Journée d'étude sur les outils d'exploration de corpus numériques à Paris

Date : 17 juin 2022, 8h30h-18h00

Lieu : Maison de la recherche, 28 rue Serpente, 75006 Paris

Contact : kimberly.oger [att] sorbonne-universite.fr & eva.lacroix [att] sorbonne-universite.fr

English version below

Le Club Corpus de Sorbonne Université propose une journée d'étude sur les outils d'exploration de corpus numériques. Cette journée d'étude donnera l'occasion à plusieurs spécialistes du domaine de faire une présentation d'un outil d'analyse de corpus dans une visée scientifique déterminée (p. ex. analyse de discours) et d'en faire la démonstration, avec possibilité de prise en main de l'outil par les participant|e|s au moment des pauses. À la fin de la journée d'étude, Kimberly Oger et Eva Schaeffer-Lacroix animeront une table ronde qui réunira toutes les personnes ayant fait une présentation.

Événement soutenu par CELISO (Centre de linguistique en Sorbonne), STIH (Sens Texte Informatique Histoire), CERES (Centre d'expérimentation en méthodes numériques pour les recherches en Sciences Humaines et Sociales).

Programme :

Horaire	Événement	Outil de corpus
8h30	Accueil café	
8h45	Discours d'accueil (Kimberly Oger)	
9h	Guillaume Désaguilier (Université Paris 8 Vincennes Saint Denis) Des lignes de commande aux scripts interactifs : Comment optimiser l'utilisation de R en linguistique de corpus ?	R
10h	Christophe Parisse (Université Paris Ouest la Défense, INSERM) Collecter des données orales avec CLAN: recueil, codage, dépôt, analyse	Clan

11h	Pause café	
11h30	Bénédicte Pincemin (Université de Lyon, CNRS) Présentation du logiciel TXM : un point de vue utilisateur	TXM (données multimodales)
12h30	Buffet offert	
14h	Marc Kupietz (IDS Mannheim, Allemagne) The Corpus Analysis Platform KorAP: its Philosophy, Features, Pros & Cons	KorAP
15h	Daniel Henkel (Université Paris 8 Vincennes Saint Denis) Techniques d'alignement pour la création de mémoires de traduction ou corpus parallèles	LF Aligner
16h	Pause café	
16h30	Maria Zimina-Poirot (Université de Paris) Le Trameur/iTrameur : naviguer dans les corpus pluritextuels avec des outils textométriques	Le Trameur/iTrameur
17h30	Table ronde (Kimberly Oger et Eva Schaeffer-Lacroix)	

Résumés

Guillaume Desagulier

- Des lignes de commande aux scripts interactifs : Comment optimiser l'utilisation de R en linguistique de corpus ?

Daniel Henkel

- Techniques d'alignement pour la création de mémoires de traduction ou corpus parallèles

Marc Kupietz

- The Corpus Analysis Platform KorAP: its Philosophy, Features, Pros & Cons

Christophe Parisse

- Collecter des données orales avec CLAN: recueil, codage, dépôt, analyse

Bénédicte Pincemin

- Présentation du logiciel TXM : un point de vue utilisateur

Maria Zimina-Poirot

- Le Trameur/iTrameur : naviguer dans les corpus pluritextuels avec des outils textométriques

Guillaume Desagulier

Des lignes de commande aux scripts interactifs : Comment optimiser l'utilisation de R en linguistique de corpus ?

R est un langage de programmation devenu incontournable en sciences sociales. Si sa maîtrise implique une courbe d'apprentissage abrupte, il est néanmoins possible de mettre à profit assez rapidement quelques fonctionnalités de base. Mon intervention se fonde sur un retour d'expérience de douze ans avec R et s'articule autour de deux moments. Le premier est un partage de bonnes pratiques en réponse aux questions que se posent les linguistes au moment de constituer un jeu de données avec R. Le second moment consiste en la démonstration de deux scripts interactifs rédigé sous R pour l'exploration sociolinguistique des composantes orales du British National Corpus (XML et 2014) : BNC.query() et BNC2014.query().

