Corpus de registres différents pour le développement d'un aligneur d'unités polylexicales

Claire Lemaire , Jean-Philippe Guilbaud

LIG-GETALP, Bâtiment IMAG, 700 avenue Centrale F-38400 St. Martin d'Hères, France,
(claire.lemaire@imag.fr), (Jean-Philippe.Guilbaud@imag.fr.)

Résumé : Comment trouver des données exprimant les mêmes concepts dans des registres de langue différents ? Après un essai infructueux d'extraction terminologique à partir de corpus comparables spécialisés dans le domaine du médical dans trois langues différentes (anglais, allemand, français), l'idée est d'ajouter pour chaque langue des sous-corpus du registre vulgarisé afin d'y détecter des relations de synonymie. Or ce type de ressources n'existe pas pour l'allemand de spécialité dans le domaine du médical. Nous présentons la constitution d'un corpus de 400 000 mots en allemand dans le domaine de la cancérologie, subdivisé en deux sous-corpus de même taille. À partir d'équivalents en allemand du mot-clé «cancer du sein», nous avons recueilli pour un premier sous-corpus, des textes qui s'adressent à des patientes (et des patients) ou à leurs familles, et pour un second sous-corpus, des textes qui s'adressent à des médecins ou des chercheurs en médecine.

Mots-clés : corpus de registres différents, corpus allemand, aligneur d'unités polylexicales, analyseur morphologique

1 Introduction

Le point de départ de ce travail est que, dans l'industrie de la communication multilingue, on ressent le besoin de disposer d'équivalents non seulement translingues, mais transregistres, par exemple, « carcinome mammaire » (registre scientifique) et « cancer du sein » (registre vulgarisé). L'événement déclencheur de la création de ce corpus est le développement d'un aligneur d'unités polylexicales au sein d'une même langue, qui utilise une méthode qui repose sur des données multilingues, et qui sera utilisé notamment pour de la recherche d'informations. La première idée, pour fabriquer cet aligneur d'unités polylexicales, a été de collecter des textes sur un sujet

d'expertise précis rédigés dans le sous-langage des experts dans un domaine spécialisé, et cela, dans différentes langues : anglais, français, allemand (Delpech *et al.*, 2012).

Dans une première partie, nous exposons cette première idée de connexion par lexique, puis une seconde idée plus efficace d'ajout de textes moins spécialisés. Dans une seconde partie, nous décrivons le recueil de données en allemand, fournissons des notions de volumétrie, décrivons la collecte des textes, puis présentons deux échantillons. Dans une troisième partie, nous expliquons le traitement des données, le prétraitement des fichiers puis le traitement en lui-même. Dans une quatrième et dernière partie, nous décrivons l'utilisation de notre corpus pour l'analyseur morphologique AMALD, en reprenant les principes linguistiques de l'analyseur et les outils utilisés, puis présentons un cas concret d'amélioration de l'analyse linguistique, celui du rattachement de particule, problème complexe en analyse automatique de la langue allemande. Nous concluons sur la poursuite des travaux en cours qui reposent sur ce même corpus.

2 Du corpus trilingue au corpus trilingue transregistre

2.1 Connexion par lexique et fouille de contextes

Reprenons cette première idée. Il s'agit de profiter de la connexion déjà faite par l'alignement dû à la traduction, puis de fouiller les contextes droits et gauches pour trouver des relations de synonymie. Par exemple, nous partons du terme « carcinome mammaire » en français ; ce terme correspond à *mamma carcinom* en anglais dans notre lexique.

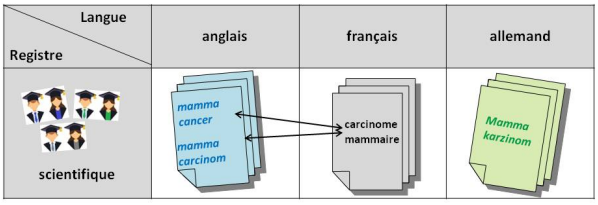


FIGURE 1: Corpus comparable trilingue, registre scientifique

Quand nous fouillons les textes en anglais cette fois à partir de *mamma carcinom*, nous trouvons par exemple dans les contextes droits et gauches anglais le terme *mamma cancer*. Ainsi, nous avons trouvé des relations de synonymie à l'intérieur d'une même langue. Ce sont ces ressources ainsi constituées qui sont ensuite très utiles pour l'industrie de la langue, afin d'effectuer de la recherche d'informations, notamment mais uniquement translingue.

