Construction of a Multilingual Corpus Annotated with Translation Relations

Yuming Zhai
LIMSI-CNRS
Univ. Paris-Sud
Univ. Paris-Saclay, France
zhai@limsi.fr

Aurélien Max
LIMSI-CNRS
Univ. Paris-Sud
Univ. Paris-Saclay, France
amax@limsi.fr

Anne Vilnat
LIMSI-CNRS
Univ. Paris-Sud
Univ. Paris-Saclay, France
anne@limsi.fr

Abstract

Translation relations, which distinguish literal translation from other translation techniques, constitute an important subject of study for human translators (Chuquet and Paillard, 1989). However, automatic processing techniques based on interlingual relations, such as machine translation or paraphrase generation exploiting translational equivalence, have not made use of these relations explicitly until now. In this work, we present a categorization of translation relations and then we annotate a parallel multilingual (English, French, Chinese) corpus of oral presentations, the *TED Talks*, with these relations. Our long-term objective will be to automatically detect these relations in order to integrate them as important characteristics for the search of monolingual segments in relation of equivalence (paraphrases) or of entailment. The annotated corpus resulting from our work will be made available to the community.

1 Introduction

Human translators have studied translation relations since a long time (Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989), which categorize different translation techniques apart from literal translations. But to the best of our knowledge, no automatic processing techniques explicitly implement these interlingual relations.

As an important natural language understanding and generation task, machine translation (MT) has been seriously improved with first phrase-based statistical machine translation (PBMT) then recently with neural machine translation (NMT). MT has also been exploited to generate paraphrases from bilingual parallel corpus, which was originally proposed by Bannard and Callison-Burch (2005). The assumption is that two segments in the same language are potential paraphrases if they share common translations in a foreign language. Currently the largest resource of paraphrases, PPDB (Paraphrase Database) (Ganitkevitch et al., 2013), has been built following this method exploiting translational equivalence. Nonetheless, the work of Pavlick et al. (2015) revealed that there exist other relations than strict equivalence (paraphrase) in PPDB (*i.e. Entailment (in two directions), Exclusion, Other related and Independent*). The existence of these other relations in PPDB reflects the lack of semantic control during the paraphrasing process.

We propose a categorization of translation relations which model human translators' choices, and we annotate a multilingual (English, French, Chinese) parallel corpus of oral presentations, the *TED Talks*[1], with these relations. The annotation is still ongoing and we are developing a classifier based on these annotations to conduct automatic detection. Our further goal is to integrate this information as important characteristics for the search of monolingual segments in relation of equivalence (paraphrases) or of entailment.

After presenting the related work in section 2, we describe our parallel corpus of *TED Talks* (section 3) and the translation relations (section 4). The annotation process and the statistics follow in section 5. Contrastive study between target languages is presented in section 6. Finally, we conclude in section 7.

[1] https://www.ted.com/

Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pages 102–111
Santa Fe, New Mexico, USA, August 20, 2018.

2 Related work

Deng and Xue (2017) have studied the divergences present in English-Chinese machine translation, by using a hierarchical alignment scheme between the parse trees of these two languages. Seven types of divergences have been identified, and some of them cause important difficulties for automatic word alignment. In particular, we can underline lexical differences resulting from non-literal translations, and structural differences between languages (with or without change of types of syntagms). In order to supply a particular dataset of multiword expressions (MWE) for evaluating machine translation, Monti et al. (2015) have annotated specifically these expressions in the English-Italian *TED Talks* corpus, associated with their translation generated by an automatic system. The phenomena discussed in these two studies are included in the translation relations that we present in this article.

PARSEME (PARSing and Multiword Expressions) [2] is a European scientific network built up to elaborate universal terminologies and annotation guidelines for MWEs in 18 languages (Savary et al., 2015). Its main outcome is a multilingual 5-million word annotated corpus, which underlies a shared task on automatic identification of verbal MWEs (Savary et al., 2017). Our trilingual annotated corpus focuses on bilingual relation between translations, and we annotate all words in the corpus, including continuous and discontinuous MWEs.

