

Manner-of-movement verbs

Is the movement action expressed in the English source correctly rendered in French? (Manner-of-movement verbs with path argument may need to be rephrased in French).		
S10a	Source	John would like to **swim across** the river.
	Ref	John aimerait **traverser** la rivière **à la nage**.
	PBMT-1	John aimerait nager dans la rivière. ✗
	NMT	John aimerait nager à travers la rivière. ✗
	Google	John aimerait nager à travers la rivière. ✗
S10b	Source	They **ran into** the room.
	Ref	Ils **sont entrés** dans la chambre **à la course**.
	PBMT-1	Ils ont couru dans la chambre. ✗
	NMT	Ils ont couru dans la pièce. ✗
	Google	Ils coururent dans la pièce. ✗
S10c	Source	The man **ran out of** the park.
	Ref	L'homme **est sorti du** parc **en courant**.
	PBMT-1	L'homme a manqué du parc. ✗
	NMT	L'homme s'enfuit du parc. ✗
	Google	L'homme sortit du parc. ✗
Hard example featuring spontaneous noun-to-verb derivation ("nonce verb").		
S10d	Source	John **guitared his way** to San Francisco.
	Ref	John **s'est rendu** jusqu'à San Francisco **en jouant de la guitare**.
	PBMT-1	John guitared son chemin à San Francisco. ✗
	NMT	John guitared sa route à San Francisco. ✗
	Google	John a guité son chemin à San Francisco. ✗

Overlapping subcat frames

Is the French verb for "know" correctly chosen? (Choice between "savoir"/"connaître" depends on syntactic nature of its object)

S11a	Source	Paul **knows** that this is a fact.
	Ref	Paul **sait que** c'est un fait.
	PBMT-1	Paul sait que c'est un fait. ✓
	NMT	Paul sait que c'est un fait. ✓
	Google	Paul sait que c'est un fait. ✓
S11b	Source	Paul **knows** this story.
	Ref	Paul **connaît** cette histoire.
	PBMT-1	Paul connaît cette histoire. ✓
	NMT	Paul connaît cette histoire. ✓
	Google	Paul connaît cette histoire. ✓
S11c	Source	Paul **knows** this story is hard to believe.
	Ref	Paul **sait que** cette histoire est difficile à croire.
	PBMT-1	Paul connaît cette histoire est difficile à croire. ✗
	NMT	Paul sait que cette histoire est difficile à croire. ✓
	Google	Paul sait que cette histoire est difficile à croire. ✓
S11d	Source	He **knows** my sister will not take it.
	Ref	Il **sait que** ma soeur ne le prendra pas.
	PBMT-1	Il sait que ma soeur ne prendra pas. ✓
	NMT	Il sait que ma soeur ne le prendra pas. ✓
	Google	Il sait que ma soeur ne le prendra pas. ✓
S11e	Source	My sister **knows** your son is reliable.
	Ref	Ma sœur **sait que** votre fils est fiable.
	PBMT-1	Ma soeur connaît votre fils est fiable. ✗
	NMT	Ma sœur sait que votre fils est fiable. ✓
	Google	Ma sœur sait que votre fils est fiable. ✓

NP to VP

Is the English "NP to VP" complement correctly rendred in the French translation? (Sometimes one needs to translate this structure as a finite clause).

S12a	Source	John believes **Bill to be dishonest**.
	Ref	John croit **que Bill est malhonnête**.
	PBMT-1	John estime que le projet de loi soit malhonnête. ✓
	NMT	John croit que le projet de loi est malhonnête. ✓
	Google	John croit que Bill est malhonnête. ✓
S12b	Source	He liked **his father to tell him stories**.
	Ref	Il aimait **que son père lui raconte des histoires**.
	PBMT-1	Il aimait son père pour lui raconter des histoires. ✗
	NMT	Il aimait son père pour lui raconter des histoires. ✗
	Google	Il aimait son père à lui raconter des histoires. ✗

S12c	Source	She wanted **her mother to let her go**.
	Ref	Elle voulait **que sa mère la laisse partir**.
	PBMT-1	Elle voulait que sa mère de lui laisser aller. ✗
	NMT	Elle voulait que sa mère la laisse faire. ✓
	Google	Elle voulait que sa mère la laisse partir. ✓

Factitives

Is the English verb correctly rendered in the French translation? (Agentive use of some French verbs require embedding under "faire").

