Implicitation of Discourse Connectives in (Machine) Translation

Thomas Meyer
Idiap Research Institute and EPFL
Martigny and Lausanne, Switzerland
thomas.meyer@idiap.ch

Bonnie Webber
University of Edinburgh
Edinburgh, UK
bonnie@inf.ed.ac.uk

Abstract

Explicit discourse connectives in a source language text are not always translated to comparable words or phrases in the target language. The paper provides a corpus analysis and a method for semi-automatic detection of such cases. Results show that discourse connectives are not translated into comparable forms (or even any form at all), in up to 18% of human reference translations from English to French or German. In machine translation, this happens much less frequently (up to 8% only). Work in progress aims to capture this natural implicitation of discourse connectives in current statistical machine translation models.

1 Introduction

Discourse connectives (DCs), a class of frequent cohesive markers, such as although, however, for example, in addition, since, while, yet, etc., are especially prone to ‘translationese’, i.e. the use of constructions in the target language (TL) that differ in frequency or position from how they would be found in texts born in the language. That is, ‘translationese’ makes DCs prone to being translated in ways that can differ markedly from their use in the source language. (Blum-Kulka, 1986; Cartoni et al., 2011; Ilisei et al., 2010; Halverson, 2004; Hansen-Schirra et al., 2007; Zufferey et al., 2012). For cohesive markers and DCs, Koppel and Ordan (2011) and Cartoni et al. (2011) have shown that they may be more explicit (increased use) or less explicit (decreased use) in translationese. The paper focuses on the latter case, but the same detection method can be applied in reverse, in order to find increased use (explicitation) as well.

In English about 100 types of explicit DCs have been annotated in the Penn Discourse TreeBank, or PDTB (Prasad et al., 2008) (We say more about this in Section 3.1). The actual set of markers or connectives is however rather open-ended (Prasad et al., 2010). DCs signal discourse relations that connect two spans of text and can be ambiguous with respect to the discourse relation they convey. Moreover, the same DC can simultaneously convey more than one discourse relation. For example, while can convey contrast or temporality, or both at the same time. On the other hand, discourse relations can also be conveyed implicitly, without an explicit DC.

Human translators can chose to not translate a SL DC with a TL DC, where the latter would be redundant or where the SL discourse relation would more naturally be conveyed in the TL by other means (cf. Section 2). We will use the term ‘zero-translation’ or ‘implicitation’ for a valid translation that conveys the same sense as a lexically explicit SL connective, but not with the same form. As we will show, current SMT models either learn the explicit lexicalization of a SL connective to a TL connective, or treat the former as a random variation, realizing it or not. Learning other valid ways of conveying the same discourse relation might not only result in more fluent TL text, but also help raise its BLEU score by more closely resembling its more implicit human reference text.

The paper presents work in progress on a corpus study where zero-translations of DCs have been semi-automatically detected in human reference and machine translations from English (EN) to French (FR) and German (DE) (Section 3). Two types of discourse relations that are very frequently omitted in FR and DE translations are studied in detail and we outline features on how these omissions could be modeled into current SMT systems (Section 4).
2 Implicitation of connectives in translation

Figure 1 is an extract from a news article in the newstest2010 data set (see Section 3.2). It contains two EN connectives — as and otherwise — that were annotated in the PDTB\(^1\). Using the set of discourse relations of the PDTB, as can be said to signal the discourse relation CAUSE (subtype Reason), and otherwise the discourse relation ALTERNATIVE. This is discussed further in Section 3.1.

| EN: The man with the striking bald head was still needing a chauffeur, 1. as the town was still unknown to him. 2. Otherwise he could have driven himself — 3. after all, no alcohol was involved and the 55-year-old was not drunk. |
| FR-Ref: L’homme dont le crâne chauve attirait l’attention, se laissa conduire 1. _as_ dans la ville qui lui était encore étrangère. 2. Autrement notre quinquagénaire aurait pu prendre lui-même le volant — 3. _otherwise_ il n’avait pas bu d’alcool et il n’était pas non plus ivre de bonheur. |
| DE-Ref: Der Mann mit der markanten Glatze ließ sich 1. _as_ der ihm noch fremden Stadt chauffieren. 2. Ansonsten hätte er auch selbst fahren können — Alkohol war 3. _otherwise_ nicht im Spiel, und besoffen vor Glück war der 55-jährige genauso wenig. |

Figure 1: Examples of EN source connectives translated as zero or by other means in human reference translations.