As a complement of MT evaluation metrics, which reflect imperfectly systems' performance, Isabelle et al. (2017) have introduced a challenge set based on difficult linguistic materials. The authors could hence determine some remaining difficulties for recent NMT systems. The problems include, in particular, incomplete generalizations; translating common and syntactically flexible idioms, or crossing movement verbs *e.g. swim across X* → *traverser X à la nage*. The annotated corpus that we present here could also constitute a challenge set, for the purpose of evaluating MT systems when human translators resort to different translation relations.

3 Corpus

In order to study translation relations for several pairs of languages, we have worked on a multilingual parallel corpus. This corpus is available from the Web inventory *WIT*[3] (Cettolo et al., 2012), which gives access to a collection of transcribed and translated talks, including the corpus of *TED Talks*[3]. This corpus was released for the evaluation campaign IWSLT 2013 and 2014.[4] The source language, *i.e.* the original language in which the speakers expressed themselves, is English. We have calculated the intersection of a parallel corpus with translations in French[5], Chinese, Arabic, Spanish and Russian. The translation of subtitles for *TED Talks* is controlled by volunteers and language coordinators per language[6], which generally ensures good quality translations. The corpus to be annotated contains 2 436 lines of parallel sentences for each pair of languages (the English corpus contains 51 926 tokens). For the moment, we annotate English-French and English-Chinese corpora to validate our hierarchy of translation relations, since these two target languages are very dissimilar in several linguistic aspects: morphology, grammar, expression, etc.

For English and French languages, the corpus has been tokenized by Stanford Tokenizer[7], and lemmatized by TreeTagger (Schmid, 1995) while keeping the tokenization of Stanford Tokenizer. The capital letters at the beginning of each sentence are kept only if these words always appear with capital initials elsewhere in the corpus, otherwise they are lowercased. We have used the tool THULAC (Li and Sun, 2009) for the segmentation of the Chinese corpus, and proceeded to several corrections before annotation. The words are automatically aligned by training FastAlign (Dyer et al., 2013), with its default parameters on each entire parallel corpus (*i.e.* 163 092 lines and 3 303 660 English tokens). We import

[2] `http://www.parseme.eu/`

[3] `https://wit3.fbk.eu/`

[4] We have used training corpus of 2014 (160 656 lines), development corpus (880 lines) and test corpus (1 556 lines) of 2010.

[5] The sentence boundaries have been corrected in French test corpus to calculate the intersection.

[6] `https://www.ted.com/participate/translate/get-started`

[7] `http://nlp.stanford.edu/software/tokenizer.shtml`

these automatic alignments before annotation to accelerate the process, in particular for the words which are literally translated. The annotators should correct these alignments if necessary.

4 Translation relations

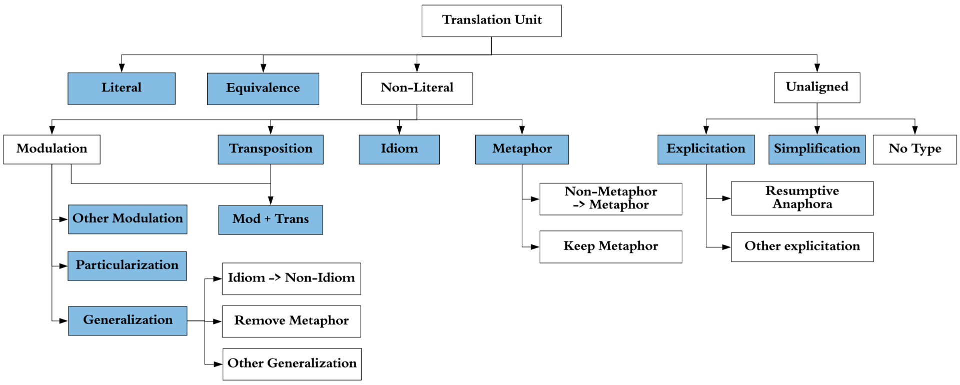


Figure 1: Hierarchy of translation relations.

