Construction of a Multilingual Corpus Annotated with Translation Relations

Yuming ZHAI, Aurélien MAX, Anne VILNAT

LIMSI-CNRS, Univ. Paris-Sud, Univ. Paris-Saclay
Orsay, France

20 August 2018

Outline

Outline

Translation relations

EN-FR

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

Translation relations

EN-FR

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

EN-ZH

well, we use that great euphemism, "trial and error"
我们普通人会做各种各样的实验不断地犯错误
(As ordinary people, we would do diverse experiments continuously and commit faults.)

Motivations

Translation relations studied by human translators

Literal translation vs. Other translation techniques (Vinay & Darbelnet, 1958; Chuquet & Paillard, 1989)

Motivations

Translation relations studied by human translators

Literal translation vs. Other translation techniques (Vinay & Darbelnet, 1958; Chuquet & Paillard, 1989)

Lack of explicit exploitation

Machine Translation (SMT, NMT) (Wu *et al.*, 2016, Lapata *et al.*, 2017)
Paraphrasing from bilingual parallel corpora (Bannard & Callison-Burch, 2005)

Motivations

Translation relations studied by human translators

Literal translation vs. Other translation techniques (Vinay & Darbelnet, 1958; Chuquet & Paillard, 1989)

Lack of explicit exploitation

Machine Translation (SMT, NMT) (Wu *et al.*, 2016, Lapata *et al.*, 2017)
Paraphrasing from bilingual parallel corpora (Bannard & Callison-Burch, 2005)

Research goal

- leverage translation relations
- better semantic control in paraphrasing from multilingual parallel corpora