Références

Desagulier, G. (2017). Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics. New York : Springer.

Desagulier, G. (2019a). "BNC.query(). An interactive R script for a sociolinguistic exploration of the spoken component of the BNC-XML," in Around the word, 08/01/2019, https://corpling.hypotheses.org/2252.

Desagulier, G. (2019b). "BNC.2014.query(). An interactive R script for a sociolinguistic exploration of the spoken component of the BNC-2014," in Around the word, 03/01/2019, https://corpling.hypotheses.org/1632.

Daniel Henkel

Techniques d'alignement pour la création de mémoires de traduction ou corpus parallèles

L'alignement est une technique qui consiste à réunir deux textes monolingues, le plus souvent un texte-source (original) et un texte-cible (traduction) afin de créer un seul document bilingue. Celui-ci contiendra des paires de segments équivalents qui correspondent le plus souvent à des phrases. Quoique ce soit possible d'aligner les segments manuellement, ce travail est fastidieux et requiert une quantité de temps incommensurable, d'où l'intérêt de commencer par un pré-alignement automatique. Un logiciel d'alignement comme LF Aligner commence par associer tous les paragraphes dans les textes source et cible, avant de les découper en phrases pour constituer les paires de segments. À la fin de ce processus environ 90-95 % des segments seront alignés correctement et le travail de correction manuelle ira beaucoup plus vite qu'un alignement 100 % manuel. Les paires de segments sont généralement encodées dans une « mémoire de traduction » au format .tmx (Translation Memory eXchange) qui rendra la base de données accessible à un logiciel de Traduction Assistée par Ordinateur. Les textes bilingues au format .tmx peuvent également être importés dans un logiciel de corpus comme TXM pour constituer un corpus bilingue parallèle. En partant de deux documents en anglais et français, au cours de cet atelier nous accomplirons chacune de ses étapes – alignement

automatique, vérification, encodage, exploitation de la base de données parallèles – à l'aide des logiciels Bitext2tmx, LF Aligner, Okapi Checkmate, OmegaT et TXM.

Marc Kupietz

(Marc Kupietz, Nils Diewald, Eliza Margaretha, Helge Stallkamp & Franck Bodmer)

The Corpus Analysis Platform KorAP: its Philosophy, Features, Pros & Cons

The corpus analysis platform KorAP has been developed mainly at the Leibniz Institute for the German Language (IDS) since 2011 (Bański et al. 2013) and has been publicly accessible since 2017. KorAP primarily serves to provide access to the German Reference Corpus DeReKo (Kupietz et al. 2010) for its 40,000 registered users – scholars of German linguistics of all kinds, from all over the world – and is in this respect the successor to the COSMAS II analysis platform (Bodmer 2005), which is still operated in parallel at the IDS. However, KorAP ist also used for the Reference Corpus of Contemporary Romanian Language CoRoLa (Tufiș et al. 2016; Cristea et al. 2019) and, since recently, also for the Hungarian National Corpus HNC (Váradi 2002; Kupietz et al. 2021).

The overall goal behind the KorAP development was to create a sustainable platform for the next 20 years that supports an in principle unlimited amount of primary data, to keep up with rapidly growing corpus sizes. Other key features of KorAP are: user definable virtual corpora; an appropriate mapping of IPR and licensing restrictions to maximize the usability of corpora without infringing on the interests of rights holders; the support of an unlimited number of potentially conflicting annotation layers, including constituency and dependency relations; and an extensible set of query languages to accommodate the broad spectrum of KorAP's user base. An important feature of KorAP in this respect is also its extensibility through user-defined functionalities at various levels from the use of the API via client libraries for R and Python (Kupietz et al. 2020) to contributions to the BSD-licensed KorAP open source project (https://github.com/KorAP, Kupietz et al. 2022).