The first attempt to establish a taxonomy of translation procedures has been carried out by Vinay and Darbelnet (1958). The hierarchy of translation relations that we propose here is based on theories clarified in the work of Chuquet and Paillard (1989), and on phenomena found during our initial study of the corpus (see figure 1).

The colored nodes represent our categories, the other nodes present the hierarchy (*i.e. Non-Literal, Unaligned, Modulation, No Type*) or the phenomena more specific but for which there is no dedicated label (*e.g. Remove Metaphor, Other Generalization*). We present below their definition and typical examples:

1. Literal translation: word-for-word translation (including insertion or deletion of determiners, changes between singular and plural forms), or possible literal translation of some idioms:

   What time is it? → *Quelle heure est-il ?*, *facts are stubborn* → *les faits sont têtus*

2. Equivalence:

   i) non-literal translation of proverbs, idioms, or fixed expressions:

   Birds of a feather flock together. → *Qui se ressemble s'assemble.*

   ii) semantic equivalence in supra-lexical level, translation of terms:

   magic trick → *tour de magie*, *hatpin* → *épingle à chapeau*

3. Modulation: consists in changing the point of view, either to circumvent a translation difficulty or to reveal a way of seeing things, specific to the speakers of the target language. This category could result in semantic shift between source text and target text. Apart from its two sub-types *Particularization* and *Generalization*, all the other phenomena are represented by the sub-type *Other Modulation*:

   this is a completely unsustainable pattern → *il est absolument impossible de continuer sur cette tendance*, *I had an assignment* → *on m'avait confié une mission*

4. Particularization: the translation is more precise or presents a more concrete sense:

   *the director **said*** → *le directeur **déclara***, *language loss* → *l'extinction du langage*

5. Generalization: this category contains three sub-types, but we annotate them with the same label:

   i) the translation is more general or neutral; in other cases, this translation technique makes the sense more accessible in the target language:

look carefully at → regardez, as we ***sit here*** *in ... → alors que nous* ***sommes*** *à ...*

ii) translation of an idiom by a non-fixed expression:

trial and error → procéder par tâtonnements

iii) removal of a metaphorical image:

ancient Tairona civilization which once ***carpeted*** *the Caribbean coastal plain → anciennes civilisations tyranniques qui* ***occupaient*** *jadis la plaine côtière des Caraïbes*

6. Transposition: translating words or expressions by using other grammatical categories than the ones used in the source language, without altering the meaning of the utterance:

   astonishingly inquisitive → dotée d'une curiosité stupéfiante

   patients ***over*** *the age of 40 → les malades* ***ayant dépassé*** *l'âge de 40 ans*

7. Modulation plus Transposition: this category can contain any sub-type of *Modulation* combined with *Transposition*:

   this is a people ***who cognitively do not distinguish*** *→ c'est un peuple* ***dont l'état des connaissances ne permet pas de faire la distinction***

8. Idiom: translate non-fixed expression by an idiom (frequently used when translating English to Chinese):

   at any given moment → à un instant "t"

   died getting old → 行将就木 *"getting closer and closer to the coffin"*

9. Metaphor: this category contains two sub-types reduced to only one label:

   i) keep the same metaphorical image by using a non-literal translation:

   the Sun begins to ***bathe*** *the slopes of the landscape → le soleil qui* ***inonde*** *les flancs de ce paysage*

   ii) introduce metaphorical expression to translate non-metaphor:

   if you ***faint*** *easily → si vous* ***tombez dans les pommes*** *facilement*

10. Unaligned - Explicitation:

    i) resumptive anaphora (Charolles, 2002): add a phrase or sentence summarizing the preceding information (which could be present in previous sentence), to help understanding the present sentence.

    ii) introduce in the target language clarifications that remain implicit in the source language but emerge from the situation; add language-specific function words:

    feel their past in the wind → ressentent leur passé ***souffler*** *dans le vent*

    an entire book → 一 本 完整 的 书 (add the Chinese classifier 本)