S13a	Source	John **cooked** a big chicken.
	Ref	John a **fait cuire** un gros poulet.
	PBMT-1	John cuit un gros poulet. ✗
	NMT	John cuit un gros poulet. ✗
	Google	John a fait cuire un gros poulet. ✓
S13b	Source	John **melted** a lot of ice.
	Ref	John a **fait fondre** beaucoup de glace.
	PBMT-1	John fondu a lot of ice. ✗
	NMT	John a fondu beaucoup de glace. ✗
	Google	John a fondu beaucoup de glace. ✗
S13c	Source	She likes to **grow** flowers.
	Ref	Elle aime **faire pousser** des fleurs.
	PBMT-1	Elle aime à se développer des fleurs. ✗
	NMT	Elle aime à cultiver des fleurs. ✓
	Google	Elle aime faire pousser des fleurs. ✓

Noun Compounds

Is the English nominal compound rendered with the right preposition in the French translation?

S14a	Source	Use the meat knife.
	Ref	Utilisez le couteau **à** viande.
	PBMT-1	Utilisez le couteau de viande. ✗
	NMT	Utilisez le couteau à viande. ✓
	Google	Utilisez le couteau à viande. ✓
S14b	Source	Use the butter knife.
	Ref	Utilisez le couteau **à** beurre.
	PBMT-1	Utilisez le couteau à beurre. ✓
	NMT	Utilisez le couteau au beurre. ✗
	Google	Utilisez le couteau à beurre. ✓
S14c	Source	Use the steak knife.
	Ref	Utilisez le couteau **à** steak.
	PBMT-1	Utilisez le steak couteau. ✗
	NMT	Utilisez le couteau à steak. ✓
	Google	Utilisez le couteau de steak. ✗

S14d	Source	Clean the water filter.
	Ref	Nettoyez le filtre **à** eau.
	PBMT-1	Nettoyez le filtre à eau. ✓
	NMT	Nettoyez le filtre à eau. ✓
	Google	Nettoyez le filtre à eau. ✓
S14e	Source	Clean the juice filter.
	Ref	Nettoyez le filtre **à** jus.
	PBMT-1	Nettoyez le filtre de jus. ✗
	NMT	Nettoyez le filtre de jus. ✗
	Google	Nettoyez le filtre à jus. ✓
S14f	Source	Clean the tea filter.
	Ref	Nettoyez le filtre **à** thé.
	PBMT-1	Nettoyez le filtre à thé. ✓
	NMT	Nettoyez le filtre de thé. ✗
	Google	Nettoyez le filtre à thé. ✓
S14g	Source	Clean the cloth filter.
	Ref	Nettoyez le filtre **en** tissu.
	PBMT-1	Nettoyez le filtre en tissu. ✓
	NMT	Nettoyez le filtre en tissu. ✓
	Google	Nettoyez le filtre en tissu. ✓
S14h	Source	Clean the metal filter.
	Ref	Nettoyez le filtre **en** métal.
	PBMT-1	Nettoyez le filtre en métal. ✓
	NMT	Nettoyez le filtre en métal. ✓
	Google	Nettoyez le filtre métallique. ✓
S14i	Source	Clean the paper filter.
	Ref	Nettoyez le filtre **en** papier.
	PBMT-1	Nettoyez le filtre en papier. ✓
	NMT	Nettoyez le filtre en papier. ✓
	Google	Nettoyez le filtre à papier. ✗

Common idioms

	Is the English idiomatic expression correctly rendered with a suitable French idiomatic expression?	
S15a	Source	Stop **beating around the bush**.
	Ref	Cessez de **tourner autour du pot**.
	PBMT-1	Cesser de battre la campagne. ✗
	NMT	Arrêtez de battre autour de la brousse. ✗
	Google	Arrêter de tourner autour du pot. ✓

