Simplification de schémas d'annotation : un aller sans retour ?

Cyril Grouin
CNRS, LIMSI
Université Paris-Saclay
F-91405 Orsay
cyril.grouin@limsi.fr

RÉSUMÉ

Dans cet article, nous comparons l'impact de la simplification d'un schéma d'annotation sur un système de repérage d'entités nommées (REN). Une simplification consiste à rassembler les types d'entités nommées (EN) sous deux types génériques (personne et lieu), l'autre revient à mieux définir chaque type d'EN. Nous observons une amélioration des résultats sur les deux versions simplifiées. Nous étudions également la possibilité de retrouver le niveau de détail des types d'EN du schéma d'origine à partir des versions simplifiées. L'utilisation de règles de conversion permet de recouvrer les types d'EN d'origine, mais il reste une forme d'ambiguïté contextuelle qu'il est impossible de lever au moyen de règles.

ABSTRACT

Annotation scheme simplification : a one way trip with no return ?

In this paper, we study the impact of annotation scheme simplification on named entity recognition (NER) performances. One simplification consists in merging all named entity (NE) types into two main types (person and location), while the other simplification relies on a better definition of all NE types. We achieved better results on the two simplified versions of the annotation scheme. We also study the ability to recover the original NE types from the simplified versions. The use of post-processing rules allows to recover a few original NE types. Nevertheless, we faced with a kind of contextual ambiguity which seems hard to process using rules.

MOTS-CLÉS : Entités nommées, schéma d'annotation, simplification.

KEYWORDS: Named entities, annotation scheme, simplification.

1 Introduction

L'annotation de corpus est un processus long et coûteux, mais utile pour construire des ressources, réaliser des analyses linguistiques, entraîner des modèles d'apprentissage statistique, ou pour évaluer les sorties de systèmes. Leech (1993) rappelle qu'un processus d'annotation comprend plusieurs étapes : écriture et mises à jour du guide d'annotation, entraînement des annotateurs humains, double annotation et phases d'adjudication. Plusieurs solutions existent pour réduire le coût, le temps et améliorer la qualité des annotations : pré-annotation automatique (Dandapat *et al.*, 2009; Fort & Sagot, 2010), apprentissage actif par Ganchev *et al.* (2007) sur du repérage d'entités nommées (REN) et par Voutilainen (2012) et Yimam *et al.* (2014) sur de l'étiquetage en parties du discours, annotation agile (Alex *et al.*, 2010), ou encore propagation contrôlée d'annotations existantes (Grouin, 2016).

Alors que ces solutions préservent le schéma d'annotation, une autre piste de simplification du processus d'annotation consiste à modifier le schéma par fusion de types ou redéfinition. Wilson & Thomas (1997, p. 54) rappellent que les schémas sémantiques relèvent d'un compromis entre le souhait de refléter l'organisation des mots dans l'esprit humain, et la nécessité d'avoir des annotations utiles pour les chercheurs, cette utilité étant guidée par la tâche. Si un schéma d'annotation n'est jamais défini en fonction de possibilités techniques, il est possible de tenir compte des propriétés des données (distribution, ambiguïté, etc.) pour améliorer la qualité des annotations. Un équilibre doit être trouvé entre simplification et possibilité de recouvrer le niveau de détail du schéma d'origine.

Dans cet article, nous évaluons l'impact de la complexité d'un schéma d'annotation sur les performances d'un système de reconnaissance d'entités nommées. Depuis un schéma d'annotation complexe (nombre élevé de catégories, ambiguïtés entre catégories, prise en compte du contexte pour décider d'annoter et déterminer l'étiquette à utiliser, etc.), nous réduisons la complexité en fusionnant des types existants ou en proposant de nouvelles définitions pour les types difficiles à traiter. Nous évaluons les résultats obtenus par des modèles entraînés sur ces versions, et la possibilité de retrouver les types du schéma d'origine après simplification.

