Société de Linguistique de Paris
École Pratique des Hautes Études (EPHE), IV[e] section

Séance du samedi 14 mai 2022 (17h-19h)

La séance aura lieu **en présence** à **l'ENS, 45 rue d'Ulm, 75005 Paris, salle Beckett (attention : lieu inhabituel)** et pourra être suivie **à distance** par GoTo Meeting: https://meet.goto.com/997229013

Les conférenciers seront à distance

Vers la Linguistique conceptuelle
- théorie et expérimentations interactives -

Hélène et André Włodarczyk

L'utilisation des technologies de l'information dans la recherche linguistique a donné lieu au traitement automatique des langues (TAL), mais sans accorder une attention suffisante à la reconstruction logique des concepts empruntés tels quels à la linguistique structurale. L'approche théorique alors disponible qu'était la Grammaire générative, malgré ses bases prétendues formelles, restreignait la portée des problèmes uniquement au traitement des structures arborescentes tout en adoptant le principe selon lequel on pourrait traiter les langues naturelles comme des langages formels. Ce que nous appelons la linguistique *interactive* profite des ressources informatiques mais sans accepter ce principe. En effet, en raison de l'ambiguïté et de l'hétérogénéité des langues, il est nécessaire de rechercher des algorithmes et de les adapter au traitement des données langagières. Le projet CASK (Computer-aided Acquisition of Semantic Knowledge) que nous avons élaboré dans les années 1990 et que nous avons ainsi appelé en 2006, est fondé sur les techniques d'exploration des bases de données (Knowledge Discovery in Databases – KDD), principalement d'analyse des concepts formels (Formal Concept Analysis – FCA de R. Wille) et de théorie des ensembles approximatifs (Rough-Set Theory de Z. Pawlak). La plateforme SEMANA réalisée et utilisée au CELTA (Paris-Sorbonne 2000-2014) était un outil de création et d'analyse de bases de données pour les linguistes sans aucune compétence de programmation.

En plus de tester nos hypothèses, de systématiser nos connaissances des problèmes linguistiques (de plusieurs langues européennes et du japonais) et de les confronter au français, ces outils nous ont conduit à imaginer des « procédures de recherche » (workflows) et à placer notre problématique *proprement* linguistique dans une perspective transdisciplinaire tenant compte des travaux en neurologie du cerveau et en neuropsychologie. Ainsi, nous avons pu reconsidérer la théorie des « structures d'arguments » (argument structure) et celle de « l'organisation informative de l'énoncé » (information structure) et construire notre théorie du *centrage méta-informatif* (Meta-informative Centering – MIC) permettant de présenter ces deux problématiques de manière synthétique en distinguant la méta-, la para- et l'ortho-information. Cela nous a conduit à utiliser le terme 'énoncé' quitte à réserver la 'phrase' à l'abstraction logique n'ayant pas de raison d`être en linguistique. De plus, il a fallu redéfinir les fonctions syntaxiques en acceptant que le même syntagme puisse recevoir plus d'une interprétation, d'une part, et que les représentations du sens d'un seul énoncé puissent ne pas être homogènes quant à leur construction, de l'autre. Ainsi, pour expliquer la forme syntaxique des énoncés, nous posons l'existence de schémas canoniques constituant une *idiomatique généralisée*. Ces schémas sont soumis à des règles de recombinaison des syntagmes servent à faire varier l'énoncé en fonction des stratégies informatives attribuant les statuts méta-informatifs *ancien* et/ou *nouveau* aux parties de l'énoncé.

Grâce au concept de *centre d'attention* et à la distinction entre la méta-information des énoncés de base (du 1er niveau) et la méta-méta-information des énoncés étendus (du 2nd niveau), la théorie MIC permet d'expliquer de manière uniforme et cohérente ce qu'ont de semblable et/ou différent le sujet et l'objet (direct) d'une part et le topique et le focus de l'autre. La distinction entre l'ortho- et la méta-information permet également de situer avec simplicité les langues ergatives par rapport aux langues actives et aux langues mixtes.

Notre programme de grammaire répartie ajoutait la para-information concernant le contraste ou la similitude par rapport à l'ontologie du contexte et du langage qui, malgré de nombreuses études, n'avait pas de formalisation théorique uniforme en linguistique générale. Pour cette raison, la grammaire répartie visait à jeter les bases théoriques de la linguistique conceptuelle. Ce programme réintègre notamment dans la linguistique

formelle la dimension paradigmatique (unités *in absentia* dans l'énoncé) négligée dans les tentatives de modélisation informatique.

