Vers l'annotation par le jeu de corpus (plus) complexes : le cas de la langue de spécialité

Karën Fort[1] Bruno Guillaume[2] Nicolas Lefèbvre[2] Laura Ramírez[3]
Mathilde Regnault[3] Mary Collins[3] Oksana Gavrilova[3] Tanti Kristanti[3]
(1) STIH - EA 4509, Université Paris-Sorbonne, France
(2) Sémagramme - Inria Nancy Grand Est, France
(3) Université Paris-Sorbonne, France
(1,3) prenom.nom@paris-sorbonne.fr, (2) prenom.nom@inria.fr

RÉSUMÉ

Nous avons précédemment montré qu'il est possible de faire produire des annotations syntaxiques de qualité par des participants à un jeu ayant un but. Nous présentons ici les résultats d'une expérience visant à évaluer leur production sur un corpus plus complexe, en langue de spécialité, en l'occurrence un corpus de textes scientifiques sur l'ADN. Nous déterminons précisément la complexité de ce corpus, puis nous évaluons les annotations en syntaxe de dépendances produites par les joueurs par rapport à une référence mise au point par des experts du domaine.

ABSTRACT

Towards (more) complex corpora annotation using a game with a purpose : the case of scientific language.

We previously showed that participants in a game with a purpose can produce quality dependency syntax annotations. We present here the results of an experiment aiming at evaluating their production on a more complex corpus of scientific texts on DNA. We precisely describe the complexity of the corpus, then we evaluate the annotations produced by the players as compared to a gold corpus, annotated and adjudicated by experts of the domain.

MOTS-CLÉS : annotation en syntaxe de dépendances, *crowdsourcing*, jeux ayant un but.

KEYWORDS: dependency syntax annotation, crowdsourcing, games with a purpose.

1 Introduction

La création de ressources langagières de qualité, indispensables au développement et à l'évaluation des outils de traitement automatique des langues (TAL), représente un coût élevé. Ce coût est rarement rendu public, mais l'un des rares cas chiffré est le corpus arboré de Prague, dont la construction a été évaluée à 600 000 dollars (Böhmová *et al.*, 2001).

Les jeux ayant un but ont montré leur efficacité dans la production de lexiques et de corpus annotés, en termes à la fois de quantité et de qualité produites (Chamberlain *et al.*, 2013; Lafourcade *et al.*, 2015). Plus intéressant encore, au-delà de la mise en œuvre des connaissances du monde et de la langue des locuteurs, comme dans `JeuxDeMots` (Lafourcade & Joubert, 2008) ou `Phrase Detectives` (Poesio *et al.*, 2013), les jeux permettent de former les participants à des tâches

complexes, par exemple le repliement de protéines dans FoldIt (Khatib *et al.*, 2011) ou l'annotation en syntaxe de dépendances dans ZombiLingo (Guillaume *et al.*, 2016).

Nous souhaitons étudier l'évolution de la qualité de la production des joueurs lorsque la complexité de la tâche est accrue. Pour ce faire, nous nous proposons de faire annoter en syntaxe de dépendances des corpus réputés plus difficiles, notamment : i) un corpus de langue de spécialité, ii) un corpus de parole transcrite et iii) un corpus de contenu généré par l'utilisateur (extrêmement bruité, de type forum). Il n'existe à notre connaissance aucune expérience de ce type relatée dans la littérature, que ce soit pour le français ou pour une autre langue.

Nous présentons ici la première expérience que nous avons menée sur un corpus de langue de spécialité, en l'occurrence des textes portant sur l'ADN, donc touchant au biomédical et aux biotechnologies. Dans un premier temps, nous détaillons la création de ce corpus, puis les conditions de l'expérience, avant de présenter les résultats obtenus et de discuter les biais potentiels de l'étude.

2 Construire un corpus de langue de spécialité

2.1 Sélection des textes

Le corpus utilisé dans cette expérience a été constitué en utilisant le moteur de recherche du site de Creative Commons [1]. Notre but était de créer un corpus i) librement disponible, ii) en langue de spécialité, iii) à partir d'échantillons suffisamment hétérogènes pour représenter la diversité de la langue de spécialité choisie et iv) d'environ 800 phrases, afin d'être comparable aux autres corpus de la plateforme.