2 État de l'art

Travailler sur les schémas d'annotation peut viser un objectif d'interopérabilité (Blache *et al.*, 2010) ou d'analyse de l'impact d'un schéma sur des outils du TAL appliqués sur ces annotations, en particulier pour démontrer l'impact des propriétés du corpus et des différences d'annotations (définitions, nombre de types, portions annotées, coordination d'entités). Plusieurs travaux ont déterminé le seuil au-delà duquel un schéma d'annotation devenait trop complexe.

Sur une tâche d'analyse de dépendances, Mille *et al.* (2012) ont graduellement ajouté des relations grammaticales au schéma d'annotation du Penn Treebank (15, 31, 44 et 60 relations) et ont comparé les performances sur quatre analyseurs de dépendances. Ils observent qu'un schéma d'annotation linguistiquement plus riche n'implique pas forcément une baisse d'exactitude (moins de 0,9) entre 15 et 44 relations alors que les différences sont plus marquées avec le schéma composé de 60 relations.

En reconnaissance d'entités nommées, Shmanina *et al.* (2013) font l'observation inverse. Les auteurs ont comparé les performances de Banner (Leaman & Gonzales, 2008) sur un corpus annoté en suivant deux schémas d'annotations des maladies. Avec une validation croisée 10-plis, l'outil obtient de moins bons résultats sur le corpus annoté avec le schéma d'annotation complexe (F-mesure de 0,7365) qu'avec un schéma d'annotation plus simple (F-mesure de 0,9164).

3 Corpus et méthodes

Nos expériences reposent sur le corpus produit pour la tâche de désidentification de la campagne d'évaluation i2b2/UTHealth 2014 NLP Challenge (Stubbs *et al.*, 2015). Ce corpus comprend 1304 documents cliniques rédigés en anglais (de 2 à 5 par patient), pour un total de 296 patients différents. Les annotations couvrent 7 types principaux (Age, Contact, Date, ID, Location, Name, et Profession) et 25 sous-types (doctor, patient, username pour Name ; street, state, city, zip, hospital, organization pour Location). Ces annotations ont ensuite été remplacées en corpus par des données fictives réalistes.

Nous nous intéressons à deux principaux types d'entités nommées : les noms de personnes en raison de leur complexité sémantique (différence entre nom et prénom, ou entre un médecin et un patient), et les noms de lieu en raison de leur composition d'éléments de types différents, en particulier sur les adresses et noms d'hôpitaux qui sont complexes à traiter pour des systèmes de REN lorsque ces noms font référence à des personnes ou à d'autres lieux.

Nos expériences reposent sur trois versions seulement d'un schéma d'annotation (voir Figure 1) :

- la version d'origine (v1) composée de différences sémantiques (distinction médecin/patient) qui peuvent être ambiguës puisque ces entités sont de type "nom de personne" composées de prénoms et de noms [1]. Cette version se fonde sur des types d'entités reposant sur des définitions humaines, ces types étant conçus pour être directement exploitables en sortie (ici, une différence entre médecins et patients) ;
- une version simplifiée (v2) composées de types d'entités globaux (personnes et lieux) qui font généralement consensus dans la communauté du REN ;
- une version plus régulière (v3) pour laquelle la fonction d'un élément ne détermine pas la catégorie :
  - les prénoms, noms de familles, de villes et d'États sont toujours étiquetés comme 'first', 'last', 'city' et 'state', même s'ils sont utilisés comme composants d'une adresse ou d'un nom d'hôpital *(86 Paris Rd ; 89 James Street ; Cooley Dickinson Hospital ; Johnsonville Family Clinic ; Southwest Texas Medical Center ; Zucker Hillside Hospital)* ;
  - tous les chiffres sont étiquetés 'number' (y compris dans les adresses et codes postaux) ;
  - les initiales de médecins et les codes internes aux hôpitaux *(XD ; MAL60 ; lc855 ; ullmann)* sont rassemblés sous l'étiquette 'code'.

