Expériences de formalisation d'un guide d'annotation : vers l'annotation agile assistée

Bruno Guillaume[1,2] Karën Fort[1,3]
(1) LORIA 54500 Vandœuvre-lès-Nancy
(2) Inria Nancy Grand-Est
(3) Université de Lorraine
bruno.guillaume@loria.fr, karen.fort@loria.fr

RÉSUMÉ

Nous proposons dans cet article une méthodologie, qui s'inspire du développement agile et qui permettrait d'assister la préparation d'une campagne d'annotation . Le principe consiste à formaliser au maximum les instructions contenues dans le guide d'annotation afin de vérifier automatiquement si le corpus en construction est cohérent avec le guide en cours d'écriture. Pour exprimer la partie formelle du guide, nous utilisons la réécriture de graphes, qui permet de décrire par des motifs les constructions définies. Cette formalisation permet de repérer les constructions prévues par le guide et, par contraste, celles qui ne sont pas cohérentes avec le guide. En cas d'incohérence, un expert peut soit corriger l'annotation, soit mettre à jour le guide et relancer le processus.

ABSTRACT

Formalizing an annotation guide : some experiments towards assisted agile annotation

This article presents a methodology, inspired from the agile development paradigm, that helps preparing an annotation campaign. The idea behind the methodology is to formalize as much as possible the instructions given in the guidelines, in order to automatically check the consistency of the corpus being annotated with the guidelines, as they are being written. To formalize the guidelines, we use a graph rewriting tool, that allows us to use a rich language to describe the instructions. This formalization allows us to spot the rightfully annotated constructions and, by contrast, those that are not consistent with the guidelines. In case of inconsistency, an expert can either correct the annotation or update the guidelines and rerun the process.

MOTS-CLÉS : annotation, guide d'annotation, annotation agile, réécriture de graphes.

KEYWORDS: annotation, annotation guide, agile annotation, graph rewriting.

1 Introduction

Il est aujourd'hui un consensus clair, non seulement que les corpus annotés sont indispensables aux outils de traitement automatique des langues (TAL) pour leur entraînement et leur évaluation, mais également que l'annotation doit être consistante pour être profitable (voir, par exemple (Reidsma et Carletta, 2008)). Or, l'obtention d'une annotation manuelle de qualité requiert l'utilisation d'un guide d'annotation suffisamment complet et cohérent (Nédellec *et al.*,

2006). La mise au point d'un tel guide est cependant, comme le soulignent Sampson (2000) et (Scott *et al.*, 2012), loin d'être triviale.

En outre, il est rare, une fois une campagne d'annotation terminée, que le guide d'annotation et le corpus annoté soient complètement cohérents, ce qui n'est pas sans poser problème pour les systèmes ou les linguistes utilisant le corpus (voir par exemple (Candito et Seddah, 2012), en ce qui concerne le corpus arboré du français).

Une solution pour remédier à ces deux difficultés consiste à développer le guide et à annoter le corpus selon des cycles courts de prototypage. Cette méthodologie est appelée *Agile Annotation* (Voormann et Gut, 2008) à l'image de l'*Agile Development* (voir figure 1). Elle n'a, à notre connaissance, été appliquée que dans un seul cas d'annotation réel (Alex *et al.*, 2010).

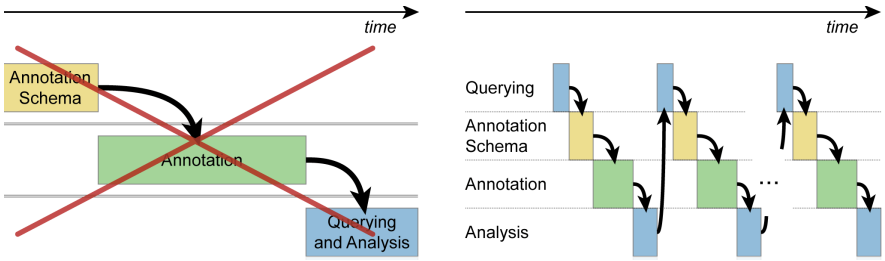


FIGURE 1 – Phases de l'annotation traditionnelle (à gauche) et cycles de l'annotation agile (à droite). Reproduction de la figure 2 de (Voormann et Gut, 2008)