Les principales innovations de la Linguistique conceptuelle sont: la sémiotique tétradique, la place des énoncés linguistiques en tant que données dans la pyramide DIK (data-information-knowledge), l'articulation entre les *catégories* linguistiques observables (syllabes, mots, syntagmes et énoncés) et les *concepts formels* qui les sous-tendent (phonèmes, morphèmes, syntactèmes et prédicèmes), la nature pragmatique (méta-informative) de la prédication langagière, l'idiomaticité des schémas syntaxiques et les règles de recombinaison en fonction de l'intention communicative.

En somme, notre *méthode interactive* a permis de créer un cadre théorique apportant ainsi des solutions plus générales et plus uniformes aux problèmes des langues de types différents (japonais, anglais, polonais, russe, allemand, français, latin, grec …) et, notamment, aux problèmes suivants: la définition du sujet et des phrases impersonnelles (à sujet anonyme), l'ordre des mots, l'aspect du verbe, l'emploi des pronoms personnels atones et accentués et les particules méta- et para-informatives.

Liens vers des travaux publiés dans le cadre théorique :

1) Méthode interactive : http://celta.paris-sorbonne.fr/anasem/papers/

2) Grammaire répartie : http://celta.paris-sorbonne.fr/DG-Biblio.html

Bibliographie

Fayyad U., Piatetsky-Shapiro G. and Smyth P., 1996, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, 17(3): 37-54.

Minsky M., 1974, *A Framework for Representing Knowledge*, MIT-AI Laboratory, Memo 306. Pawlak Z., 1981 "Information Systems –Theoretical foundations", *Information Systems* 6 (3): 205-218.

Oberauer K., 2003, "Selective attention to elements in working memory", *Experimental Psychology* 50: 257-269.

Pawlak Z., 1991 *Rough Sets. Theoretical Aspects of Reasoning about Data*, Dordrecht: Kluwer Academic Publishers.

Stacewicz P. & Włodarczyk A., 2010, "Modeling in the Context of Computer Science - a Methodological Approach", *Studies in Logic, Grammar and Rhetoric*, special issue: Philosophical, Trends in the 17th Century from the Modern Perspective, H. Święczkowska (ed.), vol. 20 (33): 155- 179.

Wille R., 1982, "Restructuring Lattice Theory: An Approach based on Hierarchies of Concepts", in: I. Rival (ed.), *Ordered Sets*, Dordrecht-Boston: Reidel: 445-470.

Włodarczyk A., 2007 ハリ・ソルホンヌ大学 理論・応用言語学研究所 (CELTA) ― CASK (Computer- aided Acquisition of Semantic Knowledge) プロシェクト― (paper in Japanese), in: Japanese Linguistics, vol 21, Tokyo: The National Institute for Japanese Language. English version at CELTA: http://celta.paris-sorbonne.fr/anasem/papers/miscelanea/CELTA-CASK-AW-E.pdf

Włodarczyk A. & Włodarczyk H., 2013, *Meta-informative centering in utterances: between semantics and pragmatics*, Companion Series in Linguistics, John Benjamins, Amsterdam.

― 2017, "Subjecthood and Topicality are both Pragmatic Issues", *Papers on and around the Linguistics of BA*, ed. Harada Y., Shudo S., Takekuro M., Institute DECODE, Waseda University, Tokyo: 1-10.

― 2019, « Qu'est-ce au juste que la prédication ? » *Bulletin de la Société de Linguistique de Paris*, t. CXIV (2019), fasc. 1, p. 1-54.

― 2019, "The Interactive Method for Language Science and Some Salient Results", *Zagadnienia Naukoznawstwa (Problems of the Science of Science),* Quarterly Review of the Polish Academy of Science, p. 73-92.

****

Séances ultérieures de la SLP :

Samedi 18 juin 2022 : Pierre LARRIVÉE et Cecilia POLETTO « Ordre des mots, changement syntaxique et micro-indicateurs dans deux langues romanes »

Samedi 19 novembre 2022 : Sebastian FEDDEN « Les langues papoues : synchronie, diachronie, diversité »

Samedi 10 décembre 2022 : Alexandre FRANÇOIS « Tectonique lexicale : Innovations locales et convergence aréale dans les structures sémantiques »