CALOR-Frame : un corpus de textes encyclopédiques annoté en cadres sémantiques

Frédéric Béchet[1] Géraldine Damnati[2] Johannes Heinecke[2] Gabriel Marzinotto[2] Alexis Nasr[1]
(1) Aix Marseille Univ, CNRS, Centrale Marseille, LIF, Marseille, France
(2) Orange Lab, Lannion, France
frederic.bechet@univ-amu.fr, alexis.nasr@univ-amu.fr
geraldine.damnati@orange.com, johannes.heinecke@orange.com,
gabriel.marzinotto@orange.com

RÉSUMÉ
Le corpus CALOR-Frame est un corpus annoté en cadres sémantiques, constitué de textes encyclopédiques dans le domaine de l'Histoire et produit conjointement par l'Université d'Aix-Marseille et Orange Labs. La constitution de cette ressource s'inscrit dans le cadre général de la recherche d'information avec pour objectif de favoriser l'accès aux contenus de la connaissance. La structuration en cadres sémantiques permet des recherches avancées dépassant le cadre de la simple recherche par mots-clés. Dans cet article est décrit le processus d'annotation en cadres sémantiques mis en place, qui utilise un outil de validation d'annotations automatiques à des fins d'optimisation. Le choix des textes et des cadres sémantiques considérés est également motivé.

ABSTRACT
CALOR-Frame : a corpus of encyclopedic texts annotated with semantic frames

CALOR-Frame is a corpus of History encyclopedic texts annotated in semantic frames, that has been jointly produced by Aix-Marseille University and Orange Labs. The constitution of this ressource has been driven by the more general context of Information Retrieval, with the purpose of enhancing access to Knowledge contents. Semantic Frame structuration enables advanced research fucntionalities, beyond key-word search. This article presents the annotation process that has been set up, using a tool to automatically validate generated annotations in an optimized way. The selection of texts and semantic frames is also motivated.

MOTS-CLÉS : Cadre sémantique, corpus, apprentissage actif, étiquetage de séquence.

KEYWORDS: Semantic Frame, corpora, active learning, sequence labelling.

1 Introduction

Les ressources linguistiques intégrant des annotations sémantiques sont de plus en plus nombreuses pour la langue anglaise, notamment grâce aux campagnes d'évaluation telles que CoNLL. Malheureusement la langue française est la plupart du temps absente de ces programmes et la constitution de telles ressources n'en est qu'au début grâce en particulier au projet ASFALDA(Candito *et al.*,

2014) [1] qui s'est attaché à produire des ressources sémantiques sur la base de FrameNet pour le français (Djemaa *et al.*, 2016). Les données annotées sont les données du French Treebank (Abeillé *et al.*, 2003) qui contenait déjà des annotations à différents niveaux d'analyse linguistique. Le corpus CALOR a pour sa part été construit sans annotations préalables, le choix des données étant motivé par une visée applicative : réaliser et évaluer des systèmes d'extraction d'information sur des données encyclopédiques pour enrichir des bases de Connaissance et mettre au point des moteurs de recherche spécifiques à visée pédagogique.

Nous proposons d'utiliser comme modèle sémantique le dictionnaire de cadres sémantiques FrameNet (Baker *et al.*, 1998). Le choix de FrameNet se justifie à la fois par son indépendance vis-à-vis de toute théorie syntaxique et la facilité avec laquelle on peut restreindre et spécialiser le modèle à nos besoins grâce à une sélection de cadre (Frame), de mots supports (Lexical Unit) et d'éléments de relations (Frame Element) ou rôles ; mais aussi par l'expertise acquise dans notre équipe sur ce modèle, notamment grâce au projet ANR ASFALDA visant à développer une version française de ce dictionnaire de cadres sémantiques.

