Traduction (automatique) des connecteurs de discours

Laurence Danlos et Charlotte Roze
ALPAGE, Université Paris Diderot (Paris 7), 175 rue du Chevaleret, F-750013 Paris
Laurence.Danlos@linguist.jussieu.fr et Charlotte.Roze@linguist.jussieu.fr

Résumé. En nous appuyant sur des données fournies par le concordancier bilingue TransSearch qui intègre un alignement statistique au niveau des mots, nous avons effectué une annotation semi-manuelle de la traduction anglaise de deux connecteurs du français. Les résultats de cette annotation montrent que les traductions de ces connecteurs ne correspondent pas aux « transpots » identifiés par TransSearch et encore moins à ce qui est proposé dans les dictionnaires bilingues.

Abstract. On the basis of data provided by the bilingual concordancer TransSearch which propose a statistical word alignment, we made a semi-manual annotation of the English translation of two French connectives. The results of this annotation show that the translations of these connectives do not correspond to the "transpots" identified by TransSearch and even less to the translations proposed in bilingual dictionaries.

Mots-clés : Traduction (automatique), TransSearch, Discours.

Keywords: (Machine) Translation, TransSearch, Discourse.

1 Introduction

Les connecteurs de discours n'appartiennent pas à une catégorie morpho-syntaxique unique : ce sont principalement des conjonctions de coordination ou de subordination (*mais, parce que*) et des adverbiaux (*ainsi, après tout*). Ils se définissent fonctionnellement comme des prédicats dont les arguments sont des « segments de discours » (pour simplifier des phrases) et qui indiquent au niveau sémantico-discursif la « relation de discours » connectant ces phrases. Ils facilitent la compréhension d'un texte par rapport aux cas où deux phrases sont simplement juxtaposées sans connecteur explicite pour les relier (ou, ce qui revient au même, avec le connecteur vide, noté ϵ) : avec le connecteur ϵ, le lecteur doit inférer la relation de discours en jeu, alors que ce travail d'inférence n'est pas nécessaire avec un connecteur explicite. Ainsi en (1) traduit de (Wilson & Sperber, 1993), si la première phrase *Pierre n'est pas idiot* est suivie de la phrase (a) avec le connecteur ϵ, on ne sait pas si (a) est relié à la première phrase par la relation de discours $Résultat$ ou par $Évidence$. En revanche, avec la phrase (b) introduite par le connecteur *du coup*, on sait que la relation de discours est $Résultat$ et avec (c) introduit par *après tout*, on sait que c'est $Évidence$.

(1) Pierre n'est pas idiot.
(a) ϵ Il peut trouver son chemin tout seul. [$Résultat$ ou $Évidence$]
(b) *Du coup*, il peut trouver son chemin tout seul. [$Résultat$]
(c) *Après tout*, il peut trouver son chemin tout seul. [$Évidence$]

Les connecteurs de discours forment une classe semi-fermée (semi-ouverte) : ainsi le français compte environ 330 connecteurs selon la base lexicale LexConn (Roze, 2009; Roze *et al.*, 2010), ce qui contraste avec les classes ouvertes (V, N, Adj, Adv) qui comptent des milliers d'éléments chacune et les classes fermées (Prép, Pro, Det, ...) qui comptent moins d'une centaine d'éléments chacune. LexConn [1] est une base lexicale des connecteurs discursifs du français, dans laquelle est renseignée, pour chaque connecteur, sa catégorie morpho-syntaxique et la ou les relation(s) de discours qu'il exprime [2]. Les connecteurs ont été identifiés grâce à des critères syntaxiques, sémantiques et discursifs, appliqués à une liste de conjonctions de subordination fournie par Eric Laporte et une

1. LexConn est librement accessible à `www.linguist.univ-paris-diderot.fr/~croze/D/Lexconn.xml`
2. Un connecteur peut exprimer plusieurs relations de discours. C'est pourquoi les 328 connecteurs donnent lieu à 428 emplois.

liste d'adverbiaux fournis par Benoît Sagot. De même, la base de connecteurs construite pour l'anglais par (Knott, 1996) compte environ 330 connecteurs [3].