Cette première idée n'a pas fourni assez de synonymes ; la raison de cet échec est qu'il n'y avait pas assez de termes différents pour exprimer le même concept dans nos corpus, puisque ceux-ci étaient extrêmement spécialisés (Hazem, 2018).

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Anglais	Français
breast cancer	carcinome mammaire
phytoestrogen	phyto-estrogène
mamma-carcinom	carcinome mammaire
homogeneity	homogénéité
mamma-carcinom	cancer du sein
mammary cancer	carcinome mammaire

FIGURE 2: Exemple de termes alignés (d'après Delpech, 2013)

2.2 Ajout de textes moins spécialisés

La seconde idée pour créer cet aligneur d'unités polylexicales est d'ajouter au corpus des textes sur le même sujet d'expertise, mais cette fois rédigés en langue générale, afin de repérer dans la même langue plusieurs termes liés par une relation paraphrastique.

Dans notre exemple, le terme « carcinome mammaire » a conduit au terme *mamma carcinom* grâce au lexique. Dans les contextes de *mamma carcinom* on a trouvé *mamma cancer*, dans le contexte de *mamma cancer*, on a trouvé *breast cancer*, qui a à son tour conduit à *cancer du sein*.

Les ressources en anglais et en français de ce type ont été réalisées pour la première fois dans le cadre de travaux sur l'amélioration d'un outil de traduction assistée par ordinateur (TAO) (Delpech, 2013).

Le problème est que cet outil, une mémoire de traductions, doit proposer des historiques de traduction, en l'occurrence des corpus parallèles. Or, suivant le domaine, il se peut que le logiciel ne propose pas de corpus parallèle. Cela arrive en particulier lorsque les textes qu'il doit produire appartiennent à un domaine émergent. C'est pourquoi les recherches se sont orientées vers l'exploitation de corpus comparables.

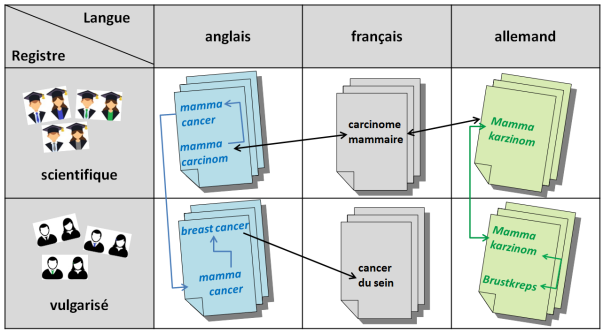


FIGURE 3: Corpus comparable trilingue et transregistre)

L'allemand est une langue bien moins dotée que le français et *a fortiori* que l'anglais ; les textes en langue de spécialité en accès libre sont donc moins nombreux (Lemaire, 2017). En revanche, il est plus facile de détecter les termes dans cette langue, car l'allemand utilise beaucoup de mots composés stricts (sans séparateur comme "-" ou "s").

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Nous avons construit la ressource nécessaire en allemand à partir de textes contenant les mots *Brustkrebs* et *Mammakarzinom*[1] (registre scientifique), équivalents du terme *cancer du sein* (registre vulgarisé), pour *carcinome mammaire*.

3 Recueil des données du corpus allemand transregistre

3.1 Volumétrie

Nous avons collecté 79 textes en allemand scientifique (pour 204 646 mots) et 162 textes en allemand vulgarisé (pour 204 634 mots), soit 278 textes (pour 409 280 mots) au total en allemand médical.