The human reference translations do not translate the first connective as explicitly. In FR there is no direct equivalent, and the reason why the man needed a driver is given with a relative clause: ..._dans la ville qui..._ (lit.: in the town that was still foreign to him). In DE _as_ is realized by means of a preposition, _wegen_ (lit.: because of). The second EN connective _otherwise_, maintains its form in translation to the target connective _autrement_ in FR and _ansonsten_ in DE.

On the other hand, baseline SMT systems for EN/FR and EN/DE (Section 3.2) both translated the two connectives _as_ and _otherwise_ explicitly by the usual target connectives, in FR: _comme, sinon_ and in DE _wie, sonst_.

3 Semi-automatic detection of zero-translations

3.1 Method

The semi-automatic method that identifies zero- or non-connective translations in human references and machine translation output is based on a list of 48 EN DCs with a frequency above 20 in the Penn Discourse TreeBank Version 2.0 (Prasad et al., 2008). In order to identify which discourse relations are most frequently translated as zero, we have assigned each of the EN DCs the level-2 discourse relation that it is most frequently associated with in the PDTB corpus. The total list of EN connectives is given in Table 1.

For every source connective, we queried its most frequent target connective translations from the online dictionary Linguee\(^2\) and added them to dictionaries of possible FR and DE equivalents.

With these dictionaries and Giza++ word alignment (Och and Ney, 2003), the SL connectives can be located and the sentences of its translation (reference and/or automatic) can be scanned for an aligned occurrence of the TL dictionary entries. If more than one DC appears in the source sentence and/or a DC is not aligned with a connective or connective-equivalent found in the dictionaries, the word position (word index) of the SL connective is compared to the word indexes of the translation in order to detect whether a TL connective (or connective-equivalent from the dictionaries) appears in a 5-word window to its left and right.\(^3\). This also helps filtering out cases of non-connective uses of e.g. _separately or once_ as adverbs. Finally, if no aligned entry is present and the alignment information remains empty, the method counts a zero-translation and collects statistics on these occurrences.

After a first run where we only allowed for actual connectives as translation dictionary entries, we manually looked through 400 cases for each, FR and DE reference translations, that were output

---

\(^1\)The excerpt contains a third possible connective _after all_ that was not annotated in the PDTB, and our data as a whole contains other possible connectives not yet annotated there, including _given that_ and _at the same time_. We did not analyse such possible connectives in the work described here.

\(^2\)http://www.linguee.com

\(^3\)The method extends on the ACT metric (Hajlaoui and Popescu-Belis, 2013) that measures MT quality in terms of connectives in order to detect more types of DCs and their equivalents.
Figure 2: Percentage of zero-translations in newstest2010+2012 for EN/FR per discourse relation and translation type: human reference (Ref) or MT output (MT).

as zero-translations (in the newstest2012 data, see Section 3.2). We found up to 100 additional cases that actually were not implications, but conveyed the SL connective’s meaning by means of a paraphrase, e.g. EN: if – FR: dans le cas où (lit.: in case where) – DE: im Falle von (lit.: in case of). For example, the EN connective otherwise ended up with the dictionary entries in Figure 3.

Figure 3: Dictionary entries of FR and DE connectives and equivalents for the EN connective otherwise.

Table 2: Total counts (TC) of English discourse connectives (2,906 tokens) from the newstest2010+2012 corpora, whose majority sense conveys one of the 10 PDTB level-2 discourse relations (Rel.) listed here.

To produce machine translations of the same data sets we built EN/FR and EN/DE baseline phrase-based SMT systems, by using the Moses decoder (Koehn et al., 2007), with the Europarl corpus v7 (Koehn, 2005) as training and newstest2011 as tuning data. The 3-gram language model was built with IRSTLM (Federico et al., 2008) over Europarl and the rest of WMT’s news data for FR and DE.

