Un analyseur morphologique étendu de l'allemand traitant les formes verbales à particule séparée

Jean-Philippe Guilbaud[1] Christian Boitet[2] Vincent Berment[2]
(1) CNRS, LIG-campus, 38041 Grenoble Cedex 09
(2) UJF, Université de Grenoble, LIG-campus, 38041 Grenoble Cedex 09
{Jean-Philippe.Guilbaud,Christian.Boitet,Vincent.Berment}@imag.fr

RÉSUMÉ

Nous décrivons l'organisation et l'état courant de l'analyseur morphologique de l'allemand AMALD de grande taille couvrant (près de 103000 lemmes et 500000 formes fléchies simples, en croissance) développé dans le cadre du projet ANR-Émergence Traouiero. C'est le premier lemmatiseur de l'allemand capable de traiter non seulement les mots simples et les mots composés, mais aussi les verbes à particules séparables quand elles sont séparées, même par un grand nombre de mots (ex : *Hier schlagen wir eine neue Methode für die morphologische Analyse vor*).

ABSTRACT

An extended morphological analyzer of German handling verbal forms with separated separable particles

We describe the organisation and the current state of the large-scale (nearly 103000 lemmas and 500000 simple inflected forms, growing) morphological analyzer AMALD developed in the framework of the ANR-Émergence Traouiero project. It is the first lemmatizer of German able to handle not only simple and compound words, but also verbs with separable particles when they are separated, even by many words (e.g. *Hier schlagen wir eine neue Methode für die morphologische Analyse vor.*).

MOTS-CLÉS : analyse morphologique, lemmatisation, allemand, verbes à particule séparable
KEYWORDS : morphological analysis, lemmatization, German, verbs with separable particles.

1 Introduction

En 2008, dans le cadre du projet ANR OMNIA, nous nous sommes réintéressés à l'analyse morphologique (AM) de l'allemand, pour pouvoir faire de la RI (recherche d'information) translingue sur des collections d'images (comme FlickR, Belga News, Picassa ou PanImages) accompagnées chacune d'un petit texte compagnon écrit de façon spontanée dans la langue de l'auteur. Constatant qu'il n'y avait pas d'AM de l'allemand de bonne qualité, libre de droits et assez couvrante, le premier auteur a alors entrepris d'en construire une, en partant du prototype construit pour sa thèse. Le besoin d'une telle AM est apparent dans de nombreuses applications qui exigent plus que de la lemmatisation ou de l'étiquetage morphosyntaxique, et l'allemand est une langue particulièrement importante. De plus, sa morphologie est particulièrement intéressante : système de flexions et de dérivations assez riche et fort ambigu, constructions compositionnelles non bornées, et abondance de formes verbales discontinues (comme *er kommt nach… an,* pour *il arrive après…)*.

Nous discutons d'abord des résultats qu'on attend d'une AM, et des méthodes qu'on peut utiliser pour les produire, sachant qu'il n'y a pas consensus sur ces deux points. Nous faisons ensuite le point sur les AM de l'allemand existantes. Nous présentons ensuite très brièvement les LSPL[1] utilisés pour construire les trois phases de notre nouvelle AM de l'allemand (AMALD), puis décrivons les aspects et les composants principaux de cet analyseur, avant de l'évaluer et de conclure.

[1] LSPL = Langage Spécialisé pour la Programmation Linguistique.

2 Buts d'une analyse morphologique et méthodes possibles

Un *lemme* est une *forme de citation* dans les dictionnaires, représentant un ensemble de *formes* constituant sa *flexion.* Pour les verbes, c'est l'infinitif dans beaucoup de langues, mais c'est la 3° personne du singulier de la conjugaison subjective en hongrois. La *lemmatisation* est l'opération qui, à chaque *occurrence* (mot typographique simple ou composé) d'un texte, associe le ou les lemmes possibles, éventuellement en utilisant le contexte. Cela suffit pour faire de l'annotation sémantique, mais pas pour traduire ou faire de la correction grammaticale. *L'étiquetage morphosyntaxique* associe à chaque mot (ou partie de mot composé) une ou plusieurs *parties du discours* (POS) dites aussi *catégories morphosyntaxiques* (nom, verbe…), choisies dans un ensemble plus ou moins riche et éventuellement structuré. Une *analyse morphologique* (AM) *complète* doit produire non seulement les lemmes et les parties du discours, mais aussi les autres *variables grammaticales* (genre, nombre, cas, mode, temps, personne, degré, etc.), et souvent, en particulier quand on veut pouvoir traiter la néologie dérivationnelle[2], ou reconnaître des équivalences paraphrastiques[3], une *unité lexicale* (UL) plus abstraite, la *famille dérivationnelle.* Elle doit aussi pouvoir produire toutes les solutions possibles, par un *treillis de possibilités,* comme le font NooJ et Chasen (NAIST, Nara), ou par un *arbre avec disjonctions,* comme le fait notre outil ATEF, de façon à pouvoir représenter les ambiguïtés et à laisser les traitements suivants tenter de les réduire. Un système de *TAO experte* a besoin d'une AM complète.