3.1.1 Unité de mesure trompeuse

Bien que l'unité choisie en linguistique de corpus soit traditionnellement le mot, elle est particulièrement mal adaptée pour l'allemand. En effet, l'allemand recourt très fréquemment aux mots composés, y compris dans le registre vulgarisé. À cause de la fréquence de longs mots composés, le caractère ou la page est souvent utilisé comme unité pour l'allemand, par exemple dans les entreprises de traduction (1400 caractères par page, une page en français équivalant à 250 mots).

Le tableau suivant (Figure 4) donne des exemples de termes issus de notre corpus ainsi que leurs correspondants en anglais, puisque l'anglais est la langue de base de l'aligneur terminologique en question, et en français, afin de donner au lecteur francophone un ordre d'idée du rapport des longueurs des termes dans ces trois langues).

anglais		français		allemand	
family members	15	**membres de la famille**	22	*Familienangehörigen*	20
quality controls	17	**contrôles de qualité**	21	*Qualitätskontrollen*	20
risk of disease	16	**risque de maladie**	18	*Erkrankungsrisiko*	18
breast cancer cases	20	**cas de cancer du sein**	22	*Brustkrebserkrankungen*	23
size difference	16	**différence de taille**	21	*Größendifferenz*	16
early detection concept	24	**concept de dépistage précoce**	29	*Früherkennungskonzept*	22
probability of developing the disease	38	**probabilité de développer la maladie**	37	*Erkrankungswahrscheinlichkeit*	30
20 mots (146 caractères)		**27 mots** (170 caractères)		**7 mots** (149 caractères)	

FIGURE 4: Les mots composés en allemand

3.1.2 Collecte de longue haleine

Dans un premier temps, nous avons recherché des textes en allemand médical rédigés par des médecins, des journalistes médicaux, et des chercheurs en médecine. Au bout de 6 mois, nous avions écumé tout ce qui était publié gratuitement en ligne dans l'espace germanophone représenté par l'Allemagne, l'Autriche et la Suisse ; nous avons ensuite enrichi le corpus au fur et à mesure de la sortie de nouveaux articles.

1. On trouve aussi l'orthographe plus rare « *Mammacarcinom* ».

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

3.2 Les sources des deux registres

3.2.1 Textes en allemand scientifique

Voici la liste des 7 sites germanophones sur lesquels nous avons collecté les textes pour le registre scientifique.

Allemand scientifique	Nom du site Web	Nombre de mots typographiques	Nombre de textes
	*aerzteblatt.de*	114 406	67
	senologie.org	81 192	1
	uni-frauenklinik-tuebingen.de	5 816	7
	iwenv.de	1 584	1
	wiralle.de	1 204	1
	krebsinformationsdienst.de	326	1
	aerztezeitung.de	118	1
Total		204 646	79

FIGURE 5: Composition du corpus en allemand scientifique

Les trois principaux sites sont :

— *Deutsches Ärzteblatt*[2], un magazine hebdomadaire médical publié en Allemagne par l'éditeur *Deutscher Ärzte-Verlag* à 370 000 exemplaires.
— Les articles publiés en ligne par la *Deutsche Gesellschaft für Senologie*[3], l'institut de recherche en sénologie de l'Allemagne.
— Les articles publiés en ligne par *Universitäts-Brustzentrum Tübingen*, l'institut de recherche en sénologie de l'Université de Tübingen.

3.2.2 Textes en allemand vulgarisé

Voici maintenant la liste des 15 sites germanophones sur lesquels nous avons collecté les textes en allemand vulgarisé.

Allemand vulgarisé	Nom du site Web	Nombre de mots typographiques	Nombre de textes
	*netdoktor.de*	45 145	80
	mammamia-online.de	39 628	1
	krebsgesellschaft-nrw.de	25 697	3
	brustkrebsdeutschland.de	20 743	16
	onkosupport.de	18 559	4
	mammakarzinom-info.de	17 765	36
	mammo-programm.de	8 881	6
	de.wikipedia.org	7 759	1
	brustkrebs-sprechstunden.de	4 663	1
	ago-online.de	4 335	1
	experten-sprechstunde.de	3 820	1
	brca-netzwerk.de	2 602	8
	krebshilfe.de	2 547	2
	sportaerztebund-	1 286	1
	medical.siemens.com	1 204	1
Total		204 634	162

