Ioana Galleron, Francesca Frontini, Carmen Brando, Motasem Alrahabi « Les Guides d'annotation des entités nommées : similitudes, différences, difficultés »

- proposition de communication –

DH Humanistica, Bordeaux, 13-14 mai 2020

Au cours des dernières années, un intérêt de plus en plus soutenu s'est manifesté pour l'identification et l'analyse des entités nommées dans les textes littéraires. Si de nombreux outils existent et fonctionnent de façon satisfaisante sur des corpus modernes, notamment de presse (Stanford NER, Unitex, Parallel Dots, Spacy NER, SEM, Dariah Entity Fishing Service, etc.), les performances de ces outils sont bien plus décevantes quand ils sont appliqués à de la fiction. Des campagnes de tests et d'amélioration de ces performances sont actuellement en cours, aussi bien au niveau européen (voir action COST « Distant reading »), qu'au niveau national (par exemple, le projet que le labex OBVIL de Sorbonne Université mène depuis un an sur l'annotation manuelle des entités nommées dans les textes littéraires).

L'objectif de cette communication est de se pencher sur un instrument principal de ces campagnes, à savoir le guide d'annotation des entités nommées. En effet, que ce soit pour l'entraînement des outils ou la validation des résultats, toutes ces campagnes doivent créer de nouvelles données en mobilisant des annotateurs et en établissant des consignes claires. Nous présenterons les principaux points communs et différences entre ces guides, en insistant sur la façon dont ils gèrent l'épineux problème du traitement de l'information que l'on peut inférer à partir des unités annotées.

Celle-ci est, en effet, multiple :

1° ces unités créent une représentation mentale et, en tant que telles, renvoient à des objets du monde réel avec lesquels elles partagent des caractéristiques. En lisant, on prélève donc des informations mimétiques de nature très diverses, qui ont donné lieu à différentes tentatives de systématisation (voir, par exemple, Galleron, 2017), et que l'on peut souhaiter expliciter dans le cadre d'une campagne d'annotation. Cependant, pour des raisons de faisabilité ainsi qu'en fonction d'objectifs scientifiques propres, chaque campagne organise différemment le codage de ces informations.

2° les mêmes unités renvoient à d'autres unités similaires, du même texte ou d'autres textes, s'inscrivant ainsi au cœur de complexes réseaux. Il s'agit ici non seulement des chaînes de co-référence qui parcourent les textes, et qui posent de nombreux problèmes d'annotation (voir Schnedeker et Landragin, 2018), ou des réseaux de personnes que l'on peut inférer à partir d'un texte, mais également des différents plans sur lesquels se place l'information (p. ex., personnages agissants/ personnages évoqués). Cette information de positionnement est traitée fort différemment dans les guides, et relevée avec plus ou moins de précision.

3° enfin, les unités relevées sont porteuses de multiples informations linguistiques, dont le relevé entre parfois en contradiction avec les informations mimétiques et réticulaires mentionnées plus haut. La présentation donnera des exemples de ces conflits, et montrera de quelle façon les guides proposent de les régler.

Nous évoquerons enfin la façon dont les guides d'annotation cherchent à prévenir d'autres problèmes liés à l'annotation : la divergence entre les annotateurs, le traitement des

annotation incomplètes ou erronées, le traitement de la redondance, l'équilibre entre exhaustivité et efficacité.

Bibliographie sommaire

Foppiano, Luca; Romary, Laurent, "entity-fishing: a DARIAH entity recognition and disambiguation service". *Digital Scholarship in the Humanities*, Sep 2018, Tokyo, Japan. <hal-01812100>

Galleron, Ioana, « Conceptualisation of theatrical characters in the digital paradigm: needs, problems and foreseen solutions », *Human and Social studies*, De Gruyter, vol. 6, issue 1, 2017 (Published Online: 2017-04-18 | DOI: https://doi.org/10.1515/hssr-2017-0007)

Ehrmann, Maud *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, thèse en Informatique et langage, Paris Diderot University, 2008 <tel-01639190>

Frontini, Francesca ; Brando Escobar, Carmen ; Ganascia, Jean-Gabriel, "REDEN ONLINE: Disambiguation, Linking and Visualisation of References in TEI Digital Editions", Jagiellonian University Pedagogical University, Digital Humanities 2016, Jul 2016, Kraków, Poland; http://dh2016.adho.org/, 2016, Digital Humanities 2016: Conference Abstracts. <http://dh2016.adho.org/>; <hal-01395125>

Schnedecker, Catherine ; Landragin, Frédéric « Les chaînes de référence : présentation », Langages 2014/3 (N° 195), p. 3-22, DOI 10.3917/lang.195.0003