S15b	Source	You are **putting the cart before the horse**.
	Ref	Vous **mettez la charrue devant les bœufs**.
	PBMT-1	Vous pouvez mettre la charrue avant les bœufs. ✓
	NMT	Vous mettez la charrue avant le cheval. ✗
	Google	Vous mettez le chariot devant le cheval. ✗
S15c	Source	His comment proved to be **the straw that broke the camel's back**.
	Ref	Son commentaire s'est avéré être **la goutte d'eau qui a fait déborder le vase**.
	PBMT-1	Son commentaire s'est révélé être la goutte d'eau qui fait déborder le vase. ✓
	NMT	Son commentaire s'est avéré être la paille qui a brisé le dos du chameau. ✗
	Google	Son commentaire s'est avéré être la paille qui a cassé le dos du chameau. ✗
S15d	Source	His argument really **hit the nail on the head**.
	Ref	Son argument a vraiment **fait mouche**.
	PBMT-1	Son argument a vraiment mis le doigt dessus. ✓
	NMT	Son argument a vraiment frappé le clou sur la tête. ✗
	Google	Son argument a vraiment frappé le clou sur la tête. ✗
S15e	Source	It's **no use crying over spilt milk**.
	Ref	**Ce qui est fait est fait**.
	PBMT-1	Ce n'est pas de pleurer sur le lait répandu. ✗
	NMT	Il ne sert à rien de pleurer sur le lait haché. ✗
	Google	Ce qui est fait est fait. ✓
S15f	Source	It is **no use crying over spilt milk**.
	Ref	**Ce qui est fait est fait**.
	PBMT-1	Il ne suffit pas de pleurer sur le lait répandu. ✗
	NMT	Il ne sert à rien de pleurer sur le lait écrémé. ✗
	Google	Il est inutile de pleurer sur le lait répandu. ✗

Syntactically flexible idioms

		Is the English idiomatic expression correctly rendered with a suitable French idiomatic expression?
S16a	Source	The cart has been put before the horse.
	Ref	La **charrue a été mise devant les bœufs**.
	PBMT-1	On met la charrue devant le cheval. ✗
	NMT	Le chariot a été mis avant le cheval. ✗
	Google	Le chariot a été mis devant le cheval. ✗
S16b	Source	With this argument, **the nail has been hit on the head**.
	Ref	Avec cet argument, **la cause est entendue**.
	PBMT-1	Avec cette argument, l'ongle a été frappée à la tête. ✗
	NMT	Avec cet argument, l'ongle a été touché à la tête. ✗
	Google	Avec cet argument, le clou a été frappé sur la tête. ✗

Syntactic

Yes-no question syntax

Is the English question correctly rendered as a French question?		
S17a	Source	**Have the kids** ever watched that movie?
	Ref	**Les enfants ont-ils** déjà vu ce film?
	PBMT-1	Les enfants jamais regardé ce film? ✗
	NMT	Les enfants ont-ils déjà regardé ce film? ✓
	Google	Les enfants ont-ils déjà regardé ce film? ✓
S17b	Source	**Hasn't your boss denied you** a promotion?
	Ref	**Votre patron ne vous a-t-il pas refusé** une promotion?
	PBMT-1	N'a pas nié votre patron vous un promotion? ✗
	NMT	Est-ce que votre patron vous a refusé une promotion? ✓
	Google	Votre patron ne vous a-t-il pas refusé une promotion? ✓
S17c	Source	**Shouldn't I attend** this meeting?
	Ref	**Ne devrais-je pas assister** à cette réunion?
	PBMT-1	Ne devrais-je pas assister à cette réunion? ✓
	NMT	Est-ce que je ne devrais pas assister à cette réunion? ✓
	Google	Ne devrais-je pas assister à cette réunion? ✓