FIGURE 6: Composition du corpus en allemand vulgarisé

2. https ://www.aerzteblatt.de/
3. https ://www.senologie.org/

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Les deux principaux sites sont :

— *Netdoktor*[4], un site scandinave d'information en matière de santé, de pharmacie et de médecine.
— *Mamma Mia !*[5] le site du magazine bimestriel *Das Brustkrebsmagazin* publié par *GeKo Verlag*.
— *Krebsgesellschaft Nordrhein-Westfalen*[6], un site d'information et de soutien aux malades du cancer.

3.2.3 Échantillon de textes de chacun des deux registres

Voici un extrait du corpus allemand.

	Extrait
Allemand scientifique	*Etwa 5 % aller Mammakarzinome entstehen aufgrund einer erblichen Disposition. Frauen mit Keimbahnmutationen in einem der prädisponierenden Gene haben ein hohes Risiko, im Laufe ihres Lebens an Brustkrebs zu erkranken. Frauen, die ein deutlich erhöhtes familiäres Brustkrebsrisiko haben, sollten einer genetischen Beratung oder einer für das familiäre Mammakarzinom spezialisierten Einheit zugewiesen werden.*
Allemand vulgarisé	*„Krebs" ist der Überbegriff für bösartige Tumoren, die aus entarteten Zellen entstehen. In ihnen ist die Erbinformation, die in den Genen gespeichert ist, verändert. In den allermeisten Fällen werden diese defekten Gene nicht an die Kinder weiter vererbt. Mit wenigen Ausnahmen: Zum Beispiel entstehen zwischen fünf und zehn Prozent aller Brustkrebserkrankungen durch eine familiäre Vorbelastung. Auch bei einigen anderen Krebsarten ist ein kleiner Teil erblich bedingt.*

FIGURE 7: Extrait du corpus allemand

4 Traitement des données

4.1 Prétraitement des fichiers

Après une première collecte de tout ce qui se trouvait déjà en accès libre sur le Web, nous n'avions qu'une centaine de milliers de mots. Nous avons ensuite souscrit des abonnements à des revues spécialisées et avons dû attendre les sorties mensuelles des nouveaux articles. Les articles avaient des formats très différents, et il aurait été plus long de développer des outils de TAL que d'effectuer le traitement manuellement pour chaque article, lorsqu'il arrivait.

Pour un travail à plus grande échelle, un développement à partir d'expressions régulières serait tout à fait adapté. Le niveau de qualité serait vraisemblablement très bon pour traiter les résumés, les références et les biographies, moyen pour les notes et pour l'anonymisation. L'anonymisation de fin d'article est facilement repérable, mais celle à effectuer au sein de l'article est plus compliquée car pas toujours explicite.

4. https ://www.netdoktor.de/
5. https ://mammamia-online.de/
6. http ://www.krebsgesellschaft-nrw.de/

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Unités	Approximation de la qualité attendue
résumés	haute
références	haute
biographies	haute
notes	moyenne
entités nommées	mauvaise
images	haute
balises html	haute

FIGURE 8: La qualité du filtrage par méthodes de TAL

4.2 Traitement des fichiers

Nous avons lu chaque texte intégralement et retiré manuellement de chacun d'entre eux les références bibliographiques, les noms d'auteur, les dates, les publicités, les résumés desdits textes, les biographies et les notes sur les intérêts et influences des auteurs. Les images ont également été retirées, ainsi que toutes les balises HTML. Voici un tableau récapitulatif du nombre de mots en allemand recueillis avant et après le traitement des fichiers. Nous avons obtenu en tout 409 280 « mots » avant traitement [7], qui représentent un peu plus de 3 millions de caractères (soit 2143 pages standard).