Bilingual pivoting paraphrasing

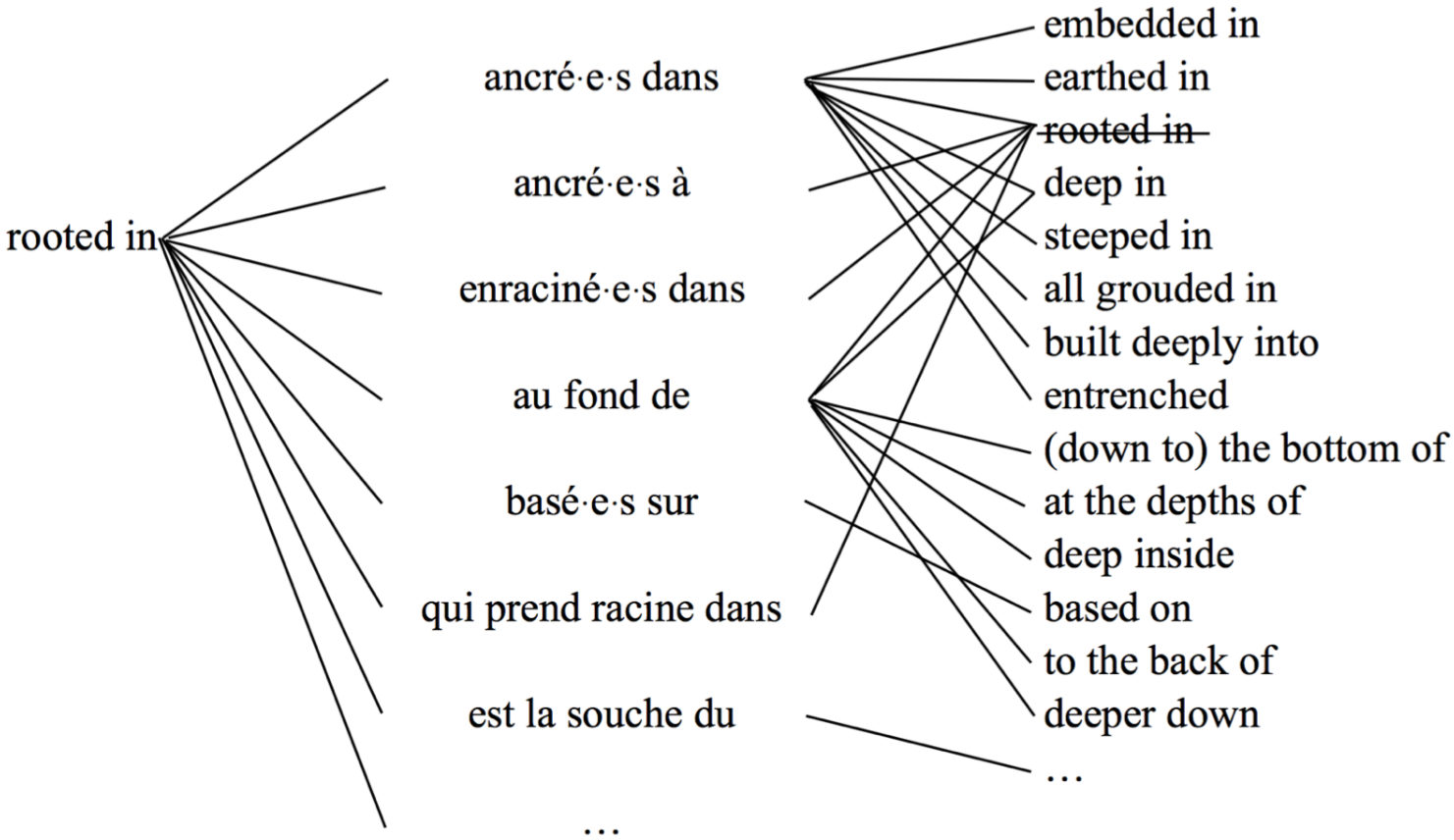


Figure: Obtaining paraphrases for "rooted in" via French pivot translations

Bilingual pivoting paraphrasing

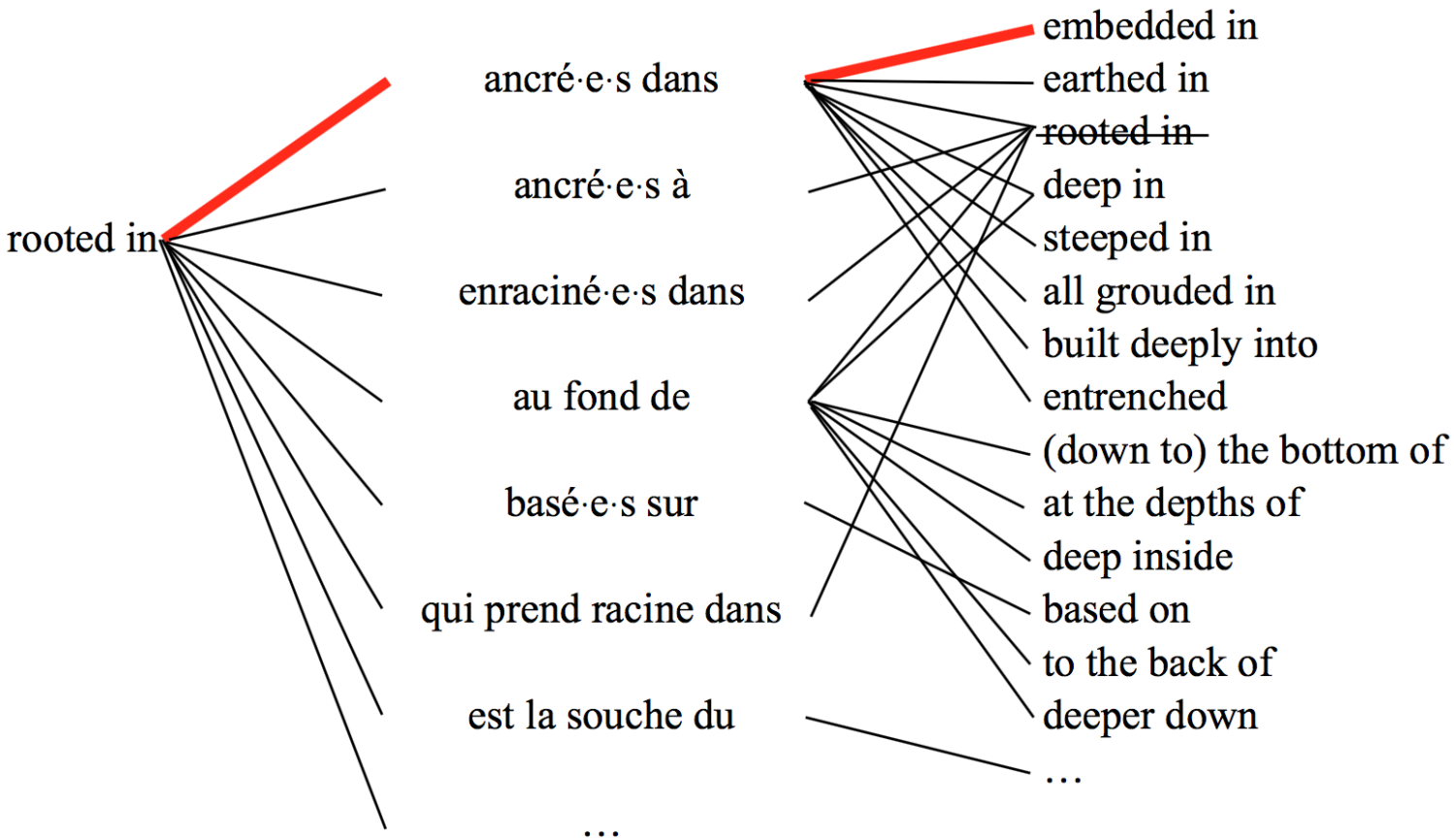


Figure: Obtaining paraphrases for "rooted in" via French pivot translations

Bilingual pivoting paraphrasing

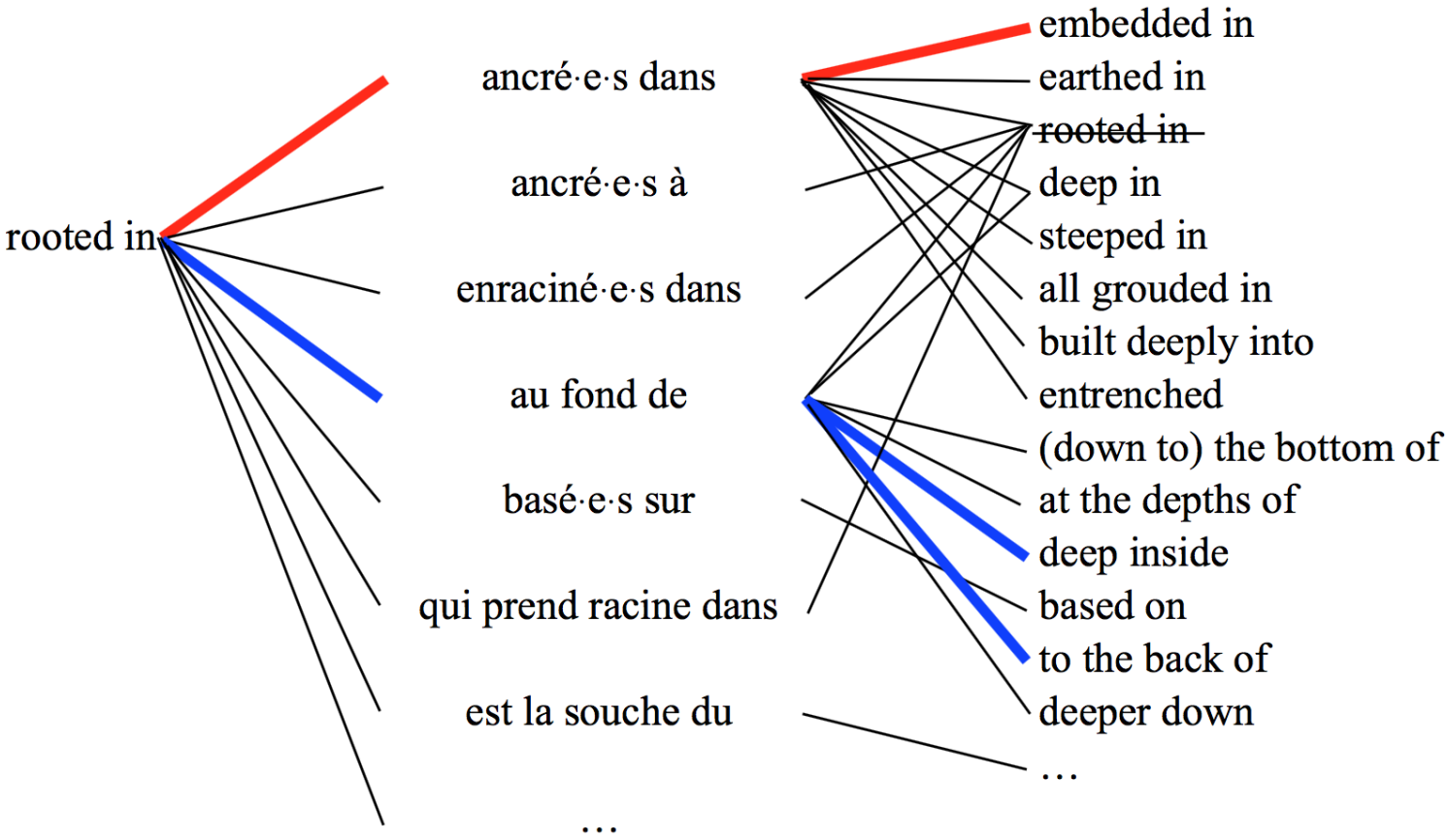


Figure: Obtaining paraphrases for "rooted in" via French pivot translations