Le corpus ADN comprend au final 22 textes en français sur le thème de l'ADN, soit en tout 792 phrases et 25 690 tokens. Il est constitué principalement d'extraits ou de résumés d'articles de recherche (18 textes, soit 21 132 tokens), complétés par deux articles de vulgarisation, un cours et un texte didactique. Le corpus est encodé en UTF-8 : les erreurs d'encodage et les césures ont été corrigées. Il est balisé en XML et librement disponible, sous licence CC BY-NC-SA [2].

2.2 Évaluation de la complexité du corpus

2.2.1 Méthodologie

Si le corpus créé semble d'évidence complexe, encore faut-il s'en assurer et tenter de préciser en quoi. De nombreuses études linguistiques ont porté sur la complexité de la langue, cependant, bien peu fournissent des métriques réellement utilisables dans notre cas, c'est-à-dire sans information sur le temps de lecture (Vasishth, 2003), sur les mouvements oculaires (Lee *et al.*, 2007) et dans un but qui n'est pas de simplifier le texte (Seretan, 2012; Brouwers *et al.*, 2014). Au vu des informations à notre disposition, nous avons décidé de nous concentrer sur la complexité lexicale et syntaxique des phrases en faisant l'hypothèse qu'elles sont un prédicteur suffisamment fiable de la complexité au sens large.

Pour la syntaxe, nous nous sommes largement inspirés de (Blache, 2010). Parmi les mesures listées

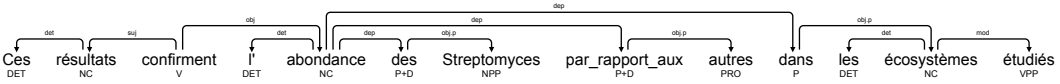
1. Disponible ici : https://search.creativecommons.org/.
2. Voir : https://zombilingo.org/informations#export.

dans cet article, nous nous inspirons de deux qui peuvent être calculées automatiquement à partir de structures en dépendances syntaxiques : i) la profondeur et ii) la longueur moyenne des dépendances. L'idée de considérer la profondeur des structures en dépendances était déjà présente dans les travaux de Gibson (Gibson, 1998). Morrill et Gavarró (Morrill & Gavarró, 2004) ont montré la corrélation de cette mesure avec les difficultés de compréhension liées à l'aphasie.

Nous comparons, selon ces dimensions, le corpus ADN à trois autres corpus utilisés dans le même dispositif, dont un est un extrait du *Corpus journalistique issu de l'Est Républicain* (ATILF, 2011) et deux corpus issus de la Wikipédia. Un de ces deux corpus a pour thème le football, l'autre les Pokémon. Ils ont été mis en jeu à l'occasion d'événements liés à ces sujets (respectivement, Euro 2016 et sortie en France de Pokémon Go).

Techniquement, nous calculons les indices de complexité syntaxique comme suit :

Profondeur. La profondeur quantifie la complexité en suivant l'*hypothèse de la dépendance incomplète*. Pour une phrase donnée, on calcule le nombre maximum de dépendances incomplètes quand on parcourt la phrase. Intuitivement, cela correspond au nombre d'éléments que l'on doit garder en mémoire en lisant la phrase pour pouvoir la comprendre. Pour calculer cette mesure, on ne tient ici pas compte des relations de ponctuation. Ainsi, dans l'analyse en dépendances ci-dessous, la profondeur est de 3 (maximum atteint deux fois entre *abondance* et *des* et entre *des* et *Streptomyces*).



Les phrases sont analysées par deux analyseurs syntaxiques, Talismane (Urieli, 2013) et FrDep-Parse (Guillaume & Perrier, 2015), et nous sélectionnons le maximum des valeurs obtenues sur chacune des deux analyses. Pour finir, nous calculons une moyenne de cette mesure pour l'ensemble des phrases du corpus.

Longueur moyenne. Dans (Blache, 2010), le *coût d'intégration* est défini comme le nombre de référents du discours entre deux têtes rattachées par une dépendance. Ici, nous donnons une approximation de cette mesure en considérant la distance moyenne (en nombre de tokens) entre la tête et le gouverneur pour une relation donnée.