	Allemand scientifique	Allemand vulgarisé	Total
Textes	79	162	241
Mots typographiques de la ressource brute	204 646	204 634	409 280
Mots typographiques après traitement	197 187	201 760	398947

FIGURE 9: Le corpus avant et après le formatage

Cependant, si le corpus a été formaté, il n'a volontairement pas été normalisé. D'une part, les paragraphes ont été conservés, et d'autre part la casse n'a pas été modifiée. La conservation de la casse est en fait très importante pour le traitement de l'allemand. En effet, quelle que soit sa position dans la phrase, un substantif commence toujours par une majuscule en allemand. Quel que soit l'objectif des recherches, il est donc important de conserver la casse.

	Anglais	Français	Allemand
Scientifique	198 244 (49%)	373 127 (49%)	197 187 (49%)
Vulgarisé	218 336 (49%)	373 127 (49%)	201 760 (49%)
Total	416 580	451 684	398 947

FIGURE 10: Composition et taille du corpus trilingue (d'après Delpech, 2013)

7. Il s'agit ici de mots typographiques comptés par l'utilitaire de Word. Parmi eux, 398 947 seront utilisés pour l'aligneur d'unités polylexicales.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Dans le cas de l'alignement d'unités polylexicales, les textes sont ensuite traités à l'aide de l'analyseur morphosyntaxique XELDA, en enchaînant une segmentation, une lemmatisation et un étiquetage.

Voici la composition et la taille du corpus total trilingue et transregistre pour l'aligneur après ajout de notre contribution indiquant le nombre de mots typographiques retenus *in fine*.

5 Utilisation du corpus allemand pour l'analyseur morphologique AMALD de Jean-Philippe Guilbaud

Actuellement, une nouvelle expérimentation est en cours. À des fins de test et de débogage, nous avons exploité notre corpus pour augmenter la couverture d'un analyseur morphologique développé par Jean-Philippe Guilbaud et présenté ci-dessous. En particulier, il a été possible d'améliorer la qualité de rattachement des particules séparables.

5.1 L'analyseur morphologique AMALD

Il s'agit d'un analyseur morphologique de l'allemand de grande taille couvrant plus de 102 000 lemmes et près de 500 000 formes fléchies simples, capable de traiter les mots composés et les verbes à particules séparables, même lorsqu'il y a de nombreux mots entre le verbe et sa particule.

5.1.1 Contexte

Jean-Philippe Guilbaud s'est intéressé à l'analyse morphologique de l'allemand afin d'effectuer de la recherche d'information translingue. Ne trouvant pas d'analyseur morphologique libre de droits et de bonne qualité, il a entrepris d'en construire un, en partant du prototype construit pour sa thèse (Guilbaud & al., 2013).

Pour qu'une analyse morphologique soit complète, elle doit produire les lemmes, les parties du discours, les autres variables grammaticales (genre, nombre, cas, mode, temps, personne, degré, etc.). Si l'on veut pouvoir reconnaître des équivalences paraphrastiques (par exemple : « phase transitoire » et « phase de transition ») et également traiter la néologie dérivationnelle (en utilisant des fonctions lexico-sémantiques à la Mel'čuk), l'analyseur doit également produire une unité lexicale (UL).

Ainsi pour « phase transitoire » et « phase de transition », l'UL commune à « transition » et « transitoire » est transiter-V, tête de la famille dérivationnelle.

5.1.2 Outils utilisés

Cet analyseur est construit à l'aide de trois langages spécialisés pour la programmation linguistique : ATEF, EXPANS et ROBRA, utilisés pour écrire trois phases successives de traitement.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Phase ATEF : traitement des mots typographiques.

Le linguiste commence par déclarer des attributs, et pour chaque attribut, des combinaisons de valeurs (appelées « formats »). Certaines combinaisons de valeurs sont souvent utilisées, comme des classes morphologiques ou linguistiques. Un article d'un dictionnaire d'unités (poly)lexicales contient deux combinaisons de valeurs, l'une syntaxique, l'autre morphologique et contient également une UL. Ainsi, en ATEF, on produit des lemmes à partir des formes fléchies, et on leur attache les informations morphosyntaxiques.

Phase EXPANS : traitement des lemmes composés (verbe + particule).