Indépendamment de la notion d'annotation agile, nous avions utilisé la réécriture de graphes pour rechercher des erreurs récurrentes dans le corpus Sequoia [1] (Candito et Seddah, 2012). Cette application directe de la réécriture à la détection d'erreurs a permis d'identifier une centaine d'erreurs d'annotation et a conduit à la publication d'une nouvelle version (3.3) du corpus en juillet 2012.

Nous présentons ici les expériences que nous avons menées plus récemment dans le cadre de la correction d'annotations syntaxiques, pour laquelle nous avons transformé les instructions d'un guide d'annotation existant en règles de réécriture appliquées sur le corpus annoté. Ces expériences ont montré l'intérêt d'une telle formalisation et nous proposons donc son intégration dans le processus d'annotation manuelle, ce qui conduirait à la mise en place d'une annotation agile assistée.

2 Formaliser un guide d'annotation

La méthode que nous proposons consiste à travailler de façon systématique à partir du guide d'annotation. En effet, pour chaque type d'annotation (pour chaque relation de dépendance syntaxique dans l'exemple utilisé plus loin) le guide énumère les cas où cette annotation doit être réalisée. On utilise alors la réécriture de graphes pour repérer les occurrences des annotations correspondant à chacun des cas énumérés dans le guide. Dans un deuxième temps, on liste

1. https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia

les annotations qui n'ont pas été repérées lors de la première phase. En théorie, pour chaque annotation identifiée par cette méthode, se présente l'un des deux cas suivants :
– l'annotation est incorrecte, on doit alors corriger le corpus ;
– l'annotation est correcte et elle correspond à un cas d'usage qui n'a pas été identifié par le rédacteur du guide, on doit alors mettre le guide à jour.

Évidemment, dans le cas où le guide est mis à jour, il faut relancer le processus pour mettre en évidence d'éventuelles nouvelles incohérences entre le guide et le corpus. La principale difficulté dans la mise en place de cette méthode réside dans le passage de la version usuelle du guide à sa version formalisée en terme de réécriture de graphes.

2.1 Guide d'annotation et implicite

Un guide d'annotation est rédigé par des humains pour des lecteurs humains. Plus précisément, il est rédigé par des experts pour des lecteurs plus ou moins spécialistes en fonction de la tâche d'annotation. Le guide repose donc souvent sur des informations implicites. L'introduction décrit généralement le cadre théorique dans lequel l'annotation est réalisée. Ce cadre permet de donner les principes généraux qui s'appliquent à l'ensemble du guide. Il faut, dans la suite du document, qui décrit des parties plus spécifiques de l'annotation, connaître ces éléments généraux pour interpréter les informations correctement.

Dans le guide (Candito *et al.*, 2009), il est expliqué, d'une part que la fonction A-OBJ (figure 2) concerne des objets indirects en « *à* » et, d'autre part que cette fonction peut être réalisée par un pronom clitique. Tout lecteur francophone sait que, dans le cas de la réalisation clitique, la préposition n'est pas présente. Cette information n'est pas dans le guide mais elle doit être rendue explicite dans la règle. Dans cet exemple, il est facile de construire la bonne règle, mais en général l'information implicite est plus compliquée à formaliser.

