” Je pense que ça traite d’expérience de lecture, à voir ... ”: retour sur une expérience d’annotation collaborative

François Vignale, Guillaume Le Noé Bienvenu, Guillaume Gravier, Pascale Sébillot

” Je pense que ça traite d’expérience de lecture, à voir ... ” : retour sur une expérience d’annotation collaborative.

François Vignale*[1], Guillaume Le Noé Bienvenu[2], Guillaume Gravier[3], and Pascale Sébillot[4]

[1]Langues, Littératures, Linguistique des universités dÁngers et du Mans – Le Mans Université : EA4335, Université d’Angers : EA4335 – France
[2]irisa – CNRS : UMR6074 – France
[3]IRISA (IRISA) – CNRS : UMR6074 – Rennes, France
[4]Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) – Université Rennes I, INSA-Rennes – France

Résumé

Cette communication a pour but de présenter une expérience d’annotation collaborative d’expériences de lecture menée avec des étudiants dans le cadre du projet READ-IT.
Au cours des dernières décennies, les connaissances sur l’histoire des pratiques de lecture ont considérablement augmenté au sujet des usages et des habitudes mais des questions fondamentales demeurent, telles que le ”pourquoi” et le ”comment” on lit. Grâce à l’exploration de sources numériques à la recherche de témoignages d’expériences de lecture, le projet READ-IT (Reading Europe Advanced Data Investigation Tool, https://readit-project.eu) vise à mieux comprendre ces phénomènes. Ce projet financé par le Joint Programming Initiative for Cultural Heritage (2018-2021) associe 5 partenaires de 4 pays (France, Royaume-Uni, Pays-Bas, République Tchèque).

En combinant différentes conceptions (Jauss 1982 ; Iser 1978) et en nous inscrivant dans une démarche fondée sur les sources, nous avons créé un modèle théorique et une ontologie (Reading Experiences Ontology, REO, https://github.com/eureadit/reading-experience-ontology) proposant un description minimale où l’expérience de lecture est définie comme un phénomène temporel précédé de prémisses et suivis d’effets dans lesquels une personne interagit avec un contenu écrit par l’intermédiaire d’un médium (Antonini et al. 2019).

Cependant, afin de dépasser le stade théorique et de commencer à disposer des outils de détection semi-automatique des expériences de lecture, il est apparu nécessaire de lancer une campagne d’annotation collaborative à partir d’une interface spécialement conçue pour le projet (https://readit-it.hum.uu.nl). Cette étape est essentielle pour le développement des algorithmes que READ-IT doit produire d’ici à son terme. Les résultats doivent donc être confrontés à l’expertise humaine afin d’améliorer leur performance, ceci dans un processus itératif. Elle a donc pour objectifs principaux : 1) de valider définitivement le modèle de données ; 2) de fournir suffisamment de sources annotées pour entraîner convenablement les détecteurs reposant sur des méthodes de traitement automatique du langage naturel ;

*Intervenant

3) de tester et d'améliorer l'ergonomie de la plate-forme d'annotation ainsi que d'évaluer la définition et la clarté des concepts repris dans le guide d'annotation dans l'objectif d'une dissémination vers le grand public.

Cette campagne s'est déroulé sur 10 semaines de mars à mai 2020 avec la participation de 10 étudiants stagiaires de M2 Lettres et langues rémunérés à hauteur de 200 heures chacun. Les sources ont constitué en des commentaires de lecteurs en ligne extraits des réseaux sociaux de lecture francophone et anglophones Babelio et Goodreads (Rebora et al. 2019). Le corpus a été construit aléatoirement et les données ont été anonymisées. Les métadonnées relatives aux titres des ouvrages ont été masquées. Le corpus a été subdivisé en unités de 100 à 150 commentaires dont 30 % étaient partagés entre plusieurs annotateurs afin de calculer l'accord inter-annotateurs qui sert à mesurer la cohérence des annotations (Bayerl et al. 2011). Les stagiaires ont bénéficié de deux séances de formation en présentiel puis, en raison du confinement, de deux autres à distance. Ils se sont également appuyés sur des supports spécialement conçus pour cette campagne et sur un guide d'annotation. Des réunions hebdomadaires ont permis de suivre l'avancement du travail et de répondre aux difficultés d'organisation, de compréhension des consignes ou aux problèmes techniques sur la plateforme d'annotation. Des rendez-vous plus restreints (par groupes de 3 annotateurs) ont également eu lieu afin de mieux comprendre les raisons de certaines annotations.

