23e Colloque International du CerLiCO: « L'Exemple et le Corpus, Quel Statut ? »
4-6 Juin 2009, Université de Poitiers

Florent Perek – Compte-rendu

1 La conférence

Le CERcle LInguistique du Centre et de l'Ouest (CerLiCO) est une association Loi 1901 qui a pour objectif de promouvoir et d'organiser la recherche linguistique entre les universités du centre et de l'ouest de la France (Angers, Bordeaux, Brest, Caen, La Rochelle, Le Mans, Limoges, Lorient, Nantes, Orléans, Poitiers, Rennes, Tours). L'association organise un colloque international annuel sur un thème spécifique, ouvert à toute la communauté scientifique.

Comme l'indique son titre, la conférence de 2009 avait pour objectif de susciter une réflexion méthodologique et théorique sur la place accordée aux corpus dans l'analyse linguistique. L'appel à contributions décrivait l'objectif de la conférence en ces termes :

> « En un peu plus de 15 ans, les corpus ont pris une place grandissante voire envahissante dans de nombreux domaines de la linguistique. Ce rapide succès doit s'accompagner d'interrogations et de questionnements qui portent notamment sur :
>
> - le statut de l'exemple et du corpus (comment est envisagée la relation entre ces données et l'analyse linguistique ?)
> - la constitution des corpus (sont-ils conçus comme représentatifs et si oui, de quoi et comment peut-on s'assurer que cet objectif est atteint ?)
> - les utilisations (quelles exploitations des corpus sont envisagées et en quoi l'expérience acquise infléchit-t-elle la pratique et la réflexion ?)
>
> Le colloque permettra de réunir et de confronter les pratiques et les conceptions de chercheurs de divers domaines. Davantage que sur les aspects techniques de la linguistique de corpus, l'accent sera mis sur les questions épistémologiques que suscitent la constitution et l'exploitation des corpus, ainsi que sur la problématique du statut à accorder aux données sur lesquelles s'appuient les linguistes. »

2 Programme

Jeudi 4 Juin : Présentation de corpus et démonstrations

- Corpus oraux
  - Découvrir les langues de France : le site Corpus de la parole (Olivier Baude)
  - Nous comprenons-nous toujours bien ? (Paul Cappeau)
- Corpus d'apprenants
  - Apprendre de ses erreurs (Freiderikos Valetopoulos)
- Corpus multilingues
  - Pourra-t-on se passer un jour des traducteurs ? (Manuel Torrellas Castillo)
  - Explorer la presse dans deux langues… ou plus (Franck Zumstein & Hélène Chuquet)
- Textes du XVIe siècle
  - Les curieuses recettes de la Renaissance (Marie-Hélène Lay & Marie-Luce Demonet)

Vendredi 5 Juin :

- Emilie Caratini & Sandra Döring, « Les exemples non-prototypiques face à la (aux) théorie(s) et au(x) corpus »
- Laure Lansari & Agnès Leroux, « Quel(s) corpus pour l'analyse contrastive? »
- Raphaël Micheli & Vincent Capt, « Le statut de l'exemple en analyse textuelle des discours »
- Christophe Benzitoun, « Le linguiste, l'ordinateur et le corpus »
- Philippe Planchon, « La fonction méthodologique des corpus dans l'analyse des formes schématiques »
- Sylviane Granger, « Vingt ans d'analyse des corpus d'apprenants : leçons apprises et perspectives »
- Florent Perek, « Identification de constructions grammaticales en corpus »

Samedi 6 Juin :

- Paul Cappeau « Qu'est-ce qu'un bon exemple (oral)? »
- Olivier Baude, « Les Eslos, un corpus variationniste représentatif d'une communauté d'auditeurs ? »
- Bernard Laks, « La phonologie des usages: de l'exemplum au datum »

- Jérémie Segouat, Annick Choisier, Annelies Braffort, « Modélisation informatique de la langue des signes française à partir de corpus : la question de la représentativité »
- Anne-Marie Parisot, Julie Rinfret, Suzanne Villeneuve, « Les spécificités propres aux langues des signes dans la constitution et l'exploitation de corpus vidéo »
- Paolo Frassi, « L'exemple et le corpus en lexicographie: le cas du Trésor de la Langue Française »
- Jacques François, « L'évolution de la polysémie verbale documentée à partir des corpus textuels et des exemples lexicographiques »
- Cendrine Pagani-Naudet, « Le statut de l'exemple dans quelques grammaires du XVIe siècle »

Présentation de posters – « Corpus, didactique et apprentissage » :

- Véronique Quanquin, « Description et analyse des textes d'album : la nécessité d'un travail à partir d'un corpus »
- Martine Marquilló Larruy, « Les corpus didactiques existent-ils? (si oui… quelle est la finalité de leurs exemples?) »
- Freiderikos Valetopoulos, « Quelques réflexions sur les critères d'élaboration d'un corpus d'apprenants »

3 Contribution: « Identification de constructions grammaticales en corpus »

Rompant avec la thèse générativiste, des travaux récents en linguistique cognitive ont mis en évidence l'influence de la fréquence d'utilisation des structures linguistiques sur l'acquisition et l'évolution de la langue : l'usage modèle la grammaire (*cf.* Langacker 1987, Tomasello 2003, Bybee 2006). Ce changement de paradigme vers des modèles basés sur l'usage (*usage-based models*) a inévitablement conduit ses partisans à donner une place plus importante aux données empiriques, telles que celles fournies par un corpus. L'augmentation considérable de la taille des corpus disponibles et l'automatisation informatique des traitements ouvrent de nouvelles perspectives aux linguistes en leur permettant d'accéder à des données naturelles nombreuses et diversifiées.