La toute première AM de l'allemand par règles semble avoir été écrite par Klaus Brockhaus à Heidelberg (Brockhaus 1971, 1976), en utilisant des grammaires hors-contexte. Les AM du système de TA METAL commandité par Siemens à J. Slocum (Austin, Texas) en 1981 utilisent le même LSPL basé sur les grammaires hors-contexte augmentées que celui utilisé pour l'analyse syntaxique (multiple). Le traitement de l'allemand en METAL a progressé depuis 30 ans[4], mais on n'en connait pas les détails (système propriétaire). Le modèle hors-contexte a aussi été utilisé pour l'AM du japonais par Tomita à CMU (GLR, 1986). L'utilisation du modèle de TEF (transducteur d'états finis) remonte à ATEF, créé à Grenoble par Jacques Chauché en 1971-72. D'autres LSPL fondés sur les TEF ont suivi, comme INTEX et son clone UNITEX, NooJ, Kimmo, XFST, etc. Mais il faut des modèles au moins hors-contexte pour analyser de nombreuses langues. C'est le cas de l'allemand, à cause de la composition récursive des mots composés. Pour les langues où le pluriel se marque par la répétition, comme les langues malaises, le modèle hors-contexte n'est pas non plus assez puissant[5]. Notons aussi que toutes les AM cherchent à réduire les ambiguïtés en utilisant le contexte, ce qui les fait inévitablement "déborder vers la syntaxe". Il s'agit bien sûr de syntaxe très "surfacique", mais ça en est. Désambiguïser des parties du discours sur la base de fréquences de n-grammes est bien de la syntaxe[6], même si le résultat est morphologique.

Demander à une AM de reconnaître les verbes à particules séparables même quand leur particule est séparée et à longue distance est bien un des buts souhaitables d'une AM. Il ne s'agit en effet pas de produire en résultat des constituants, même élémentaires (chunks), ni des relations de dépendance syntaxique. Il s'agit seulement de reconnaître que deux mots typographiques distants ne forment qu'un lemme, et de dire lequel. En français, cela revient à identifier *ne… pas* comme un tout. On pourrait même dire que c'est nécessaire en allemand, car la syntaxe et la sémantique des composés Verbe + Particule ne sont la plupart

[2] en utilisant des fonctions lexico-sémantiques à la Mel'tchuk, dites *dérivations productives.*

[3] ex. : phase transitoire ≈ phase de transition : l'UL est transiter-V, *tête* de la famille dérivationnelle.

[4] METAL a connu une histoire compliquée et est maintenant l'outil de TA de LucySoftware (Allemagne).

[5] Le langage des mots doublés sur un vocabulaire Σ, D = { ww | w ∈ Σ* }, n'est pas hors-contexte… et donc pas non plus les langages de programmation exigeant que les variables soient déclarés avant d'être utilisées. XML ne l'est pas non plus.

[6] Syntaxe = « [la] mise ensemble ». Ici c'est seulement de la parataxe (coordination), pas de l'hypotaxe (subordination).

du temps pas compositionnelles. Simplement, jusqu'ici, personne n'avait apparemment pensé qu'on pourrait le faire dans une application autonome. Comment le faire ? Assez simplement dans notre cas : on enchaîne trois *phases*, la première en ATEF[7], la seconde en EXPANS (dictionnaires transformant des arbres par expansion de chaque nœud), qui produit pour tout verbe simple un arbre prédisant tous ses composés à particule possibles, et la troisième, en ROBRA (un outil très puissant de grammaires transformationnelles), qui utilise le contexte pour déterminer si une occurrence est une particule verbale (il y a souvent ambiguïté), et si oui avec quelle occurrence verbale simple elle peut ou doit être regroupée.

En 2008, nous n'avions pas trouvé d'AM de l'allemand utilisable dans le cadre du projet OMNIA et avions donc entrepris de développer la nôtre, à partir de la maquette de (Guilbaud 1981). Au moment d'intensifier cet effort et de viser le passage à l'échelle, dans le cadre du projet ANR Traouiero, nous avons réexaminé la situation. Elle n'a apparemment pas progressé[8]. On trouve d'abord des références déjà anciennes à *Morphy* (Lezius & al. 1998, 2000 ; Rapp & Lezius 2001). Morphy est installable sous Windows, et il y a sur le Web un fichier (http://www.danielnaber.de/morphologie/morphy-export-20110722.tar.gz) de 368175 formes (pas 431000 comme écrit sur la page Web), soit environ 76000 lemmes, qui donne les lemmes résultant de l'analyse des formes (et aucun autre attribut). Malheureusement, près de 90% des lemmes proposés sont faux. Voici un exemple.