Enfin, pour estimer la complexité lexicale des corpus considérés, nous comptabilisons les tokens n'apparaissant pas dans un lexique standard des formes fléchies du français (*token inconnus*). Nous utilisons pour cela les résultats produits par MElt (Denis & Sagot, 2010), qui préfixe d'une étoile les lemmes quand la forme fléchie correspondante n'existe pas dans le lexique Le*fff* (Sagot, 2010) [3]. Nous calculons également le nombre de tokens inconnus différents.

2.2.2 Analyse des résultats

Les tableaux 1 et 2 présentent les tailles des corpus, ainsi que les résultats des mesures de ces indices de complexité syntaxique et lexicale.

Les indices de complexité syntaxique utilisés montrent que le corpus ADN présente une plus grande profondeur moyenne des dépendances (4,20, Pokémon atteignant 4,06) et des relations pour la plupart plus longues que pour les autres corpus, ou de longueurs proches du maximum (pour A_OBJ

3. Nous avons retiré les cardinaux (en chiffres et en lettres) du décompte, car MElt les préfixe d'une étoile.

	Prof.	Longueur moyenne								
	moy	mod	mod.rel	suj	obj	de_obj	a_obj	p_obj.o	coord	dep.coord
ADN	**4,20**	**3,92**	**7,09**	2,70	**2,90**	**2,70**	2,40	**2,41**	**6,44**	2,23
Foot	3,92	3,30	5,61	**4,19**	2,83	1,50	2,35	1,87	4,59	2,29
Pokémon	4,06	3,43	5,81	3,92	2,42	2,10	1,77	1,76	5,69	**2,94**
Est Rép.	3,40	3,43	5,79	3,55	2,59	2,16	**2,51**	2,00	4,32	2,22

TABLE 1 – Indices de complexité syntaxique.

et DEP.COORD). Seule exception, les sujets, pour lesquels ADN présente la plus petite longueur de relation [4].

	Nb de phrases	Nb de mots	Tokens par phrases	% Tokens inconnus	Nb de tokens inconnus différents
ADN	771	25 160	**32,63**	4,92 %	**624**
Foot	825	21 439	25,99	3,17 %	413
Pokémon	804	24 280	30,20	**6,19 %**	485
Est Rép.	3 536	73 174	20,69	4,55 %	~~3 326~~ 540 (800 phrases)

TABLE 2 – Indices de complexité lexicale.

En ce qui concerne la complexité lexicale, les résultats sont là encore assez nets, avec une longueur moyenne de phrase supérieure pour ADN (32,63 tokens par phrase). La proportion de tokens inconnus par phrase (4,92 %) n'est supérieure que pour Pokémon (6,19 %), ce qui s'explique par le nombre très élevé de noms de Pokémons (inconnus du lexique) dans le corpus. Par ailleurs, le corpus ADN présente le plus grand nombre de tokens inconnus différents (624). Pour le corpus Est Rép. le nombre de tokens inconnus différents est bien supérieur (3 326) du fait de la taille du corpus. Si on sélectionne aléatoirement 800 phrases dans ce corpus pour obtenir une taille comparable, ce nombre devient alors inférieur (540). Ces mesures de complexité confirment donc notre intuition de départ : le corpus ADN est plus complexe que les autres corpus annotés précédemment par le jeu.

3 Conditions de l'expérience

3.1 ZombiLingo

Le jeu utilisé dans cette expérience, ZombiLingo, est présenté en détail dans (Guillaume *et al.*, 2016). Il a permis de faire produire plus de 300 000 annotations en syntaxe de dépendances à plus de 1 000 joueurs [5] sur des corpus en français, qui sont librement disponibles sur le site du jeu [6].

Afin de remobiliser les joueurs, nous lançons régulièrement des défis, lors desquels les participants jouent (donc annotent) pendant deux semaines des corpus sur un thème (le football, les Pokémons, etc.). Un classement spécifique est affiché et une publicité adaptée est réalisée sur les réseaux sociaux.