Tag questions

Is the English "tag question" element correctly rendered in the translation?		
S18a	Source	Mary looked really happy tonight, **didn't she**?
	Ref	Mary avait l'air vraiment heureuse ce soir, **n'est-ce pas**?
	PBMT-1	Marie a regardé vraiment heureux de ce soir, n'est-ce pas elle? ✗
	NMT	Mary s'est montrée vraiment heureuse ce soir, ne l'a pas fait? ✗
	Google	Mary avait l'air vraiment heureuse ce soir, n'est-ce pas? ✓
S18b	Source	We should not do that again, **should we**?
	Ref	Nous ne devrions pas refaire cela, **n'est-ce pas**?
	PBMT-1	Nous ne devrions pas faire qu'une fois encore, faut-il? ✗
	NMT	Nous ne devrions pas le faire encore, si nous? ✗
	Google	Nous ne devrions pas recommencer, n'est-ce pas? ✓
S18c	Source	She was perfect tonight, **was she not**?
	Ref	Elle était parfaite ce soir, **n'est-ce pas**?
	PBMT-1	Elle était parfait ce soir, elle n'était pas? ✗
	NMT	Elle était parfaite ce soir, n'était-elle pas? ✗
	Google	Elle était parfaite ce soir, n'est-ce pas? ✓

WH-MVT and stranded preps

Is the dangling preposition of the English sentence correctly placed in the French translation?		
S19a	Source	The guy **that** she is going out **with** is handsome.
	Ref	Le type **avec qui** elle sort est beau.
	PBMT-1	Le mec qu'elle va sortir avec est beau. ✗
	NMT	Le mec qu'elle sort avec est beau. ✗
	Google	Le mec avec qui elle sort est beau. ✓

S19b	Source	**Whom** is she going out **with** these days?
	Ref	**Avec qui** sort-elle ces jours-ci?
	PBMT-1	Qu'est-ce qu'elle allait sortir avec ces jours? ✗
	NMT	À qui s'adresse ces jours-ci? ✗
	Google	Avec qui sort-elle de nos jours? ✓
S19c	Source	The girl **that** he has been talking **about** is smart.
	Ref	La fille **dont** il a parlé est brillante.
	PBMT-1	La jeune fille qu'il a parlé est intelligent. ✗
	NMT	La fille qu'il a parlé est intelligente. ✗
	Google	La fille dont il a parlé est intelligente. ✓
S19d	Source	**Who** was he talking **to** when you left?
	Ref	**À qui** parlait-il au moment où tu est parti?
	PBMT-1	Qui est lui parler quand vous avez quitté? ✗
	NMT	Qui a-t-il parlé à quand vous avez quitté? ✗
	Google	Avec qui il parlait quand vous êtes parti? ✓
S19e	Source	The city **that** he is arriving **from** is dangerous.
	Ref	La ville **d'où** il arrive est dangereuse.
	PBMT-1	La ville qu'il est arrivé de est dangereuse. ✗
	NMT	La ville qu'il est en train d'arriver est dangereuse. ✗
	Google	La ville d'où il vient est dangereuse. ✓
S19f	Source	**Where** is he arriving **from**?
	Ref	**D'où** arrive-t-il?
	PBMT-1	Où est-il arrivé? ✗
	NMT	De quoi s'agit-il? ✗
	Google	D'où vient-il? ✓

Adverb-triggered inversion

Is the adverb-triggered subject-verb inversion in the English sentence correctly rendered in the French translation?

S20a	Source	Rarely **did the dog** run.
	Ref	Rarement **le chien courait-il**.
	PBMT-1	Rarement le chien courir. ✗
	NMT	Il est rare que le chien marche. ✗
	Google	Rarement le chien courir. ✗
S20b	Source	Never before **had she been** so unhappy.
	Ref	Jamais encore **n'avait-elle** été aussi malheureuse.
	PBMT-1	Jamais auparavant, si elle avait été si malheureux. ✗
	NMT	Jamais auparavant n'avait été si malheureuse. ✗
	Google	Jamais elle n'avait été aussi malheureuse. ✓

S20c	Source	Nowhere **were the birds** so colorful.
	Ref	Nulle part **les oiseaux n'étaient** si colorés.
	PBMT-1	Nulle part les oiseaux de façon colorée. ✗
	NMT	Les oiseaux ne sont pas si colorés. ✗
	Google	Nulle part les oiseaux étaient si colorés. ✗

Middle voice

Is the generic statement made in the English sentence correctly and naturally rendered in the French translation?