FORME	LEMME	FORME	LEMME	
zufriedengestelltem	zufriedengestellt	zufriedengestellter	zufriedengestellt	
zufriedengestellten	zufriedengestellt	zufriedengestelltestem	zufriedengestellt	
Zufriedengestellten	Zufriedengestellte	zufriedengestelltesten	zufriedengestellt	
zufriedengestellterem	zufriedengestellt	zufriedengestelltester	zufriedengestellt	
zufriedengestellteren	zufriedengestellt	zufriedengestelltestes	zufriedengestellt	
Zufriedengestellteren	Zufriedengestelltere	zufriedengestelltester	zufriedengestellt	
zufriedengestellterer	zufriedengestellt	zufriedengestelltes	zufriedengestellt	
zufriedengestellteres	zufriedengestellt	zufriedengestellte	zufriedengestellt	
zufriedengestelltere	zufriedengestellt	zufriedengestellt	zufriedenstellen	-> seul résultat correct

Le lemme de toutes ces formes devrait être le verbe *"zufriedenstellen"* (satisfaire), et certainement pas le participe passé passif (*"zufriedengestellt"* et sa flexion). Il faudrait aussi indiquer que la "particule" est ici *"zufrieden-"* et pas "zu-" dans la notation du lemme, faute de quoi on pourrait reconnaître la forme impossible ***zugefriedenstellt*. Avec *"Zufriedengestellte"*,** c'est comme si on disait en français que *"Comprise"* est un nom[9] et que c'est le lemme de "comprises" (alors que c'est "comprendre-V" ou "compris-Adj"). Enfin, *"Zufriedengestelltere"* donné comme lemme pour la forme *"Zufriedengestellteren"*, est doublement impossible, car (1) ça veut dire "plus satisfait", au féminin singulier, ou, en forme "forte" au nominatif et à l'accusatif pluriel aux 3 genres, alors qu'un lemme adjectival est au nominatif masculin singulier, et surtout, (2) la majuscule à l'initiale est impossible dans la dénotation d'un lemme non nominal. Sans doute ce mot est-il apparu dans le corpus en début de phrase. Il y a une AM écrite en Kimmo (http://www2.lingsoft.fi/doc/gercg/NODALIDA-poster.html). Elle aussi n'est vraiment pas satisfaisante. De plus, il est dit qu'elle traite bien les mots composés, mais c'est inexact. Voici le résultat pour *"*wortformerkennung"* = "reconnaissance de la forme d'un mot". L'étoile '*' est une bascule minuscule↔majuscule, comme dans notre transcription.

"<*wortformerkennung>"			
"*wort#form#er|kenn~ung"	S FEM SG NOM	"*wort#form~er#kenn~ung"	S FEM SG NOM
"*wort#form#er|kenn~ung"	S FEM SG AKK	"*wort#form~er#kenn~ung"	S FEM SG AKK
"*wort#form#er|kenn~ung"	S FEM SG DAT	"*wort#form~er#kenn~ung"	S FEM SG DAT
"*wort#form#er|kenn~ung"	S FEM SG GEN	"*wort#form~er#kenn~ung"	S FEM SG GEN

[7] ATEF traite une suite d'occurrences, avec un contexte de 4 occurrences avant et 1 après, et sa sortie est un arbre décoré.
[8] Sauf peut-être avec NooJ, mais nous n'avons pas pu avoir de détails.
[9] En allemand, les noms communs ont une majuscule à l'initiale.

Cette AM ne produit rien sur les parties du mot composé, même pas le lemme du dernier morceau (c'est *"Erkennung"* = *reconnaissance* ou *"erkennen"* = *reconnaître*, selon qu'on veut comme UL un lemme ou une famille dérivationnelle). La couverture annoncée de 60000 formes parait très petite (pas plus de 20000 lemmes)[10]. Il y a aussi SMOR, proposé par l'université de Stuttgart et construit sur `Stuttgart fst lib` (www.ims.uni-stuttgart.de/projekte/gramotron/PAPERS/LREC04/smor.pdf). Mais sa limitation à un seul préfixe est rédhibitoire : il faut en permettre plusieurs pour analyser correctement l'allemand. De plus, son traitement des mots composés est présenté par ses auteurs comme non satisfaisant.