2.2 Limites de la formalisation

Il est bien évidemment impossible de formaliser complètement le guide sous forme de règles. En effet, dans le cas contraire, cela signifierait que l'annotation peut-être faite de façon complètement automatique sans avoir recours à un jugement humain. Par exemple, dans le cas de la fonction A-OBJ (cf. figure 2), le guide indique qu'un objet indirect introduit par la préposition « *à* » peut-être annoté par une relation A-OBJ entre le verbe et la préposition, mais peut aussi dans certains cas être annoté comme un locatif (avec la relation P-OBJ_LOC). Le choix entre l'une des deux annotations se fait à l'aide d'un test basé sur la cliticisation ou sur la forme interrogative. On ne peut donc pas automatiquement détecter une erreur d'annotation qui consiste à utiliser la relation A-OBJ au lieu de P-OBJ_LOC ou l'inverse.

3 Expériences

Nous décrivons ici une première expérience d'application de notre méthodologie sur un corpus et le guide associé.

3.1 Corpus Sequoia

Il existe peu de ressources annotées syntaxiquement pour le français. Le corpus arboré du français (Abeillé *et al.*, 2003), ou French Treebank (FTB), existe depuis une dizaine d'années mais il n'est pas librement accessible et redistribuable. L'an dernier, un corpus comparable au FTB mais librement accessible a été proposé, le corpus Sequoia (Candito et Seddah, 2012). Celui-ci contient environ 3 000 phrases provenant de quatre sources différentes (Wikipédia, Parlement européen, Est Républicain et Emea). Ces phrases ont été annotées en constituants. L'annotation en constituants a ensuite été convertie en une annotation en dépendances. L'annotation en dépendances visée est décrite dans le guide [2] (Candito *et al.*, 2009).

3.2 Réécriture de graphes

Pour formaliser les informations du guide, nous utilisons GREW (Guillaume *et al.*, 2012), un outil de réécriture de graphes spécialisé pour les applications en TAL. En effet, GREW propose un langage de description riche qui permet de repérer automatiquement un motif de graphe dans un ensemble de phrases. Dans un motif, on peut exprimer des combinaisons complexes de contraintes sur les nœuds, sur les traits et sur les relations de dépendances. De plus, un motif peut être sous-spécifié et peut également exprimer des contraintes négatives sur le contexte.

La réécriture de graphes permet, une fois qu'un motif est repéré, de modifier la structure du graphe. Ici, on n'utilisera cette fonctionnalité que pour marquer chaque occurrence reconnue (à l'aide de suffixes `ok` ou `fail` sur les étiquettes de dépendances).

GREW dispose également d'un mécanisme de modules qui permet d'appliquer successivement plusieurs ensembles de règles de réécriture. Dans notre application, on utilisera deux modules : le premier pour repérer les occurrences correctes des dépendances et un second pour mettre en évidence les dépendances restantes et donc considérées comme incorrectes.

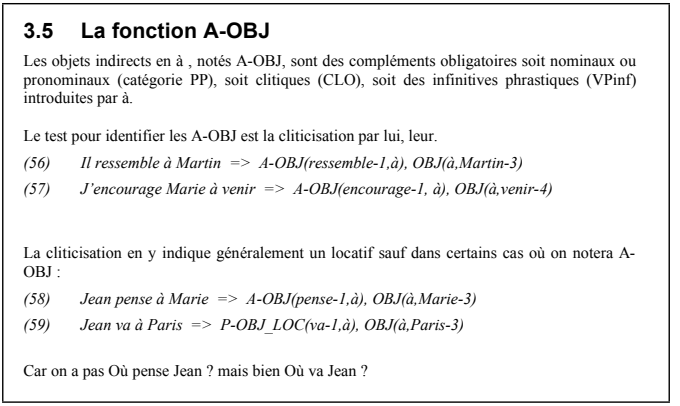
3.3 Un exemple de formalisation : la fonction A-OBJ

La section du guide spécifique à la relation A-OBJ est reproduite dans la figure 2, ci-dessous.

En général, quelques itérations sont nécessaires pour coder correctement les parties implicites ou les parties décrites ailleurs dans le guide.