Cependant, même en l'absence de la distinction compétence-performance, l'utilisation des corpus en linguistique cognitive n'échappe pas aux écueils méthodologiques objectés par la tradition générativiste, notamment que (i) en tant que fraction de l'ensemble infini des phrases d'une langue, un corpus est nécessairement déséquilibré et non représentatif et (ii) un corpus est incapable de présenter des preuves négatives (i.e. des phrases non grammaticales). Stefanowitsch et Gries (2003) soutiennent que ces limites sont uniquement dues à l'utilisation de fréquences brutes et donc biaisées, et pour remédier à ces problèmes, présentent l'analyse collostructionnelle, un ensemble de procédures statistiques permettant d'utiliser les fréquences brutes de manière appropriée. Grâce à cette méthode, les corpus peuvent être utilisés avec toute la rigueur scientifique nécessaire.

Dans cette présentation, nous partons de l'idée que les corpus devraient jouer un plus grand rôle dans l'analyse grammaticale. Suivant Stefanowitsch (2006), notre hypothèse de départ est qu'il est possible de dériver une grammaire à partir d'un corpus grâce à l'analyse collostructionnelle. Nous avons testé cette hypothèse dans le cadre théorique des grammaires de constructions, en particulier sur le cas des constructions argumentales, i.e. une approche constructionnelle de la structure argumentale des verbes. Nous avons conçu et implémenté trois indices statistiques de la forme de l'analyse collostructionnelle basés sur les caractéristiques distributionnelles des constructions argumentales prédites par le modèle de Goldberg (1995). Nous avons testé ces indices sur le corpus ICE-GB (1 million de segments) afin de vérifier à quel point ils permettent d'identifier trois constructions de l'anglais préalablement annotées.

Bien que les résultats affichent une tendance positive, ces indices s'avèrent à eux seuls insuffisants pour identifier des constructions argumentales en corpus. Il semble vain de baser une technique d'identification des constructions sur des critères purement statistiques et nécessaire de réintégrer des informations sémantiques si nous voulons identifier les paires forme-sens que sont les constructions. L'analyse collostructionnelle à elle seule est donc inadéquate et nécessite d'être adaptée. Nous présenterons quelques perspectives de développements futurs vers une nouvelle technique d'exploration de corpus.

Enfin, en dépit de ce succès mitigé, cette étude nous a conduit à aborder des points théoriques en confrontant explicitement les prédictions de la théorie aux données concrètes. Ces expérimentations nous ont ainsi livré plusieurs cas qui semblent entrer en conflit avec la définition théorique des constructions argumentales. Sur cette base empirique, nous suggérons que le modèle de Goldberg nécessite lui aussi des ajustements.

Bibliographie

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language* 82 (4), 711-733.

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Langacker, R. W. (1987). Foundations of Cognitive Grammar. Volume 1: Theoretical Prerequisites. Stanford University Press.

Stefanowitsch, A. (2006). Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 2.

Stefanowitsch, A. and S. T. Gries (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8 (2), 209-243.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

4 Bilan

D'un point de vue général, la conférence a réussi avec succès à réunir des acteurs très divers de la linguistique de corpus. Des disciplines très variées de la linguistique ont été représentées: didactique des langues, traduction, traitement automatique du langage, phonologie, lexicographie, linguistique historique, etc. L'audience a ainsi pu réaliser que les corpus possèdent à l'heure actuelle un vaste champ d'application. L'intégration de nouvelles modalités (son, vidéo, images, etc.) en supplément, voire remplacement, du texte transcrit, décuplent encore les possibilités offertes par les corpus, qui présentent maintenant une grande variété de formes. Une mention particulière est méritée par les langues des signes qui commencent elles aussi à disposer de leurs propres corpus vidéo, ce qui offre de nouveaux défis tant sur le point de vue technique (compilation et exploitation de données vidéo) que méthodologique. Par ailleurs, la conférence a aussi fait la part belle aux corpus de langue orale qui se développent de plus en plus dans l'hexagone, développement que beaucoup de linguistes ne sauraient qu'applaudir.

D'un point de vue personnel, la conférence fut l'occasion pour moi de prendre contact avec de nombreux chercheurs principalement français mais aussi étrangers, et de recueillir de nombreux commentaires sur mes travaux.