Si l'on regarde l'existant « académique », on ne trouve donc rien qui réponde à nos attentes, tant au niveau des résultats produits que des résultats productibles. Il existe certainement de bonnes AM de l'allemand dans les systèmes commerciaux de TA, mais ils ne sont pas utilisables librement comme des modules séparés. On ne peut estimer leur qualité qu'indirectement, par traduction, et seulement sur les mots composés. Ainsi, ProMT traduit *"Hauptbahnhofgepäckaufbewahrung"* par *"Central station checkroom"*, et pour ça il a fallu ne pas segmenter en les plus petits atomes, soit *"Haupt-bahn-hof-gepäck-auf-bewahrung"* (sans séparer les préfixes). Il a fallu considérer des morceaux plus longs, comme *"bahnhof"*, sans considérer le possible *"Hauptbahn"* (*voie principale*), et *"gepäckaufbewahrung"* (*consigne à bagages*). *"Bahnhofgepäck"* serait aussi possible… Pour bien traiter les mots composés allemands, il faut au moins pouvoir produire plusieurs solutions, et éliminer des sous-découpages, et/ou calculer des scores pour ne produire que les solutions assez bien notées.

3 Brève présentation des outils utilisés

Faute de place, il n'est pas possible de décrire précisément le fonctionnement des LSPL utilisés (ATEF, EXPANS, ROBRA). Décrivons seulement ce que sont les *composants* des *linguiciels* qu'ils permettent d'écrire. Les *variables* sont les attributs déclarés par le linguiste. Une combinaison de valeurs des variables déclarées est une *décoration*. Les *formats* sont des décorations constantes, souvent utilisées comme des *classes* morphologiques ou syntaxiques. Les *tournures* sont des suites d'occurrences, le seul séparateur d'occurrences étant le blanc. Un *article* d'un dictionnaire de bases ou de tournures contient 2 formats dits morphologique (FTM) et syntaxique (FTS), ces termes n'étant que mnémoniques), et une UL. Un article d'un dictionnaire d'affixes ne contient pas d'UL. 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. Enfin, les règles transformationnelles de ROBRA permettent de reconnaître des schémas de sous-arbres et de transformer leurs occurrences.

L'AM de l'allemand de départ, écrite en ATEF seul, était parfaite à 100% sur sa couverture. AMALD utilise 3 phases d'Ariane-G5 abrégées en AM (AM1/atef), AX (AM2/expans), et AS (AM3/robra). AM produit des lemmes à partir des formes fléchies, et leur attache les informations morphosyntaxiques. AX produit les UL classiques (familles dérivationnelles), et leur attache les informations syntaxo-sémantiques. En AS, 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), et on peut ainsi regrouper les mots ou expressions composés non connexes.

4 Principes linguistiques de l'analyseur morphologique AMALD

Nous convenons que toute *occurrence* (mot typographique) d'un texte est une *forme fléchie* d'un *lemme* (ce qui impose l'existence d'une désinence nulle). Une occurrence est constituée

[10] Mais il y a peut-être une erreur de terminologie dans la description, puisque par ailleurs il est dit qu'il y a "tous les mots du Collins" (das komplette Sprachmaterial des Deutsch-Englischen Wörterbuchs von Collins (The Collins German Dictionary, Neubearbeitung 1991, Copyright HarperCollins Publishers)). Or le Collins a au moins 50000 à 70000 entrées.

d'une suite ordonnée d'affixes ou infixes (*morphes*[11] grammaticaux) et d'une ou, dans le cas des mots composés, de plusieurs *bases lexicales* (ou *radicaux*). Le *moteur* d'ATEF cherche à les reconnaître en consultant des dictionnaires qui contiennent toutes les bases lexicales et tous les morphes grammaticaux de la langue. La consultation des dictionnaires, régie par une grammaire (compilée vers un transducteur fini étendu), permet de découper les mots du texte et de les interpréter en leur affectant des valeurs de classe, cas, genre, nombre, personne, etc. 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*. S'il en a plusieurs, il s'agit d'allomorphes qui sont en variation libre ou en distribution complémentaire. Chaque radical relève d'un *paradigme flexionnel* (morphème). *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* (1WAERE), *-en* (1WAEREN), *-est* (1WAERST), *-et* (1WAERET), *-st* (1WAERST)

Morphème 1WAERE :	1ère ou 3ème personne du singulier du subjonctif II ;
Morphème 1WAEREN :	1ère ou 3ème personne du pluriel du subjonctif II ;
Morphème 1WAERET :	2ème personne du pluriel du subjonctif II ;
Morphème 1WAERST :	2ème personne du singulier du subjonctif II.