1. Une traduction naïve des informations du guide nous amène à définir 4 règles : une pour chacune des réalisations possibles de l'objet indirect : un nom, un pronom clitique, un pronom non-clitique ou une proposition infinitive.
2. Si l'objet indirect est un clitique, la préposition n'est pas présente (« *Il lui parle.* ») ; il faut donc modifier la règle correspondante.
3. En cas d'élision « *au* », le lemme reste bien « *à* » mais la catégorie est P+D et non pas P ; il faut généraliser les règles.
4. Par contre, en cas d'élision « *auquel* » ; le lemme n'est « *à* » mais « *auquel* » et la catégorie P+PRO ; il faut une cinquième règle.

2. `http://alpage.inria.fr/statgram/frdep/Publications/FTB-GuideDepSurface.pdf`



3.5 La fonction A-OBJ

Les objets indirects en à , notés A-OBJ, sont des compléments obligatoires soit nominaux ou pronominaux (catégorie PP), soit clitiques (CLO), soit des infinitives phrastiques (VPinf) introduites par à.

Le test pour identifier les A-OBJ est la cliticisation par lui, leur.

(56) Il ressemble à Martin => A-OBJ(ressemble-1,à), OBJ(à,Martin-3)

(57) J'encourage Marie à venir => A-OBJ(encourage-1, à), OBJ(à,venir-4)

La cliticisation en y indique généralement un locatif sauf dans certains cas où on notera A-OBJ :

(58) Jean pense à Marie => A-OBJ(pense-1,à), OBJ(à,Marie-3)

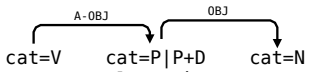
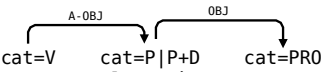
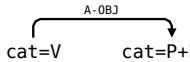
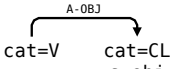
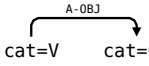
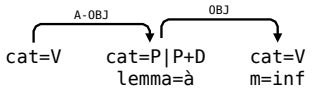
(59) Jean va à Paris => P-OBJ_LOC(va-1,à), OBJ(à,Paris-3)

Car on a pas Où pense Jean ? mais bien Où va Jean ?

FIGURE 2 – Extrait du guide d'annotation : la fonction A-OBJ

5. Pour les clitiques, le guide demande la catégorie clitique objet (avec le trait `s=obj`), mais le corpus contient des relations A-OBJ dont le dépendant est un clitique réfléchi (avec le trait `s=refl`) : *« je me pose des questions »* ; cette annotation est correcte ; il faut donc mettre à jour le guide et ajouter une règle pour ce cas.

Au final, on obtient donc les six motifs suivants :

nominal	pronominal *« à »* ou *« au »*	pronominal *« auquel »*
		
clitique objet	**clitique réflexif**	**infinitif**
		

L'application de ces motifs sur les 3 203 phrases de Sequoia donne les résultats ci-dessous [3] :

	nominal	pronominal *« à »* ou *« au »*	pronominal *« auquel »*	clitique objet	clitique réflexif	infinitif
nb d'occurrences	476	17	3	84	16	87

Il reste alors cinq occurrences de la relation A-OBJ qui ne correspondent à aucune des règles ci-dessus. Trois de ces occurrences correspondent à une erreur d'annotation :

– une erreur de POS : *« [. . .] on ne condamne pas à mort [. . .] »* avec *« mort »* adjectif ;
– dans la construction *« répondre à côté de la question »*, le groupe prépositionnel *« à côté de la question »* est un complément circonstanciel de manière, on doit donc avoir la relation MOD ;
– utilisation de la préposition *« auprès du »* dans la construction *« se renseigner auprès du comité »* : l'argument du verbe est un P-OBJ introduit par le préposition *« auprès de »* ;