4. On peut faire l'hypothèse que cela est lié au style scientifique, mais cela reste à vérifier et n'est pas l'objet de notre expérience.

5. Données à la date du 17 avril 2017.

6. Voir : https://zombilingo.org/informations#export.

Ainsi, le défi ADN a eu lieu du 14 au 30 novembre 2016, mobilisant 232 joueurs, qui ont produit 25 039 annotations.

3.2 Évaluation

Un sous-corpus ADN−EVAL du corpus ADN a été constitué pour évaluer les annotations produites par le jeu. ADN−EVAL contient 39 phrases parmi les 792 du corpus ADN, soit 1 245 tokens (environ 5 % du corpus). Ces phrases ont été annotées sur l'outil WebAnno (Yimam *et al.*, 2013) par des étudiantes en Master 2 de linguistique et informatique à l'université Paris-Sorbonne. La majorité a été annotée par deux annotatrices, puis les annotations divergentes ont fait l'objet d'une adjudication par un chercheur spécialiste de la syntaxe (B. Guillaume). Certaines annotations n'ont été réalisées que par une seule annotatrice (par manque de temps) et ont été directement corrigées par ce même chercheur.

Dans le jeu, les corpus de référence sont utilisés de trois façons différentes (voir figure 1) :

— Pour la **formation** : avant de commencer à annoter une relation, un joueur doit suivre une formation et s'entraîner sur des phrases de référence issues du corpus $REF_{Form.\&Ctrl}$ pour lesquelles on lui donne la bonne réponse en cas d'erreur.

— Pour le **contrôle** : en cours de jeu, de temps en temps, une phrase de référence issue du même corpus est proposée au joueur. S'il se trompe, la correction lui est indiquée. Cela permet de contrôler que les joueurs n'ont pas oublié les consignes.

— Pour l'**évaluation** : les phrases d'évaluation sont traitées comme le reste du corpus pendant le déroulement du jeu. C'est *a posteriori* que le corpus ADN−EVAL est utilisé pour l'évaluation.

On peut utiliser la même partie du corpus pour la formation et le contrôle mais bien évidemment, l'évaluation se fait sur une partie différente du corpus de référence.

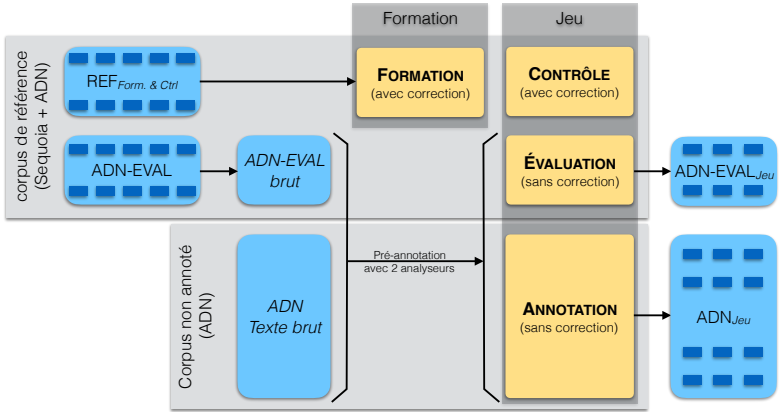


FIGURE 1 – Méthodologie d'évaluation.

	# jouable	densité	FrDep-Parse	Talismane	ZombiLingo
Tout sauf PONCT	292	8,7	0,772	0,858	**0,859**
MOD	155	1,6	0,623	0,759	**0,773**
COORD	28	4,7	0,521	0,474	**0,718**
OBJ	13	13,5	0,857	0,872	**0,901**
OBJ.P	13	17,6	0,931	**0,978**	0,944
DEP.COORD	12	9,5	0,838	0,877	**0,883**
DET	12	31,3	0,946	**0,978**	0,958
P_OBJ.O/P_OBJ.AGT	11	12,0	0,720	**0,815**	0,774

TABLE 3 – F-mesure pour le corpus ADN−EVAL complet, puis par relation.