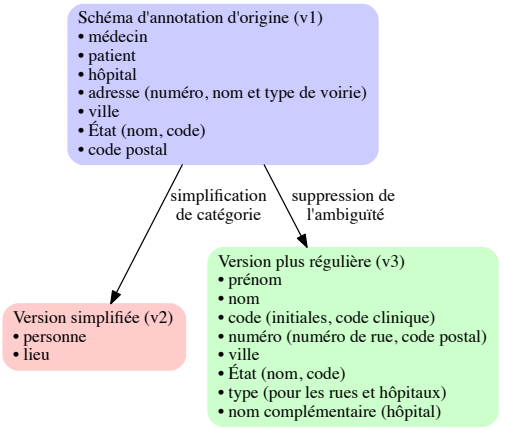

FIGURE 1: Évolution du schéma d'annotation, de la version d'origine (v1) vers les versions simplifiée (v2) et plus régulière (v3)

Type	Entraînement (521 fichiers)			Test (514 fichiers)		
	v1	v2	v3	v1	v2	v3
Person	–	2809	–	–	2791	–
Doctor	1932	–	–	1913	–	–
Patient	879	–	–	879	–	–
First	–	–	1661	–	–	1670
Last	–	–	2663	–	–	2586
Code	–	–	308	–	–	217
Location	–	1695	–	–	1610	–
Hospital	928	–	–	875	–	–
Address	144	–	–	136	–	–
City	259	–	702	264	–	721
State	222	–	243	190	–	205
Zip	139	–	–	140	–	–
Number	–	–	280	–	–	273
Name	–	–	692	–	–	624
Type	–	–	507	–	–	467

TABLE 1: Nombre d'annotations par type dans les corpus d'entraînement et de test pour chaque version du schéma d'annotation

1. Il n'est pas possible de savoir si *Paul Martin* est un médecin ou un patient sans analyser le contexte (sauf en cas de médecin ou de patient connu).

Les trois versions du schéma d'annotation sont présentées sur la phrase 1 pour les noms de personnes et la phrase 2 pour les lieux. Les boîtes extérieures (en rouge) renvoient à la v2, les boîtes intermédiaires (en bleu) à la v1, et les boîtes intérieures (en vert) à la v3.

(1) Ms. person patient first Michelle last Klein was seen in general neurology clinic today following her recent admission for complex migraine. Dr. person doctor first Remigio L. last Allison was present for all salient aspects of the history and physical exam.

(2) Internal Medicine.
loc street number 86 city Paris type Rd
loc city city Washington , loc state state DC loc zip number 20006 .

Afin de produire les versions 2 (simplifiée) et 3 (plus régulière), nous avons converti la version d'origine du schéma d'annotation au moyen des heuristiques suivantes :

— pour produire la v2, nous avons rassemblé les catégories 'Doctor' et 'Patient' sous 'Person', et de manière similaire pour 'Street', 'City', 'State' et 'Zip' sous 'Location' ;

— pour produire la v3, nous avons appliqué des règles définies empiriquement pour réintroduire les prénoms et noms de famille à la place de la distinction 'Patient' et 'Doctor' [2], et scindé les adresses en 'Number', 'Name', et 'Type' [3]. Un contrôle manuel de ces modifications automatiques a été réalisé en deux temps (4h12 pour la première vérification et 1h06 pour la deuxième).

Le tableau 1 présente le nombre total d'annotations par type dans chaque version du schéma d'annotation sur les corpus d'entraînement (521 fichiers) et de test (514 fichiers).

Méthodes Pour chaque version du schéma d'annotation, nous avons identifié les entités nommées grâce à un système d'apprentissage statistique, puis appliqué des règles de conversion pour retrouver les types d'entités nommées de la version d'origine. Le schéma 2 résume les différentes étapes suivies.

Nous avons utilisé l'outil NeuroNER fondé sur des réseaux de neurones récurrents (bi-LSTM) pour identifier les entités nommées. Cet outil prend en entrée un corpus annoté au format BRAT ou ConLL et produit une sortie annotée en entités nommées. Nous avons conservé la configuration d'origine de l'outil et renvoyons à Dernoncourt *et al.* (2017) pour de plus amples détails sur l'architecture du réseau de neurones.

2. Dans les documents traités, le prénom précède toujours le nom (« Michelle Klein ») sauf si le premier élément est suivi d'une virgule, auquel cas le nom précède le prénom (« Glenn, Olivia »). Les règles visent à identifier la frontière entre prénom et nom en fonction du nombre de tokens dans la portion et en tenant compte de la présence d'initiales (« Remigio L. Allison »).

3. Ce découpage repose sur des listes (hôpitaux, États), des déclencheurs (types d'hôpitaux et de voiries) et des règles (éléments numériques).

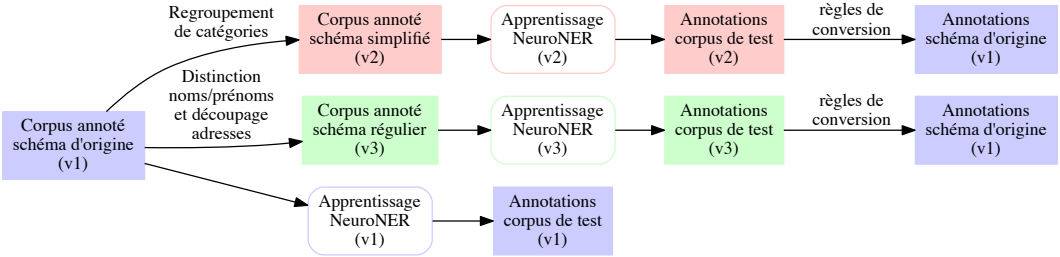

FIGURE 2: Étapes suivies : conversion du schéma d'annotation d'origine en versions simplifiées (v2) et régulière (v3), construction de modèles statistiques pour ces trois versions, application des modèles sur le corpus de test, et conversion des annotations v2 et v3 vers le schéma d'origine

La tokénisation a été réalisée avec l'outil spaCy [4] pour l'anglais, nous avons repris les plongements lexicaux entraînés avec GloVe (Pennington *et al.*, 2014) fournis avec l'outil NeuroNER (200 dimensions pour les plongements lexicaux et 15 dimensions pour les plongements de tokens), l'optimisation est fondée sur un SGD (stochastic gradient descent), et nous avons conservé le seuil d'apprentissage par défaut de 0.005. La construction du modèle a été réalisée en utilisant 8 CPU (<12 minutes par itération).

Nous avons créé des règles de conversion pour retrouver les types d'entités nommées de la version d'origine du schéma d'annotation depuis les sorties produites en versions simplifiée (v2) et régulière (v3). Les règles se fondent sur des déclencheurs et une analyse du contexte. Elles visent uniquement à convertir les types d'EN, sans chercher à identifier de nouvelles entités.

Depuis la version simplifiée (v2), nous avons produit des règles de conversion uniquement pour traiter les cas les plus fréquents (soit une dizaine d'heures de travail). Pour les cas non couverts par nos règles, nous avons utilisé des valeurs par défaut, même si cela engendre une baisse de précision pour les types d'entités utilisés comme valeurs par défaut. Ces valeurs reposent sur les types les plus utilisés en corpus (soit 'Doctor' pour un nom de personne et 'Hospital' pour un nom de lieu). Sur les noms de personnes, nous utilisons les indices de conversion suivants :

— vers 'Doctor' : déclencheurs en contexte gauche *'Att :' 'Attending :' 'CC :' 'Dr' 'Dr.' 'Fellow :' 'Intern :' 'MD :' 'RefMD :'*, droit *'MD' 'M.D.'*, et expressions introductrices *'as per by' 'Dear' 'follow-up X months'* , etc.

— vers 'Patient' : déclencheurs *'Mr' 'Mr.' 'Mrs' 'Mrs.' 'Ms' 'Ms.'* et expressions introductrices *'Patient :' 'Pt :' 'Patient Name :' 'HPI :' 'RE :' 'Impression :' 'to see' 'of seeing' 'I saw''w/ sister' 'your patient'*, etc.

Concernant les lieux, notre approche repose sur des règles simples pour retrouver les types d'origines : séquence de 5 chiffres pour 'Zip', liste d'États américains pour 'State', une chaîne de caractères commençant par une majuscule suivie d'un État et d'un code postal pour 'City', la combinaison d'un numéro, d'un nom et d'un type de voirie (*'Avenue' 'Circle' 'Ct' 'Drive' 'Dr' 'Lane' 'Place' 'Road' 'Street' 'STREET' 'St' 'Terrace' 'Way'*) pour 'Street'. Depuis la version régulière (v3), nous avons rassemblé dans la même portion les noms et prénoms des personnes et avons repris les mêmes déclencheurs et phrases introductives que ci-dessus.

4. https://spacy.io

4 Résultats

La figure 3 présente l'évolution de la F-mesure de NeuroNER sur le test, pour chaque version du schéma. Le tableau 2 présente les résultats par catégorie à l'issue de la construction du modèle. L'apprentissage prend fin lorsque les dix dernières itérations n'ont présenté aucune amélioration. Le tableau 3 présente les résultats par application de NeuroNER et des règles de conversion des types d'EN produits par NeuroNER. Puisque les types produits sur la version d'origine du schéma d'annotation (v1) constituent déjà les types d'EN attendus, les résultats sur la v1 sont identiques entre les tableaux 2 et 3. La tableau 4 présente les résultats lorsque la conversion est appliquée sur les entités nommées de référence. Cette évaluation met en évidence la qualité des règles de conversion indépendamment de l'identification des entités nommées par NeuroNER.

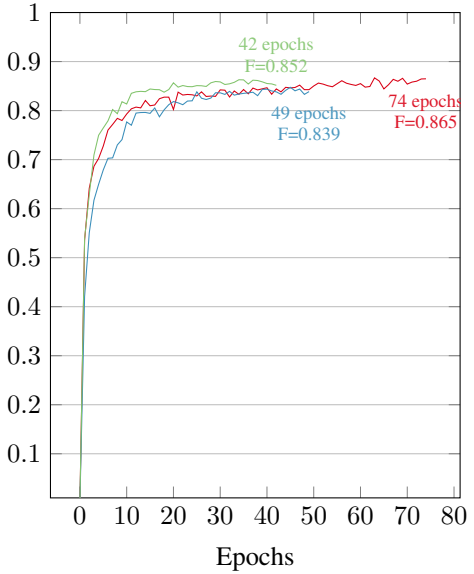

FIGURE 3: Évolution de la F-mesure sur le test par itération pour les versions d'origine (v1), simplifiée (v2) ou régulière (v3)

Type	v1 (origine)			v2 (simple)			v3 (régulier)		
	R	P	F	R	P	F	R	P	F
Pers.	–	–	–	.885	.938	.911	–	–	–
Doc.	.859	.929	.893	–	–	–	–	–	–
Pat.	.846	.863	.854	–	–	–	–	–	–
First	–	–	–	–	–	–	.887	.919	.903
Last	–	–	–	–	–	–	.917	.927	.922
Code	–	–	–	–	–	–	.657	.888	.755
Loc.	–	–	–	.737	.836	.784	–	–	–
Hosp.	.655	.830	.732	–	–	–	–	–	–
Street	.838	.826	.832	–	–	–	–	–	–
City	.633	.788	.702	–	–	–	.633	.761	.691
State	.790	.802	.796	–	–	–	.756	.791	.773
Zip	.914	.928	.921	–	–	–	–	–	–
Num.	–	–	–	–	–	–	.834	.931	.883
Name	–	–	–	–	–	–	.488	.743	.589
Type	–	–	–	–	–	–	.842	.879	.860
TOUS	.800	.881	.839	.831	.902	.865	.818	.888	.852

TABLE 2: Évaluation (rappel, précision, F-mesure) du repérage d'entités nommées selon le schéma d'annotation utilisé pour créer le modèle