Dans la théorie à la base de ce système, tout lemme appartient à une famille dérivationnelle appelée *UL (unité lexicale),* notée le plus souvent en combinant la chaîne du lemme *source* de cette famille et sa catégorie. Un lemme est ainsi caractérisé par une valeur d'UL, sa classe morphosyntaxique (nom, verbe, etc.), et aussi par une valeur de dérivation s'il est dérivé d'un autre ou considéré comme tel. Une des caractéristiques essentielles de l'allemand est de pouvoir créer très facilement de nouveaux mots par agglutination de lemmes simples ou déjà composés. Chaque locuteur peut ainsi créer librement de nouveaux mots. Dans un mot composé, le déterminant précède toujours le déterminé (ordre centripète des éléments de signification). Il peut y avoir parfois un infixe spécial ('s', 'e', 'en') entre les composants. Dans notre analyseur, un mot composé, quand il n'est pas trouvé dans les dictionnaires comme tel, parce qu'il n'est pas considéré comme un terme ou qu'il n'a pas encore été indexé[12], est découpé en ses éléments constitutifs simples trouvés dans les dictionnaires. AMALD travaille sur une *transcription minimale*[13]*,* utilisant seulement les lettres majuscules, et des *séquences spéciales* pour la mise en majuscule ('*' pour la lettre suivante, '**' pour la suite du mot jusqu'à un autre '**') et pour les diacritiques (!1 : ´, !2 : `, !3 : ^, !4 : ¨, !5 : cédille).

En allemand, comme en français, les verbes peuvent avoir des préfixes inséparables (ex. allemand : *empfangen ≠ fangen (recevoir ≠ attraper)* ; français : *détourner ≠ tourner*). Mais, à la différence du français, ils peuvent aussi avoir des particules séparables. Ces préfixes, dans certains contextes, peuvent être soit séparés de la base par des infixes tout en restant agglutinés, soit être déplacés en fin de proposition ou de phrase (ex: *auffangen, aufgefangen, aufzufangen, fängt....auf, [r]attraper).* On ne peut alors plus identifier correctement en AM le composé verbal, et c'est la phase (pré-syntaxique) AS (AM3/robra) ultérieure qui le fait.

[11] Un morphe est une chaîne de caractères pertinente pour l'analyse de la langue considérée.
[12] Les premiers chercheurs en TA ont utilisé « indexer » pour signifier « inclure dans un dictionnaire d'un module de TALN ».
[13] Cela permet de diviser par 3 ou 4 la taille des dictionnaires de bases et préfixes, et de traiter le *Umlaut* en tant que tel. Un transcripteur permet de partir d'un texte écrit normalement, de façon transparente. Des transcriptions similaires, réversibles et prononçables, ont été définies et utilisées pour le russe, l'arabe, le chinois, le thaï, le vietnamien…, et même pour le japonais.

Exemple de résultat :

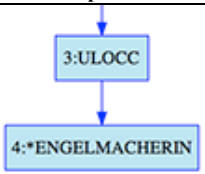
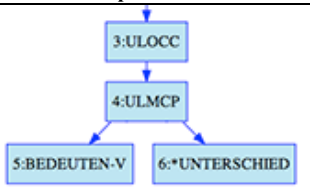
mot composé indexé	mot composé non indexé
	
4 '***ENGELMACHERIN**': UL('*ENGELMACHERIN'), KMS(NM), SUBN(IP), PSG(3), CSFB(NOM,ACC,DAT,GEN), TYPO(MJ1), GNR(F) (faiseuse d'anges)	5 '***BEDEUTUNGS**': UL('BEDEUTEN-V'), KMS(NM), SUBN(IP), PSG(3), VERB(GE,UN), CSFB(GEN), TYPO(MJ1), GNR(F), DRV (VN2), SUBV (HAB), VAL1(ACC), VAL2A(ACC), VAL2B(FUER) 6 '**UNTERSCHIED**': UL('*UNTERSCHIED'), KMS(NM), SUBN(IP), PSG(3), CSMB(NOM,ACC,DAT), TYPO(MJ1), GNR(M) (différence de signification)

FIGURE 1 : sous-arbres produits pour des mots composés

La particule séparable est toujours agglutinée au verbe lorsqu'il est gouverneur d'une subordonnée, et seulement à l'infinitif, au participe passé ou au participe présent dans les autres cas. Nous traitons toutes les formes verbales agglutinées en leur affectant immédiatement l'UL définitive, celle du composé. Nous indexons donc dans les dictionnaires de bases lexicales toutes les bases préfixées.