Diverse semantic relations in PPDB 2.0

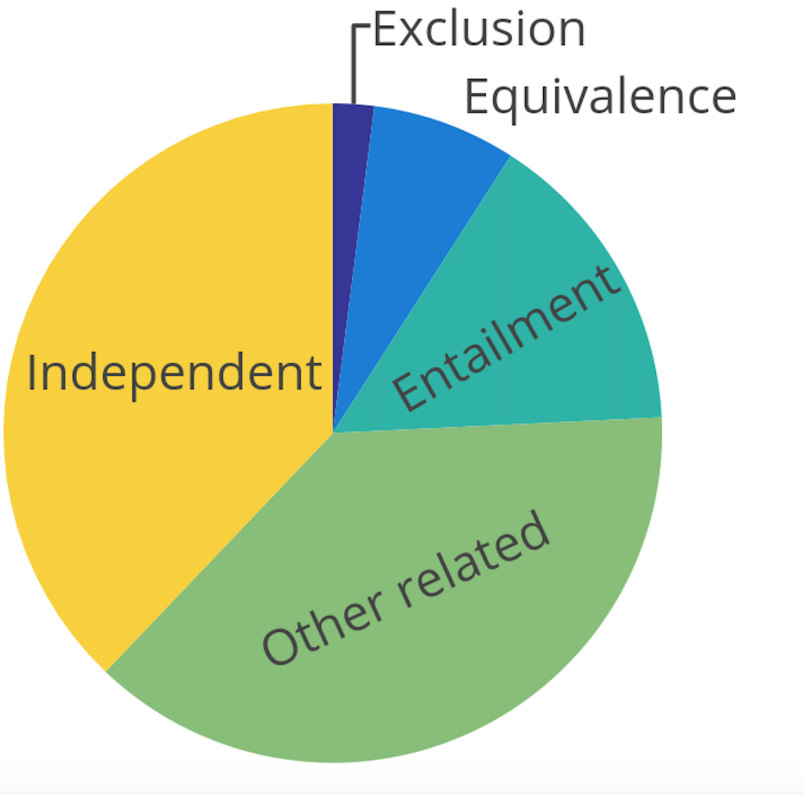


Figure: Estimated semantic relations in PPDB 2.0 XXXL (English)

Paraphrase resource

PPDB (Paraphrase Database)
(Ganitkevitch & Callison-Burch, 2013)

Diverse semantic relations in PPDB 2.0

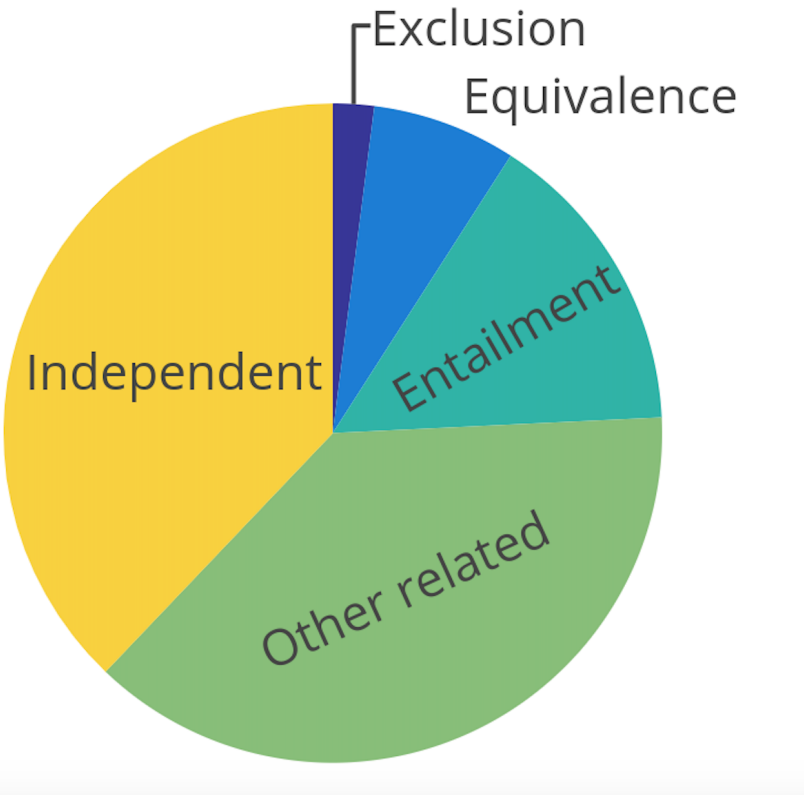


Figure: Estimated semantic relations in PPDB 2.0 XXXL (English)

Paraphrase resource

PPDB (Paraphrase Database)
(Ganitkevitch & Callison-Burch, 2013)