Un article d'un dictionnaire d'expansion lexicale (EXPANS) permet de transformer un nœud de l'arbre en entrée en un sous-arbre de l'arbre produit en sortie. Ainsi, en EXPANS, on produit les UL classiques (familles dérivationnelles), et on leur attache les informations syntaxico-sémantiques, ainsi que, dans le cas d'un verbe de base (sans particule séparable), les informations de tous les verbes composés partageant ce verbe de base.

Phase ROBRA : identification et regroupement des formes verbales composées.

Les règles transformationnelles de ROBRA permettent de reconnaître des schémas de sous-arbres et de transformer leurs occurrences. Ainsi, en ROBRA, on écrit des règles qui ont accès au contexte de toute la phrase (et même de tout le texte traité comme une unité de traduction). Il est ainsi possible de regrouper les mots ou expressions composés non connexes.

5.1.3 Principes linguistiques de l'analyseur morphologique AMALD

Dans AMALD, chaque mot typographique d'un texte est appelé occurrence. Une occurrence est donc constituée d'une suite ordonnée d'affixes ou d'infixes (morphes grammaticaux) et d'une ou, dans le cas des mots composés, de plusieurs bases lexicales (ou radicaux).

Le moteur d'ATEF cherche à reconnaître les occurrences en consultant des dictionnaires qui contiennent toutes les bases lexicales et tous les morphes grammaticaux de la langue, et en appliquant (de façon non déterministe multiple) des « règles » qui sont des transitions d'un automate sous-jacent (comme dans NooJ). Ces règles « ouvrent » et « ferment » des dictionnaires, et combinent les valeurs « courantes » des attributs avec celles provenant des dictionnaires.

Les affixes sont des préfixes, des suffixes de dérivation ou des désinences de flexion ; les infixes sont des morphes qui relient l'une à l'autre les bases lexicales d'un mot composé (ex. *Handlungsfreiheit*, liberté d'action).

Un lemme a un ou plusieurs radicaux. Chaque radical relève d'un paradigme flexionnel (morphème).

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

L'extension du paradigme est la liste des désinences possibles pour le radical. Chaque désinence est un morphe qui renvoie à un ou plusieurs morphèmes, selon la stratégie d'analyse choisie.

Exemple :	lemme « *singen* », chanter
Radicaux :	*sing-, sang-, säng-, gesungen-*
Paradigmes flexionnels :	FCPPA (*gesungen-*), WGAEB (*säng-*), WGAB (*sang-*), WSING (*sing-*)
Les désinences de WGAEB et leurs morphèmes associés sont :	*-e* (1 WAERE), *-en* (1 WAEREN), *-est* (1 WAERET), *-st* (1 WAERST)
Morphème 1 WAERE :	1ère ou 3ème personne du singulier du subjonctif II ;
Morphème 1 WAEREN :	1ère ou 3ème personne du pluriel du subjonctif II ;
Morphème 1 WAERET :	2ème personne du pluriel du subjonctif II ;
Morphème 1 WAERST :	2ème personne du singulier du subjonctif II ;

FIGURE 11: Le paradigme flexionnel du verbe *singen* (d'après Guilbaud *et al.*, 2013)

5.2 Cas concret d'amélioration de rattachement de particule

Notre corpus a permis d'améliorer l'analyseur morphologique AMALD grâce, entre autres, à sa forme, qui respectait la casse d'origine. Par exemple, voici une phrase que l'analyseur n'était auparavant pas capable d'analyser.

> *Passen Sie gegebenenfalls Ihre Übungen so an, dass der Arm \ der betroffenen Seite nicht zu stark belastet wird.* [8]

Le verbe à particule séparable, *an/passen* (adapter) qui est ici en début de phrase, commence par une majuscule. Avant le travail présenté ici, l'analyseur morphologique AMALD n'était pas été capable de traiter le cas de ce verbe. En effet, le candidat à avoir une particule séparable est lui-même ambigu entre verbe et nom, à cause de la majuscule, due à sa position en début de phrase.

De ce fait, le mot *Passen* doit avoir ici 3 solutions :

(1) un nom (*die Passe*),
(2) un verbe conjugué (*passen*), et
(3) un verbe substantivé (*das Passen*).