11. Unaligned - Simplification: remove deliberately certain content words in translation:

    and you'll ***suddenly*** *discover what it would be like → et vous découvrirez ce que ce serait*

12. Unaligned and no type attributed: function words necessary in one language but not in the other; segments not translated but which don't impact the meaning; segments giving repeated information in context; translated segments which don't correspond to any source segment:

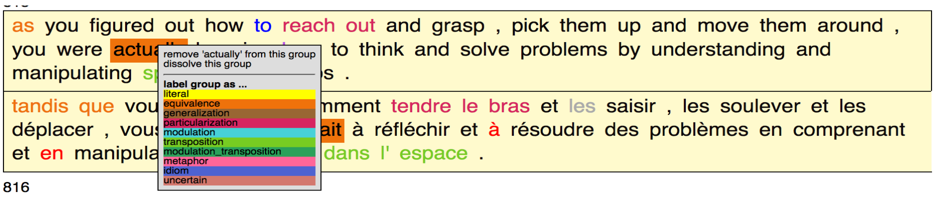
    minus 271 degrees, colder than → moins 271 degrés, ***ce qui est*** *plus froid*

    the last example I have time to → le dernier exemple ***que*** *j'ai le temps de*

5 Annotation

5.1 Annotation tool and configuration

We have used the Web application Yawat[8] (Germann, 2008), which allows us to align words or segments (continuous or discontinuous), and then to assign labels adapted to our task on monolingual or bilingual units (see figure 2).



as you figured out how to reach out and grasp , pick them up and move them around , you were actua … to think and solve problems by understanding and manipulating sp … s .

tandis que vou … mment tendre le bras et les saisir , les soulever et les déplacer , vous … ait à réfléchir et à résoudre des problèmes en comprenant et en manipula … dans l' espace .

816

Figure 2: Annotation interface of Yawat.

Here is a trilingual example from our corpus:

well, we use that great euphemism, "trial and error", which is exposed to be meaningless.

eh bien, nous employons cet euphémisme, procéder par tâtonnements, qui est dénué de sens.

我们 *"we"* 普通人 *"ordinary people"* 会 *"particle for future tense"* 做 *"do"* 各种各样 *"diverse"* 的 *"particle for attribute"* 实验 *"experience"* 不断 *"continuously"* 地 *"particle for adverb"* 犯错误 *"make a mistake"* 结果 *"consequently"* 却 *"however"* 一无所获 *"have gained nothing"*

The segments *well* and *we use that great euphemism* are translated literally in French (with "*great*" omitted), but they are omitted in Chinese. The idiom *trial and error* is translated by a generalization in the two languages. The segment *which is exposed to be* is translated by a generalization in French (*est "is"*) and by a modulation in Chinese (结果 *"consequently"* 却 *"however"*). The adjective *meaningless* is translated by a transposition in French (*dénué de sens "lacking meaning"*) and by an idiom of four characters in Chinese (一无所获 *"have gained nothing"*).

The spelling errors in the original corpus have not been corrected, because on one hand, there are not many of them, on the other hand, generally they don't prevent us from assigning labels. Nonetheless, we have included a category *Uncertain* for pairs for which the annotators don't know which label to assign, or for those which contain obvious translation errors.

Three annotators [9] have participated in the annotation. The training of annotators relies on an annotation guide which defines all categories illustrated by giving characteristic examples. The hierarchy of relations provides a general view of relations between these categories. To better understand the context, the annotators could watch the corresponding video of talks[10] before annotating.

5.2 Control study

We have evaluated in a conventional way the feasibility of our annotation task, by measuring the inter-annotator agreement on a control corpus, which contains 100 pairs of trilingual parallel sentences (3 055 English tokens, 3 238 French tokens and 4 195 Chinese characters). For each pair of languages (EN-FR and EN-ZH), two annotators have independently annotated the corpus.