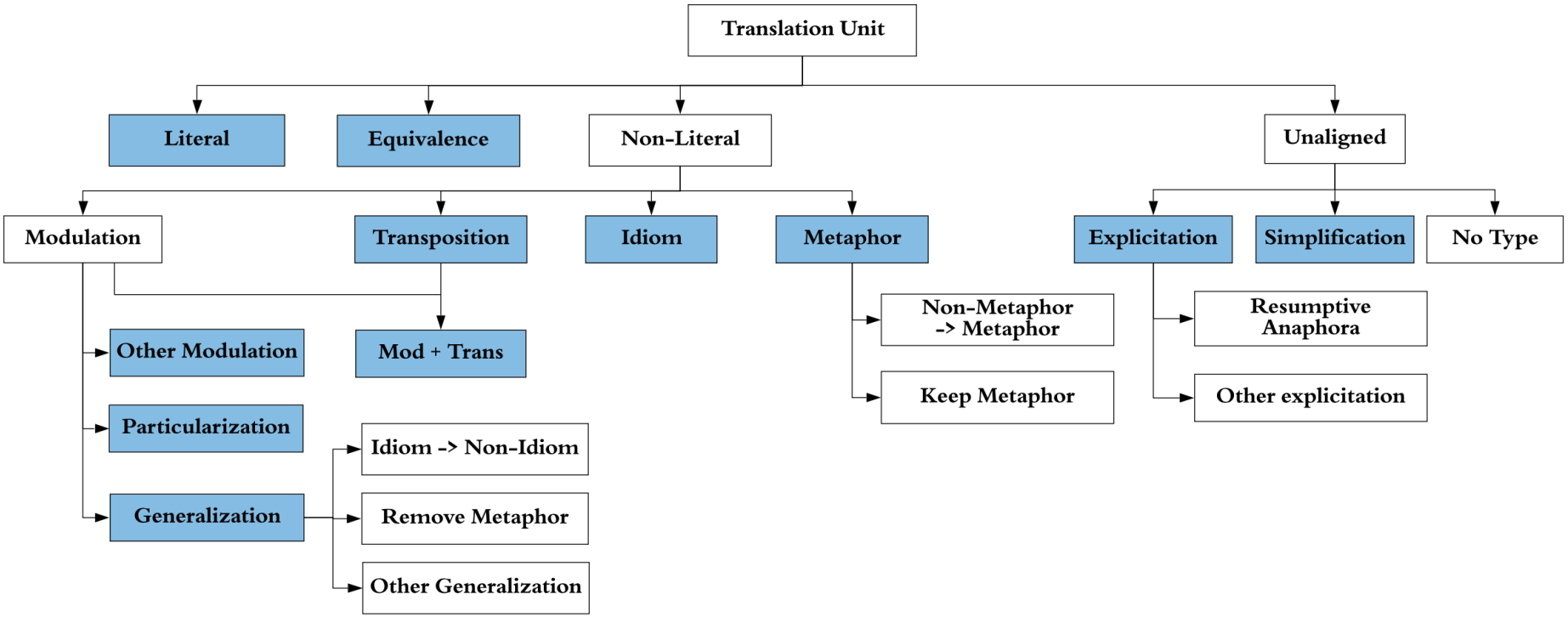
Lack of semantic control

Equivalence: *illegal entry / smuggling*
Exclusion: *close / open*
Other related: *husband / marry to*
Independent: *found / party*
Entailment: *tower / building*
(Pavlick *et al.*, 2015)

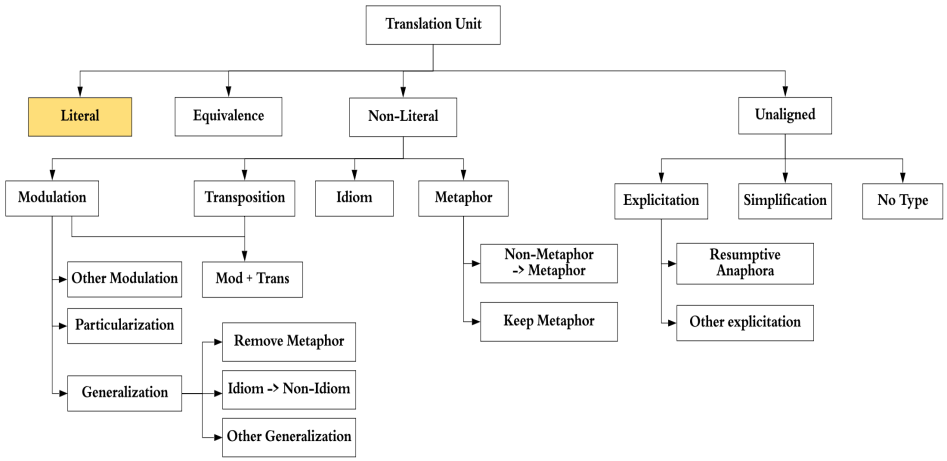
Outline

Hierarchy of translation relations

Our proposition based on theories of translation studies clarified in (Chuquet & Paillard, 1989) and on corpus analysis.



Literal translation

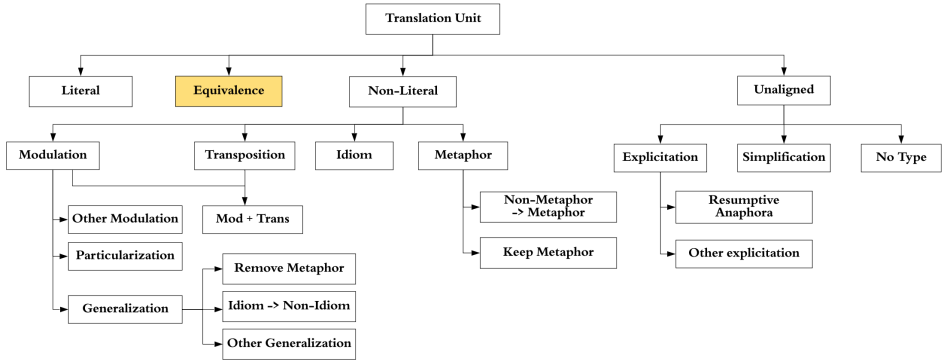


Word-for-word translation (including insertion or deletion of determiners), or possible literal translation of some idioms.

What time is it? $\rightarrow$ *Quelle heure est-il ?*

facts are stubborn $\rightarrow$ *les faits sont têtus*

Equivalence



- Non-literal translation of proverbs, idioms, or fixed expressions.

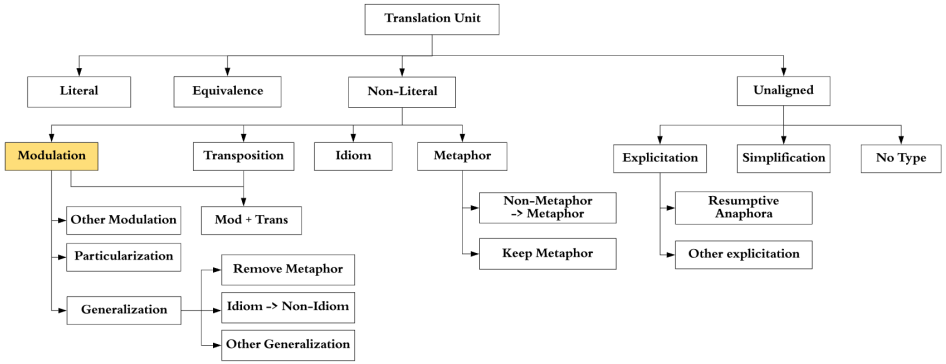
  Birds of a feather flock together. $\rightarrow$ *Qui se ressemble s'assemble.*

- Semantic equivalence in supra-lexical level, translation of terms.