3.3 Results

In order to group the individual counts of zero-translations per DC according to the discourse relation they signal, we calculated the relative frequency of zero-translations per relation as percentages, see Figures 2 for EN/FR, and 4 for EN/DE.
The total percentage of zero-translations in the references and the baseline MT output is given in Table 3.

A first observation is that an MT system seems to produce zero-translations for DCs significantly less often than human translators do. Human FR translations seem to have a higher tendency toward omitting connectives than the ones in DE. Figures 2 and 4 also show that the discourse relations that are most often rendered as zero are dependent on the TL. In the FR reference translations, SYNHRONY, ALTERNATIVE and CONCESSION account for most implicitations, while in the DE reference translations, CONDITION, ALTERNATIVE and CONCESSION are most often left implicit.

The results are to some extent counterintuitive as one would expect that semantically dense discourse relations like CONCESSION would need to be explicit in translation in order to convey the same meaning. Section 4 presents some non-connective means available in the two TLs, by which the discourse relations are still established.

We furthermore looked at the largest implicitation differences per discourse relation in the human reference translations and the MT output. For EN/FR for example, 13.8% of all CONDITION relations are implicitated in the references, by making use of paraphrases such as dans le moment où (lit.: in the moment where) or dans votre cas (lit.: in your case) in place of the EN connective if. The MT system translates if in 99.4% of all cases to the explicit FR connective si. Similarly, for INSTANTIATION relations and the EN connective for instance in the references, the translators made constrained use of verbal paraphrases such as on y trouve (lit.: among which we find). MT on the other hand outputs the explicit FR connective par exemple in all cases of for instance.

For EN/DE, there is the extreme case, where ALTERNATIVE relations are, in human reference translations, quite often implicitated (in 23.3% of all cases), whereas the MT system translates all the instances explicitly to DE connectives: wenn (unless), sonst (otherwise) and statt, stattdessen, statt (instead). The translators however make use of constructions with a sentence-initial verb in conditional mood (cf. Section 4.2) for otherwise and unless, but not for instead, which is, as with MT, always explicitly translated by humans, most often to the DE connective statt. The very opposite takes place for the RESTATEMENT relation.
and the EN connective *in fact*. Here, MT leaves implicit just as many instances as human translators do, i.e. 14.3% of all cases. Translators use paraphrases such as *in Wahrheit* (lit.: in truth) or *übrigens* (lit.: by the way), while the translation model tends to use *im Gegenteil* (lit.: opposite), which is not a literal translation of *in fact* (usually *in der Tat* or *tatsächlich* in DE), but reflects the contrastive function this marker frequently had in the Europarl training data of the baseline MT system.

4 Case studies

4.1 Temporal connectives from EN to FR

The most frequent implicitated discourse relation for EN/FR translation is SYNCHRONY, i.e. connectives conveying that their arguments describe events that take place at the same time. However, since the situations in which SYNCHRONY relations are implicitated are similar to those in which CONTRAST relations are implicitated, we discuss the two together.

We exemplify here cases where EN DCs that signal SYNCHRONY and/or CONTRAST are translated to FR with a ‘en/Preposition + Verb in Gerund’ construction without a TL connective. The EN source instances giving rise to such implicitations in FR are usually of the form ‘DC + Verb in Present Continuous’ or ‘DC + Verb in Simple Past’, see sentences 1 and 2 in Figure 5.

Out of 13 cases of implicitations for *while* in the data, 8 (61.5%) have been translated to the mentioned construction in FR, as illustrated in the first example in Figure 5, with a reference and machine translation from newstest2010. The DC *while* here ambiguously signals SYNCHRONY and/or CONTRAST, but there is a second temporal marker (*at the same time*, a connective-equivalent not yet considered in this paper or in the PDTB), that disambiguates *while* to its CONTRAST sense only or to the composite sense SYNCHRONY/CONTRAST. The latter is conveyed in FR by *en méprisant*, with CONTRAST being reinforced by *tout* (lit.: all).