5 Discussion

Concernant le repérage d'entités nommées, la figure 3 met en évidence deux conclusions : (i) un schéma d'annotation complexe (v1 avec 7 types) obtient de moins bons résultats que la version simplifiée (v2 avec 2 types) : F=0.839 vs. F=0.865 ; (ii) un schéma plus régulier (v3 avec 8 types) permet à NeuroNER d'obtenir plus rapidement de meilleurs résultats (la courbe verte augmente plus rapidement que les autres). Les conclusions que nous pouvons tirer du tableau 2 sont plus contrastées. Comme attendu, la distinction entre patient et médecin produit de moins bons résultats (F=0.854 and 0.893 respectivement) que celle entre prénom et nom (F=0.903 and 0.922), ce qui prouve sa complexité pour un système statistique. Malgré la réduction de l'ambiguïté dans le schéma

Cat.	v1 (origine)			v2 (simple)			v3 (régulier)		
	R	P	F	R	P	F	R	P	F
Doc.	.859	.929	**.893**	.848	.800	.823	.788	.673	.726
Patient	.846	.863	**.854**	.713	.896	.794	.473	.447	.460
Hosp.	.655	.830	**.732**	.591	.611	.601	.459	.634	.532
Street	.838	.826	**.832**	.596	.853	.701	.588	.370	.455
City	.633	.788	**.702**	.617	.593	.605	.481	.418	.447
State	.790	.802	.796	.795	.853	.823	.816	.856	**.836**
Zip	.914	.928	.921	.900	.984	**.940**	.900	.984	**.940**
TOUS	.800	.881	**.839**	.747	.774	.760	.640	.607	.623

TABLE 3: Rappel, précision, et F-mesure des prédictions de NeuroNER avec application de post-traitement. Les meilleurs résultats sont en gras

v1 (origine)			v2 (simple)			v3 (régulier)		
R	P	F	R	P	F	R	P	F
.859	.929	**.893**	.950	.834	.888	.879	.702	.780
.846	.863	**.854**	.735	.896	.808	.507	.471	.488
.655	.830	.732	.830	.724	**.774**	.753	.747	.750
.838	.826	**.832**	.721	.961	.824	.794	.480	.598
.633	.788	**.702**	.761	.638	.694	.883	.581	.701
.790	.802	.796	.842	.856	.849	.953	.924	**.938**
.914	.928	.921	.943	.971	.957	.957	.971	**.964**
.800	.881	**.839**	.860	.815	.837	.664	.783	.718

TABLE 4: Rappel, précision, et F-mesure des conversions sur les entités nommées de référence. Les meilleurs résultats sont en gras

d'annotation plus régulier, nous obtenons des résultats plus faibles pour les types d'EN utilisés dans plusieurs contextes (F=0.691 vs. 0.702 pour les villes, F=0.773 vs. 0.796 pour les États) par rapport au schéma d'origine. Puisque les systèmes statistiques sont sensibles aux variations contextuelles, une plus grande variété de contextes dans lesquelles apparaissent des EN de ces types a un impact négatif sur les performances du système. De manière non intentionnelle, nous avons remplacé l'ambiguïté de définition par une ambiguïté de contexte. La conversion des types d'EN depuis les deux versions simplifiées se révèle complexe dans la mesure où elle revient à réintroduire de l'ambigüité dans un jeu d'annotations simplifiées.

6 Conclusion

Dans cet article, nous avons comparé l'impact de deux versions d'un schéma d'annotations sur un système de REN, une version simplifiant les types d'EN en deux types consensuels (nom et lieu) et une version proposant des définitions régulières. Dans les deux cas, cette simplification améliore les résultats. Nous avons étudié la possibilité de recouvrer les types d'EN du schéma d'origine à partir des versions simplifiées. Les règles de conversion ne suffisent pas pour retrouver le niveau de détail du schéma d'annotation d'origine. Si l'intérêt d'une simplification est réel, nous estimons que d'autres solutions doivent être identifiées pour retrouver les types d'EN d'origine. Dans le cadre de travaux futurs, nous envisageons de reproduire ces expériences au moyen de CRF de chaîne linéaire, ainsi que l'application de méthodes symboliques. La comparaison des résultats obtenus avec ces approches permettra de vérifier l'impact de la simplification de schémas d'annotation avec d'autres approches.