4 Résultats obtenus

4.1 Annotations réalisées

Le tableau 3 présente le résultat de l'évaluation des phrases annotées par les deux parseurs FrDep-Parse (Guillaume & Perrier, 2015) et Talismane (Urieli, 2013), puis par ZombiLingo. La densité est le nombre moyen de joueurs qui ont donné leur avis par item jouable. Nous observons que le jeu produit des annotations de qualité similaire au parseur Talismane (0,859 de F-mesure). Afin d'interpréter plus finement ce résultat, nous le détaillons par relation (pour les relations où le nombre d'items jouables est supérieur à 10) [7]. Les scores sont plus élevés pour les coordinations, en particulier en ce qui concerne la relation COORD (+0,197) pour la partie gauche de la coordination. On observe également un meilleur score sur la relation MOD, mais la différence (+0,014) n'est pas significative compte-tenu de la taille du corpus de référence. Cela correspond à l'intuition que l'annotation des coordinations ou des rattachements prépositionnels est difficile pour les analyseurs automatiques, moins pour les joueurs.

Dans le format FTB/Sequoia, sur lequel est basé le jeu, la relation MOD est utilisée dans des contextes différents (entre noms et adjectifs, entre verbes et adverbes, entre noms, entre verbes et prépositions, etc.). En revanche, le lien entre noms et prépositions dans le cas d'un groupe prépositionnel modifiant un nom est codé avec la relation DEP [8]. Il est important de noter que la relation MOD représente plus de la moitié des items jouables (155 sur 292), le score sur cette relation représente donc une part importante du score global.

Dans (Guillaume *et al.*, 2016), nous avions utilisé une partie du corpus Sequoia comme corpus d'évaluation. Cette expérience avait montré que, sur un corpus non-spécialisé, la qualité des annotations obtenues avec le jeu était meilleure que celles obtenues avec les analyseurs, dès lors que les annotations étaient jouées par un nombre assez grand de joueurs. Sur un corpus de langue de spécialité, et bien que la densité d'annotation soit nettement supérieure à 1 partout, nous ne retrouvons pas des résultats aussi nets.

7. La relation P_OBJ.AGT n'est pas gérée par Talismane, nous l'avons donc regroupée avec la relation P_OBJ.O pour avoir des données comparables entre les deux parseurs.

8. Dans les données, il y a 13 cas où l'export contient une relation MOD entre un nom et une préposition alors que la bonne solution est la relation DEP. Nous avons considéré ces cas comme corrects car c'est un changement systématique lié à un choix d'annotation dans le format FTB/Sequoia.

4.2 Discussion

L'analyse de la complexité du corpus réalisée pour cette expérience est limitée par les indices dont nous disposions. Les indices choisis nous semblent cependant suffisants pour confirmer l'intuition selon laquelle ce corpus est plus difficile à annoter que les corpus `Wikipédia` utilisés précédemment. Une analyse plus fine de la difficulté d'annotation devrait prendre en compte plus de paramètres et notamment la complexité de la tâche d'annotation telle que définie dans (Fort *et al.*, 2012).

D'autre part, l'expérience présente un biais lié à la petite taille du corpus de référence utilisé (39 phrases, 1 245 tokens). Un travail d'annotation est en cours pour obtenir une référence sur un corpus de plus grande taille pour permettre une analyse plus fiable de la qualité produite.

En ce qui concerne le profil des joueurs, une première étude générale (non spécifique à ce corpus) a été réalisée (Fort *et al.*, 2017), qui montre que ceux-ci ont un niveau d'études élevé, puisque 75 % des gros joueurs ont au moins un niveau Master. Pour autant, ils ne sont probablement pas spécialistes du biomédical ou des biotechnologies (40 % des gros joueurs viennent du domaine du TAL ou de la linguistique).

5 Conclusion

Cette étude présente des résultats partiels concernant l'annotation en syntaxe de dépendances à l'aide d'un jeu ayant un but dans le cas d'un corpus complexe. Si pour l'instant, le jeu n'a pas permis d'obtenir de meilleurs annotations que les analyseurs, on observe que certaines relations comme la coordination sont mieux annotées par les joueurs, ce qui est très prometteur.