Since there exist disagreements on the boundary of certain segments, we have calculated the agreement of Cohen's Kappa (Cohen, 1960) only for the segments annotated with exactly the same boundaries by two annotators. The value 0.672 signifies a substantial agreement for the pair EN-FR, while EN-ZH has a lower agreement (0.61, the minimum value to be considered as substantial). The number of tokens

[8] Yet Another Word Alignment Tool, which is available for research under the license GNU Affero General Public License v3.0.

[9] One French annotator and one Chinese annotator for the whole corpus, and another Chinese annotator for only the English-Chinese control corpus.

[10] The corpora to be annotated are transcriptions of *TED Talks* and their translations.

annotated in segments with same boundaries represent 72.60% of English tokens for EN-FR, but only 52.76% for EN-ZH.

We also calculate another inter-annotator agreement in a more flexible way, by including the pairs with different but compatible boundaries (i.e. without overlap of boundaries), and with a common category.[11] In this scenario, while the coverage of English tokens increase to 85.56% for EN-FR and 74.10% for EN-ZH, the Kappa decreases to 0.617 for EN-FR and to 0.60 (moderate) for EN-ZH. The remaining tokens belong to segments with incompatible boundaries (see table 1 and table 2).

	κ	%EN tokens
strict	0.672	72.60%
flexible	0.617	85.56%

Table 1: EN-FR inter-annotator agreement.

	κ	%EN tokens
strict	0.61	52.76%
flexible	0.60	74.10%

Table 2: EN-ZH inter-annotator agreement.

We have compared the annotations of two annotators in a confusion matrix. There exist certain disagreements between *Literal* and these categories: *Equivalence (e.g. in this way → de cette façon), Modulation (e.g. this entire time → tout ce temps), Particularization (e.g. snuff → tabac)* and *Transposition (e.g. their prayers **alone** → **seulement** leurs prières).* However, our categories *Equivalence* and *Literal* are very close, and *Particularization* is a sub-type of *Modulation*, consequently we could consider these confusions as acceptable in a more flexible measure. *Modulation* presents the majority of confusions with *Literal* and *Transposition (e.g. from the forest floor → tombées par terre)*, which indicates that it is necessary to better explain their differences when training annotators. *Mod+Trans* is a combined type for which certain annotators perceive sometimes only one of these two types (*e.g. a great distance → de loin*). *Metaphor* is the origin of several disagreements (*e.g. at the **base** of glaciers → aux **pieds** des glaciers*), which could mainly be explained by the difficulty of the annotation task for a non-native annotator of the target language.

5.3 Annotation scheme with several passes

The calculation of inter-annotator agreement enables certain standard interpretations on the control corpus, and the confusion matrix helps us to identify the difficulties of the task. For the purpose of converging on boundaries of segments and on attributions of labels, we have adopted an annotation scheme with several passes to ensure a better annotation quality. For each subcorpus[12], the first annotator conducts a first annotation of all categories, then the second annotator takes over, which allows him/her to modify the alignments and/or the categories in case of disagreement. Each annotation file is saved at the end of each pass to document the differences in annotation. This alternation can be repeated until the convergence of all annotations. In practice, we limit ourselves to 3 passes, the third one being accomplished by the first annotator. We observe that the number of modifications in the third pass drop gradually with the documents annotated, which reflects a progressive and fast adaptation of annotators to the task. This annotation scheme is cost-intensive, but has been made necessary by the quality intended, as well as by the inherent difficulties of segmentation observed in the control corpus.

5.4 Statistics of annotated subcorpora

Apart from two control corpora, we have annotated six English-French subcorpora and three English-Chinese subcorpora. Their statistics are presented in table 3. Table 4 shows a distribution of number of tokens annotated in different translation relations for English-French corpus.

[11] For example, *I was asked by* and *I was asked by my professor at Harvard* have both been annotated by *Modulation*, and *my professor at Harvard* has been annotated with *Literal* by the first annotator. Here we consider that the two boundaries are compatible, and there is an agreement between the two annotators on one segment (the longest one), because of the common category *Modulation*.

