Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Démo de AMALD-serveur et AMALD-corpus, dédiés à l'analyse morphologique de l'allemand

Christian Boitet[1], Vincent Berment[1,3], Jean-Philippe Guilbaud[1], Claire Lemaire[1,2]
(1) LIG-GETALP, UGA, IMAG, 700 av. Centrale, 38058 Grenoble cedex 9, France
(2) LAIRDIL, IUT, Univ. Paul Sabatier, 115 B rte de Narbonne, 31077 Toulouse, France
(3) INaLCO, 65 rue des Grands Moulins, 75214 Paris cedex 13, France
{Prénom.Nom}@imag.fr

RÉSUMÉ

Le projet AMALDarium vise à offrir sur la plateforme lingwarium.org (1) un service d'analyse morphologique de l'allemand (AMALD-serveur), à grande couverture et de haute qualité, traitant la flexion, la dérivation et la composition, ainsi que les verbes à particule séparable séparée (ou agglutinée), (2) un corpus de référence de haute qualité donnant tous les résultats possibles de l'analyse morphologique, avant filtrage par une méthode statistique ou syntaxique, et (3) une plateforme (AMALD-éval) permettant d'organiser des évaluations comparatives, dans la perspective d'améliorer les performances d'algorithmes d'apprentissage en morphologie. Nous présentons ici une démonstration en ligne seulement de AMALD-serveur et AMALD-corpus. Le corpus est un sous-ensemble anonymisé et vérifié d'un corpus en allemand formé de textes sur le cancer du sein, contenant de nombreux mots composés techniques.

ABSTRACT

Demonstration of AMALD-serveur and AMALD-corpus, dedicated to the morphological analysis of German

The AMALDarium project aims to offer on the lingwarium.org platform (1) a large-coverage and high-quality morphological analysis service for German (AMALD-server), handling flexion, derivation and composition, as well as verbs with separated (or agglutinated) separable particles, (2) a high-quality reference corpus giving all possible results of the morphological analysis, before filtering by a statistical or syntactic analysis, and (3) a platform (AMALD-eval) to organize comparative evaluations, with a view to improve the performance of morphology learning algorithms. We present an online demonstration of AMALD-server and AMALD-corpus only. The corpus is an anonymized and verified subset of a German corpus of texts on breast cancer, containing many technical compound words. The parser accepts as input a text of any length.

MOTS-CLES : Allemand, analyse morphologique, corpus de référence, services web gratuits.
KEYWORDS: German language, morphological analysis, reference corpus, free web services.

1 Motivations

Dans une publication à TALN-2013 (Guilbaud & al. 2013), nous avons décrit l'AM (analyseur morphologique) AMALD construit par J.P. Guilbaud. Depuis, sa couverture est passée de 103.000 à 209.325 lemmes simples et composés, soit plus de 1M de formes, et le traitement des balises XML (sans liste attributs-valeurs) a été introduit. Il est construit par une méthode « experte » et non pas par une méthode empirique, statistique ou neuronale. En 2013, nous l'avions comparé à d'autres AM en ligne, comme DEMorphy (Altinok 2018), *Disambiguator* (Lezius, Rapp et Wettler, 1998), SMOR et Morphisto (Piskorski *et al.*, 2009). Nous avions montré qu'il était de très loin meilleur, en particulier pour la lemmatisation des formes verbales et des noms composés, et 2

HAL : hal-02768493.