Exemple *zuträgst* :

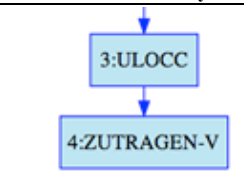
Dictionnaire				Résultat d'analyse
base lexicale	**FTM**	**FTS**	**UL**	
ZUGETRAGEN	==FCPPA	(HA3	*,ZUTRAGEN-V).*	
ZUTRA!4G	==YGRAEB	(HA3	*,ZUTRAGEN-V).*	
ZUTRAG	==YGRAB	(HA3	*,ZUTRAGEN-V).*	
ZUTRU!4G	==YGAEB	(HA3	*,ZUTRAGEN-V).*	
ZUTRUG	==YGAB	(HA3	*,ZUTRAGEN-V).*	
ZUZUTRAG	==FCINFZU	(HA3	*,ZUTRAGEN-V).*	

FIGURE 2 : sous-arbre pour un verbe à particule séparable à formes préfixées indexées dans le dictionnaire

Dans les exemples ci-dessus et ci-dessous, les formes préfixées du lemme *zutragen* (*se faire, arriver*) sont indexées, mais pas celles du lemme *emportragen* (*élever, porter en haut*), pour lesquelles l'AM fabrique donc un mot composé.

Exemple *emporträgst* :

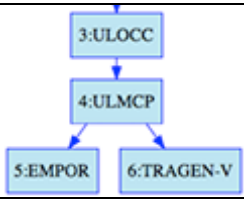
Dictionnaire				Résultat d'analyse
base lexicale	**FTM**	**FTS**	**UL**	
EMPOR	==PARTSEP	(VID	*,EMPOR).*	
EMPOR	==FERD7E	(BNIP	*,*EMPORE).*	
TRA!4G	==WGRAEB	(HABA	*,TRAGEN-V).*	
TRAG	==WGRAB	(HABA	*,TRAGEN-V).*	
TRU!4G	==WGAEB	(HABA	*,TRAGEN-V).*	
TRUG	==WGAB	(HABA	*,TRAGEN-V).*	

FIGURE 3 : sous-arbre pour un verbe à particule séparable à formes préfixées non indexées dans le dictionnaire

Lorsque la particule n'est pas agglutinée au verbe (par exemple *"auf"* du verbe "*aufstehen*" dans la phrase *"Er steht jeden Morgen um fünf Uhr auf"*), une règle (écrite en AS = AM3/robra) permet d'aller la chercher en fin de proposition, comme dernier mot de celle-ci, avant une

éventuelle subordonnée, ou bien devant un syntagme généralement introduit par les conjonctions *"wie"* ou *"als"* ou encore par une préposition.

Pour la regrouper avec le verbe simple, on passe d'abord par un dictionnaire (écrit en AX = AM2/expans), dans lequel on accumule les UL des verbes simples qui acceptent une particule séparable. À chacune de ces UL, le dictionnaire associe, en partie droite, un sous-arbre ayant autant de feuilles qu'il y a de combinaisons possibles de type *particule + verbe simple.* Chaque feuille a dans sa décoration la valeur d'UL qui correspond à la combinaison, et une valeur de variable codant la particule. En AS, l'exécution d'une règle trouvant une particule séparable met le sous-arbre correspondant à la place du verbe simple et efface les nœuds correspondant aux autres particules candidates.

Exemple d'entrée de dictionnaire EXPANS (en AX = AM2/expans) :

```
'TRAGEN-V' ==/0(01,02,03,05,06,07,08,09)/
       01 : 'ABTRAGEN-V', +AB; 02 : 'AUFTRAGEN-V', +AUF;
       03 : 'AUSTRAGEN-V', +AUS; 04 : 'EINTRAGEN-V', +EIN;
       05 : 'EMPORTRAGEN-V', +EMPOR; 06 : 'NACHTRAGEN-V', +NACH;
       07 : 'VORTRAGEN-V', +VOR; 08 : 'ZUTRAGEN-V', +ZU;
       09 : 'ZUSAMMENTRAGEN-V', +ZUSAMMEN.
```

FIGURE 4 : entrée d'un dictionnaire EXPANS (en AX) pour le verbe *'tragen' (porter)* et ses verbes composés

Exemple de règle ROBRA (en AS = AM3/robra) :

```
R1PSEP: (FRA,&NIV=1) FRA(VC($L1,PSEP1(NEWVC),$L2),$L,PSEP2,*,PONC,*)
    / FRA: $ULFRA; VC: $VCONJ; PONC: $PONC; PSEP1: $PSEP; PSEP2: $PSEP
    / $IDUL(PSEP1,PSEP2) ** Particule prédite = trouvée: garder le composé.
 == FRA(VC,$L,PONC)  /*<--PSEP1,PSEP2,NEWVC/ VC:NEWVC. ** Effacer le reste.
```