In Example 2, from newstest2012, the sentence-initial connective *when*, again signaling SYNCHRONY, is translated to the very same construction of ‘en/Preposition + Verb in Gerund’ in the FR reference.

In the baseline MT output for Example 1, neither of the two EN DCs is deleted, *while* is literally translated to *alors que* and *at the same time* to *dans le même temps*. While the MT output is not totally wrong, it sounds disfluent, as *dans le même temps* after *alors que* is neither necessary nor appropriate.

In the baseline MT output for Example 2, the direct lexical equivalent for *when – lorsque* is generated, which is correct, although the translation has other mistakes such as the wrong verb *semblait* and the untranslated *weather-beaten*.

To model such cases for SMT one could use POS tags to detect the ‘DC + Present Continuous/Simple Past’ in EN and apply a rule to translate it to ‘Preposition + Gerund’ in FR. Furthermore, when two DCs follow each other in EN, and both can signal the same discourse relations, a word-deletion feature (as it is available in the Moses decoder via sparse features), could be used to trigger the deletion of one of the EN connectives, so that only one is translated to the TL. We
will examine in future work whether there are systematic patterns in the translation of such ‘double’ connectives in SL and TL. Another possibility would be to treat cases like while at the same time as a multi-word phrase that is then translated to the corresponding prepositional construction in FR.

### 4.2 Conditional connectives from EN to DE

Out of the 41 cases involving a condition relation (10.5% of all DE implicitations), 40 or 97.6% were due to the EN connective if not being translated to its DE equivalents wenn, falls, ob. Instead, in 21 cases (52.5%), the human reference translations made use of a verbal construction which obviates the need for a connective in DE when the verb in the if-clause is moved to sentence-initial position and its mood is made conditional, as in Figure 6, a reference translation from newstest2012, with the DE verb wäre (lit.: were) (VMFIN=modal finite verb, Konj=conditional). This construction is also available in EN (Were you here, I would...), but seems to be much more formal and less frequent than in DE where it is ordinarily used across registers. In the baseline MT output for this sentence, if was translated explicitly to the DE connective wenn, which is in principle correct, but the syntax of the translation is wrong, mainly due to the position of the verb tun, which should be at the end of the sentence.

The remaining 19 cases of EN if were either translated to DE prepositions (e.g. bei, wo, lit.: at, where) or the condition relation is not expressed at all and verbs in indicative mood make the use of a conditional DE connective superfluous.

Of the 21 tokens of if whose reference translations used a verbal construction in DE, 14 (66.7%) were tokens of if whose argument clause explicitly referred to the preceding context – e.g., if they were, if so, if this is true etc. These occurrences could therefore be identified in EN and could be modeled for SMT as re-ordering rules on the verbal phrase in the DE syntax tree after constituent parsing in syntax-based translation models.

### 5 Conclusion

This study showed that human translators do not translate explicit EN discourse connectives as FR or DE discourse connectives in up to 18% of all cases. In MT output this happens about 3 times less often. We thus plan to examine how to produce higher-scoring translations without a target language connective but with some other syntactic pattern that conveys the same source language discourse relation. Depending on the features identified, movements of syntactical constituents or re-ordering of POS tags at the phrase and/or sub-tree level will be implemented for hierarchical syntactic or phrase-based SMT models.

### Acknowledgments

We are grateful to the Swiss National Science Foundation (SNSF) for partially funding this work and the research visit to Edinburgh with the COMTIS Sinergia project, n. CRSI22 127510 (see www.idiap.ch/comtis/) and to the three anonymous reviewers for their helpful comments.

### References


Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an Open Source Toolkit for
Handling Large Scale Language Models. In *Proceedings of Interspeech*, Brisbane, Australia.


Table 1: English connectives with a frequency above 20 in the PDTB. Also listed are the level-2 majority relations with the number of tokens out of the total tokens of the connective in the PDTB (counts including the majority relation being part of a composite sense tag). *For some connectives there is no level-2 majority because some instances have only been annotated with level-1 senses. We did not consider the connectives and and or (too many non-connective occurrences for automatic detection).