In my presentation, I will give an overview of KorAP, from its background, philosophy and design principles, to sample applications and queries, to the extended possibilities through the use of its R library and possibilities of contrastive research on virtual comparable corpora (Kupietz et al. 2020b). I will discuss the strengths of KorAP, but also its weaknesses and the functionalities that are still under development.

References

Bański, Piotr/Bingel, Joachim/Diewald, Nils/Frick, Elena/Hanl, Michael/Kupietz, Marc/Pęzik, Piotr/Schnober, Carsten/Witt, Andreas (2013): KorAP: the new corpus analysis platform at IDS Mannheim. In: Vetulani, Zygmunt/Uszkoreit, Hans (eds.): Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference. Poznań: Fundacja UAM. 586-587.

Bodmer, Franck (2005): COSMAS II. Recherchieren in den Korpora des IDS. In: Sprachreport 3/2005. Mannheim: Institut für Deutsche Sprache. 2-5.

Cristea, Dan/Diewald, Nils/Haja, Gabriela/Mărănduc, Cătălina/Barbu Mititelu, Verginica/Onofrei, Mihaela (2019): How to find a shining needle in the haystack. Querying CoRoLa: solutions and perspectives.

In: Cosma, Ruxandra/Kupietz, Marc (Hrsg.), On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo, Revue Roumaine de Linguistique, 64(3). Editura Academiei Române, Bucharest, Romania.

Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (eds): Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010). Paris/La Valetta: ELRA. 1848-1854.

Kupietz, Marc/Diewald, Nils/Trawiński, Beata/Cosma, Ruxandra/Cristea, Dan/Tufiş, Dan/Váradi, Tamás/Wöllstein, Angelika (2020): Recent developments in the European Reference Corpus EuReCo. In: Granger, Sylviane/Lefer, Marie-Aude (eds.): Translating and Comparing Languages: Corpus-based Insights. (= Corpora and Language in Use, Proceedings 6). Louvain-la-Neuve: Presses universitaires de Louvain. 257-273.

Kupietz, Marc/Diewald, Nils/Margaretha, Eliza (2020b): RKorAPClient: An R Package for Accessing the German Reference Corpus DeReKo via KorAP. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), Marseille/Paris: ELRA. 7016-7021.

Kupietz, Marc/Diewald, Nils/Margaretha, Eliza (2022): Building Paths to Corpus Data - A multi-level least effort and maximum return approach. In: Fišer, Darja/Witt, Andreas (eds.): CLARIN. The Infrastructure for Language Resources. DeGruyter (to appear 10/2022).

Kupietz, Marc/ Trawiński, Beata / Diewald, Nils (2021): DeutUng in the Context of the European Reference Corpus EuReCo. Presentation in the DeutUng-project closing workshop. Publication in preparation.

Tufiș, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ș., D., Boroș, T. (2016): The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Paris/Portoroz: ELRA.

Váradi, T. (2002): The Hungarian National Corpus. In Rodríguez, M. & Araujo, C. (eds) Proceedings of LREC 2002, Las Palmas / Paris: ELRA, 385–389.

Christophe Parisse

Collecter des données orales avec CLAN: recueil, codage, dépôt, analyse

CLAN est un outil informatique utilisable sur Windows et MacOS permettant de transcrire, éditer, et analyser des données de langage oral. CLAN est associé très étroitement à un format (CHAT) et à un site de dépôt (CHILDES/TALKBANK) qui forment un triptyque parfait pour recueillir, transcrire, déposer, diffuser, analyser, et même réutiliser, des données de corpus oral. Ces étapes seront décrites à partir d'un exemple de données, celles du corpus COLAJE qui cible l'acquisition du langage chez le jeune enfant. Elles peuvent s'appliquer à tout type de corpus de langue orale, enfant, adulte, mono- ou multi-locuteur, comme on pourra le montrer avec d'autres exemples issus de la base TALKBANK.