  magic trick $\rightarrow$ *tour de magie*

  hatpin $\rightarrow$ *épingle à chapeau*

Modulation

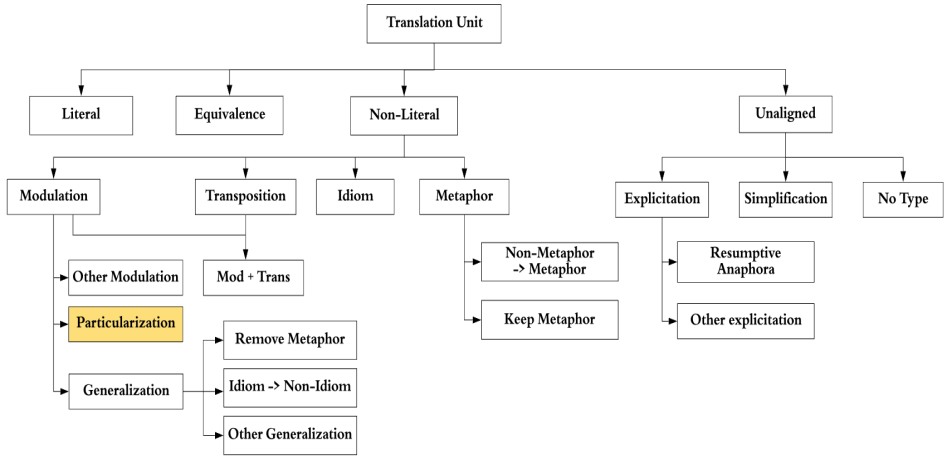


- circumvent translation difficulties
- use natural expression in target language
- could change the point of view
- could result in **semantic shift**
- sub-types: *Particularization*, *Generalization*, *Other Modulation*

we're looking at $\rightarrow$ *on a devant les yeux*

that scar has stayed with him $\rightarrow$ *il a souffert de ce traumatisme*

Particularization

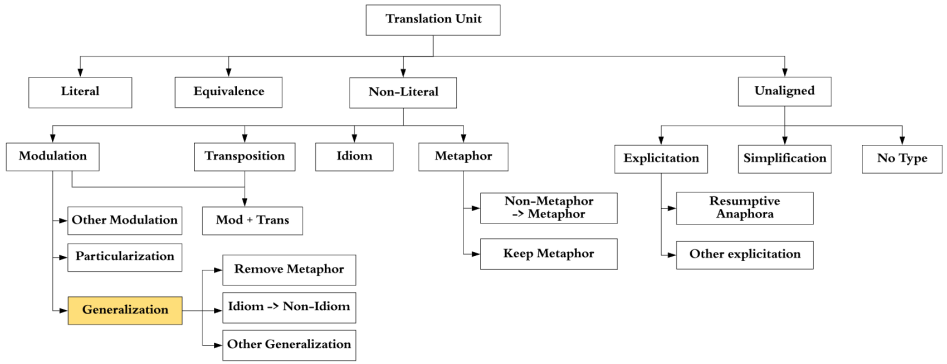


The translation is more precise or presents a more concrete sense.

the director said $\rightarrow$ *le directeur déclara*

language loss $\rightarrow$ *l'extinction du langage*

Generalization



- The translation is more general or neutral.

  get queasy easy $\rightarrow$ *êtes sensibles*

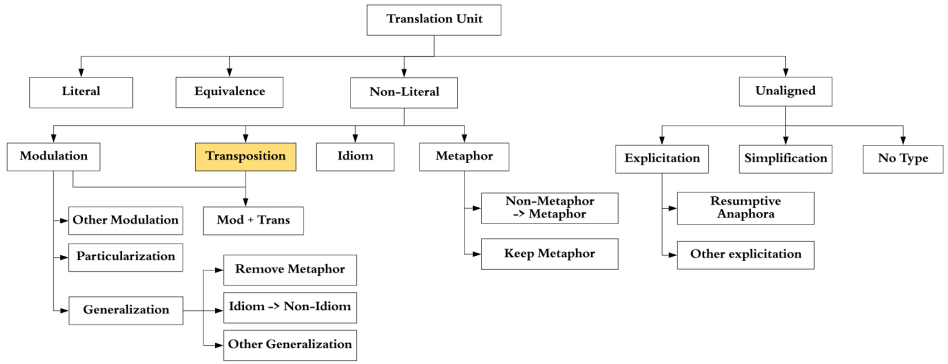
- Translate an idiom by a non-fixed expression.

  trial and error $\rightarrow$ *procéder par tâtonnements*

- Remove a metaphorical image.

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

Transposition

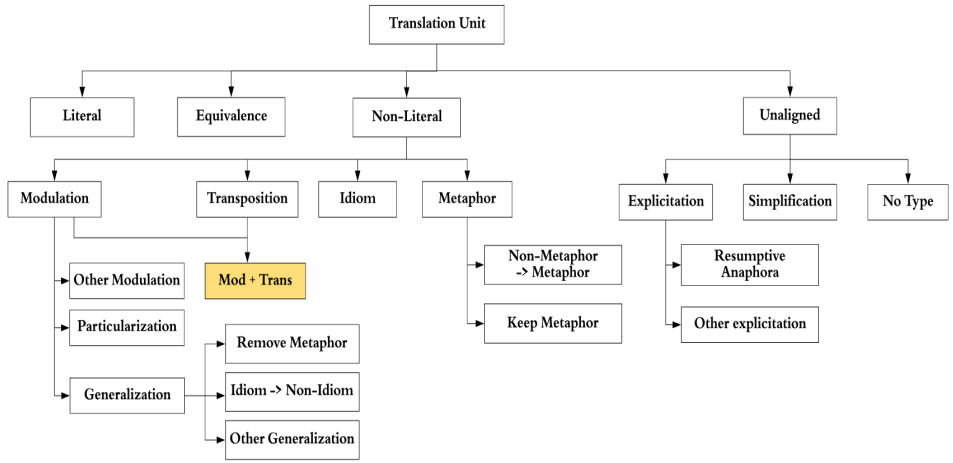


- use **other grammatical categories** than the source segment
- without altering the meaning of the utterance