Par exemple, si l'on s'intéresse à l'annotation de relations sémantiques liées au commencement d'un évènement historique, on peut utiliser la définition FrameNet du cadre « *Activity-Start* ». Ce cadre contient, entre autre, les rôles :

- Activity : l'activité qui débute
- Agent : l'entité débutant l'activité
- Place : l'endroit où se déroule l'activité
- Time : le moment où se déroule l'activité
- Purpose : le but visé par l'activité

FrameNet contient d'autres rôles optionnels, que l'on peut décider ou pas de rajouter dans nos annotations. En spécifiant les éléments lexicaux déclencheurs (ou *Lexical Units* – LU) de chaque Frame spécifiques aux corpus cibles, on peut pré-annoter chaque corpus avec des hypothèses de présence de relations sémantiques. Le but de l'annotation (manuelle ou automatique) consiste alors à valider ou infirmer la présence d'une relation déclenchée par un LU potentiellement ambigu, et à déterminer quels sont les rôles de cette relation si celle-ci est validée. Par exemple, en considérant le LU *commencer* pour la relation *Activity-Start*, on peut annoter le texte de la figure 1 avec 5 rôles (*Frame elements*).

Le nombre de relations et d'éléments à annoter dépendra à la fois du cadre d'utilisation et du temps disponible pour réaliser ces annotations.

FIGURE 1 – Exemple de phrase annotée au moyen d'étiquettes associées à un cadre sémantique provenant de FrameNet

La section 2 présente les motivations qui ont conduit au choix des textes à annoter et des cadres sémantiques considérés. La section 3 présente le processus mis au point pour l'annotation de ce

1. https ://sites.google.com/site/anrasfalda/

affecter	agir	aide	aliment	alimenter	appeler
apprendre	arrêter	arrivée	arriver	assassinat	attaque
attaquer	atteindre	boire	boisson	bombardement	bouger
cacher	chasse	chasser	chercher	choisir	choix
colonisation	coloniser	combat	combattre	commandement	commander
commencement	commencer	compréhension	comprendre	connaissance	connaître
construction	construire	contenir	contre-attaque	créer	datation
date	dater	début	débuter	décès	décider
déclaration	déclarer	décocher	découverte	découvrir	défaite
déguster	demande	demander	départ	déplacement	déplacer
désignation	désigner	devenir	dire	diriger	discours
don	donner	écrire	élection	élever	élire
enseignant	enseignement	enseigner	envoyer	étude	étudier
exécution	exiger	existence	exister	expression	exprimer
fabrication	fabriquer	fouille	fouiller	génocide	identification
identifier	influence	influencer	inspiration	inspirer	installation
installer	inventer	invention	localisation	manger	message
mesurer	mort	mourir	nomination	nommer	offensive
offrir	organiser	origine	participation	participer	pêche
pêcher	perdre	perte	peser	peuplement	peupler
poids	promotion	provenir	quitter	réalisation	réaliser
recherche	rechercher	repas	requête	retrouver	sélectionner
soutenir	soutien	succès	taille	tirer	traque
traquer	trouver	tuer	usage	utilisation	utiliser
victoire					

TABLE 1 – Liste des Lexical Units (LU) pouvant déclencher l'apparition d'une frame dans le corpus CALOR

corpus de façon à optimiser le temps d'annotation et la charge cognitive des annotateurs. La section 4 présente enfin la description quantitative du corpus CALOR annoté obtenu, corpus ayant vocation à être distribué publiquement prochainement.

2 Choix des corpus et du modèle d'annotation

Le corpus CALOR est constitué de documents issus de 4 sources différentes :

- textes issus du portail Wikipédia sur l'Archéologie (WA) : 201 documents
- textes issus du portail Wikipédia sur la Première Guerre Mondiale (WGM) : 355 documents
- textes issus de Vikidia (VKH), l'encyclopédie en ligne pour enfants, à partir de deux portails (Préhistoire et Antiquité) : 183 documents
- textes historiques issus de ClioTexte [2] sur la Première Guerre Mondiale (CTGM) : 16 documents

En choisissant ces corpus nous avons cherché à avoir à la fois une diversité de styles et une diversité de domaines. La présence des données issues de l'encyclopédie pour enfants doit nous permettre d'étudier l'influence du niveau de langage sur la complexité de la tâche. Par ailleurs, les textes issus de Cliotexte sont des documents historiques (déclarations, essais, discours, lettres) qui sont formulés différemment des textes déclaratifs présents dans les pages Wikipedia. L'étude de ces textes relatifs à la période de la Première Guerre Mondiale pourra permettre d'étudier la dépendance des systèmes à la nature des données pour un domaine commun. Enfin, le fait de retenir des textes issus de deux portails de Wikipedia, permettra d'étudier la dépendance au domaine : typiquement, est-ce que pour un cadre sémantique donné un modèle appris sur des données relatives à l'archéologie peut s'appliquer avec

2. https ://clio-texte.clionautes.org/

Accomplishment	Activity-start	Age	Appointing	Arrest
Arriving	Assistance	Attack	Awareness	Becoming
Becoming-aware	Buildings	Change-of-leadership	Choosing	Colonization
Coming-to-believe	Coming-up-with	Conduct	Contacting	Creating
Death	Deciding	Departing	Dimension	Education-teaching
Existence	Expressing-publicly	Finish-competition	Giving	Hiding-objects
Hostile-encounter	Hunting	Inclusion	Ingestion	Installing
Killing	Leadership	Locating	Losing	Making-arrangements
Motion	Objective-influence	Origin	Participation	Request
Scrutiny	Seeking	Sending	Shoot-projectiles	Statement
Subjective-influence	Using	Verification		

TABLE 2 – Liste des cadres sémantiques retenus dans le corpus CALOR

succès sur des données relatives à la Première Guerre Mondiale ?

Afin de déterminer les cadres sémantiques à annoter en priorité, nous nous sommes d'abord intéressés aux déclencheurs de frames, les *Lexical Units* (LU). Nous nous sommes concentrés en premier lieu sur les verbes, pour sélectionner parmi les verbes les plus fréquents de chaque corpus ceux susceptibles de conduire à des annotations utiles pour le cadre applicatif choisi. Ce cadre étant celui de la recherche d'information, nous nous sommes concentré sur des verbes pouvant intervenir dans des requêtes liées aux thèmes des corpus sélectionnés (par exemple *combattre* ou *déclarer* pour les textes historiques). Nous avons ensuite associé à ces verbes les noms déverbaux et les noms dérivés fréquemment observés dans le corpus pour arriver à une sélection de 145 lemmes (70 noms, 75 verbes), présentés dans le tableau 1.

Le fait de ne choisir que des déclencheurs verbaux ou nominaux est un parti pris applicatif. Nous ne cherchons pas à décrire un texte de façon exhaustive, mais à extraire des informations ciblées sur un certain nombre de cadres sémantiques. Contrairement à la façon dont a été élaboré le corpus ASFALDA, où l'objectif était de couvrir de façon la plus précise possible plusieurs domaines notionnels, nous avons ici priorisé les lemmes les plus fréquents dans notre corpus. Le nombre moyen de déclencheurs possibles par cadre est donc bien inférieur.

A partir des lemmes choisis, nous avons sélectionné un ensemble de cadres sémantiques inspirés du FrameNet anglais. Au total, 53 cadres ont été retenus, associés à 150 *Frame Elements*. Comme pour les LU, les cadres ont été choisis en fonction des domaines des corpus, en ne retenant que les cadres susceptibles d'apparaître dans les textes sélectionnés. Nous n'avons rajouté aucun nouveau cadre par rapport à ceux contenu dans FrameNet. Si, lors de la phase d'annotation de nos corpus, un LU déclenche un cadre sémantique non prévu dans notre modèle, il reçoit l'étiquette `other`.

Enfin pour chaque cadre sémantique retenu nous avons sélectionné un ensemble de rôles (ou *Frame Elements*) parmi ceux décrits dans la ressource FrameNet pour l'anglais.

Parmi les lemmes retenus, certains peuvent déclencher plusieurs cadres différents. Le tableau 3 résume les principales sources d'ambiguïté rencontrées parmi les cadres de notre modèle.

Enfin un certain nombre de lemmes peuvent déclencher d'autres cadres que ceux que nous avons sélectionné dans notre modèle. Dans ce cas leurs occurrences sont annotées avec le symbole `other`. Les 10 lemmes les plus fréquemment rencontrés dans des cadres non modélisés sont : *trouver, appeler, atteindre, dire, agir, nommer, arrêter, comprendre, organiser et arriver*.

LU	Cadre 1	Cadre 2	Cadre 3
affecter	Objective-influence	Appointing	
compréhension	Coming-to-believe	Awareness	
comprendre	Coming-to-believe	Awareness	
désigner	Choosing	Change-of-leadership	Appointing
élire	Choosing	Change-of-leadership	
exécution	Killing	Creating	
fabrication	Creating	Buildings	
installation	Installing	Colonization	
installer	Installing	Colonization	
perdre	Losing	Finish-competition	
réalisation	Creating	Accomplishment	
réaliser	Creating	Coming-to-believe	Accomplishment
victoire	Losing	Finish-competition	

TABLE 3 – Liste des lemmes ambiguës selon le dictionnaire de cadre sémantique retenu

3 Processus d'annotation

Le processus d'annotation est décrit dans la figure 2. Les corpus à annoter sont découpés automatiquement en phrases, puis pré-traités par la suite d'outils Macaon (Nasr *et al.*, 2010) (étiquetage en POS, lemmatisation, analyse syntaxique en dépendance). Chaque lemme du dictionnaire de *Lexical Unit* déclencheurs apparaissant dans une phrase du corpus génère un exemple à annoter. Il y a donc au plus un cadre à annoter dans chaque exemple, et une même phrase peut se retrouver dans plusieurs exemples si elle contient plusieurs LU. Cette annotation syntaxique n'est pas visible pour les annotateurs, elle sert uniquement à obtenir les lemmes pour les LU et à fournir des traits syntaxiques pour l'analyseur automatique en cadre sémantique utilisé pour les pré-annotations.

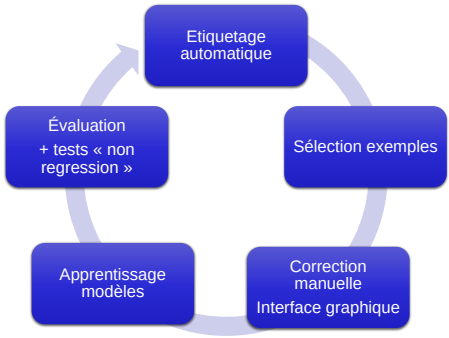

FIGURE 2 – Processus d'annotation

L'annotation est effectuée selon un processus circulaire composé de 5 étapes effectué jusqu'à annotation complète du corpus :

1. *Etiquetage automatique* : un analyseur automatique en cadre sémantique, présenté dans (Gabriel Marzinotto, 2017), est appris sur la partie du corpus déjà étiquetée, puis appliqué sur les données encore brutes. Bien sûr, lors de la première itération, l'analyseur n'est pas appliqué et aucune hypothèse n'est présente dans les données à étiqueter.

2. *Sélection exemples* : les exemples qui n'ont pas encore été validés manuellement sont triés par similarité et découpés en batch qui seront présentés aux annotateurs humain pour validation. La mesure de similarité prend en compte tout d'abord les LUs, puis les constructions syntaxiques (obtenues par Macaon) dans lesquelles se trouvent les LU, puis les contextes lexicaux. L'idée est de présenter aux annotateurs des batch d'exemples très proches afin de limiter l'effort cognitif d'annotation.

3. *Correction manuelle* : une interface graphique permet aux annotateurs d'annoter les batchs choisis dans l'étape précédente, en ne visualisant que le texte et les annotations automatique en cadre sémantique (aucune annotation syntaxique n'est affichée). Cette annotation consiste à corriger ces annotations en cadre sémantique si elles sont présentes et à ajouter les annotations manquantes.

4. *Apprentissage des modèles* : après le traitement de chaque batch par les annotateurs humains, les modèles de l'analyseur automatique en frame sont mis à jour en étant entraînés sur le corpus composé de toutes les annotations validées manuellement.

5. *Evaluation* : à chaque nouvel apprentissage des modèles, des tests de non-régression permettent de tester la cohérence des nouvelles données étiquetées. Ces tests consistent à vérifier que les nouveaux modèles améliorent (ou tout du moins ne dégradent pas) les annotations automatiques effectuées sur un petit corpus de validation, ou obtenues grâce à une évaluation de type *k-fold*. Si le nouveau batch annoté dégrade les résultats, il est mis en *quarantaine* afin d'être vérifié par un autre annotateur. Sinon l'analyseur automatique avec les nouveaux modèles est appliqué sur les données non-validées et l'on repart à l'étape 1.

Ce processus d'annotation a été appliqué sur l'ensemble du corpus CALOR. En faisant collaborer annotation automatique et validation manuelle, nous avons réussi à optimiser le temps d'annotation, les performances du système automatique s'améliorant à chaque itération. La sélection d'exemples proche pour la constitution des batchs a également permis de faciliter le travail des annotateurs en leur permettant de traiter les exemples par lots présentant des caractéristiques similaires. Il est cependant difficile de donner des mesures précises quantifiant le gain apporté par ces méthodes dans la mesure où les annotateurs eux-même devenaient de plus en plus expert au cours du temps. Cependant les retours subjectifs des annotateurs ont tous été très positifs, à la fois sur le fait que la pré-annotation automatique permettait de gagner du temps mais aussi sur le traitement par lot *similaires*.

Une estimation de la cohérence des annotations entre annotateurs est en cours de traitement.

4 Le corpus CALOR-Frame

Le tableau 4 présente la répartition des données annotées en fonction de l'origine des textes. Les deux ensembles issus de Wikipedia dominent le corpus. A ce jour, 21399 LU ont été annotés parmi lesquels 3403 ont été annotés avec le symbole `other` (soit 13.7%), signifiant que leur sens n'est pas couvert par notre modèle. La colonne **lexique** précise la taille des vocabulaires de chacun des corpus ; la colonne **#frame** contient le nombre d'instance de frames annotés manuellement dans chaque corpus ; enfin les deux dernières colonnes précisent à la fois le pourcentage de phrases contenant au moins une instance de frame (**phr+fr**) ainsi que le ratio de frame par phrase. Au total 46896 instances de rôles (*Frame Elements*) ont été segmentées et labellisées, soit en moyenne 2.2 rôles par instance de cadre sémantique.

corpus	#phrases	#mots	lexique	#frames	phr+fr	fr/phr
WGM (Wikipedia 1ère Guerre Mondiale)	30994	686355	42635	13008	31.95%	0.42
WA (Wikipedia Archéologie)	27023	540653	41418	5869	18.59%	0.22
CTGM (Cliotexte 1ère Guerre Mondiale)	3523	67736	10844	905	21.37%	0.26
VKH (Vikidia Préhistoire et Antiquité)	5841	85034	11649	1617	21.88%	0.28
all	67381	1379778	72127	21399	25.73%	0.32

TABLE 4 – Description du corpus CALOR-Frame

Le corpus CALOR-Frame contient l'ensemble des corpus annotés en cadre sémantique avec validation manuelle de toutes les annotations. Il contient également les annotations syntaxiques automatiques fournies par l'outil MACAON, sans aucune correction manuelle, donc contenant un certain nombre d'erreurs. La figure 3 présente la distribution des occurrences de cadres sémantiques annotés dans le corpus. Si certains cadres ne sont que peu représentés, il est intéressant de noter que la représentation des cadres ne décroît pas de façon trop rapide. 22 cadres sur les 53 représentés ont ainsi un nombre d'occurrences supérieur à 400 et 8 ont un nombre d'occurrences supérieur à 800. Les cadres les plus représentés sont `Leadership` (commander, diriger, commandement), `Attack` (attaquer, attaque, offensive, bombardement, contre-attaque) `Locating` (retrouver, trouver, localisation) et `Becoming_Aware` (découvrir, decouverte).

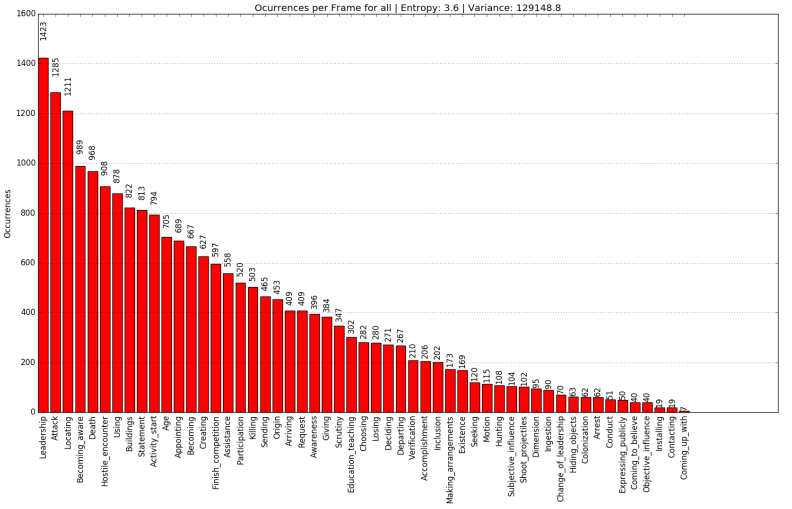

FIGURE 3 – Distribution des occurrences de cadres sémantiques dans le corpus CALOR-Frame

5 Conclusion

Le corpus CALOR-Frame est une ressource de textes encyclopédiques annotés en cadres sémantique, dont la constitution a été guidée par des choix applicatifs. L'objectif étant de développer et d'évaluer des outils d'analyse sémantique pour alimenter des systèmes de recherche d'information. A ce jour, 53 cadres sémantiques différents, associés à 150 rôles, ont été annotés avec un total de 46896 occurrences. Nous nous sommes concentré sur les déclencheurs verbaux et nominaux les plus fréquents de façon à optimiser la quantité de données annotées pour chacun d'eux, sans chercher à garantir une couverture lexicale des cadres sémantiques considérés. Le corpus produit constitue en cela une ressource complémentaire au corpus ASFALDA qui vise pour sa part à optimiser la couverture linguistique de domaines notionnels.

Remerciements

Les annotateurs ayant participé à cette ressource sont Laure Dupont, Matthieu Stali, Manon Scholivet, Sebastien Delecraz. L'interface d'annotation a été réalisée par Camille Gobert.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. *Treebanks*, p. 165–187.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, p. 86–90, Stroudsburg, PA, USA : Association for Computational Linguistics.

CANDITO M., AMSILI P., BARQUE L., BENAMARA F., CHALENDAR G., DJEMAA M., HAAS P., HUYGHE R., MATHIEU Y., MULLER P., SAGOT B. & VIEU L. (2014). Developing a french framenet : Methodology and first results. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

DJEMAA M., CANDITO M., P. M. & L. V. (2016). Corpus annotation within the french framenet : methodology and results. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, p. 3794–3801.

GABRIEL MARZINOTTO, GÉRALDINE DAMNATI F. B. (2017). Analyse automatique framenet : une étude sur un corpus français de textes encyclopédiques. In *soumis à TALN 2017* : ATALA.

NASR A., BÉCHET F. & REY J.-F. (2010). Macaon : Une chaîne linguistique pour le traitement de graphes de mots. In *Traitement Automatique des Langues Naturelles - session de démonstrations*, Montréal.