Nous nous intéressons à la traduction (automatique) des connecteurs qui pose des problèmes particuliers comme en témoignent de nombreux articles dédiés à la traduction de connecteurs particuliers d'une langue à une autre, par exemple traduction de *or* en russe (Iordanskaja & Mel'čuk, a paraitre) et en espagnol (Rey, 1999), traduction de *de fait* et *en fait* en italien (Rossari, 1993), traduction français ↔ hollandais des connecteurs causaux (Degand & Maat, 2003). Notre étude se situe dans cette optique, mais à la différence des travaux cités, il repose sur des données bilingues français-anglais fournies par l'outil de « transpotting » (alignement sous-phrastique) TransSearch développé par l'équipe RALI de l'Université de Montréal (Huet *et al.*, 2009; Bourdaillet *et al.*, 2010). Cet outil est présenté à la Section 2. Ces données bilingues concernent deux connecteurs du français, *en effet* (adverbial) et *alors que* (conjonction de subordination), sur lesquels nous avons effectué une annotation semi-manuelle en prenant en compte la langue originale, selon une méthodologie expliquée à la Section 2. Cette annotation permet de donner des résultats chiffrés sur la traduction ou la source anglaise de ces connecteurs, Sections 3 et 4. La dernière section est consacrée à l'analyse de ces résultats qui ne correspondent guère à ce qu'indiquent les dictionnaires et qui présentent des scores de transpotting en dessous de ceux annoncés par TransSearch.

2 Données bilingues fournies par TransSearch et méthodologie

TransSearch est un concordancier bilingue qui s'appuie sur un corpus aligné au niveau des phrases (un bitexte), et notamment les textes des débats parlementaires français-anglais discutés au Canada (les Hansards). Lorsque l'utilisateur soumet une requête, par exemple *en effet*, TransSearch ramène les paires d'unités alignées (souvent des phrases) dont la partie française contient la séquence *en effet*. La version récente de TransSearch (avec laquelle nous avons travaillé ici) intègre un alignement statistique au niveau des mots et ramène les paires d'unités alignées en indiquant le « transpot » de la requête, c'est-à-dire pour *en effet* par exemple, la traduction anglaise de ce connecteur si la langue originale est le français ou la séquence de mots pouvant être considérée comme la source de ce connecteur si la langue originale est l'anglais (Huet *et al.*, 2009; Bourdaillet *et al.*, 2010).

Pour mener à bien notre travail sur la traduction des connecteurs, Stéphane Huet nous a fourni, pour chaque connecteur, un fichier XML composé d'une suite d'alignements où chaque alignement indique le transpot du connecteur identifié par TransSearch ainsi que la langue originale [4]. Ci-dessous deux exemples d'alignement. Dans le premier, dont la langue originale est le français, TransSearch a identifié à juste titre que la traduction de *en effet* est *in fact*.

(2) <alignment . . . > <part lang="fr" original="yes">
<s> <match>En effet</match>, je les appelle des « mesures-spectacle » parce qu'on en fait de beaux spectacles. </s> </part>
<part lang="en">
<s> <match pattern="">In fact</match>, I call them "measures for show" because they are used to put on a good show. </s> </part> </alignment>

Dans l'alignement suivant, présenté de façon plus lisible et dont la langue originale est l'anglais, TransSearch propose *talking* (en gras dans l'exemple) comme source de *en effet*, ce qui est erroné. Cette erreur vient de ce que le traducteur a introduit *en effet* dans le texte français sans qu'aucun élément correspondant ne figure dans le texte source, ce que nous notons Ø → *en effet* [5] . Nous verrons que ce cas de figure (et l'inverse soit *en effet* → Ø) s'observe avec une fréquence élevée. Il conduit à des erreurs systématiques de TransSearch qui ne peut pas prendre en compte l'alignement d'un mot avec Ø.

(3) <part lang="en" original="yes">

3. Néanmoins, seuls 100 connecteurs sont pris en compte dans le corpus annoté du Penn Discourse Tree Bank (PDTB Group, 2008).

4. Un alignement comporte aussi le contexte discursif du connecteur, soit les phrases précédant et suivant celle où il apparaît. Ce contexte discursif est nécessaire pour comprendre le texte et donc l'emploi du connecteur. Néanmoins, nous l'omettons dans nos exemples (faute de place) quand il n'est pas pertinent pour la compréhension.

5. La notation Ø est aussi utilisée dans l'article de (Rey, 1999) consacré à la traduction du connecteur *or* en espagnol, voir *Tous les hommes sont mortels. Or Socrate est un homme. Donc Socrate est mortel.* → *Todos los hombres son mortales. Ø Socrates es un hombre. Ø Socrates es mortal.*

[fr] C'est pour cette raison que les jeunes scientifiques canadiens sont de plus en plus nombreux à chercher des emplois verts à l'étranger.
En effet, nos cerveaux verts vont de plus en plus souvent chercher les emplois verts à l'étranger à cause des politiques du gouvernement ; de tels emplois sont créés par milliers aux États-Unis grâce au projet ReEnergize du président Obama.
[en] That is why young Canadian scientists are having increasingly to go to other countries to find jobs, green jobs.
We are **talking** here about green jobs for the green brains that are increasingly having to leave our country because of the government's policies ; green jobs like the thousands being created by Obama through project re-energize.

A partir de ces données, notre méthodologie d'annotation a été la suivante pour un connecteur donné :

1. séparer le fichier XML en deux selon la langue originale,
2. repérer les transpots les plus courants (par exemple, *in fact* et *indeed* pour *en effet*), vérifier que TransSearch les a bien identifiés et si c'est le cas les annoter comme bons transpots,
3. annoter manuellement le reste.

Ainsi à partir de l'alignement (2), l'annotation est celle donnée en (4), dans laquelle le transpot calculé par TransSearch est indiqué en gras et la traduction semi-manuellement identifiée est soulignée. Cette annotation incrémente le nombre de fois où on a *en effet* → *in fact* et le nombre de fois où TransSearch identifie correctement un transpot.

(4) <part lang="fr" original="yes">
[fr] **En effet** , je les appelle des « mesures-spectacle » parce qu'on en fait de beaux spectacles.
[en] **In fact**, I call them "measures for show" because they are used to put on a good show.

L'annotation de l'alignement (3) incrémente le nombre de fois où on a Ø → *en effet* et le nombre de fois où TransSearch identifie un transpot erroné. Enfin, l'annotation en (5) incrémente le nombre de fois où on a *because* → *en effet* et le nombre de fois où TransSearch identifie un transpot erroné.

(5) <part lang="en" original="yes">
[fr] Une fois l'accord signé, nous avons été désavantagés. **En effet**, comme il l'a expliqué, au moment où nous avons signé l'accord, presque tous les marchés avaient déjà été attribués
[en] When the agreement was signed, we ended **up getting the short end of** the stick because, as he explained, by the time we signed the agreement, almost all of the available business was already spoken for.

3 Résultats de l'annotation pour *en effet*

Notre corpus contenait 2157 exemples pour *en effet* avec le français comme langue originale et 1977 exemples avec l'anglais comme langue originale. Les dictionnaires bilingues (Robert & Collins, Reverso, Dictionnaire Google, Wordreference) donnent généralement comme traduction de *en effet* : *indeed* ou *in fact*. Nous avons effectivement trouvé ces deux traductions parmi les trois plus fréquentes, la troisième étant la séquence vide Ø. Plus précisément, les résultats de notre annotation sont détaillés dans la Table 1 pour ces trois traductions (sources) de *en effet* selon la langue originale et sans tenir compte de la langue originale (colonne français + anglais). Cette table indique aussi le pourcentage couvert par ces trois traductions ainsi que le nombre de fois où TransSearch a identifié le bon transpot quel qu'il soit (parmi les trois traductions plus fréquentes ou non).

Langue originale	français	anglais	français + anglais
indeed	713 (33%)	532 (26.9%)	1245 (30%)
Ø	574 (26.6%)	570 (28.8%)	1144 (27%)
in fact	669 (31%)	410 (20.7%)	1079 (26%)
Pourcentage couvert	90%	76%	83%
Transpots corrects	1470 (68%)	1134 (57%)	2604 (62%)

TABLE 1 – Les trois traductions les plus fréquentes de *en effet* et nombre de transpots corrects

En dehors de ces trois traductions les plus fréquentes, les résultats sont les suivants avec le français comme langue originale : *actually* = 41 (1.9%), *the fact is that* = 25 (1.1%), *as a matter of fact* = 18 (0.8%), *yes* = 14 (0.6%), *it is true that* = 9 (0.4%), *basically* = 7 (0.3%) ... Avec l'anglais comme langue originale, les sources de *en effet* sont : *because* = 83 (4.1%), *yes* = 77 (3.8%), *in effect* = 45 (2.2%), *as a matter of fact* = 15 (0.7%), *actually* = 12 (0.6%), *effectively* = 11 (0.5%), *certainly*= 9 (0.4%), *this is because* = 9 (0.4%) ...

4 Résultats de l'annotation pour *alors que*

Soulignons en premier lieu que la séquence *alors que* ne correspond pas toujours à la conjonction de subordination, voir *Luc comprit alors que Marie ne viendrait pas*. TransSearch ne peut identifier de tels cas que nous avons donc dû écarter manuellement. Notre corpus contenait 623 exemples de la conjonction *alors que* avec le français comme langue originale et 807 avec l'anglais comme langue originale.

Les dictionnaires bilingues donnent généralement comme traduction de *alors que* : *while, whereas, when* ou *as*. Nous n'avons trouvé que *when* parmi les quatre traductions les plus fréquentes, les trois autres étant *even though*, Ø et *at a time when* en ne prenant pas en compte la direction de traduction. Si on prend en compte la direction de traduction, les quatre traductions les plus fréquentes de *alors que* avec le français comme langue originale sont *even though*, Ø, *at a time when* et *given that*, *when* n'arrivant qu'en cinquième position. Avec l'anglais comme langue originale, les quatre sources les plus fréquentes de *alors que* sont Ø, *when, at a time when* et *and*, *even though* n'arrivant qu'en sixième position après *when in fact*. Les résultats de notre annotation sont détaillés dans la Table 2 pour ces quatre traductions les plus fréquentes de *alors que*.

Langue originale	français	anglais	français + anglais
even though	302 (48.4%)	26 (3.2%)	328 (22.9%)
Ø	60 (9.6%)	167 (20.6%)	227 (15.8%)
when	27 (4.3%)	115 (14,2%)	142 (9.9%)
at a time when	35 (5.6%)	96 (11.9%)	131 (9.1%)
Pourcentage couvert	68%	54%	58%
Transpots corrects	312 (50%)	82 (10.7%)	394 (27.5%)

TABLE 2 – Les quatre traductions les plus fréquentes de *alors que* et nombre de transpots corrects

En dehors de ces quatre traductions les plus fréquentes, les résultats sont les suivants avec le français comme langue originale, en gras les traductions données dans les dictionnaires : *given that* = 31 (5%), **while = 19** (3.04%), *but* = 18 (2.88%), *compared to* = 11 (1.76%), *despite the fact that* = 8 (1.28%), *and* = 8, *when in fact* = 7 (1.12%), *however* = 7, *yet* = 5 (0.8%), **as = 4** (0.64%) ... **whereas = 2** (0.32%) ... et 6.4% d'hapax.
Avec l'anglais comme langue originale, les sources les plus fréquentes sont : *and* = 29 (3.6%), *when in fact* = 28 (3.5%), *even though* = 26 (3.22%), **while = 24** (2.97%), **as = 19** (2.35%) ... **whereas = 5** (0.62%) ... et 8.4% d'hapax.

5 Conclusion et perspectives

Une conclusion s'impose d'emblée : les traductions (ou sources) de *en effet* et *alors que* observés sur corpus ne correspondent guère à ce qu'indiquent les dictionnaires bilingues ou un système d'alignement au niveau des mots comme TransSearch. Ce système annonce un score d'environ 70% de bons transpots (Bourdaillet *et al.*, 2010) : nous sommes légèrement en dessous pour *en effet* (62%) et carrément en dessous pour *alors que* (27.5%).

Une première raison de cette baisse des scores est la traduction (ou source) Ø qui est généralement ignorée des dictionnaires bilingues ; elle est parfois prise en compte dans les systèmes d'alignement au niveau des mots mais pas dans TransSearch. Il est connu que certains mots (simples ou composés) appartenant à des classes fermées correspondent à des traductions Ø, e.g. certaines prépositions régies par un verbe *Je doute de ta sincerité* ↔ *I doubt Ø your sincerity*, certains déterminants *Les baleines sont des mammifères* ↔ *Ø Whales are Ø mammals*, certains pronoms *Je t'aime* ↔ *Ø Te quiero*. A rebours, les mots (simples ou composés) appartenant à des classes ouvertes

sont la plupart du temps traduits explicitement. Pourquoi et quand observe-t-on un transpot Ø des connecteurs — qui appartiennent à une classe semi-fermée mais qui ne sont pas vides de sens puisqu'ils indiquent explicitement, rappelons-le, la relation sémantico-discursive liant leurs arguments) ? Nous pouvons avancer les hypothèses suivantes, qui sont en rapport l'une avec l'autre.

– la première consiste à dire que certaines langues utilisent plus de connecteurs que d'autres. Il faut en effet rappeler que les connecteurs sont souvent en compétition monolinguement avec le connecteur vide ϵ sans que le sens de l'énoncé soit fondamentalement changé, voir *Luc est tombé parceque/ϵ Marie l'a poussé* ou *Luc entra dans le salon. Puis/ϵ il s'assit sur le sofa*. Le connecteur ϵ demande au lecteur d'inférer la relation de discours, ce qui n'est pas le cas d'un connecteur explicite (qui demande au plus un travail de désambiguation lorsqu'il lexicalise plusieurs relations de discours). On peut avancer l'hypothèse que certaines langues sont plus implicites — demandent plus de travail d'inférence au lecteur — que d'autres, tout du moins en ce qui concerne les connecteurs [6].
– la seconde concerne l'absence de traduction(s) stable(s) d'un connecteur. Cette hypothèse, que nous allons développer ci-dessous pour des traductions non vides, peut entraîner la stratégie suivante pour un traducteur humain : utiliser le connecteur ϵ en laissant au lecteur le soin d'inférer la relation de discours plutôt que d'avancer une traduction erronée.

L'absence de traduction(s) stable(s) d'un connecteur est particulièrement criante (et imprévue) au vu des nos résultats pour *alors que*. Ainsi, avec le français comme langue originale, rappelons que les quatre traductions les plus fréquentes de ce connecteur ne correspondent à aucun élément donné dans les dictionnaires. Ceux-ci mettent en tête *while* et *whereas*, qui ne sont utilisés que dans 3.3% des cas, et ne mentionnent pas *even though* qui est la traduction la plus fréquente (48.4%). De plus, les traductions les plus fréquentes, *even though* et Ø, ne couvrent que 58% des cas, les autres faisant appel à une myriade de traductions s'observant au mieux dans 5.5% des cas (*at a time when*) avec 6.4% d'hapax.

Certes, notre étude est limitée à deux connecteurs pour une seule paire de langues et elle ne repose que sur un corpus d'un genre particulier. Elle demande donc à être prolongée dans plusieurs directions. Néanmoins, nous pensons qu'il y a un problème spécifique de traduction pour bon nombre de connecteurs, ce qui explique la multitude d'études contrastives qui leurs sont dédiées, voir Section 1. Ainsi (Iordanskaja & Mel'čuk, a paraitre) et (Rey, 1999) argumentent qu'il n'y a pas de bonne(s) traduction(s) de *or* respectivement en russe et en espagnol et nous pensons qu'il n'y en a pas non plus en anglais, de même pour *certes, au fur et à mesure que, quitte à, ...* (Cartoni *et al.*, 2011) arrivent au même type de conclusion pour les connecteurs causaux.

Les systèmes de traduction automatique fonctionnent actuellement phrase par phrase. Pour les connecteurs adverbiaux, ils ne peuvent donc pas prendre en compte le contexte discursif de la phrase où apparaît le connecteur, ce qui serait pourtant nécessaire pour identifier son emploi. Il serait toutefois important que les systèmes de traduction automatique développent des stratégies spécifiques pour les connecteurs, qui rappelons-le, jouent un rôle crucial pour la compréhension d'un texte puisqu'ils indiquent comment s'articule sémantiquement et rhétoriquement la succession des phrases d'un texte.

Remerciements

Nous remercions Stéphane Huet, Philippe Langlais et Guy Lapalme du RALI sans lesquels cette étude n'aurait pu avoir lieu, ainsi que Bruno Cartoni, Thomas Meyer et Sandrine Zufferey de l'Université de Genève avec lesquels nous avons eu des discussions fructueuses.

Références

BOURDAILLET J., HUET S., LANGLAIS P. & LAPALME G. (2010). TransSearch : from a bilingual concordancer to a translation finder. *Machine Translation*, **24**(3-4), 241–271.

6. Ainsi, le français est plus implicite que le coréen dans les discours causaux mettant en jeu un verbe causatif (Pak, 1997).

CARTONI B., ZUFFEREY S., MEYER T. & POPESCU-BELIS A. (2011). How comparable are parallel corpora ? measuring the distribution of general vocabulary and connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora (BUCC 2011)*, Portland, USA.

DEGAND L. & MAAT H. P. (2003). A contrastive study of dutch and french causal connectives on the speaker involvement scale. In A. VERHAGEN & J. M. VAN DE WEIJER, Eds., *Usage-based approaches to Dutch*, p. 175–19. Utrecht : LOT.

HUET S., JULIEN B. & LANGLAIS P. (2009). Intégration de l'alignement de mots dans le concordancier bilingue TransSearch. In *Actes de la 16è Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France.

IORDANSKAJA L. & MEL'ČUK I. (à paraitre). Cet OR mystérieux. *Festschrift Professeur B. Oguibénine*.

KNOTT A. (1996). *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh.

PAK M.-G. (1997). Les relations causales directes en français et en coréen. *Linguisticae Investigationes*, **XXI :1**, 139–162.

PDTB GROUP (2008). *The Penn Discourse Treebank 2.0 Annotation Manual*. Rapport interne, Institute for Research in Cognitive Science, University of Philadelphia.

REY J. (1999). Approche argumentative des textes scientifiques : la traduction de *or* en espagnol. *Meta*, **44 :3**, 411–428.

ROSSARI C. (1993). Problèmes posés par la traduction français-italien des connecteurs *de fait* et *en fait*. *Actes du XXème Congrès International de Linguistique et Philologie Romanes*, p. 69–80.

ROZE C. (2009). LEXCONN : Base lexicale des connecteurs discursifs du français. Master's thesis, Université Paris 7 Denis Diderot, Paris, France.

ROZE C., DANLOS L. & MULLER P. (2010). LEXCONN : a French Lexicon of Discourse Connectives. In *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010)*, Moissac, France.

WILSON D. & SPERBER D. (1993). Linguistic form and relevance. *Lingua*, **90**, 1–25.