astonishingly inquisitive → *dotée d'une curiosité stupéfiante*

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

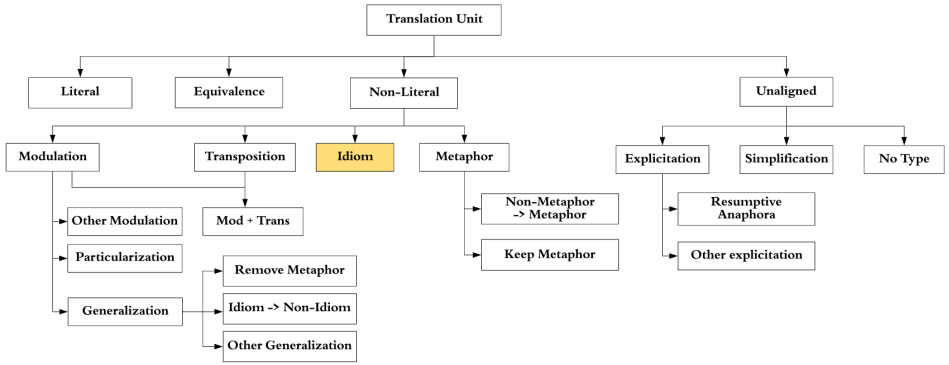
Mod+Trans



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

Idiom

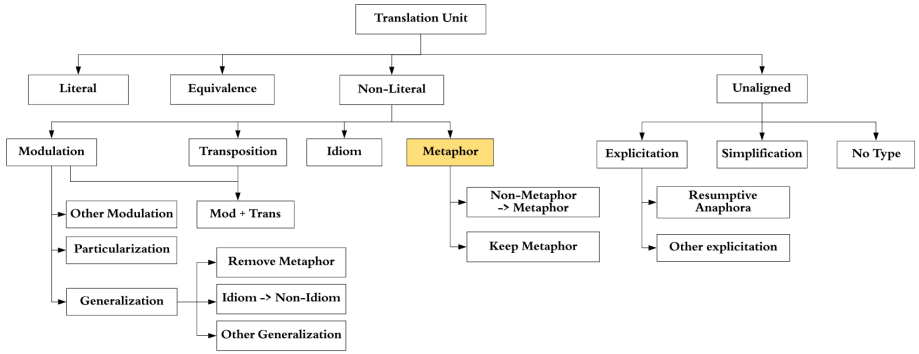


Translate **non-fixed expression** by an idiom (frequently used when translating English to Chinese).

at any given moment → à un instant "t"

died getting old → 行将就木(getting closer and closer to the coffin)

Metaphor



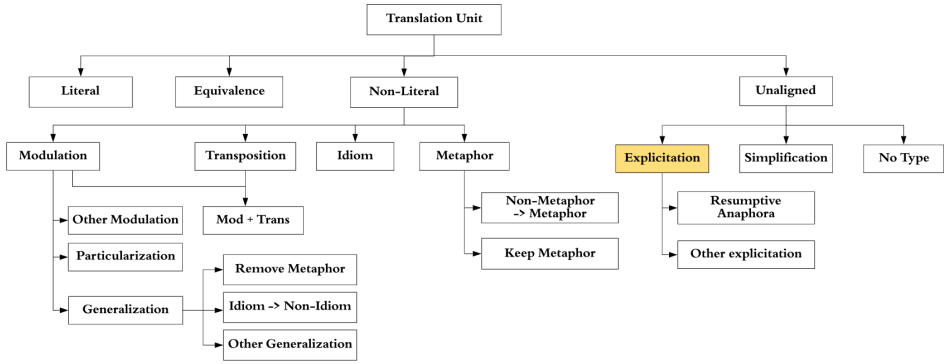
- Keep the same metaphorical image by using a non-literal translation.

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

- Introduce metaphorical expression to translate non-metaphor.

  if you faint easily $\rightarrow$ *si vous tombez dans les pommes facilement*

Unaligned - Explicitation



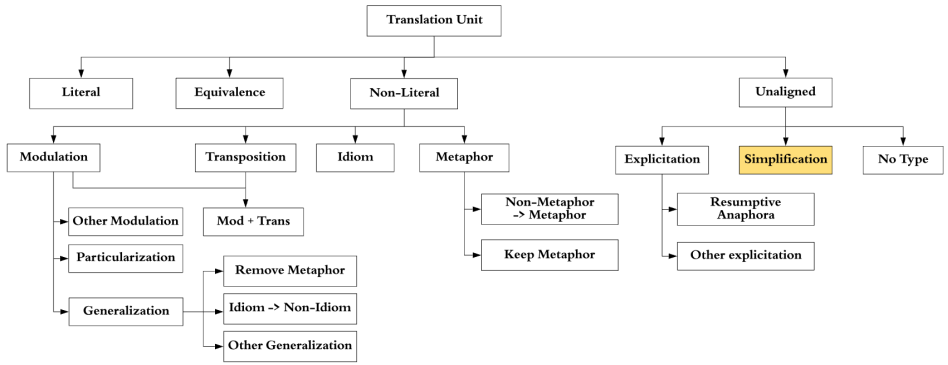
- Resumptive anaphora (Charolles, 2002): add a phrase or sentence summarizing the preceding information.

- Introduce 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)*

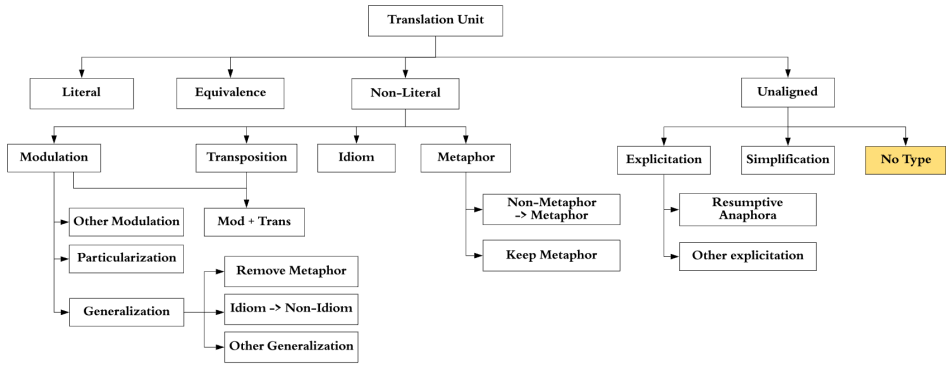
Unaligned - Simplification



Deliberately remove certain content words in translation.