Deux tâches ont été confiées : 1) le balisage des limites extrêmes des expériences de lecture contenues ou non dans les commentaires en ligne ; 2) l'identification des composantes de l'expérience de lecture selon les classes définies dans l'ontologie REO.

Le balisage consiste à repérer le début et la fin de l'action de lecture à l'intérieur du commentaire. Les éléments décisifs sont le sujet lecteur (pronoms personnels), les œuvres citées (entités nommées), les verbes d'action liés à la lecture (lire, dévorer ...), le vocabulaire lié au médium (livre, bouquin...) et des éléments contextuels sans que toutes ces conditions soient nécessairement réunies en même temps.

Les composantes de l'expérience de lecture à identifier correspondent à des classes existant dans le modèle de données et sont donc alignées avec l'ontologie REO. Schématiquement, elles appartiennent à 3 grands groupes :

Reading agent (informations sur le lecteur)

Reading resource (informations sur le texte lu)

Reading process (informations sur la nature, le déroulé, les circonstances et les conséquences de l'expérience de lecture)

Au total, plus de 5 000 commentaires ont donné lieu à des annotations.

Les résultats de cette campagne sont contrastés. Le volume des données nécessaire pour entraîner les modèles de détection a été atteint et ces derniers ont donné des résultats déjà très encourageants (précision très satisfaisante mais rappel encore un peu faible) et l'ergonomie de la plate-forme s'est considérablement améliorée à la suite des remarques des stagiaires. En revanche, la qualité et la précision des annotations sont très hétérogènes. Sur ce point, la question de la motivation des annotateurs se pose et l'on ne peut exclure la présence de turkers (Cardon 2015) dans le groupe, ce qui a affecté la cohérence des annotations. Enfin, le taux d'accord inter-annotateurs a été très faible (kappa de Fleiss inférieur à 0,3). Ceci s'explique pour plusieurs raisons. Le concept d'expérience de lecture est encore flou et n'a pas de définition universellement admise pour le moment ce qui laisse une part très importante à l'interprétation personnelle et à la subjectivité, à la différence de la plupart des expériences d'annotation qui ont pour sujet des concepts clairs (entités nommées) ou la reconnaissance d'objets ou de formes dans des images par exemple.

Au total, ce bilan montre la nécessité de reformuler et de préciser la plupart des concepts de

base en étant nettement plus explicite et en donnant des consignes claires si l'on veut ouvrir la plate-forme d'annotation à un public plus large et non spécialiste. Dans le même temps la formation des futurs utilisateurs, ainsi que leur accompagnement par des animateurs – qui ont ici été fortement perturbés par les exigences du confinement - que ce soit en présentiel ou à distance, doivent être au centre de l'organisation de ce type d'opération.

Références :

Antonini, Alessio, François Vignale, Guillaume Gravier et Brigitte Ouvry-Vial. 2019. " The Model of Reading: Modelling principles, Definitions, Schema, Alignments ". https://hal-univ-lemans.archives-ouvertes.fr/hal-02301611.

Jauss, Hans Robert. 1982. *Towards an Aesthetic of Reception.* Minneapolis : University of Minnesota Press.

Iser, Wolfgang. 1978. *The act of reading: a theory of aesthetic response.* London : Routledge.

Bayerl, Petra Saskia et Karsten Ingmar Paul. 2011. " What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation ". *Computational Linguistics* 37 (4) : 699-725. https://doi.org/10.1162/COLI_a_00074.

Fort, Karen. 2016. *Collaborative annotation for reliable natural language processing: technical and sociological aspects.* London : John Wiley & Sons: ISTE.

Rebora, Simone, Peter Boot, Federico Pianzola, Brigitte Gasser, J. Berenike Herrmann, Maria Kraxenberger, Moniek Kuijpers, et al. 2019. " Digital Humanities and Digital Social Reading ". OSF Preprints. https://doi.org/10.31219/osf.io/mf4nj.

Cardon, Dominique, Antonio A. Casilli. 2015. *Qu'est-ce que le digital labor?* Bry-sur-Marne: INA éditions.