8. Adaptez le cas échéant vos exercices, de façon à ce que le bras du côté touché ne soit pas trop sollicité.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

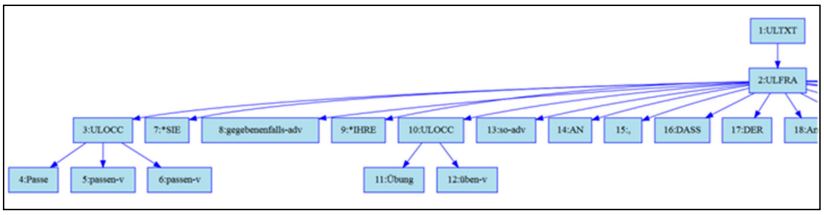


FIGURE 12: Un extrait de l'analyse morphologique de *Passe ... an*

Dans la version précédente de l'analyseur morphologique, la règle de transformation d'arbre utilisée pour « raccrocher » la particule séparable *an* au verbe conjugué (solution 2) excluait la possibilité d'avoir une ambiguïté avec un nom différent du verbe substantivé. C'est pourquoi, dans cet exemple, la particule n'avait pas été regroupée avec le verbe simple : on aurait dû avoir le lemme *an/passen* sur la deuxième solution.

Cet exemple a permis d'améliorer la résolution d'ambiguïté, car il faut raccrocher la particule uniquement s'il n'y a pas d'ambiguïté possible. C'est désormais le cas : la particule est raccrochée et la possibilité de traitement des autres solutions est conservée, c'est-à-dire qu'on conserve le cas échéant une ambiguïté externe à un lemme.

6 Conclusion

Dans l'industrie de la communication multilingue, notamment dans le cadre de la recherche d'informations, il existe un besoin en ressources transregistres. Ces ressources sont très difficiles à trouver lorsqu'il s'agit de l'allemand. Nous avons constitué un corpus transregistre dans le domaine médical d'environ 400 000 mots (3 millions de caractères).

Les données ont été recueillies manuellement à partir de sites Web de vulgarisation et à partir de revues scientifiques. Les données ont subi un prétraitement également manuel, car il s'agit d'un petit corpus avec des formats très variés. Cependant, la constitution d'un corpus similaire de taille plus importante (plus de 1 millions de mots) justifierait le développement d'outils de TAL.

Ce corpus a été d'abord utilisé pour développer un aligneur d'unités polylexicales, et ensuite pour permettre d'améliorer la couverture d'un analyseur morphologique. Nous poursuivons actuellement les travaux sur cet analyseur en lui soumettant l'intégralité du corpus, et en l'améliorant au fur et à mesure, en visant la production d'un corpus annoté de très grande qualité.

Bibliographie

Delpech, Estelle, *Traduction assistée par ordinateur et corpus comparables : contributions à la traduction compositionnelle,* thèse de doctorat, Université de Nantes, 2013.

Delpech, Estelle ; Daille, Béatrice ; Morin, Emmanuel et Lemaire, Claire, *Extraction of domain-specific bilingual lexicons from comparable corpora : compositional translation and ranking,* COLING 2012, 8–12 décembre, Mumbai, Inde, 2012.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus,* Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Guilbaud, Jean-Philippe ; Boitet, Christian et Berment Vincent, *Un analyseur morphologique étendu de l'allemand traitant les formes verbales à particule séparée*, TALN 2013, 17–21 juin, Les Sables d'Olonne, France, 2013.

Hazem, Amir et Morin, Emmanuel, *Leveraging Meta-Embeddings for Bilingual Lexicon Extraction from-Specialized Comparable Corpora,* Proceedings of the 27th International Conference on Computational Linguistics, p. 937–949, Santa Fe, New Mexico, USA, August 20-26, 2018.

Lemaire, Claire, *Traductologie et traduction outillée : du traducteur spécialisé professionnel à l'expert métier en entreprise,* Thèse de doctorat, Université Grenoble Alpes, 2017.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454