3. Tous les résultats sont disponibles sur : `http://wikilligramme.loria.fr/doku.php?id=taln_2013`

Les deux autres occurrences sont correctes mais mériteraient de figurer d'une façon ou d'une autre dans le guide :
– le dépendant de la préposition a un POS inattendu : par exemple **ET** (POS pour étiqueter les mots d'origine étrangère) dans *« [. . .] délivré à The Medecine Company [. . .] »* ;
– le gouverneur de la relation A-OBJ est une coordination ;
On peut facilement imaginer le type de précisions qu'il est nécessaire d'apporter à cette partie du guide et donc le type de modifications qu'il faudra apporter aux règles à la prochaine étape pour tenir compte des deux derniers points.

4 Méthodologie proposée

L'expérience décrite ci-dessus a été réalisée sur un corpus et son guide figé : le guide n'a pas été mis à jour depuis plusieurs années et l'annotation du corpus Sequoia est terminée depuis plus d'un an. Par ailleurs, le guide n'est pas complet et il reste des sections qui ne sont pas complètement rédigées, notamment à propos de la coordination. Le corpus n'est donc pas toujours annoté de manière consistante, notamment en ce qui concerne les phénomènes non finalisés dans le guide. Le corpus Sequoia, auquel nous nous intéressons, est annoté en dépendances syntaxiques, mais l'annotation de départ et celle sur laquelle les développeurs du corpus travaillent est une annotation en constituants, et la conversion des constituants vers les dépendances est réalisée de manière automatique. Cela ajoute une difficulté dans la tâche de corrections du corpus : quand une erreur d'annotation est détectée dans les dépendances, il faut retrouver l'origine de l'erreur dans les constituants ou dans la conversion.

4.1 Intégration dans le processus d'annotation manuelle

Les expériences que nous avons menées nous ont convaincus que notre outil de réécriture de graphes peut être un allié précieux dans la recherche de cohérence entre le guide et le corpus. S'il est intéressant de l'utiliser sur des données statiques, nous pensons qu'il a un rôle encore plus important à jouer sur des données en construction. Nous proposons donc d'utiliser ce type d'outil très tôt dans le processus d'annotation, notamment au moment de la création du guide.

Dans l'idéal, chaque application de la réécriture de graphes permet de repérer des erreurs d'annotation et des erreurs, des manques ou des imprécisions dans le guide. On peut donc imaginer un processus comme celui décrit dans le schéma de la figure 3 qui représente un pas du cycle de développement menant de la version i du guide et du corpus à la version $i+1$ de ceux-ci. Par souci de simplicité, le schéma ci-dessous ne fait pas intervenir de façon explicite le travail de conversion du guide en règle de réécriture. Ce travail n'est pour autant pas trivial, comme nous l'avons vu sur notre exemple d'annotation syntaxique.

4.2 Mise en œuvre

La méthodologie d'annotation agile décrite ci-dessus coûte cher et ne peut probablement pas être appliquée tout au long d'une campagne de grande envergure. Cependant, il est possible (et souhaitable) de la mettre en œuvre lors de la phase de préparation de la campagne, en particulier

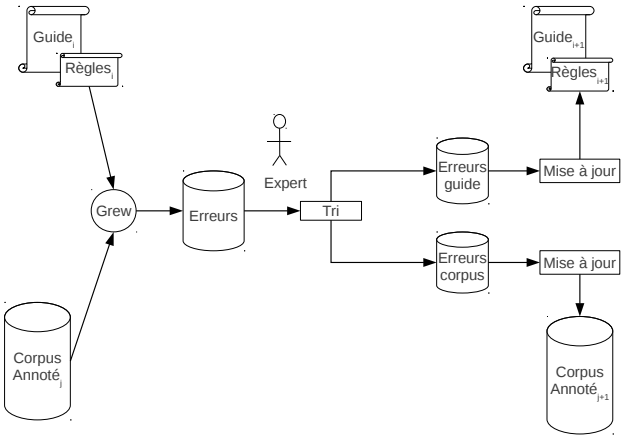


FIGURE 3 – Une itération du processus d'annotation agile

lors de la mise au point du guide, réalisée en parallèle de l'annotation d'une mini-référence (Fort, 2012). Si la mini-référence se doit d'être représentative du corpus, sa taille va largement dépendre des contraintes pratiques de la campagne (coût, disponibilité des experts). Il en va de même pour le nombre d'itérations du cycle d'annotation agile.