FIGURE 5 : règle ROBRA (en AS) pour la recherche de particule et l'association au verbe

5 Couverture et qualité

AMALD a été portée sur Héloïse (version en C/C++ des LSPL d'Ariane-G5 et moniteur Web, créés par Vincent Berment). Elle comprenait (au 17/5/2013) 149155 lignes de composants linguiciels (variables, formats, dictionnaires, grammaires), avec 102243 unités lexicales (des lemmes dans ce système), dont 11147 verbes, 85038 noms, 5368 adjectifs, 156 mots-outils, 235 particules séparables, et 29 tournures figées connexes. Cela correspond à environ 480000 formes fléchies simples[14].

Le traitement des mots simples et composés est presque parfait, par rapport aux objectifs classiques[15]. Notre nouvel objectif, au niveau linguistique, est maintenant de réaliser une analyse syntaxique interne des mots composés, en ajoutant un traitement (en ROBRA) sur les sous-arbres de racine 'ULMCP'. On atteindrait alors la limite d'une AM de l'allemand.

Le traitement des formes à particule séparée des verbes à particules séparables est correct sur les nombreux exemples que nous avons traités, même quand la distance entre le verbe et la particule est grande, et même si la particule est homographe d'une préposition et si la phrase comporte une ou des occurrences de cette préposition, avant et/ou après la particule.

Voici un exemple artificiel *(il arrive avec les cartes collées au mur pour se venger de nous)*, où il y a trois occurrences du mot *an*. La seconde est correctement reconnue comme la particule séparable du verbe *kommen*. Pour cela, on n'a pas dû construire de groupes ni de relations syntaxiques, il a suffi d'examiner le contexte formé par les nœuds frères et leurs fils).

[14] Il y a en moyenne 7 formes par verbe, 4 par nom, 12 par adjectif. L'ensemble des mots composés reconnus est ouvert.
[15] On évalue la qualité sur la couverture courante par échantillonnage pour la mise au point, puis exhaustivement. On l'a aussi testée sur de grandes parties de la liste des formes du dictionnaire Morphy, et on travaille à une évaluation exhaustive.

FIGURE *6* : **AS : Er kommt mit den an den Wänden angeklebten Karten an, um sich an uns zu rächen.**

Un service Web utilisable librement a été créé par V. Berment à l'url: http://www.taranis-software.com/Heloise/ALD/Heloise.htm. Un exemple d'écran de ce serveur est donné en annexe. On voit que *kommt…an* a été regroupé dans un nœud, avec le lemme *ANKOMMEN-V*.

6 Conclusion et perspectives

Nous avons donc atteint deux des trois buts fixés au départ (qualité, et traitement des particules séparables séparées). Quant à la couverture, il nous reste à passer de près de 103000 à 200000 entrées (celles du Duden, cf. *http://www.duden.de/)*. Ce travail devrait être achevé fin 2013. En parallèle, (1) l'amélioration de la grammaire d'AM continue, surtout pour le traitement des substantifs et des verbes, et (2) le service Web AMALD est disponible. Nous espérons qu'il sera utilisé par les chercheurs et par d'autres, comme base d'expérimentation permettant de passer à l'échelle. Nous avons aussi commencé à travailler sur la production d'une structure arborescente interne des mots composés.

Remerciements

Nos remerciements vont à l'ANR, pour son soutien au projet Émergence Traouiero qui a motivé la reprise et l'opérationnalisation de notre analyseur morphologique de l'allemand.

Références

BROCKHAUS K. (1971) Automatische Übersetzung. Untersuchungen am Beispiel des Sprachen Englisch und Deutsch. Ed. Braunschweig.

BROCKHAUS K. (1976) *Das Übersetzungssystem SALAT. Teilprojekt A 2 Automatische Übersetzung (Universität Heidelberg).* Forschungsbericht 1. 11.1973—31.3.1976. I. und II. SFB 99 Linguistik, Universität Konstanz, 1976.

DUDEN (2012) *Duden online,* http://www.duden.de/

GUILBAUD J.-P. (1981) *Analyse morphologique de l'allemand en vue de la traduction par ordinateur de textes techniques spécialisés.* Thèse de 3ème cycle, Université de Paris III, juin 1981, 240 p. (recherche menée au GETA, Grenoble).

GUILBAUD J.-P. (1984) *Principles and results of a German-French MT system.* In "Machine Translation today: the state of the art" (Proc. third Lugano Tutorial, 2–7 April 1984), M. King, ed., Edinburgh University Press (1987).

