Christophe Parisse

Acquisition du langage & morphologie/syntaxe » : outils et données

L'analyse morphologique considère les modifications internes aux mots, modifications qui recouvrent des fonctions lexicales ou syntaxiques. L'analyse syntaxique du langage porte à la fois sur la nature et les propriétés des combinaisons de mots et de marqueurs syntaxiques internes aux mots (marques de genre, de nombre, de temps, de personne, etc). Comme les marqueurs internes aux mots relèvent de la morphologie, cette analyse est souvent qualifiée de « morphosyntaxique ».

Ce cours présentera les techniques de recueil de corpus et d'analyse morphologique, morphosyntaxique ou syntaxique à travers une description détaillée des méthodes utilisées pour travailler sur le langage oral, et en particulier celui des enfants.

L'analyse morphologique et morphosyntaxique est relativement aisée à réaliser de manière automatique ou semi-automatique. En effet, l'utilisation de systèmes probabilistes ou semi-probabilistes se révèle efficace et relativement facile à implémenter [1-2]. En particulier, on peut utiliser des systèmes à apprentissage qui utilisent, soit des données préalablement étiquetées, soit des données brutes, pour les généraliser à de nouvelles analyses. Ces outils ne sont pas parfaits mais ont permis de nombreux travaux liant étude de corpus et syntaxe (par exemple [3-4]).

Ce type d'analyse est particulièrement adapté au langage des enfants de 1 an et demi à 4 ans car ceux-ci se révèlent particulièrement sensibles aux régularités phonologiques et syntaxiques de langue qui les environne, et ceci depuis leur plus jeune âge [5]. L'analyse du langage produit par les enfants est très enrichissante car elle permet de découvrir comment les processus syntaxiques automatiques (régularités, accords, marques morphosyntaxiques) apparaissent au cours du développement du langage (indépendamment du développement sémantique ou pragmatique).

Pour cela il faut transcrire le langage spontané d'enfants, ce qui pose des problèmes spécifiques, surtout si l'on cherche ensuite à en faire une analyse syntaxique automatique. En effet, il faut faire un tri souvent subjectif entre accidents de production et caractéristiques du langage oral ou du langage du petit enfant, tri dont le résultat doit être utilisable pour des analyses automatiques sans pour autant déformer les productions en redressant de manière excessive ou incorrectes les énoncés. Enfin, ce tri est le plus souvent dépendant du type de théorie linguistique défendue (par exemple, [6] ou [7]), ce qui rend les comparaisons entre études parfois délicates et qui montre bien l'importance de méthodes claires et bien argumentées dans le domaine de la transcription et de l'analyse de corpus de langage d'enfant ou de langage oral.


[1] Manning, C. D., & Schütze, H. (1999). Fondations of statistical language processing. Cambridge, MA: MIT Press.

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

[3] Leistyna, P. & Meyer, C. F. (2003). Corpus analysis - Language structure and language use. Amsterdam: Rodopi.

[4] Blanche-Benveniste, C. (1990). Le français parlé : études grammaticales. Paris: Editions du CNRS.

[5] Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274(5294), 1926-1928.

[6] Pinker, S. (1999). Words and rules: The ingredients of language. New York: Basic Books.

[7] Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge: MA: Harvard.