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

Unaligned and no type attributed



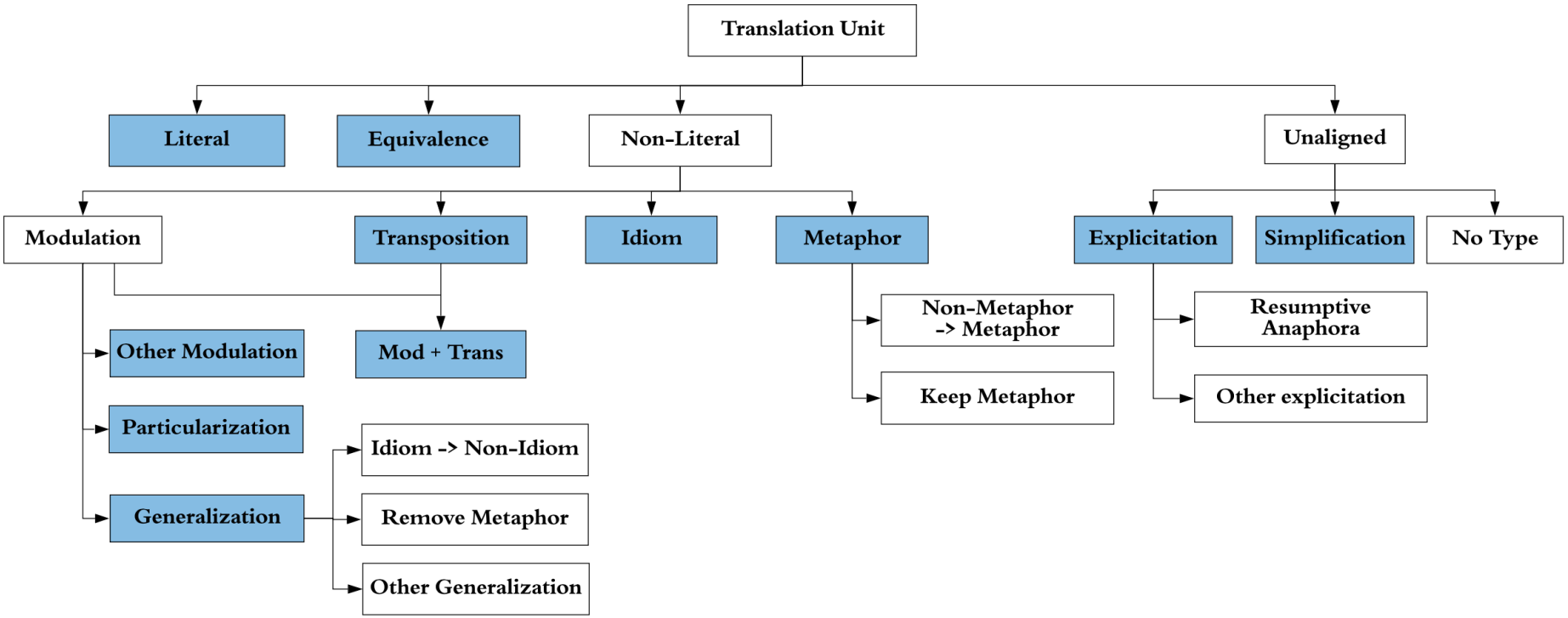
- function words only necessary in one language
- segments not translated but which don't impact the meaning
- segments giving repeated information in context
- translations not corresponding to any source segment

minus 271 degrees, colder than $\rightarrow$ *moins 271 degrés, ce qui est plus froid*

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

Hierarchy of translation relations

Applicable for English-French and English-Chinese pairs.



Outline

Corpus and annotation

- *TED Talks* (Technology, Entertainment, Design and many other topics)
- talks transcribed and translated, provided by the website *WIT*^{3}
- multilingual parallel corpus: source in **English**; translations in **French**, **Chinese**, Arabic, Spanish and Russian
- 2 436 lines of parallel sentences for each pair of languages

Corpus and annotation

- *TED Talks* (Technology, Entertainment, Design and many other topics)
- talks transcribed and translated, provided by the website *WIT*^{3}
- multilingual parallel corpus: source in **English**; translations in **French**, **Chinese**, Arabic, Spanish and Russian
- 2 436 lines of parallel sentences for each pair of languages

Annotation tool: web application *Yawat* (Germann, 2008)

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 .

Control study

- control corpus: 100 lines of parallel sentences
- inter-annotator agreement: Cohen's Kappa (Cohen, 1960)

Control study

- control corpus: 100 lines of parallel sentences
- inter-annotator agreement: Cohen's Kappa (Cohen, 1960)

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

Table: EN-FR inter-annotator agreement

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

Table: EN-ZH inter-annotator agreement

Control study

- control corpus: 100 lines of parallel sentences
- inter-annotator agreement: Cohen's Kappa (Cohen, 1960)

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

Table: EN-FR inter-annotator agreement

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

Table: EN-ZH inter-annotator agreement

- compatible segmentation: (blue: *Modulation*, red: *Literal*)
  Annotator 1: *I was asked by my professor at Harvard*
  Annotator 2: *I was asked by my professor at Harvard*

Control study

- control corpus: 100 lines of parallel sentences
- inter-annotator agreement: Cohen's Kappa (Cohen, 1960)

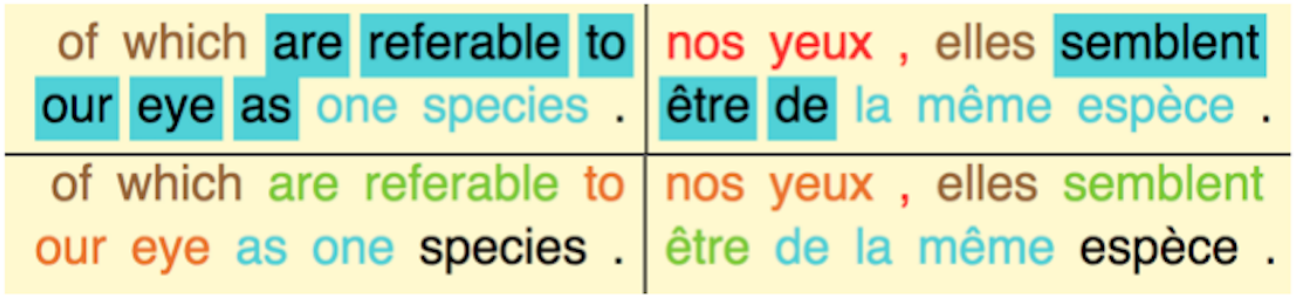
	κ	%EN tokens
strict	0.672	72.60%
flexible	0.617	85.56%

Table: EN-FR inter-annotator agreement

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

Table: EN-ZH inter-annotator agreement

- compatible segmentation: (blue: *Modulation*, red: *Literal*)
  Annotator 1: *I was asked by my professor at Harvard*
  Annotator 2: *I was asked by my professor at Harvard*
- incompatible segmentation: overlap of boundaries



Annotation scheme with three passes

- converge on boundaries of segments and on attributions of type
- more time-consuming but necessary for the targeted quality
- for each subcorpus, limited to 3 passes in practice:
  annotator 1 $\rightarrow$ annotator 2 $\rightarrow$ annotator 1

Annotation scheme with three passes

- converge on boundaries of segments and on attributions of type
- more time-consuming but necessary for the targeted quality
- for each subcorpus, limited to 3 passes in practice:
  annotator 1 $\rightarrow$ annotator 2 $\rightarrow$ annotator 1

Annotation guide

- typical examples
- hierarchy of translation relations
- control study: annotation confusion matrix
- three passes: differences between passes