Pendant la phase de production, durant laquelle les annotateurs travaillent sur l'ensemble du corpus, cette méthodologie peut sans doute continuer à être utilisée, mais avec une durée de cycle beaucoup plus longue. Le repérage d'erreurs par réécriture de graphes est alors un outil supplémentaire (en complément d'une évaluation régulière, voir, là encore, (Fort, 2012)) pour le gestionnaire de la campagne, qui lui permet d'être alerté au plus tôt en cas de problème dans l'annotation.

5 Conclusion et perspectives

Nous avons proposé une méthodologie permettant d'assister l'annotation agile lors d'une campagne d'annotation, à l'aide d'un outil de réécriture de graphes. Si nous avons obtenu des résultats intéressants lors des expériences présentées ici, il reste à vérifier l'utilisabilité du système dans le cadre d'une campagne d'annotation réelle, c'est-à-dire de l'intégrer dans un cycle d'annotation.

Nous comptons donc appliquer cette méthodologie dans les mois qui viennent, pour la création de la mini-référence et la mise au point du guide, dans le cadre d'une campagne d'annotation en dépendances syntaxiques profondes du corpus Sequoia.

Pour d'autres types de campagnes d'annotation (par exemple, sémantique ou discursive), la réécriture de graphes n'est sans doute pas l'outil le plus adapté. Pour autant, une assistance à l'aide d'outils TAL, même frustres, pourrait profiter à l'annotation agile, dont le principal écueil est le coût.

Remerciements

Nous tenons à remercier Florian Besnard, étudiant à l'École des Mines de Nancy, qui a participé lors de son stage à la conversion d'une partie du guide en règles de réécriture.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for French. *In* Abeillé, A., éditeur : *Treebanks*, pages 165 –187. Kluwer, Dordrecht.

Alex, B., Grover, C., Shen, R. et Kabadjov, M. (2010). Agile corpus annotation in practice : An overview of manual and automatic annotation of CVs. *In Proceedings of the Fourth Linguistic Annotation Workshop (LAW)*, pages 29–37, Uppsala, Suède. Association for Computational Linguistics.

Candito, M., Crabbé, B. et Falco, M. (2009). Dépendances syntaxiques de surface pour le français. Rapport technique, Université Paris 7.

Candito, M. et Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France.

Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Thèse de doctorat, Université Paris XIII, LIPN, INIST-CNRS.

Guillaume, B., Bonfante, G., Masson, P., Morey, M. et Perrier, G. (2012). Grew : un outil de réécriture de graphes pour le TAL. *In Actes de Conférence annuelle sur le Traitement Automatique des Langues (TALN)*, Grenoble, France.

Nédellec, C., Bessières, P., Bossy, R., Kotoujansky, A. et Manine, A.-P. (2006). Annotation guidelines for machine learning-based named entity recognition in microbiology. *In* et C. Nédellec, M. H., éditeur : *Proceedings of the Data and text mining in integrative biology workshop*, pages 40–54, Berlin, Allemagne.

Reidsma, D. et Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Sampson, G. (2000). The role of taxonomy in language engineering. *Philosophical Transactions of the Royal Society of London. Series A :Mathematical, Physical and Engineering Sciences*, 358(1769): 1339–1355.

Scott, D., Barone, R. et Koeling, R. (2012). Corpus annotation as a scientific task. *In International Conference on Language Resources and Evaluation*, Istanbul, Turquie.

Voormann, H. et Gut, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251.