Nous pensons cependant qu'il est possible d'aller plus loin et d'améliorer ces résultats, notamment en ce qui concerne les relations MOD et DEP. Plusieurs joueurs nous ont en effet fait part de leur difficulté à jouer la relation MOD. Nous avons donc prévu de modifier la présentation de ces relations dans le jeu pour les rendre plus intuitives pour les joueurs. Nous comptons séparer les différents usages de la relation MOD et regrouper les relations MOD et DEP quand elles introduisent un groupe prépositionnel. Ces modifications sont en cours et une nouvelle version du jeu va voir le jour en juin 2017.

Références

ATILF (2011). Corpus journalistique issu de l'est républicain. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BLACHE P. (2010). Un modèle de caractérisation de la complexité syntaxique. In *Traitement Automatique des Langues Naturelles*, p. 1–10, Montréal, Canada.

BÖHMOVÁ A., HAJIČ J., HAJIČOVÁ E. & HLADKÁ B. (2001). The prague dependency treebank : Three-level annotation scenario. In A. ABEILLÉ, Ed., *Treebanks : Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

BROUWERS L., BERNHARD D., LIGOZAT A.-L. & FRANÇOIS T. (2014). Syntactic sentence simplification for french. In *The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014)*.

CHAMBERLAIN J., FORT K., KRUSCHWITZ U., LAFOURCADE M. & POESIO M. (2013). Using games to create language resources : Successes and limitations of the approach. In I. GUREVYCH & J. KIM, Eds., *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, p. 3–44. Springer Berlin Heidelberg.

DENIS P. & SAGOT B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada.

FORT K., GUILLAUME B. & LEFÈBVRE N. (2017). Who wants to play Zombie ? A survey of the players on ZOMBILINGO. In *Games4NLP 2017 - Using Games and Gamification for Natural Language Processing*, Symposium Games4NLP, Valence, Espagne.

FORT K., NAZARENKO A. & ROSSET S. (2012). Modeling the complexity of manual annotation tasks : a grid of analysis. In *Actes de International Conference on Computational Linguistics (COLING)*, p. 895–910, Mumbai, Inde.

GIBSON E. (1998). Linguistic complexity : locality of syntactic dependencies. *Cognition*, **68**, 1–76.

GUILLAUME B., FORT K. & LEFEBVRE N. (2016). Crowdsourcing complex language resources : Playing to annotate dependency syntax. In *Actes de International Conference on Computational Linguistics (COLING)*, Osaka, Japon.

GUILLAUME B. & PERRIER G. (2015). Dependency Parsing with Graph Rewriting. In *Proceedings of IWPT 2015, 14th International Conference on Parsing Technologies*, p. 30–39, Bilbao, Espagne.

KHATIB F., COOPER S., TYKA M. D., XU K., MAKEDON I., POPOVIĆ Z., BAKER D. & PLAYERS F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France.

LAFOURCADE M., LEBRUN N. & JOUBERT A. (2015). *Jeux et intelligence collective : résolution de problèmes et acquisition de données sur le web*. Collection science cognitive et management des connaissances. ISTE.

LEE Y., LEE H. & GORDON P. C. (2007). Linguistic complexity and information structure in korean : Evidence from eye-tracking during reading. *Cognition*, **104**(3).

MORRILL G. & GAVARRÓ A. (2004). On aphasic comprehension and working memory load. In *Categorial Grammars*, p. 259–287, Montpellier, France.

POESIO M., CHAMBERLAIN J., KRUSCHWITZ U., ROBALDO L. & DUCCESCHI L. (2013). Phrase detectives : Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, **3**(1), 3 :1–3 :44.

SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

SERETAN V. (2012). Acquisition of syntactic simplification rules for french. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II le Mirail, France.

VASISHTH S. (2003). Quantifying processing difficulty in human sentence parsing : The role of decay, activation, and similarity-based interference. In *Proceedings of Eurocogsci 03 : The European Cognitive Science Conference*.

YIMAM S. M., GUREVYCH I., ECKART DE CASTILHO R. & BIEMANN C. (2013). Webanno : A flexible, web-based and visually supported system for distributed annotations. In *Actes de Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 1–6, Sofia, Bulgarie : Association for Computational Linguistics.