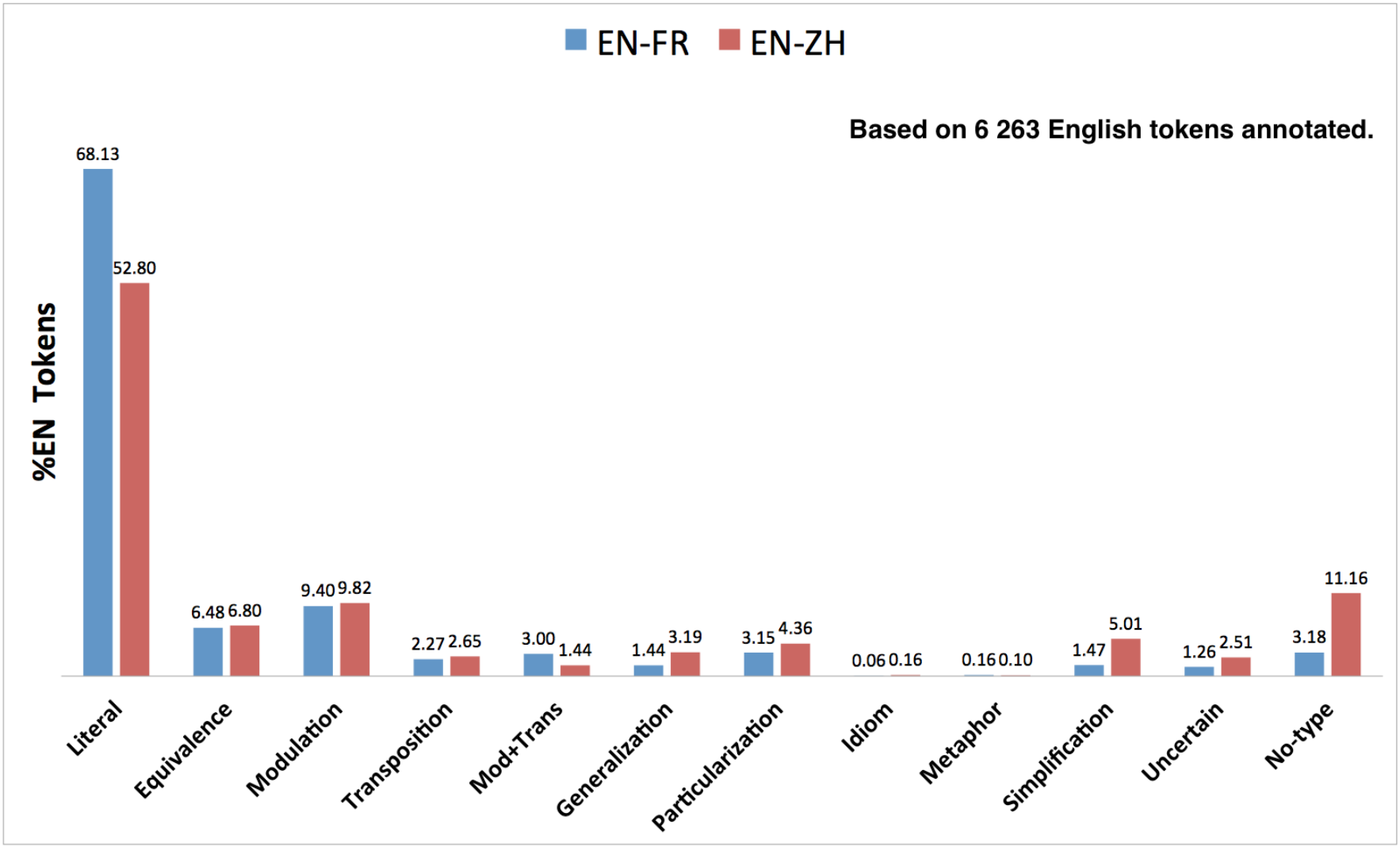
Outline

Statistics of annotated subcorpora

Subcorpus	Nb lines	Nb EN Tokens	Nb FR Tokens	Nb ZH Characters
control corpus	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	120	2 691	2 919	-
5	92	1 381	1 489	-
6	126	3 424	3 690	-
7	133	2 566	2 766	-
8	52	1 597	1 696	-
Total	925	20 977	22 474	13 814

Final goal: annotate 2 436 lines of trilingual parallel sentences.

Contrast between translations towards French and Chinese (number of tokens)



Outline

Conclusion and perspectives

- categorization of translation relations
- annotation of a parallel multilingual corpus of *TED Talks*
- annotation scheme with three passes to guarantee a better annotation quality

Conclusion and perspectives

- categorization of translation relations
- annotation of a parallel multilingual corpus of *TED Talks*
- annotation scheme with three passes to guarantee a better annotation quality

- finer annotation on blocs of type *Modulation*, *Transposition* and *Mod+Trans* to learn patterns
  e.g. they're able to be [moved around] $\rightarrow$ *on peut les [déplacer]*
- automatic detection of translation relations
- integration of these linguistic information to provide a better semantic control during bilingual pivoting paraphrasing

Thanks for your attention, any question?

Mail: zhai@limsi.fr

Examples of annotation confusion

Literal with :
Equivalence (e.g. in this way → de cette façon)
Modulation (e.g. this entire time → tout ce temps)
Particularization (e.g. snuff → tabac)
Transposition (e.g. their prayers alone → seulement leurs prières)
Modulation presents most of the confusions with *Literal* and *Transposition (e.g. from the forest floor → tombées par terre)*

Mod+Trans is a combined type for which certain annotators perceive sometimes only one type (*e.g. a great distance → de loin*)
(e.g. *this is a completely unsustainable pattern → il est absolument impossible de continuer sur cette tendance*)

Generalization: there are few confusions (*e.g. because they're denatured → étant dénaturés*), but it's less the case for *Particularization*.
Metaphor: origin of several disagreements (e.g. *at the base of glaciers → aux pieds des glaciers*)

Statistics of English-French annotations (number of tokens)

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

Contrast between target languages

	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: Contrast between translations towards French and Chinese (number of tokens).

References

- Jean-Paul Vinay and Jean Darbelnet. 1958. Stylistique comparée du français et de l'anglais : méthode de traduction. Bibliothèque de stylistique comparée. Didier.
- Hélène Chuquet and Michel Paillard. 1989. Approche linguistique des problèmes de traduction anglais-français. Ophrys.
- Yonghui Wu *et al.* 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. CoRR abs/1609.08144.
- Jonathan Mallinson, Rico Sennrich and Mirella Lapata. 2017. Paraphrasing Revisited with Neural Machine Translation. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.
- Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In ACL 2005, 43rd Annual Meeting of 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.
- Michel Charolles. 2002. La référence et les expressions référentielles en français. Ophrys.

References

- Ellie Pavlick *et al.* 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.
- Ulrich Germann. 2008. Yawat: Yet Another Word Alignment Tool. In ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics.