# A LINGUÍSTICA EM DIÁLOGO

## VOLUME COMEMORATIVO DOS 40 ANOS DO CENTRO DE LINGUÍSTICA DA UNIVERSIDADE DO PORTO

COMISSÃO ORGANIZADORA
João Veloso
Joana Guimarães
Purificação Silvano
Rui Sousa-Silva


TÍTULO A Linguística em diálogo
Volume comemorativo dos 40 anos do Centro de Linguística da Universidade do Porto

COORDENAÇÃO João Veloso
Joana Guimarães
Purificação Silvano
Rui Sousa-Silva

EDITOR Centro de Linguística da Universidade do Porto

ANO DE EDIÇÃO 2018

CONCEÇÃO GRÁFICA Invulgar - Artes Gráficas, S.A.

TIRAGEM 200 exemplares

ISBN 978-989-54104-3-9

DEPÓSITO LEGAL **443246/18**

A publicação deste volume contou com o apoio financeiro da Fundação para a Ciência e a Tecnologia, através do financiamento atribuído ao Centro de Linguística da Universidade do Porto ao abrigo do Fundo de Reestruturação de Unidades 2016 - Refª UID/LIN/0022/2016.

# SYSTÈMES DE TRADUCTION AUTOMATIQUE ET LEVÉE D'AMBIGUÏTÉ : ÉTUDE COMPARÉE DE SYSTÈMES DE TABR, TAS ET TAN

Françoise Bacquelaine
franba@letras.up.pt
Faculdade de Letras da Universidade do Porto (Portugal)
Centro de Linguística da Universidade do Porto (Portugal)

RÉSUMÉ. La traduction automatique (TA) est un domaine en pleine effervescence depuis l'invention de l'ordinateur. L'aventure a commencé par la traduction à base de règles (TABR) dans les années 1940-1950. La traduction automatique statistique (TAS) s'est imposée quelques décennies plus tard et la traduction automatique neuronale (TAN) a vu le jour au XXIe siècle. Cette distinction n'est pas stricte puisque la plupart des systèmes de TA sont aujourd'hui hybrides, mais l'ambiguïté reste un piège bien connu, tant dans le cadre de la traduction humaine que de la traduction automatique.

Le mot anglais courant *issue* soulève deux types d'ambiguïté : « ambiguïté grammaticale » (nom ou verbe ?) et « ambiguïté homographique et polysémique » lorsqu'un mot a plusieurs sens dans la langue source (Hutchins 2005 : 17). Cette recherche se limite à trois sens du nom *issue* et à deux sens du verbe. Un échantillon de phrases comportant au moins une occurrence du mot *issue* dans un de ces cinq sens a été sélectionné dans le *British National Corpus* afin de comparer quatre systèmes de traduction automatique anglais-français : SYSTRAN (TABR, accès gratuit en ligne), Google Translate (TAS, accès gratuit en ligne), MT@EC (TAS, accès limité) et le système de TAN de LISA (Université de Montréal). Les résultats ont été comparés à un modèle de traduction humaine fondé sur des mémoires de traduction afin d'évaluer les faiblesses et les atouts de chaque système de TA, de comparer leurs performances et de proposer des possibilités d'amélioration grâce à l'hybridation des systèmes.

MOTS-CLÉ: Traduction automatique, linguistique informatique, levée d'ambiguïté.

ABSTRACT. Machine Translation (MT) has been a lively field of research ever since the invention of the computer. Rule-Based Machine Translation (RBMT) was the first option back in the 1940s–1950s; Statistical Machine Translation (SMT) appeared a few decades later and Neural Machine Translation (NMT) in the 21st century. This distinction is not strict since most MT systems are now hybrid, but natural language ambiguity is a well-known pitfall, be it in human or machine translation.

Two types of ambiguity can arise when using the rather common English word *issue*: "grammatical ambiguity" (noun or verb?), on the one hand, and "homographic and polysemic ambiguity (one word form with different senses in the source language)" (Hutchins 2005: 17), on the other hand. The scope of this research is limited to three senses of the noun *issue* (1. An important topic or problem for debate or discussion; 2. The action of supplying or distributing an item for use, sale, or official; 3. (*formal or law*) Children of one's own) and two senses of the verb *to issue* (1. [WITH OBJECT] Supply or distribute (something) for use or sale; 2. [NO OBJECT] (*issue from*) Come, go, or flow out from). A sample of sentences containing at least one example of usage was selected from the *British National Corpus* in order to test and compare four English-French MT systems: SYSTRAN (free online RBMT), Google Translate (free online SMT), MT@EC (restricted access SMT) and free online Neural Machine Translation by LISA (University of Montreal). The outputs were compared to a human translation model based on translation memories (parallel corpora) in order to evaluate weaknesses and strengths of each system, compare the results and find out possible ways of improving MT output through hybridisation.

Research results in cases like these are not just useful for theoretical linguistics but can also be used to heighten awareness in human translators and demonstrate that translators who are trained in computational linguistics can also work together with experts in artificial intelligence and machine translation.

KEYWORDS: Machine translation, computational linguistics, ambiguity resolution.

L'idée de la mécanisation de la traduction remonte au XVIIe siècle (Descartes, Leibniz, Wilkins) et plusieurs modèles de « machines à traduire » ont été conçus dans les années 1930 (Artsouni en France et Smirnov-Troyanskii en URSS), mais c'est indiscutablement l'invention de l'ordinateur qui a provoqué l'essor de la traduction automatique (TA) après la Seconde Guerre mondiale (Somers 2003 : 4). On distingue aujourd'hui

trois types de système. L'approche théorique a ouvert la marche avec la traduction automatique à base de règles (TABR). À partir des années 1980, l'approche empirique a démontré les atouts de la traduction automatique statistique (TAS). Depuis le milieu des années 2010, la traduction automatique neuronale (TAN) ouvre de nouvelles perspectives. Chaque système présente des atouts et des faiblesses et la plupart d'entre eux sont aujourd'hui hybrides. Cette hybridation résulte des efforts mis en œuvre pour résoudre les problèmes persistants de la TA, tous systèmes confondus. Malgré les progrès incontestables de la TA, l'ambiguïté constitue encore aujourd'hui un obstacle de taille au même titre que les phrases complexes ou les « unités de construction préformées » ou « UCP » (Schmale 2013) telles que les collocations ou les expressions idiomatiques.

Il s'agit simplement ici d'explorer les limites de la TA en termes de levée d'ambiguïté dans le cas particulier d'un objet lexical dont la polysémie et la fréquence d'emploi élevée constituent un défi tant pour le traducteur humain que pour la TA : le mot anglais *issue*. L'approche est celle d'un linguiste se prêtant à une petite expérience de TA de deux formes du mot anglais *issue* (*issue* et *issues*) par quatre systèmes différents pour en tirer quelques conclusions quant aux atouts et aux faiblesses de chacun.

Dans un premier temps, le matériel technologique et linguistique nécessaire à l'expérience est décrit : (1) systèmes de TA, (2) ambiguïté du mot anglais *issue*, (3) corpus, (4) modèle de référence (traduction humaine) et (5) critères de comparaison des systèmes de TA. Les résultats de la TA sont ensuite comparés et commentés.

## 1 – Systèmes de TA

Quatre systèmes de TA ont été choisis en fonction de l'approche dont ils relèvent, de la disponibilité de la paire de langues anglais-français et de leur accès gratuit : SYSTRANet, Google Translate, MT@EC[1] et NMT (Neural Machine Translation) by LISA. Même s'il est généralement admis

[1] Accès gratuit mais réservé aux utilisateurs autorisés.

que tous les systèmes sont aujourd'hui hybrides, les informations sur l'architecture de ces divers systèmes sont plutôt rares. SYSTRANet résulte de l'évolution des systèmes de la première génération (TABR) vers des systèmes hybrides tandis que les trois autres relèvent essentiellement de l'approche empirique de la deuxième génération (TAS) ou de la troisième génération (TAN).

Traditionnellement, SYSTRAN se fonde sur la TABR, dont l'architecture plus ou moins complexe[2] repose sur des données lexicales (dictionnaires bilingues) et des règles de transformation pour l'analyse en langue source, le transfert de la langue source à la langue cible et la génération du texte en langue cible (Koehn 2017). SYSTRANet est un outil gratuit, disponible en ligne et destiné à un usage général. La qualité de la traduction en termes de congruence (*adequacy*) et de fluidité (*fluency*) est toujours inférieure à celle que peut atteindre un outil sur mesure, destiné à un usage restreint à un domaine (très) particulier et à un type de texte bien précis tel que les bulletins météorologiques : plus le domaine couvert est réduit, meilleurs sont les résultats de la TA, toutes générations confondues. La TABR a régné pendant plusieurs décennies, elle a montré ses limites et est passée à l'arrière-plan : aujourd'hui la TAS et la TAN occupent le devant de la scène où se joue l'avenir de la TA (Koehn 2017).

Google Translate et MT@EC relèvent de l'approche empirique et reposent sur l'entraînement des systèmes à partir d'un modèle linguistique pour chaque langue cible (corpus monolingues) et d'un modèle de traduction pour chaque paire de langues (corpus aligné). Google Translate se sert des ressources en ligne (Internet) et de mémoires de traduction (Ryan McDonald, Lisbonne 2016, communication personnelle). MT@EC se fonde sur le logiciel libre Moses (Koehn *et al.* 2007) et est entraîné à partir des mémoires de traduction de divers services de traduction de la Commission européenne, d'autres institutions de l'Union européenne et de l'administration publique des États-membres. Ces systèmes sont programmés pour établir des correspondances entre le texte source et le

[2] Le fameux triangle de Vauquois donne une idée précise des divers niveaux de complexité de l'analyse et de la génération dans les systèmes de TABR, de la traduction directe à la traduction via l'interlangue en passant par le transfert syntaxique et le transfert sémantique.

texte cible à partir du modèle de segments alignés que sont les mémoires de traduction et autres corpus alignés. Ils établissent des correspondances mot à mot ou 'n-gramme à n-gramme', un n-gramme étant un groupe de n mots contigus récurrent dans un corpus. Au-delà de 3-grammes, la tâche se complique : « [the SMT] model of a single language is a trigram model because moving up to even one item longer (i.e. a quadgram model) would be computationally prohibitive » (Wilks 2009 : 106). Il est toutefois possible d'enregistrer des segments alignés plus longs tels que les expressions idiomatiques complètement figées (Koehn 2017). Si l'hybridation de la TABR consiste à ajouter des données empiriques aux dictionnaires et aux règles de transformation, celle des systèmes de TAS résulte par exemple de l'ajout de l'analyse morphologique aux données empiriques ou du regroupement automatique de mots en n-grammes (Koehn 2017). Comme SYSTRANet, Google Translate se destine au public en général tandis que MT@EC est conçu pour répondre aux besoins de l'UE et des administrations publiques des États-membres. La qualité de la TA est étroitement liée au(x) domaine(s) couvert(s). Il serait vain de demander à MT@EC de traduire une phrase comme « I love you ». Nous en avons tenu compte lors du choix du corpus de test.

La TAN utilise le même genre de données empiriques que la TAS (corpus monolingues et corpus alignés), mais au lieu de calculer la correspondance la plus probable entre mots ou n-grammes du texte source et du texte cible, la TAN calcule le mot suivant le plus probable en langue cible en se fondant sur un réseau de neurones récurrents s'inspirant du système nerveux du cerveau humain (Koehn 2017). La priorité va à la fluidité au détriment de la congruence, contrairement aux systèmes des générations précédentes. 'Neural Machine Translation by LISA' est alimenté par des mémoires de traduction des Nations unies et du Parlement européen. Les domaines couverts sont donc proches de ceux de MT@EC, mais le volume de données est moindre car l'entraînement de ces nouveaux systèmes est plus complexe et plus lent (Koehn 2017).

## 2 – Ambiguïté du mot anglais *issue*

Un énoncé ambigu donne lieu à plusieurs interprétations et donc à plusieurs candidats à la traduction. L'ambiguïté peut être lexicale, si elle se situe au niveau du mot, ou structurelle, si elle se situe au niveau de la syntaxe. Seule l'ambiguïté lexicale nous intéresse ici. Le mot anglais *issue* soulève deux types d'ambiguïté : « ambiguïté grammaticale » puisqu'il peut appartenir à la classe des verbes (V) ou des noms (N) et « ambiguïté homographique et polysémique » puisqu'il a plusieurs sens dans la langue source (Hutchins 2005 : 17). Selon la terminologie de Polguère (2002 : 128), *issue* est un vocable polysémique. Un « vocable » correspond à une « entrée » de dictionnaire (*idem* : 42) et un vocable polysémique comporte plusieurs lexies, « chaque lexie (lexème ou locution) [étant] associée à un sens donné » (*idem* : 41). En nous fondant sur le vocable *issue* de l'*Oxford English Dictionary* (OED), nous avons sélectionné les deux lexies verbales et trois des cinq lexies nominales[3] répertoriées. Les exemples et les définitions proviennent tous de l'OED.

### 2.1 – Lexie $N_1$

La lexie $N_1$ est un nom comptable qui correspond à la définition « important topic or problem for debate or discussion » comme dans l'exemple « the issue of racism ». Au pluriel, il peut signifier « personal difficulties or problems » (ex. : « You will learn to resolve personal issues without using guns, lawyers or therapists. ») ou « problems or difficulties, especially with a service or facility » (ex : « connectivity issues »). Cette lexie au sens très général remplit parfois une fonction que l'on peut qualifier d'anaphorique, ce qui complique encore la tâche de la TA, notamment lorsque les relations anaphoriques sont interphrastiques.

[3] Les sens « result or outcome of something » et « action of flowing or coming out » ont été exclus car ils n'étaient pas attestés dans le corpus de référence (cf. *infra*).

### 2.2 – Lexie $N_2$

La lexie $N_2$ est un nom massif signifiant « action of supplying or distributing an item for use, sale or official purposes » (ex. : « the issue of notes by the Bank of England »). C'est aussi un nom comptable dans deux sous-sens : « number or set of items distributed at one time » (ex. : share / bond issues ») et « each of a regular series of publications » (ex. : « the December issue of the magazine »).

### 2.3 – Lexie $N_3$

La lexie $N_3$ est un terme juridique ou appartenant au registre soutenu. Il s'agit d'un nom massif désignant l'ensemble des « children of one's own » (ex. : « the earl died without male issue »).

### 2.4 – Lexie $V_1$

La lexie $V_1$ est un verbe « with object » (donc transitif) qui correspond au sens de $N_2$ : « supply or distribute (sth) for use or sale » (ex. : « Christmas stamps to be issued in November »). L'OED distingue en outre deux emplois particuliers : « supply someone with (something) » (ex. : « everyone was issued with a gas mask ») et « formally send out or make known » (ex. : « the minister issued a statement »). Seuls les deux emplois particuliers entrent dans le cadre de cette étude.

### 2.5 – Lexie $V_2$

La lexie $V_2$ est un verbe « without object[4] » signifiant « come, go, or flow out of » (ex. : « the many springs which issue from the highlands above the wood ») dans un sens concret et « result or be derived from » (ex. : « the struggles of history issue from the divided heart of humanity ») dans un sens abstrait.

---

[4] Qui admet la voix passive, ce qui est impossible en français.

On le voit, le mot anglais *issue* est propice à l'étude de la levée d'ambiguïté en traduction. Seules les formes *issue* et *issues* présentent à la fois une ambiguïté lexicale et grammaticale. Les autres formes verbales ont donc été exclues d'emblée.

## 3 – Corpus

Pour mener à bien cette étude, deux corpus sont nécessaires : un corpus de référence permettant de dégager un modèle de traduction humaine et un corpus destiné à tester les quatre systèmes de TA.

### 3.1 – Corpus de référence

Le corpus de référence provient d'Europarl v7, un corpus aligné issu des mémoires de traduction du Parlement européen. Il a été choisi en raison de la qualité que l'on peut attendre des traducteurs de l'UE et parce qu'il compte parmi les données empiriques de MT@EC et NMT by LISA. Ce corpus a été exploré grâce à l'interface de recherche du projet OPUS (Tiedemann 2012). Les recherches lancées sur *issue* et *issues* (avec minuscule et majuscule) donnent 69.713 concordances alignées anglais-français. Nous avons donc sélectionné quelques n-grammes pour constituer un corpus de référence et déterminer le modèle de traduction humaine à partir des segments alignés anglais-français.

Pour la lexie $N_1$, de loin la plus fréquente, nous avons choisi le 3-grammes *human rights issue*(*s*) (68 occurrences au singulier et 338 au pluriel) et les trois « *phrases*[5] » mentionnées par l'OED : le 4-grammes *what is at issue* (148 occurrences) plutôt que le 2-grammes *at issue* (547 occurrences), le 4-grammes *make/made an issue of* (14 occurrences, 0 occurrence avec *makes* et *making*) et le 3-grammes *take*/*takes*/*taking issue with* (80 occurrences, 1 segment mal aligné, 0 occurrence avec *took* et

[5] (1) *at issue* : under discussion ; in dispute ; (2) *make an issue of* : treat too seriously or as a problem ; (3) *take issue with* : disagree with ; challenge. (OED)

*taken*). Nous avons ainsi analysé 648 segments alignés contenant la lexie $N_1$.

Le 4-grammes *the issue of* N [banknotes, coins, Eurobonds] (12 occurrences) illustre l'emploi de la lexie $N_2$ en tant que N massif. Quant aux deux N comptables, nous avons retenu le 2-grammes *bond*/*SDR*[6] *issue* (3 occurrences), le 5-grammes *issues of shares and bonds* (1 occurrence) et le 2-grammes *recent*/*latest*/*next*/*April*/… *issue* (16 occurrences), soit 32 segments alignés analysés.

Le terme juridique *issue* n'est pas attesté dans ce corpus. Il a néanmoins été conservé car il est susceptible d'être reconnu par MT@EC et NMT by LISA contrairement à SYSTRANet et Google Translate. Le modèle de traduction provient de la banque de données terminologiques de l'UE : Interactive Terminology for Europe ou IATE.

Pour le verbe transitif, nous avons choisi une série d'objets fréquents du $V_1$ dans le sens de « formally send out or make known » et retenu toutes les occurrences ou une partie représentative lorsqu'elles étaient trop nombreuses. Ainsi, l'analyse a porté sur 159 segments alignés comportant les objets *statement* (20 occurrences sur 92), *opinion* (20 occurrences sur 106), *directive*(*s*) (20 occurrences sur 26), *recommendation*(*s*) (20 occurrences sur 39), *decision*(*s*) (9 occurrences), *warning*(*s*) (30[7] occurrences sur 100), *report*(*s*) (20 occurrences sur 39) et *instructions* (20 occurrences sur 24).

Quant à la construction du $V_1$ *issue someone with something*, l'échantillon analysé se limite aux cas où *something* est *statement* (3 occurrences), *reports* (4 occurrences) et *list* (2 occurrences).

Enfin, la lexie $V_2$ est très rare dans ce corpus. Nous n'avons détecté que 10 occurrences du sens concret et 4 occurrences du sens abstrait.

---

[6] Special drawdown rights.

[7] Nous en avons retenu 30 au lieu de 20 en raison de la diversité des traductions proposées avec cet objet.

### 3.2 – Corpus de test

Le corpus de test provient du British National Corpus (BNC). Il comporte quarante-deux exemples d'emplois de formes *issue* et *issues*. La priorité a été donnée aux collocations attestées dans le corpus de référence.

Ainsi, nous avons sélectionné

- *human rights issue*(*s*) (4), *civil rights issues* (1), *what is at issue* (1), *make an issue of* (4), *take issue with* (5), soit quinze exemples d'emploi de $N_1$ ;
- *issue of banknotes* (massif, 1), *rights issue*(*s*) (comptable, 3), *recent issue* (*of a magazine*) / *May issue* (comptable, 2), soit six exemples d'emploi de $N_2$ ;
- trois exemples d'emploi de $N_3$ dans deux phrases contiguës ;
- *issue a directive*, *an opinion*, *recommendations*, *a decision*, *a warning*, *a statement*, *a report*, *instructions* (9, car il y a 2 exemples de *statement*), *issue someone with something* (4), soit treize exemples d'emploi de $V_1$ ;
- *issue from* (2 dans le sens concret et 3 dans le sens abstrait), soit cinq exemples d'emploi de $V_2$.

## 4 – Modèle de traduction humaine

L'analyse du corpus de référence révèle une grande diversité d'équivalents français pour $N_1$ et $V_1$, ce qui ne facilite pas l'élaboration du modèle de traduction. Les résultats de l'analyse des traductions humaines de ce mot anglais polysémique sont présentés sous forme de tableaux commentés.

| $N_1$ | Équivalent le plus fréquent | Autres équivalents |
|---|---|---|
| the human rights issue | la question des droits de l'homme (73,7%) | problématique (13,2%), [NUL](0,6%), problème (0,3%), s'agissant de (0,3%), thème (0,3%) |
| a human rights issue | une question de droits de l'homme (63,3%) | problème (13,3%), [NUL] (6,7%), débat (6,7%), thème (3,3%), sujet (3,3%), affaire(3,3%) |
| human rights issues | questions de(s) droits de l'homme (54,7%) | [NUL] (16%), problèmes (11%), question (7,4%) |
| what is at issue (is) | il s'agit de/ce dont il s'agit (31,8%) | question (19,6%), enjeu/en jeu (18,9%), ce qui est en cause (4,7%) |
| make an issue of | faire une montagne/une affaire de, monter en épingle, remettre en question, ... | |
| take issue with | désaccord (23,75%) | ne pas être d'accord avec (10%), contester qch (10%), contredire qn (6,25%) |

TABLEAU 1 – Modèle de traduction de $N_1$

L'équivalent français *question* se distingue nettement, mais lorsque *human rights issue* est précédé de l'article indéfini, les traducteurs ont tendance à ajouter *qui touche* / *touchant* / *liée aux* (*droits de l'homme*). Il en va de même au pluriel où les variantes sont beaucoup plus nombreuses : *relatives*, *portant sur*, *concernant*… Dans le cas de *human rights issues*, seuls les équivalents apparaissant plus de vingt fois ont été retenus (y compris *question* au singulier), mais il y en a d'autres qui sont attestés plusieurs fois : *affaires*, *sujets*, *aspects*, *dossiers* au pluriel et *problématique*, *domaine*, *situation*, *sujet*, mais aussi *défense* ou *respect* au singulier. Ainsi,

$N_1$ admet de nombreux équivalents, y compris l'omission (NUL), surtout au pluriel. Par exemple, *report on human rights issues* est traduit par *rapport sur les droits de l'homme*.

Lorsque *issue* dans *what is at issue* (*is*) est traduit par *question*, on trouve plusieurs formules : *il est question de*, *la question est de savoir* et *ce dont il est question*. *Enjeu* et *en jeu* ont été considérés comme le même équivalent utilisé vingt-huit fois pour traduire ce 4-grammes : *l'enjeu* (11/28), *ce qui est en jeu* (15/28), *la question en jeu* et *les éléments qui sont en jeu*. Parmi les 30 % restants, on trouve l'omission pure et simple, *problème*, *sujet*, *objectif*, *but*, *idées défendues*, voire le pronom *cela* qui suggère une valeur anaphorique de $N_1$.

*Make an issue of* (14) n'a pas été traduit deux fois de la même façon. Les quatre expressions présentées dans le tableau constituent chacune un hapax parmi d'autres.

Enfin, on remarque quelques contresens ou atténuations dans la traduction de *take issue with* (*débattre de*, *revenir sur*, *aborder la question*,…). *Désaccord* est l'équivalent le plus fréquent de *issue*, mais ce, dans diverses formules : *être* / *rester en désaccord*, *exprimer* / *afficher* / *marquer* /*ne pouvoir cacher son désaccord* (*avec* quelqu'un et parfois quelque chose, *sur* quelque chose). Parmi les autres équivalents, on trouve *être* / *s'élever* / *prendre parti contre*, *s'opposer à*, *ne pas partager l'avis de*, …

$N_2$ pose moins de problèmes, comme le montre le tableau 2 :

| $N_2$ | Équivalent le plus fréquent | Autres équivalents |
|---|---|---|
| issue of (bank notes, coins, Eurobond, shares and bonds) | émission (100%) | |
| (SDR, bond) issue | émission (100%) | |
| (recent, latest, next) issue of (a magazine, newspaper, etc.) | numéro (83,3%) | édition (16,7%) |
| (date) issue of (a magazine, newspaper, etc.) | nul (100%) | |
| the April issue of (an information sheet) | parution (hapax) | numéro, édition (Internet) |

TABLEAU 2 – Modèle de traduction de $N_2$

Comme *human rights issue*, il s'agit ici de collocations et la traduction en français du collocatif $N_2$ dépend de la base (*bank notes*, *newspaper*,…). Avec une date, *issue* est systématiquement omis en français comme dans l'exemple *in the 14 September issue of one of the major Paris daily papers* qui se traduit par *dans un grand quotidien parisien daté du 14 septembre*. Le corpus ne contient qu'une seule occurrence avec un nom de mois, or *numéro d'avril* ou *édition d'avril* sont plutôt plus fréquents en français[8]. Nous admettons donc les trois solutions dans notre modèle.

$N_3$ n'a pu être identifié dans Europarl v7. L'équivalent proposé par IATE est *descendance*. Ce terme correspond bien à la définition de l'OED.

À l'instar du collocatif $N_2$, le choix du collocatif $V_1$ dépend de la base nominale (N objet direct ou précédé de *with* dans la structure figée). Le tableau 3 présente les résultats de l'analyse du corpus :

[8] Comme le confirment les résultats d'une recherche lancée sur Google (sites .fr) : « numéro d'avril » : 75.300 résultats ; « édition d'avril » : 27.900 résultats ; « parution d'avril » : 7.770 résultats.

| $V_1$ | Équivalent le plus fréquent | Autres équivalents |
|---|---|---|
| issue a statement | faire une déclaration (25%) | NUL (25%), déclarer (10%), publier une déclaration (10%),... |
| issue an/POSS opinion | se prononcer (35%) | rendre un/son avis (25%), émettre un avis (20%), ... |
| issue [(a) directive(s)] | émettre (25%) | publier (10%), rédiger (10%), présenter (10%), élaborer (10%), adopter (5%), ... |
| issue (a/POSS) recommendation(s) | émettre (40%) | produire (15%), formuler (5%), ... |
| issue [(a) decision(s)] | prendre (33,3%) | rendre (22,2%) |
| issue (a) warning(s) | lancer un/des avertissement(s) (26,7%) | mettre en garde (23,3%), adresser un/ des avertissement(s) (10%), avertir (10%) |
| issue [(a) report(s)] | publier (45%) | rédiger (10%), fournir (10%), NUL (10%), produire (5%), ... |
| issue (an) instruction(s) | donner (60%) | hapax: NUL, publier, ... |
| issue s.o. with (statement, list, mandate, reports) | faire une déclaration, fournir (une liste, des rapports), donner (un mandat) | |

TABLEAU 3 – Modèle de traduction de $V_1$

Dans le corpus de référence, le verbe « without object » *issue from* présente certaines particularités. L'origine (*from* N) peut être un lieu concret ou une quelconque origine ou cause humaine ou autre. À cela s'ajoute la métonymie habituelle (*Bruxelles* pour *la Commission européenne*, etc.) et

la possibilité d'emploi de ce verbe « without object » à la voix active et à la voix passive sans changer de sujet (*Grandiose ideas issued from on high* (…) *will always fail*), ce qui est impossible en français. Le modèle doit donc aussi refléter les choix des traducteurs selon que $V_2$ est employé à la voix active ou passive. C'est ce dont rend compte le tableau 4.

| $V_2$ | Équivalents |
|---|---|
| issue from (cause humaine ou non ; lieu = métonymie) | émaner de (2) ; venir de ; découler de ; dériver de ; résulter de ; provenir de ; de la part de (6 hapax) |
| issued from (lieu) | formulé (en coulisses) ; lancé depuis (2 hapax) |
| issued from (cause humaine ou non ; lieu = métonymie) | émanant de ; diffusé à partir de ; imposé par (3 hapax) |

TABLEAU 4 – Modèle de traduction de $V_2$

Le nombre réduit d'exemples d'emploi de $V_2$ empêche toute généralisation : le traducteur ne peut se fier aux mémoires de traduction et choisit l'équivalent qui lui semble le plus adéquat au cas par cas. On imagine l'obstacle que cela constitue pour la TA. Retenons simplement que *émaner de* est ici l'équivalent le plus fréquent puisque toutes les autres solutions retenues par les traducteurs sont des hapax.

## 5 – Critères de comparaison

La question de la qualité de la TA du mot anglais *issue* en français ne se pose pas de la même façon selon que *issue* fonctionne comme un N ou un V collocatif (*issue of banknotes*, *issue a statement*) ou qu'il fait partie d'une des quatre UCP dont le degré de figement est supérieur à celui d'une collocation (*what is at issue*, *make an issue of*, *take issue with*, *issue someone with something*). Les critères de comparaison des systèmes de TA diffèrent donc sensiblement.

La qualité de la TA des quatorze exemples d'emploi de ces quatre

UCP se mesure en termes de reconnaissance et traduction en bloc des trois premières, et en termes de reconnaissance et transformation syntaxique dans le cas de la quatrième. Ces critères concernent à la fois la congruence de la traduction et le degré de fluidité du style.

Quant à la qualité de la traduction du collocatif *issue*, le premier critère concerne la levée de l'ambiguïté grammaticale, c'est-à-dire la distinction entre N et V. Le deuxième sert à évaluer la TA en classant les résultats selon le candidat proposé : (1) le plus fréquent, (2) une variante adéquate, (3) un faux sens, (4) un non-sens. La variante adéquate peut être attestée ou non dans le corpus de référence. L'équivalent le plus fréquent et la variante adéquate peuvent donner lieu ou non à une solution digne d'un être humain, ce qui sera également signalé. Le faux sens signifie que le texte cible est fluide mais ne correspond pas au texte source. L'omission à mauvais escient (puisqu'elle est parfois naturelle en français) et la méconnaissance du mot (UNK ou non-traduction) sont considérées comme des non-sens.

## 6 – Résultats de la TA

Le premier test porte sur les 14 UCP dont le degré de figement est très élevé. Comme le montre le tableau 5, la TAS (GT et surtout MT@EC) se montre bien plus performante que la TABR (SYSTRANet) ou la TAN (NMT by Lisa) dans la reconnaissance de l'UCP et sa traduction en bloc ou sa transformation syntaxique.

| Figement | SYSRANet | GT | MT@EC | NMT by LISA |
|---|---|---|---|---|
| Reconnaissance de l'UCP | 5/14 | 5/14 | 9/14 | 1/14 |
| Traduction en bloc | 5/10 | 3/10 | 6/10 | 4/10 |
| Transformation syntaxique | 0/4 | 2/4 | 3/4 | 0/4 |

TABLEAU 5 – Traduction automatique de quatre UCP
(unités de construction préformées)

MT@EC reconnaît et traduit correctement au moins une fois chaque UCP (64 % de traduction réussie, 6 traductions en bloc et 3 transformations syntaxiques). GT ne reconnaît que deux UCP sur quatre : *take issue with* (3 traductions en bloc / 5 occurrences traduites) et *issue someone with* (2 transformations syntaxiques / 4 occurrences traduites). SYSTRANet ne reconnaît que l'UCP *take issue with* qu'il traduit systématiquement en bloc par *contester*. NMT by LISA propose une traduction en bloc de qualité humaine de *what is at issue is* (*il s'agit de savoir*) et trois traductions en bloc jugées inadéquates de *take issue with* (*aborder*, *s'interroger sur* et *se pencher sur*). Les exemples (1) à (5) illustrent ces différences en matière de TA d'UCP.

(1) *What is at issue is whether …* (BNC)
*Il s'agit (donc) de savoir si…* (MT@EC et NMT by LISA)
**Ce qui est à la question est si…* (SYSTRANet)
**Quel est l'enjeu est de savoir si…* (GT)

Ce premier exemple révèle que l'UCP va au-delà du 4-grammes *what is at issue*. Cette UCP est fréquente dans les corpus de MT@EC et NMT by LISA, ce qui explique leur succès par rapport à GT et à SYSTRANet, conçus pour un usage général. Cette UCP passe inaperçue dans le système de TAS (GT) et n'est manifestement pas répertoriée comme telle dans le système de TABR (SYSTRANet).

(2) Nor will he *make an issue of the incident …* (BNC)
Il ne sera pas non plus *mettre en cause l'incident …* (MT@EC)
*Ni il *fera une question de l'incident …* (SYSTRANet)
*Il ne pourra d'en *faire un problème l'incident …* (GT)
*Il ne *fera* pas non plus *la question de l'incident …* (NMT by LISA)

Seul *mettre en cause* peut ici prétendre au statut de candidat à la traduction de *make an issue of*. La traduction de *Nor will he* sort du cadre de cette étude, mais notons que la solution de NMT by LISA est incontestablement la plus fluide.

(3) Pesh Framjee *takes issue with* the way …
Pesh Framjee *conteste* la manière/façon … (SYSTRANet, GT et MT@EC)
UNK UNK *se penche sur* la façon dont… (NMT by LISA)

Tous les systèmes traduisent l'UCP en bloc. Cependant la TAN propose une solution fluide mais inadéquate au texte source, puisque la nuance de désaccord est omise.

(4) … their employers should *issue them with briefing documents* … (BNC)
… leurs employeurs doivent / leur employeur devrait *leur délivrer des documents d'information* … (GT et MT@EC)
… *leurs employeurs devraient *les publier avec des documents de briefing* … (SYSTRANet)
… *leur employeur *les UNK avec des documents d'information* … (NMT by LISA)

Dans l'exemple (4), les systèmes de TAS reconnaissent l'UCP et apportent les transformations syntaxiques nécessaires contrairement aux deux autres.

Ces quatre exemples montrent la supériorité de MT@EC. L'exemple (5) illustre la systématicité de la TABR par rapport à la TAS et la traduction en bloc par la TAN :

(5) … I must *take issue with* one or two of his conclusions … (BNC)
… je dois *contester* un ou deux de ses conclusions … (SYSTRANet)
*… je dois *prendre problème avec* un ou deux de ses conclusions … (GT)
*…nécessaire d'*aborder la question sous* un ou deux de ses conclusions… (MT@EC)
?… je dois *aborder* un ou deux de ses conclusions … (NMT by LISA)

Le deuxième test porte sur l'échantillon de 28 occurrences de collocations où le N ou le V *issue* joue le rôle de collocatif. Le tableau 6 rend compte de la levée de l'ambiguïté grammaticale (N ≠ V) et des cas où l'omission du collocatif équivalent à *issue* (Ø), la méconnaissance du mot ou de la collocation (UNK) ou la non-traduction (*issue* conservé en français) ne permettent pas de trancher sur ce point.

| | SYSTRANet | GT | MT@EC | NMT by LISA |
|---|---|---|---|---|
| N ≠ V | 27 | 25 | 22 | 19 |
| Ø – UNK - *issue* | 0 | 1 | 3 | 9 |

TABLEAU 6 – Levée de l'ambiguïté grammaticale

SYSTRANet parvient à distinguer le N du V vingt-sept fois sur 28, mais il tombe une fois dans le piège de l'ambiguïté grammaticale de $N_1$. GT le talonne, omet une fois $N_3$ et tombe deux fois dans le piège tendu par $V_2$. MT@EC tombe trois fois dans le piège tendu par $V_2$, omet deux fois $N_2$ et conserve une fois $V_2$ (non-traduction). Les problèmes de la TAS concernent les emplois les moins fréquents du N ($N_2$ et $N_3$) et du V ($V_2$) tandis que l'analyseur morpho-syntaxique de SYSTRANet remplit sa mission à un cas près. Les résultats de la TAN sont complètement différents. Ce système ne se trompe jamais de catégorie grammaticale. Il lui arrive sept fois de ne pas reconnaître le mot dans son contexte (UNK) et deux fois d'omettre le collocatif indispensable correspondant à $N_2$.