S21a	Source	Soup **is eaten** with a large spoon.
	Ref	La soupe **se mange** avec une grande cuillère
	PBMT-1	La soupe est mangé avec une grande cuillère. ✗
	NMT	La soupe est consommée avec une grosse cuillère. ✗
	Google	La soupe est consommée avec une grande cuillère. ✗
S21b	Source	Masonry **is cut** using a diamond blade.
	Ref	La maçonnerie **se coupe** avec une lame à diamant.
	PBMT-1	La maçonnerie est coupé à l'aide d'une lame de diamant. ✗
	NMT	La maçonnerie est coupée à l'aide d'une lame de diamant. ✗
	Google	La maçonnerie est coupée à l'aide d'une lame de diamant. ✗
S21c	Source	Champagne **is drunk** in a glass called a flûte.
	Ref	Le champagne **se boit** dans un verre appelé flûte.
	PBMT-1	Le champagne est ivre dans un verre appelé une flûte. ✗
	NMT	Le champagne est ivre dans un verre appelé flûte. ✗
	Google	Le Champagne est bu dans un verre appelé flûte. ✗

Fronted "should"

Fronted "should" is interpreted as a conditional subordinator. It is normally translated as "si" with imperfect tense.

S22a	Source	**Should** Paul leave, I would be sad.
	Ref	**Si** Paul **devait** s'en aller, je serais triste.
	PBMT-1	Si le congé de Paul, je serais triste. ✗
	NMT	Si Paul quitte, je serais triste. ✗
	Google	Si Paul s'en allait, je serais triste. ✓
S22b	Source	Should he become president, she would be promoted immediately.
	Ref	**S'**il devait devenir président, elle recevrait immédiatement une promotion.
	PBMT-1	S'il devait devenir président, elle serait encouragée immédiatement. ✓
	NMT	S'il devait devenir président, elle serait immédiatement promue. ✓
	Google	Devrait-il devenir président, elle serait immédiatement promue. ✗
S22c	Source	**Should** he fall, he would get up again immediately.
	Ref	**S'** il venait à tomber, il se relèverait immédiatement.
	PBMT-1	S'il devait tomber, il allait se lever immédiatement de nouveau. ✓
	NMT	S'il tombe, il serait de nouveau immédiatement. ✗
	Google	S'il tombe, il se lèvera immédiatement. ✗

Clitic pronouns

Are the English pronouns correctly rendered in the French translations?		
S23a	Source	She had a lot of money but he did not have **any**.
	Ref	Elle avait beaucoup d'argent mais il n'**en** avait pas.
	PBMT-1	Elle avait beaucoup d'argent mais il n'en avait pas. ✓
	NMT	Elle avait beaucoup d'argent, mais il n'a pas eu d'argent. ✓
	Google	Elle avait beaucoup d'argent mais il n'en avait pas. ✓
S23b	Source	He did not talk **to them** very often.
	Ref	Il ne **leur** parlait pas très souvent.
	PBMT-1	Il n'a pas leur parler très souvent. ✗
	NMT	Il ne leur a pas parlé très souvent. ✓
	Google	Il ne leur parlait pas très souvent. ✓
S23c	Source	The men are watching **each other**.
	Ref	Les hommes **se** surveillent **l'un l'autre**
	PBMT-1	Les hommes se regardent les uns les autres. ✓
	NMT	Les hommes se regardent les uns les autres. ✓
	Google	Les hommes se regardent. ✗
S23d	Source	He gave **it** to the man.
	Ref	Il **le** donna à l'homme.
	PBMT-1	Il a donné à l'homme. ✗
	NMT	Il l'a donné à l'homme. ✓
	Google	Il le donna à l'homme. ✓
S23e	Source	He did not give **it** to **her**.
	Ref	Il ne **le lui** a pas donné.
	PBMT-1	Il ne lui donner. ✗
	NMT	Il ne l'a pas donné à elle. ✗
	Google	Il ne lui a pas donné. ✗