à 3 fois plus couvrant. Par exemple, le corpus de Morphy (annoté par ses résultats), qui en fait n'est pas un corpus mais une liste de mots-formes, montrait 90% de résultats faux au niveau des lemmes sur les formes verbales du corpus associé. D'autre part, AMALD était le seul à traiter convenablement les mots composés connexes, c'est-à-dire à bien les segmenter et à fournir pour chaque morceau un lemme et les informations lexicales associées. AMALD était également le seul AM à regrouper un verbe composé d'un verbe simple et d'une particule séparable, quand la particule n'est pas agglutinée, mais séparée par un nombre quelconque de mots. Nous sommes en train d'actualiser notre étude de l'état de l'art, en y incluant Fips-de du LATL (Scherrer 2008) et phpMorphy. Cela devrait faire l'objet d'une publication ultérieure liée à AMALD-eval.
La construction manuelle d'un AM par un ou des experts est un très gros travail, et on ne peut sans doute pas le poursuivre indéfiniment. Nous avons inclus (comme annoncé en 2013) les ≈210.000 lemmes simples ou composés (agglutinés, comme *Onkogen,* oncogène) trouvés dans le Duden, mais, pour pouvoir monter à 6 ou 7 millions d'entrées (comme dans le système japonais↔anglais ATLAS-2.v13 de Fujitsu), il semble nécessaire de mettre en œuvre des techniques d'apprentissage. Or, pour donner de bons résultats, il faut qu'elles disposent d'un corpus annoté (par les résultats souhaités) de grande taille et de très haute qualité, exactement comme pour la TA ou d'autres applications. Nous avons donc entrepris de construire un tel corpus, avec dans l'idée (1) de partir d'un corpus réel, appartenant à un sous-langage justifiant de nombreuses applications, (2) d'utiliser AMALD pour obtenir une première version des annotations, puis (3) de les évaluer (au niveau des lemmes, des catégories morphosyntaxiques, de la décomposition pour les composés, et des variables d'actualisation comme cas, genre, nombre, personne, temps, mode), et d'améliorer AMALD. Le corpus actuel est un sous-ensemble anonymisé et vérifié de 20 textes tirés d'un corpus formé de textes en allemand sur le cancer du sein, contenant de nombreux mots composés techniques. Nous le présentons plus bas (pas très en détail, à cause de la limite à 4 pages).
Depuis 2013, la couverture de l'AM a été beaucoup augmentée, de sorte qu'il n'y a plus de « mots inconnus » dans ce corpus (à part les balises Xml et les emprunts à l'anglais), et on ne trouve pratiquement pas d'erreur dans les résultats. Nous nous sommes rendu compte qu'AMALD devrait être utilisé par d'autres pour que des erreurs ou incomplétudes puissent nous être signalées, et que nous puissions continuer à corriger ou compléter. Enfin, il est apparu qu'il était très difficile de définir des mesures de qualité pour les AM. C'est dans ce contexte que nous avons défini le projet AMALDarium. Ce projet vise à offrir sur la plate-forme lingwarium.org[1] (1) un service d'analyse morphologique de l'allemand (AMALD-serveur), à grande couverture et de haute qualité, traitant la flexion, la dérivation et la composition, ainsi que les verbes à particule séparable séparée (ou agglutinée), (2) un corpus de référence de haute qualité donnant tous les résultats possibles de l'analyse morphologique, avant filtrage par une méthode statistique ou syntaxique, et (3) une plateforme (AMALD-éval) permettant d'organiser des évaluations comparatives, dans la perspective d'améliorer les performances d'algorithmes d'apprentissage en morphologie. Nous présentons ici une démonstration en ligne uniquement de AMALD-serveur et de AMALD-corpus.

2 AMALD-serveur

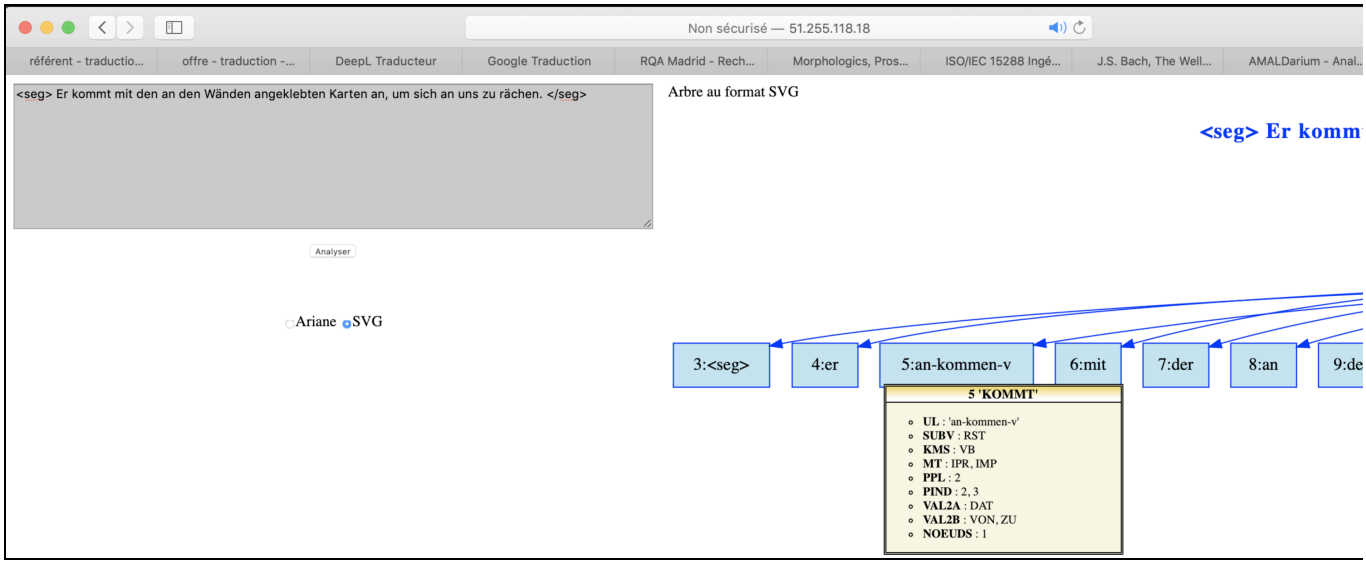
Il est accessible à http://51.255.118.18/TestAnaALD/. On peut entrer un texte complet, codé en UTF-8, en format texte simple ou Xml/TEI simplifié pour l'instant : les balises ouvrantes ne doivent pas contenir de liste d'attributs-valeurs. D'autre part, les balises doivent être séparées du texte (pas '`<h1>Titel`' mais '`<h1> Titel`'). Exemple :

```
<txt>  <!-- <txt> et pas: <txt id = "1_cle200200206"> -->      <title> 1 </title>
<body> <h1> Risikoabschätzung für das familiäre Auftreten von Brustkrebs </h1> <p>
... <seg> Die Autoren legen Tabellen vor, aus denen das Erkrankungsrisiko bei
einer bestimmten familiären Vorgeschichte abhängig vom Diagnosealter von ein oder
zwei Verwandten ersten oder zweiten Grades entnommen werden kann. </seg>
```

Figure 1 : exemple de texte soumis à l'AM (XML/TEI simplifié)

[1] Plateforme générale : http://www.lingwarium.org, AMALD-serveur : http://51.255.118.18/TestAnaALD
AMALD-corpus : http://51.255.118.18/AMALDarium

Le résultat est un arbre décoré, visible en format texte ou graphique (SVG).
Voici une image d'écran, sur la phrase : **Er kommt mit den an den Wänden angeklebten Karten an, um sich an uns zu rächen.** On voit ici que la forme composée "kommt… an" a été reconnue, et que l'information donnée sur "kommt" est celle attachée au lemme "an-kommen".



La définition des noms et valeurs de variables (comme KMS pour catégorie morphosyntaxique, etc.) est donnée dans les « fichiers de déclarations de variables » accessibles sur le site. Notons que l'analyseur accepte en entrée un texte de longueur quelconque.

3 AMALD-corpus

Nous avons considéré plusieurs corpus médicaux en partie ou entièrement en allemand : UMLS (Browne, C., Divita, G, Aronson, R., et al., 2003), UFL medical corpus, DiK (Korte, 2019), et TIGER (Brants, S., Dipper, S., Eisenberg, P., et al., 2004). Nous n'avons pas pu les utiliser, car ils sont souvent annotés sémantiquement, dans le but de développer des techniques d'extraction de concepts, mais pas morphologiquement. Ils sont aussi prétraités, avec élimination de la casse, qui introduit des parasites dans l'AM[2]. Il y a d'autres problèmes : par exemple (Scherrer 2008) note que TIGER n'a qu'une étiquette pour les adjectifs et les adverbes, ce qui n'est pas très satisfaisant, puisque les adjectifs se déclinent et pas les adverbes, et qu'il y a des lemmes seulement adverbiaux (sehr, zu, dann…) — alors que tout adjectif allemand peut être adverbe.

Nous sommes partis d'un corpus médical (non morphologiquement annoté) d'environ 400.000 mots concernant le cancer du sein, collecté et anonymisé vers 2013 pour étudier les différences entre deux genres, la langue de spécialité et la langue de vulgarisation, et comprenant une moitié dans chaque genre (les sources sont dans la Table 2). Le langage médical comprend beaucoup de termes composés qui sont difficiles à analyser morphologiquement, ce qui le rend très intéressant.

Nous avons extrait une partie de ce corpus et l'avons annotée morphologiquement, de façon à obtenir un premier état qu'on peut déjà considérer comme un corpus de référence. AMALD-corpus est anonymisé et vérifié. Il est pour l'instant composé de 20 fichiers (Table 1) balisés en XML/TEI (simplifié), qui sont très lisibles sous tout navigateur.

[2] Les noms communs prennent la majuscule à l'initiale, ce qui permet, en n+1-ième position, de distinguer par exemple 'Auftritt' (nom) de 'auftritt' (verbe).

Table 1 : 20 fichiers de AMALD-corpus

Répertoire		#occurrences textuelles		#occurrences avec balises		#occurrences de balises	% de balises	#pages standard
Fichier	1_ald	713	txt	808	xml	95	11,8%	2,85
Fichier	2_ald	836	txt	966	xml	130	13,5%	3,34
Fichier	3_ald	490	txt	588	xml	98	16,7%	1,96
Fichier	4_ald	525	txt	628	xml	103	16,4%	2,10
Fichier	5_ald	458	txt	546	xml	88	16,1%	1,83
Fichier	6_ald	691	txt	815	xml	124	15,2%	2,76
Fichier	7_ald	470	txt	558	xml	88	15,8%	1,88
Fichier	8_ald	731	txt	851	xml	120	14,1%	2,92
Fichier	9_ald	492	txt	588	xml	96	16,3%	1,97
Fichier	10_ald	812	txt	961	xml	149	15,5%	3,25
Fichier	11_ald	792	txt	930	xml	138	14,8%	3,17
Fichier	12_ald	901	txt	1093	xml	192	17,6%	3,60
Fichier	13_ald	534	txt	618	xml	84	13,6%	2,14
Fichier	14_ald	644	txt	760	xml	116	15,3%	2,58
Fichier	15_ald	1368	txt	1602	xml	234	14,6%	5,47
Fichier	16_ald	641	txt	741	xml	100	13,5%	2,56
Fichier	17_ald	742	txt	906	xml	164	18,1%	2,97
Fichier	18_ald	538	txt	652	xml	114	17,5%	2,15
Fichier	19_ald	607	txt	725	xml	118	16,3%	2,43
Fichier	20_ald	857	txt	1038	xml	181	17,4%	3,43
Total		13842		16374		2532	15,5%	55,37
Mots/page	250							

Perspective. Comme nous avons déjà soigneusement relu la totalité des 409.280 mots de ces 241 textes (de 1461 mots par texte en moyenne), il devrait être possible d'augmenter rapidement AMALD-corpus, de 20 à 240 textes, en les convertissant dans notre format d'entrée (XML/TEI simplifié) et en les soumettant à AMALD-serveur.

Table 2 : Sources de AMALD-corpus

Source web	Nombre d'articles	Nombre de mots
aerzteblatt.de	67	114406
senologie.org	1	81192
uni-frauenklinik-	7	5816
iwenv.de	1	1584
wiralle.de	1	1204
krebsinformationsdienst.de	1	326
aerztezeitung.de	1	118
netdoktor.de	80	45145
mammamia-online.de	1	39628
krebsgesellschaft-nrw.de	3	25697
brustkrebsdeutschland.de	16	20743
onkosupport.de	4	18559
mammakarzinom-info.de	36	17765
mammo-programm.de	6	8881
de.wikipedia.org	1	7759
brustkrebs-sprechstunden.de	1	4663
ago-online.de	1	4335
experten-sprechstunde.de	1	3820
brca-netzwerk.de	8	2602
krebshilfe.de	2	2547
sportaerztebund-niedersachsen.de	1	1286
medical.siemens.com	1	1204
TOTAL	241	409280

Références

ALTINOK, D. *DEMorphy, German Language Morphological Analyzer*. arXiv preprint:1803.00902, 2018.

BERMENT V., BOITET CH., GUILBAUD J.PH., KAPOCIUTE-DZIKIENE J. *Several Ways to Use the Lingwarium.org Online MT Collaborative Platform to Develop Rich Morphological Analyzers*. Computational Linguistics and Intelligent Text Processing - CICLing 2017, Revised Selected Papers, Part I, Apr 2017, Budapest, pp.81--86.

BRANTS, S., DIPPER, S., EISENBERG, P., *et al*. TIGER: *Linguistic interpretation of a German corpus*. Research on language and computation, 2004, vol. 2, no 4, p. 597-620.

BROWNE, C., DIVITA, G, ARONSON, R., *et al*. *UMLS language and vocabulary tools: AMIA 2003 open source expo*. In : AMIA annual symposium proceedings. American Medical Informatics Association, 2003. p. 798.

DELPECH, E., DAILLE, B., MORIN, E. et LEMAIRE, C., *Extraction of domain-specific bilingual lexicons from comparable corpora: compositional translation and ranking,* COLING 2012, 8–12 Dec. Mumbai, 2012.

DELPECH, E., DAILLE, B., MORIN, E. et LEMAIRE, C., *Identification of Fertile Translations in Medical Comparable Corpora: a Morpho-Compositional Approach*. In Proceedings of the 10th biennial conference of the Association for Machine Translation in the Americas (AMTA), 28 Oct.–1 Nov., San Diego, 2012.

GUILBAUD, J.PH., BOITET, C. et BERMENT V., *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.

KORTE, L. *Online-Ressourcen zum Thema Linguistik und Medizin*. Zeitschrift für germanistische Linguistik, 2019, vol. 47, n° 1, pp. 260-268.

LEMAIRE, C., *Un nouveau besoin dans l'industrie : une aide au « rédacteur-traduisant »*. In Actes de la 24e conférence sur le Traitement Automatique des Langues Naturelles (TALN), 26–30 juin, Orléans, 2017.

LEMAIRE, C., GUILBAUD , J.PH., *Corpus de registres différents pour le développement d'un aligneur d'unités polylexicales*. 11e conférence sur la Lexicologie, Terminologie et Traduction (LTT), 25–28 septembre, Grenoble, France, 2018. (Article long accepté et présenté, à paraître en 2020).

LEZIUS, W., RAPP, R. ET WETTLER, M. *A freely available morphological analyzer, disambiguator and context-sensitive lemmatizer for German*. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2. Association for Computational Linguistics, 1998. pp. 743-748.

PISKORSKI, J., *et al*. *Morphisto, an open source morphological analyzer for German*. In: Finite-state Methods and NLP: Postproceedings of the 7th International Workshop FSMNLP. 2009. p. 224.

SCHERRER, Yves. *Part-of-Speech Tagging with a Symbolic Full Parser: Using the TIGER Treebank to Evaluate Fips*. In: Proceedings of the ACL 2008 Workshop on Parsing German. 2008. pp. 16-23.