L'exemple 6 illustre le seul cas d'échec de SYSTRANet lors de la TA de $N_1$ qui résulte dans une phrase agrammaticale :

(6) … a complete change of tack from labour education, to *civil rights issues*. (BNC)
*… un changement complet de pointe d'éducation de travail, vers *des droits civiques publie*. (SYSTRANet)
… un changement complet de cap de l'éducation du travail, les *questions de droits civiques*. (GT)
… un changement de cap, de l'éducation, du travail pour les

*émissions de droits*. (MT@EC)
… un changement complet de UNK de l'éducation au travail, aux *questions relatives*
*aux droits civils*. (NMT by LISA)

MT@EC propose un N inadéquat, mais ce critère n'entre pas ici en ligne de compte. GT propose l'équivalent le plus fréquent et NMT by LISA propose une solution comparable à la traduction humaine. En effet, « relatives aux » a été ajouté alors que ces deux mots ne correspondent à rien dans le texte source. Cet ajout améliorant la fluidité est attesté dans le corpus de référence et illustre en quoi peut consister une traduction digne d'un être humain.

Le tableau 7 rend compte des équivalents proposés par les divers systèmes testés et la capacité à proposer des solutions comparables à la TH :

| Équivalent | SYSTRANet | GT | MT@EC | NMT by LISA |
|---|---|---|---|---|
| le + fréquent | 6 | 11 | 11 | 8 |
| variante | 6 | 6 | 7 | 6 |
| faux sens | 4 | 4 | 3 | 4 |
| non-sens | 12 | 7 | 9 | 10 |
| solution TH | 0 | 0 | 2 | 3 |

TABLEAU 7 – Levée de l'ambiguïté sémantique

Seuls MT@EC et NMT by LISA proposent des solutions dignes d'un être humain telles que l'illustrent les exemples (6) ci-dessus et (9) ci-dessous. Si l'on additionne les équivalents les plus fréquents aux variantes acceptables, MT@EC se classe en tête (64,3 % de levée d'ambiguïté sémantique) et est suivi de près par GT (60,7 %). La TAN occupe la troisième position (50 %) et la TABR se retrouve lanterne rouge (42,9%). En termes de faux sens, les résultats sont très proches, mais MT@EC sort vainqueur avec seulement trois faux sens tandis que c'est GT qui s'en sort

le mieux en termes de non-sens avec seulement sept non-sens. Le poids des non-sens est équivalent à celui des traductions acceptables dans le cas de SYSTRANet, mais il équivaut à 50% du poids des traductions acceptables proposées par MT@EC. Quant à NMT by LISA, la plupart des non-sens correspondent à « *UNK* », seules deux omissions du collocatif résultent dans un non-sens. Les exemples (7) à (10) illustrent chacun un cas de TA de $N_2$, $N_3$, $V_1$ et $V_2$.

(7) Eurotunnel is expected to make *a new rights issue* next year ... (BNC)
Eurotunnel devrait faire *une nouvelle émission de droits* l'année prochaine ... (GT)
Eurotunnel devrait procéder à *une nouvelle émission de droits* l'année prochaine ... (MT@EC)
*On s'attend à ce que fasse *une question de nouvelles droites* l'année prochaine ...(SYSTRANet)
?L'on s'attend à ce que l'on UNK *une nouvelle question de droits* l'année prochaine … (NMT by LISA)

L'exemple (7) illustre la levée de l'ambiguïté sémantique de $N_2$ par les systèmes de TAS, le non-sens de la TABR et le faux sens de la TAN.

(8) … if the wife or husband *dies without issue* the tenant ... (BNC)
?... si l'épouse ou le mari *meurt sans question* le locataire ... (SYSTRANet)
?... si la femme ou le mari *meurt sans* le locataire ... (GT)
?... si l'épouse ou l'époux *décède sans problème* le preneur … (MT@EC)
?... si la femme ou le mari *meurt sans problème*, le locataire ... (NMT)

Dans l'exemple (8), $N_3$ (*descendance*) n'a pas été identifié. Seul GT propose d'omettre purement et simplement ce terme juridique appartenant à un registre soutenu pour aboutir à un non-sens tandis que tous les autres proposent un équivalent de $N_1$ et donc un faux sens.

(9) … four months in which to *issue a final decision*. (BNC)
?… quatre mois (…) pour *publier une conclusion définitive*. (SYSTRANet)
… quatre autres mois pour *rendre une décision définitive*. (GT)
… quatre mois pour *arrêter une décision finale*. (MT@EC)
?... quatre mois pour *publier une décision finale*. (NMT by LISA)

En (9), GT propose l'un des équivalents les plus fréquents de $V_1$ tandis que MT@EC propose une collocation plus rare digne d'un être humain. L'équivalent proposé par la TABR et la TAN a été considéré comme un faux sens puisque rien ne permet de supposer que la décision sera publiée.

La TA des cinq occurrences de $V_2$ a toujours résulté dans un non-sens, que l'équivalent proposé corresponde à $N_1$ (10) ou à $V_1$ (11) dans le cas de la TAS, à $V_1$ dans le cas de la TABR ou à *UNK* dans le cas de la TAN.