Ordinal placement

Is the relative order of the ordinals and numerals correct in the French tranlation?		
S24a	Source	The **first four** men were exhausted.
	Ref	Les **quatre premiers** hommes étaient tous épuisés.
	PBMT-1	Les quatre premiers hommes étaient épuisés. ✓
	NMT	Les quatre premiers hommes ont été épuisés. ✓
	Google	Les quatre premiers hommes étaient épuisés. ✓
S24b	Source	The **last three** candidates were eliminated.
	Ref	Les **trois derniers** candidats ont été éliminés.
	PBMT-1	Les trois derniers candidats ont été éliminés. ✓
	NMT	Les trois derniers candidats ont été éliminés. ✓
	Google	Les trois derniers candidats ont été éliminés. ✓

S24c	Source	The **other two** guys left without paying.
	Ref	Les **deux autres** types sont partis sans payer.
	PBMT-1	Les deux autres mecs ont laissé sans payer. ✓
	NMT	Les deux autres gars à gauche sans payer. ✓
	Google	Les deux autres gars sont partis sans payer. ✓

Inalienable possession

Is the French translation correct and natural both in: a) its use of a particular determiner on the body part noun; and b) the presence or absence of a reflexive pronoun before the verb?

S25a	Source	He washed **his** hands.
	Ref	Ils **s'**est lavé **les** mains.
	PBMT-1	Ils se lavait les mains. ✓
	NMT	Il a lavé ses mains. ✗
	Google	Ils se lava les mains. ✓
S25b	Source	I brushed **my** teeth.
	Ref	Je **me** suis brossé **les** dents.
	PBMT-1	J'ai brossé mes dents. ✗
	NMT	J'ai brossé mes dents. ✗
	Google	Je me suis brossé les dents. ✓
S25c	Source	You brushed **your** teeth.
	Ref	Tu **t'**es brossé **les** dents
	PBMT-1	Vous avez brossé vos dents. ✗
	NMT	vous avez brossé vos dents. ✗
	Google	Tu as brossé les dents. ✗
S25d	Source	I raised **my** hand.
	Ref	J'ai levé **la** main.
	PBMT-1	J'ai levé la main. ✓
	NMT	J'ai soulevé mamain. ✗
	Google	Je levai la main. ✓
S25e	Source	He turned **his** head.
	Ref	Il a tourné **la** tête.
	PBMT-1	Il a transformé sa tête. ✗
	NMT	Il a tourné sa tête. ✗
	Google	Il tourna la tête. ✓
S25f	Source	He raised his eyes to heaven.
	Ref	Il leva **les** yeux au ciel.
	PBMT-1	Il a évoqué les yeux au ciel. ✓
	NMT	Il a levé les yeux sur le ciel. ✓
	Google	Il leva les yeux au ciel. ✓

Zero REL PRO

		Is the English zero relative pronoun correctly translated as a non-zero one in the French translation?
S26a	Source	The strangers the woman saw were working.
	Ref	Les inconnus **que** la femme vit travaillaient.
	PBMT-1	Les étrangers la femme vit travaillaient. ✗
	NMT	Les inconnus de la femme ont travaillé. ✗
	Google	Les étrangers que la femme vit travaillaient. ✓
S26b	Source	The man your sister hates is evil.
	Ref	L'homme **que** votre sœur déteste est méchant.
	PBMT-1	L'homme ta soeur hait est le mal. ✗
	NMT	L'homme que ta soeur est le mal est le mal. ✓
	Google	L'homme que votre sœur hait est méchant. ✓
S26c	Source	The girl my friend was talking about is gone.
	Ref	La fille **dont** mon ami parlait est partie.
	PBMT-1	La jeune fille mon ami a parlé a disparu. ✗
	NMT	La petite fille de mon ami était révolue. ✗
	Google	La fille dont mon ami parlait est partie. ✓