École d'été dynamique de la production écrite

Indicateurs infra-lexicaux et lexicaux des mots dans un corpus de textes narratifs collectés avec Eye and Pen

F. Chenu, B. Lété, H. Jisa & M. Fayol

Résumé

Cet atelier se propose de présenter la méthodologie utilisée pour transcrire puis traiter un corpus de textes (Maggio, Lété, Chenu, Jisa & Fayol, 2012) recueillis avec Eye and Pen (Chesnet & Alamargot, 2005). Après de brefs rappels sur les données recueillies (format Eye and Pen), nous présenterons les conventions de transcriptions que nous avons adoptées (Chenu & Jisa, 2009 ; MacWhinney, 2000a et b, 2001 & 2007 ; MacWhinney, Bird, Cieri & Martell, 2004) en les motivant (aménagement des conventions CLAN). Nous détaillerons ensuite comment enrichir les données textuelles automatiquement avec des scripts Perl (Tanguy & Hathout,2007) à l'aide d'outils disponibles tels que des bases lexicales (Manulex, Lexique3) ou des tagueurs (Cordial, Nooj, Treetagger). Enfin, nous présenterons comment extraire et traiter les données chronométriques (Perl, cf. Tanguy & Hathout,2007) pour qu'elles soient analysables dans des logiciels de statistiques tels que SPSS ou R (attention, nous ne verrons pas les analyses statistiques mais comment traiter les données pour qu'elles soient analysables). Une application concrète sera proposée.

Plan

1. Rappels Eye & Pen

Bref descriptif des procédures pour collecter les données et des données collectées (format des données brutes).

2. Conventions de transcription pour l'écrit

    Segmentation et codage du texte

    Annotations

    Choix de CLAN

3. Enrichissement des données textuelles

    Cordial, Treetagger

    Manulex, Lexique 3

Et autres

4. Extractions pour l'analyse des données chronométriques

    EP Keys & Perl

Références :

Chenu, F. & Jisa, H., 2009, "Les rapports entre méthodologie et théorie : le cas des corpus en acquisition", *Cahiers de Linguistique de Louvain*, 33:2, pp. 147-161

Chesnet, D., & Alamargot, D. 2005. Analyses en temps réel des activités oculaires et graphomotrices du scripteur: intérêt du dispositif 'Eye and Pen'. *L'Année Psychologique*, *105*(3), 477-520.

MacWhinney, B. 2000a. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ, Lawrence Erlbaum Associates. Third Edition. Vol 1: The Format and Programs.

MacWhinney, B. 2000b. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ, Lawrence Erlbaum Associates. Third Edition. Vol 2: The Database.

MacWhinney, B. 2001. From Childes to Talkbank. In M. Almgren, A. Barreña, M. Ezeizaberrena, I. Idiazabal & B. MacWhinney (Eds.), *Research on child language acquisition*. Somerville, MA: Cascadilla Press.

MacWhinney, B. 2007. Opening up video databases to collaborative commentary. In R. Goldman, R. Pea, B. Barron & S. Derry (Eds.), *Video research in the learning sciences*. Mahvah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B., Bird, S., Cieri, C., & Martell, C. 2004. TalkBank: Building an open unified multimodal database of communicative interaction. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 525-528). Lisbon: LREC.

Maggio, S., Lété, B., Chenu, F., Jisa, H. & Fayol, M., 2012, "Tracking the mind during writing: Immediacy, delayed, and anticipatory effects on pause and writing rate", *Reading and Writing*, XX

Tanguy, L., & N. Hathout. 2007. *Perl pour les linguistes. Programmes en Perl pour exploiter les données langagières* , Paris/Londres: Hermès Lavoisier.

Références pour les bases et taggers utilisés

CLAN

http://childes.psy.cmu.edu

Parisse, C., & Le Normand, M.-T. 2000. Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments and Computers,* **32**(3), 468-481.

Lexique 3 :

http://www.lexique.org/

New, B. 2006. Lexique 3 : Une nouvelle base de données lexicales. *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006), avril 2006, Louvain, Belgique.*

Manulex

http://www.manulex.org/

Lété, B., Peereman, R., & Fayol, M. 2008. Phoneme-to-grapheme consistency and word-frequency effects on spelling among first-to-fifth-grade French children: A regression-based study. *Journal of Memory and Language, 58*, 952-977.

Lété, B., Sprenger-Charolles, L., & Colé, P. 2004. Manulex: A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments, & Computers*, *36*, 156-166.

Peereman, R., Lété, B., & Sprenger-Charolles, L. 2007. Manulex-infra: Distributional characteristics of grapheme-phoneme mappings, infra-lexical and lexical units in child-directed written material. *Behavior Research Methods, 39*, 593-603.

TreeTagger

http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Helmut Schmid 1995. Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Helmut Schmid 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Nooj

http://www.nooj4nlp.net/pages/nooj.html

Silberztein, Max. 2003. NooJ manual. available at the WEB site http://www.nooj4nlp.net (200 pages).