GUILBAUD J.-P. (1986) *Variables et catégories grammaticales dans un modèle ARIANE.* Proc. COLING-86, Bonn, août 1986, IKS, ACL, ed., pp. 405—407.

RAPP, Reinhard; Lezius, Wolfgang (2001) *Statistische Wortartenannotierung für das Deutsche.* Sprache und Datenverarbeitung 25(2):5-21.

LEZIUS, Wolfgang (2000) *Morphy - German Morphology, Part-of-Speech Tagging and Applications*, in Ulrich Heid; Stefan Evert; Egbert Lehmann and Christian Rohrer, editors, Proc. 9th EURALEX International Congress, pp. 619-623, Stuttgart, Germany.

LEZIUS, Wolfgang; Rapp, Reinhard; Wettler, Manfred (1998) *A Freely Available Morphological Analyzer, Disambiguator, and Context Sensitive Lemmatizer for German,* in Proc. COLING-ACL 1998, pp. 743-747.

ANNEXE : SITE D'EXPÉRIMENTATION http://www.taranis-software.com/Heloise/ALD/Heloise.htm

FIGURE 7 : Plate-forme de démonstration de l'analyseur AMALD de l'allemand

FORME	LEMME	FORME	LEMME
zufriedengestelltem	ZUFRIEDEN-STELLEN-V	zufriedengestellter	ZUFRIEDEN-STELLEN-V
zufriedengestellten	ZUFRIEDEN-STELLEN-V	zufriedengestelltestem	ZUFRIEDEN-STELLEN-V
Zufriedengestellten	ZUFRIEDEN-STELLEN-V	zufriedengestelltesten	ZUFRIEDEN-STELLEN-V
zufriedengestelltem	ZUFRIEDEN-STELLEN-V	zufriedengestelltester	ZUFRIEDEN-STELLEN-V
zufriedengestellteren	ZUFRIEDEN-STELLEN-V	zufriedengestelltestes	ZUFRIEDEN-STELLEN-V
Zufriedengestellteren	ZUFRIEDEN-STELLEN-V	zufriedengestelltestе	ZUFRIEDEN-STELLEN-V
zufriedengestellterer	ZUFRIEDEN-STELLEN-V	zufriedengestelltes	ZUFRIEDEN-STELLEN-V
zufriedengestellteres	ZUFRIEDEN-STELLEN-V	zufriedengestellte	ZUFRIEDEN-STELLEN-V
zufriedengestelltere	ZUFRIEDEN-STELLEN-V	zufriedengestellt	ZUFRIEDEN-STELLEN-V

FIGURE 8 : lemmes produits par l'analyseur AMALD sur les exemples donnés pour Morphy

FORME	DÉCORATION
zufriedengestelltem	UL('ZUFRIEDEN-STELLEN-V'), DRV (VA3), SUBV (HAB), KMS(VB,ADJ), SUBADJ(RSTA), MT(PPA), CSMT(DAT), CSNT(DAT), IC(2), VAL1(ACC)
zufriedengestellten	UL('ZUFRIEDEN-STELLEN-V'), DRV (VA3), SUBV (HAB), KMS(VB,ADJ), SUBADJ(RSTA), MT(PPA), CSMT(ACC,GEN), CSNT(GEN), CPT(DAT), CSMB(ACC,DAT,GEN), CSFB(DAT,GEN), CSNB(DAT,GEN), CPB(NOM,ACC,DAT,GEN), IC(2), VAL1(ACC)
Zufriedengestellten	UL('ZUFRIEDEN-STELLEN-V'), DRV (VA3), KMS(ADJ), SUBADJ(NMEX), CSMT(ACC,GEN), CSNT(GEN), CPT(DAT), CSMB(ACC,DAT,GEN), CSFB(DAT,GEN), CSNB(DAT,GEN), CPB(NOM,ACC,DAT,GEN), IC(2), TYPO(MJ1)
zufriedengestellter	UL('ZUFRIEDEN-STELLEN-V'), DRV (VA3), SUBV (HAB), KMS(VB,ADJ), SUBADJ(RSTA), MT(PPA), CSMT(NOM), CSFT(DAT,GEN), CPT(GEN), IC(1,2), DEG(CP), VAL1(ACC)
zufriedengestelltestes	UL('ZUFRIEDEN-STELLEN-V'), DRV (VA3), SUBV (HAB), KMS(VB,ADJ), SUBADJ(RSTA), MT(PPA), CSNT(NOM,ACC), IC(2), DEG(SP), VAL1(ACC)

FIGURE 9 : Valeurs de tous les attributs produits par l'analyseur AMALD sur 5 de ces mêmes exemples