(10) … the River Hope which *issues from the freshwater Loch Hope*. (BNC)
*… l'espoir de rivière qui *publie de l'espoir d'eau douce de loch*. (SYSTRANet)
*… l'espérance qui *les questions de l'eau douce de la rivière Espérance Loch*. (GT)
*… l'espoir (…) la rivière *questions du Loch Hope d'eau douce*. (MT@EC)
*… la rivière UNK qui *UNK de l'eau douce UNK Hope.* (NMT by LISA)

Outre le non-sens de ces TA, la TABR propose un V (*publie*) et la TAN reconnaît son ignorance (*UNK*), mais la TAS se trompe de catégorie grammaticale (*questions*), ce qui illustre également les résultats du test sur la levée de l'ambiguïté grammaticale. Dans le dernier exemple, la TAS et la TABR traduisent $V_2$ par un équivalent de $V_1$ :

(11) … savoury smells began to *issue from the kitchen*. (BNC)
*… les odeurs savoureuses commençaient à *publier de la cuisine*. (SYSTRANet)
*… odeurs salées ont commencé à *émettre à partir de la cuisine*. (GT)
*… les odeurs salés a commencé à *délivrer de la cuisine*. (MT@EC)
*… les odeurs UNK ont commencé à *s'UNK de la cuisine*. (NMT by LISA)

Si SYSTRANet propose systématiquement *publier* comme équivalent de $V_2$, ses TA ne sont pas tout à fait systématiques dans les autres cas. Pour traduire le N *issue*, il propose une fois *numéro* pour $N_2$ et pour traduire $V_1$, une fois *émettre* (*un avis*) et deux fois *fournir* (*des instructions*, *un avertissement*).

## 7 – Conclusion

Étant donné la dimension réduite de l'échantillon de test, le choix des UCP figées et des collocations (plutôt favorable à MT@EC et NMT by LISA), les critères de comparaison retenus et le choix des outils testés, les résultats sont relatifs. Ainsi, MT@EC est naturellement plus performant que les systèmes de TA disponibles en ligne, mais cette analyse confirme que la levée de l'ambiguïté grammaticale et surtout lexicale reste un défi de taille pour la TA. Chaque système présente néanmoins des atouts et des faiblesses.

SYSTRANet se distingue en ce qui concerne la levée de l'ambiguïté grammaticale, mais ce système manque cruellement de flexibilité. Il tend à privilégier les relations sémantiques bi-univoques entre mot de la langue source et mot de la langue cible. À quelques rares exceptions près, le N anglais est traduit systématiquement par *question* et le V par *publier*, qu'il s'agisse d'UCP figées ou de collocations.

La TAS domine nettement le palmarès de la TA des UCP figées et des collocations. Ses performances sont intimement liées à la fréquence d'emploi de chacun des cas envisagés ici. On pouvait s'y attendre puisque

GT et MT@EC disposent d'un volume impressionnant de données empiriques et que l'échantillon est favorable à MT@EC, dont les données sont sans doute de meilleure qualité que celles de GT. Comme la TABR, elle semble toutefois avoir atteint ses limites.

Le système de prévision du mot suivant compte tenu d'un contexte plus large que 3-grammes permet à la TAN de surpasser tous ses concurrents en termes de fluidité et de traductions dignes d'un être humain. Ses résultats médiocres dans le cadre de ce test particulier sont dus essentiellement à la taille réduite des corpus d'entraînement de la version gratuite testée en ligne, mais cette nouvelle approche n'en est qu'à ses débuts et les perspectives sont prometteuses, comme en témoignent par exemple les résultats obtenus par le système de TAN de l'université d'Édimbourg en 2016 (Bojar *et al.* 2016 : 141-144 *et passim*).

**RÉFÉRENCES**

BNC Consortium. 2007. *British National Corpus*. Explored September 2016, from the World Wide Web: http://corpus.byu.edu/bnc/.

Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huck, M.; Jimeno Yepes, A.; Koehn, P.; Logacheva, V.; Monz, C.; Negri, M.; Neveol, A.; Neves, M.; Popel, M.; Post, M.; Rubino, R.; Scarton, C.; Specia, L.; Turchi, M.; Verspoor, K.; Zampieri, M. 2016. Findings of the 2016 Conference on Machine Translation (WMT16). In: *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*. Berlin: Association for Computational Linguistics, 131-198. Retrieved March 17, 2017, from the World Wide Web: http://www.statmt.org/wmt16/pdf/W16-2301.pdf

European Union. 1995-2014. "issue" in *Interactive Terminology for Europe* (*IATE*). Retrieved October 10, 2016, from the World Wide Web: http://iate.europa.eu/SearchByQuery.=12&matching=&start=0&next=1&targetLanguages=fr

Hutchins, J. 2005 [Corrected version of 1997 paper in: *Machine Translation*. **12 (3):**195-252]. From first conception to first demonstration: the nascent years of machine translation, 1947-1954. A chronology. Retrieved March

17, 2017, from the World Wide Web: http://www.hutchinsweb.me.uk/MTJ-1997-corr.pdf.

Koehn, P. 2017. Introduction to Neural Machine Translation (NMT). Omniscien Technologies [Webinar held on 24 January 2017]. Retrieved March 17, 2017, from the World Wide Web: https://vimeo.com/201401054.

Koehn, P. 2012. *Europarl v7*. [corpus aligné, voir Tiedemann 2012]

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation.In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague: Association for Computational Linguistics, 177–180. Retrieved March 17, 2017, from the World Wide Web: http://www.aclweb.org/anthology/P07-2045.

MacDonald, R. 2016. Communication personnelle au META-FORUM 2016: *Beyond Multilingual Europe* (Lisbonne, 4-5 juillet 2016)

Oxford University Press. 2017. "issue" in *Oxford English Dictionary*. Retrieved March 17, 2017, from the World Wide Web: https://en.oxforddictionaries.com/definition/issue.

Polguère, A. (2002). *Notions de base en lexicologie*. Montréal: OLST. [Manuel de cours polycopié]. Retrieved October 10, 2016, from the World Wide Web: https://www.fichier-pdf.fr/2013/11/18/notions-de-base-en-lexicologie/.

Schmale, G. 2013. Qu'est-ce qui est préfabriqué dans la langue ? – Réflexions au sujet d'une définition élargie de la préformation langagière. *Langages*. **189**: 27-45.

Somers, H. 2003. Introduction. In: Somers, H. (Ed.). *Computers and Translation. A translator's guide*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1-11.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. In: Calzolari, N.; Choukri, K.; Declerck, T.; Doğan, M. U.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; Piperidis, S. (Eds). *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*. Istanbul: European Language Resources Association, 2214-2218.

Wilks, Y. 2009. *Machine Translation. Its Scope and Limits*. New York: Springer.