[12] Each subcorpus represents one or several complete speeches of *TED Talks*, in order to ensure better understanding of the context.

	Nb lines	Nb EN Tokens	Nb FR Tokens	Nb ZH Characters
control	100	3 055	3 238	4 195
1	95	1 792	1 774	2 388
2	106	2 282	2 545	3 851
3	101	2 189	2 357	3 380
4	92	1 381	1 489	-
5	133	2 566	2 766	-
6	120	2 691	2 919	-
Total	747	15 956	17 088	13 814

Table 3: Statistics of annotated subcorpora.

	English	French	% EN tokens
Literal	8 701	9 086	67.44%
Equivalence	690	874	5.35%
Modulation	1 671	1 734	12.95%
Transposition	208	297	1.61%
Mod+Trans	250	301	1.94%
Generalization	198	159	1.53%
Particularization	391	560	3.03%
Idiom	4	6	0.03%
Metaphor	16	19	0.12%
Simplification	166	0	1.29%
Explicitation	0	165	0.00%
Uncertain	127	148	0.98%
All types	12 422	13 349	96.29%
No Type	479	501	3.71%
Total nb tokens	12 901	13 850	-

Table 4: Statistics of English-French annotations (number of tokens).

6 Contrast between target languages

We present comparative statistics based on three annotated trilingual subcorpora (see table 5). English and French languages are very similar in grammar and syntax, but Chinese translators should often change word or even phrase order to adapt the translation.

We can see that there are fewer English tokens translated into Chinese using literal translation. For *Equivalence, Modulation* and *Transposition*, the proportions are not so different, but the boundaries of segments could be different. Chinese translations use less complicated *Modulation+Transposition*, but they resort much often to *Generalization* and *Particularization*. In general, *Idiom* and *Metaphor* are under-represented in both target languages, but translating by using a Chinese idiom in four characters is considered as a good practice, which result in a concise language style and translations adapted to Chinese culture. *Simplification* and *Explicitation* show clear differences with French translations. Chinese translations often render only the most important information and leave some other content phrases non-translated. Concerning *Explicitation*, this includes adding necessary Chinese classifiers; inserting content words (e.g. subject noun phrase) to keep the phrase grammatical instead of translating literally word by word and some instances of resumptive anaphora (see section 4). There exist also more instances of *Uncertain*, together with *Simplification*, these phenomena reflect that the quality of Chinese translations are not as good as French translations. Many more English words do not have any type attributed because of the big differences in grammar. At the same time, Chinese translations are more concise than the transcriptions of oral English, which results in the omission of many English transition words.

Here are some examples:

1) EN: *We had to worry about the lawyers and so on.*

FR: *On a dû se préoccuper des avocats et des trucs dans le genre.*

"We had to worry about the lawyers and things like that." (Literal, Modulation+Transposition)

ZH: *要 顾及 到 很多 法律 问题* *"have to worry about many legal issues"* (Generalization)

2) EN: *pressured the people a little bit about it*

FR: *obliger le peuple à en parler* *"force the people to talk about it"* (Modulation)

ZH: *刨根问底* *"inquire into the root of the matter"* (Idiom)

3) EN: *and it doesn't matter how much information we're looking at, how big these collections are or how big the images are.*

FR: *et ce quelle que soit la quantité d'informations que l'on visionne, la taille de ces collections ou la taille des images.*

"regardless of the amount of information viewed, the size of the collections or the size of the images." (Modulation+Transposition)

ZH: 不管 所 见到 的 数据 有 多少 、 图像 集 有 多大 以及 图像 本身 有 多 大 ， *Seadragon* 都 拥有 这样 的 处理 能力 。

"it doesn't matter how much data we're looking at, how big the collections of images are, and how big the images themselves are, Seadragon possesses this processing ability"

(Explicitation for the last phrase: resumptive anaphora which refers to information in the previous sentence)