Références

MacWhinney, B. (1996). The CHILDES system. American Journal of Speech-Language Pathology, 5(1), 5-14. Disponible à https://psyling.talkbank.org/years/1996/ajslp-childes.pdf

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk (3rd ed.). Psychology Press. https://doi.org/10.4324/9781315805672, Voir aussi https://talkbank.org

Morgenstern, A., & Parisse, C.. The Paris Corpus. Journal of French Language Studies, Cambridge University Press (CUP), 2012, 22 (Special issue 1), pp.7-12. https://doi.org/10.1017/S095926951100055X, ⟨halshs-01350592⟩

Bénédicte Pincemin

Présentation du logiciel TXM : un point de vue utilisateur

Le logiciel TXM propose une implémentation open-source de la méthode textométrique, pour une approche à la fois qualitative et quantitative des corpus textuels.

La textométrie peut se caractériser par des calculs statistiques de référence (spécificités, AFC), par le caractère central du retour au texte et l'attention au document source, et par la construction de parcours d'analyse par l'utilisateur plutôt que la production d'un résultat pour le corpus. Ces orientations permettent de la situer par rapport à des pratiques de la linguistique de corpus et de discuter de sa pertinence en fonction des attentes des chercheurs. De la même façon (contrastive), nous rendrons compte de notre connaissance pratique du logiciel TXM pour pointer des caractéristiques techniques générales, que nous pourrons mettre en lien avec des conséquences pratiques pour l'utilisateur. Nous évoquerons des développements récents avancés (comme l'annotation en cours d'analyse ou l'analyse de corpus audiovisuels) à travers l'expérience concrète que nous avons pu en avoir dans le cadre de projets. TXM s'avère un logiciel très souple et très puissant, on abordera la question de ses choix et de ses limites pour discuter de sa pertinence en fonction des contextes.

Maria Zimina-Poirot

Le Trameur/iTrameur : naviguer dans les corpus pluritextuels avec des outils textométriques

L'informatisation des alignements textuels est confrontée à la complexité de l'organisation et du fonctionnement des textes et discours. L'architecture modulaire *Trame/Cadre* issue des recherches menées en textométrie facilite la navigation dans l'espace textuel grâce aux ressources incrémentales qui conservent la trace de séquences de traitements quantitatifs appliqués aux données. Au cours de notre démonstration, nous montrerons les principes de navigation textuelle à l'aide des logiciels de textométrie *Le Trameur/iTrameur* en nous appuyant sur l'analyse des corpus comportant plusieurs volets, dont une archive conservant les différents cycles de révision (de la post-édition des sorties de la traduction automatique neuronale à la livraison de la traduction entièrement révisée).

Corpus mining and analysis tools workshop, 17 June 2022, Paris

The Corpus Club of Sorbonne University is organising a workshop on digital corpus mining tools. During this workshop, several specialists in the field will present and demonstrate a corpus mining and analysis tool focusing on a specific scientific area (e.g. discourse analysis), with a hands-on session during the breaks. At the end of the workshop, Kimberly Oger and Eva Schaeffer-Lacroix will lead a round-table discussion with all the guest speakers.

17 June 2022, 8:30am-6:00pm

Maison de la recherche, 28 rue Serpente, 75006 Paris

8:30 am - 6:00pm

Invited speakers (scheduled order):

- 9am: Guillaume Desagulier (Université Paris 8 Vincennes Saint Denis): R
- 10am: Christophe Parisse (Université Paris Ouest la Défense, INSERM): Clan
- 11:30am: Bénédicte Pincemin (Ens de Lyon): TXM (multimodal data)
- 2pm: Marc Kupietz (IDS Mannheim): KorAP
- 3pm: Daniel Henkel (Université Paris 8 Vincennes Saint Denis): Alignment techniques for translation memories
- 4:30pm: Maria Zimina-Poirot (Université de Paris): Le Trameur

Event supported by CELISO (Centre de linguistique en Sorbonne), STIH (Sens Texte Informatique Histoire), CERES (Centre d'expérimentation en méthodes numériques pour les recherches en Sciences Humaines et Sociales).

A registration form will be available shortly.

Contact: koger [att] hotmail.fr & eva.lacroix [att] sorbonne-universite.fr