7 Remerciements

Ce travail a été réalisé dans le cadre du groupe de travail « Sécurité des Données Textuelles » du Labex DigiCosme (projet ANR-11-LABEX-0045-DIGICOSME) financé par l'ANR au travers du programme "Investissement d'Avenir" Idex Paris-Saclay (ANR-11-IDEX-0003-02).

Références

ALEX B., GROVER C., SHEN R. & KABADJOV M. (2010). Agile corpus annotation in practice : an overview of manual and automatic annotation of CVs. In *Proc of LAW*, p. 29–37, Uppsala, Sweden.

BLACHE P., BIGI B., PREVOT L., RAUZY S. & SEINTURIER J. (2010). Annotation schemes, annotation tools and the question of interoperability : from typed feature structures to XML schemas. In *Proc of International Conference on Global Interoperability for Language Resource*, Hong Kong, China.

DANDAPAT S., BISWAS P., CHOUDHURY M. & BALI K. (2009). Complex linguistic annotation – no easy way out ! a case from Bangla and Hindi POS labeling tasks. In *Proc of LAW*, p. 10–18, Suntec, Singapore.

DERNONCOURT F., LEE J. Y. & SZOLOVITS P. (2017). NeuroNER : an easy-to-use program for named-entity recognition based on neural networks. In *Proc of EMNLP*, Copenhagen, Denmark.

FORT K. & SAGOT B. (2010). Influence of pre-annotation on POS-tagged corpus development. In *Proc of LAW*, p. 56–63, Uppsala, Sweden.

GANCHEV K., PEREIRA F., MANDEL M., CARROLL S. & WHITE P. (2007). Semi-automated named entity annotation. In *Proc of LAW*, p. 53–56, Prague, Czech Republic.

GROUIN C. (2016). Controlled propagation of concept annotations in textual corpora. In *Proc of LREC*, Portorož, Slovenia.

LEAMAN R. & GONZALES G. (2008). Banner : an executable survey of advances in biomedical named entity recognition. In *Proc of Pacific Symposium on Biocomputing*, p. 652–63, Hawaii, USA.

LEECH G. (1993). Corpus annotation schemes. *Lit Linguist Computing*, **8**(4), 275–281.

MILLE S., BURGA A., FERRARO G. & WANNER L. (2012). How does the granularity of an annotation scheme influence dependency parsing performance ? In *Proc of COLING, posters*, p. 839–852, Mumbai, India.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). GloVe : global vectors for word representation. In *Proc of EMNLP*, volume 12, p. 1532–43.

SHMANINA T., ZUKERMAN I., YEPES A. J., CAVEDON L. & VERSPOOR K. (2013). Impact of corpus diversity and complexity on NER performance. In *Proc of Australasian Language Technology Association Wrokshop*, p. 91–95, Brisbane, Australia.

STUBBS A., KOTFILA C. & UZUNER O. (2015). Automated systems for the de-identification of longitudinal clinical narratives : Overview of 2014 i2b2/UTHealth shared task track 1. *J Biomed Inform*, **58**, S11–S19.

VOUTILAINEN A. (2012). Improving corpus annotation productivity : a method and experiment with interactive tagging. In *Proc of LREC*, p. 2097–2102, Istanbul, Turkey.

WILSON A. & THOMAS J. (1997). Semantic annotation. In R. GARSIDE, G. LEECH & T. MCENERY, Eds., *Corpus Annotation*, chapter 4, p. 53–65. Routledge.

YIMAM S. M., DE CASTILHO R. E., GUREVYCH I. & BIEMANN C. (2014). Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proc of ACL, System Demonstrations*, p. 91–96, Baltimore, MA.