	English	French	%EN tokens	English	Chinese	%EN tokens
Literal	4 267	4 423	68.13%	3 307	5 311	52.80%
Equivalence	406	514	6.48%	426	629	6.80%
Modulation	589	617	9.40%	615	863	9.82%
Transposition	142	195	2.27%	166	258	2.65%
Mod+Trans	188	225	3.00%	90	134	1.44%
Generalization	90	65	1.44%	200	208	3.19%
Particularization	197	256	3.15%	273	661	4.36%
Idiom	4	6	0.06%	10	21	0.16%
Metaphor	10	15	0.16%	6	10	0.10%
Simplification	92	-	1.47%	314	-	5.01%
Explicitation	-	58	-	-	929	-
Uncertain	79	79	1.26%	157	277	2.51%
All types	6 064	6 453	96.82%	5 564	9 301	88.84%
No type	199	223	3.18%	699	318	11.16%
Total nb tokens	6 263	6 676	-	6 263	9 619	-

Table 5: Contrasts between translations towards different target languages (number of tokens).

7 Conclusion and perspectives

In this work we are interested in translation relations, to our best knowledge, they have never been taken into account in machine translation as well as in paraphrase generation by exploiting translational equivalence. We categorized these relations and annotated them in a multilingual parallel corpus of *TED Talks*. We chose this specific genre (transcribed and translated speech) in order to obtain more diversity than with the technical corpora. The measured inter-annotator agreement is strong for segments with the same boundaries, but we have adopted a more time-consuming annotation process with three passes to ensure better annotation quality.

A parallel task with our corpus annotation is the development of a classifier to automatically detect these relations. Our long-term objective will be to have a better semantic control using this information as important characteristics to exploit translational equivalence, for the search of monolingual segments in relation of equivalence (paraphrases) or of entailment.

We think that there exist many other types of applications for this corpus, for example: evaluate machine translation or automatic word alignment for different translation relations, like the work of Isabelle et al. (2017); understand what are the characteristics of non-literal translations in order to improve machine translation to generate natural and native expressions; provide a concordancer for translation studies to investigate different translation relations. We could also supply pre-processed parallel corpora in other languages (Arabic, Spanish and Russian) for those who are interested to enrich annotation and comparative studies.

Acknowledgements

This work is founded by a scholarship from Paris-Sud University. We thank Yaqiu Liu for her annotation of English-Chinese control corpus and for her valuable insights in corpus analysis. We are grateful to our anonymous reviewers for their thoughtful and constructive comments.

References

Colin J. Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 597–604.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Michel Charolles. 2002. *La référence et les expressions référentielles en français*. Ophrys.

Hélène Chuquet and Michel Paillard. 1989. *Approche linguistique des problèmes de traduction anglais-français*. Ophrys.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Dun Deng and Nianwen Xue. 2017. Translation Divergences in Chinese-English Machine Translation: An Empirical Investigation. *Computational Linguistics*, 43(3):521–565.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648. The Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 758–764.

Ulrich Germann. 2008. Yawat: Yet Another Word Alignment Tool. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Demo Papers*, pages 20–23. The Association for Computer Linguistics.

Pierre Isabelle, Colin Cherry, and George F. Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2476–2486. Association for Computational Linguistics.

Zhongguo Li and Maosong Sun. 2009. Punctuation As Implicit Annotations for Chinese Word Segmentation. *Computational Linguistics*, 35(4):505–512, December.

Johanna Monti, Federico Sangati, and Mihael Arcan. 2015. TED-MWE: a bilingual parallel corpus with MWE annotation. Towards a methodology for annotating MWEs in parallel multilingual corpora. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it*, pages 193–197, Trento.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1512–1522.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland, Nov. Available at `https://hal.archives-ouvertes.fr/hal-01223349/document`.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Proceedings of the 13th Workshop on Multiword Expressions, MWE@EACL 2017, Valencia, Spain, April 4, 2017*, pages 31–47. Association for Computational Linguistics.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais: méthode de traduction*. Bibliothèque de stylistique comparée. Didier.