UNIVERSITÉ DE LORRAINE | crem centre de recherche EA 3476 sur les médiations communication, langue, art, culture

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

UNIVERZITA KARLOVA

10 avril 2018

JOURNÉE D'ÉTUDE FRANCO-TCHÈQUE

LINGUISTIQUE TEXTUELLE, LINGUISTIQUE DE CORPUS

Université de Lorraine
Campus du Saulcy, Metz
Salle Ferrari

crem.univ-lorraine.fr

Le but de la journée est de mettre en contact et de confronter des recherches de laboratoires spécialisés dans le traitement linguistique et informatique des textes. Il est également de questionner l'adéquation voire l'assimilation qui est souvent faite aujourd'hui entre linguistique de texte et linguistique de corpus. Le but sera alors de mettre en avant et de confronter des modèles de traitements « naturels », i.e. sémantiques, et formels, ou « assistés », qui montreront, sans les confondre, leur complémentarité, des points de vue épistémologique et méthodologique.

Responsables scientifiques

- **Guy Achard-Bayle** (Université de Lorraine, Crem*)
- **Ondřej Pešek** (Jihočeská univerzita v Českých Budějovicích, Ústav romanistiky [Université de Bohême du Sud, Institut d'études romanes]*)
  En partenariat avec l'Université Charles de Prague (Univerzita Karlova v Praze)

* Les universités de Bohême du Sud et de Lorraine, leurs équipes de recherches et départements de langue française et de sciences du langage, coopèrent étroitement depuis dix ans dans le domaine de la *linguistique textuelle* : création d'un double diplôme (*EMTex* European Master in Textology), codirections de mémoires, cotutelle de thèse... Cette coopération n'est pas le fruit du hasard : les recherches en sciences du langage des deux universités s'ancrent dans une tradition de linguistique textuelle qui se réclame de l'École de Prague (créée en 1926, donc bientôt centenaire). Par ailleurs les universités de Metz et de Nancy 2 (aujourd'hui réunies dans l'Université de Lorraine) ont joué un rôle majeur dans la diffusion des travaux de cette École en France – et au-delà à l'Ouest – dans les années 1970. Ce rappel « géo-historique » permet de justifier l'organisation partagée de ces journées. Il permet également de redire que linguistique textuelle et linguistique de corpus ne se confondent pas… Mais dès lors que les outils numériques entrent dans le traitement des données et des corpus textuels, ils viennent, en termes d'analyse et d'interprétation, enrichir ce traitement de manière notable – comme cette journée tentera de le montrer.

Orientations bibliographiques

Ablali D., Achard-Bayle G., Reboul-Touré S., Temmar M., 2018, « (Re-)Penser le texte et le discours dans le paysage actuel des sciences du langage », pp. 9-32, *in* : Ablali D., Achard-Bayle G., Reboul-Touré S., Temmar M., éds, *Texte et discours en confrontation dans l'espace européen*, Actes du Coll. Int. de Metz 19-21 sept. 2015, Berne, P. Lang (sous presse).

Achard-Bayle G., 2013, « Perspective fonctionnelle de la phrase : histoire-géographie d'une idée linguistique – Prague & l'Europe », *Verbum*, XXXV 1-2, pp. 3-10.

Pešek O., 2010, « La linguistique textuelle tchèque au seuil du XXI^e^ siècle : la genèse d'une discipline et la tradition pragoise », *Verbum* XXXII-2, pp. 263-282.

Comité d'organisation

Coresponsables du CO : Hussain BILHAJ et José Pedro FELICIANO, doctorants (Université de Lorraine)

Membres : Gisèle CASAGRANDA, doctorante (Université de Lorraine), Tarek FITOURY, doctorant (Université de Lorraine), LI Jun-Kai, doctorant (Université *Sun Yat-sen* Canton/Université de Lorraine), Klára ŽEMLIČKOVÁ, doctorante (Université de Bohême du Sud/Université de Lorraine)

Programme

09h15-09h30 : Accueil et mot du directeur du Crem, Jacques Walter

09h30-09h45 : Introduction des organisateurs
Guy Achard-Bayle (Université de Lorraine, Crem)
et Ondřej Pešek (Université de Bohême du Sud)

09h45-11h15 : **Session 1. Président : Ondřej Pešek (Université de Bohême du Sud)**

Šárka Zikánová et Eva Hajičová (Université Charles de Prague, Institut de linguistique formelle et appliquée)
Linguistique textuelle et linguistique de corpus pragoise (I)

11h15-11h30 : **Pause**

11h30-12h30 : **Session 2. Président : Guy Achard-Bayle (Université de Lorraine, Crem)**

Frédéric Landragin (CNRS/ENS Ulm/Université Paris 3 Sorbonne Nouvelle, Lattice)
Linguistique textuelle et linguistique de corpus en France (I)

12h30-14h00 : **Pause déjeuner**

14h00-15h00 : **Session 3. Président : Bilhaj Hussain (Université de Lorraine, Crem)**

Olga Nádvorniková (Université Charles de Prague, Institut d'études romanes)
Linguistique textuelle et linguistique de corpus tchèque (II) :
ouverture vers la linguistique contrastive et la traductologie

15h00-16h00 : **Session 4. Président : José Pedro Feliciano (Université de Lorraine, Crem)**

Maxime Amblard (CNRS/Inria/Université de Lorraine, Loria)
Linguistique textuelle et linguistique de corpus en France (II)

16h00-16h15 : **Pause**

16h15-17h00 : **Session 5. Animation : Klara Zemlickova, doctorante (Université de Bohême du Sud/Université de Lorraine)**

Discussion générale/clôture de la journée

Présentation des interventions et des intervenant·e·s

How a text is built: Coherence as a net of relations

Šárka ZIKÁNOVÁ (Faculty of Mathematics and Physics, Charles University, Prague)

Coherence as a feature of a text is based on a system of different types of relations. We will deal with three basic types of these relations. The flow of information is segmented in terms of *information structure,* so that an addressee of the text can differ known and new piece of information in a sentence and the basic aim why the sentence was used in the communication. In order to keep the addressee oriented about the entities in the text, rules of *coreference* and *bridging anaphora* are followed. Thoughts (sentences) as units, connected together with *discourse relations*, build a higher structure of a text. These aspects of a structure of a text will be demonstrated on the data from the Prague Discourse Treebank.

Interplay of intra- and inter-sentence structural relations in discourse analysis

Eva HAJIČOVÁ (Faculty of Mathematics and Physics, Charles University, Prague)

There are many factors making discourse an integrated whole, both inter-sentential and extra-sentential. In our contribution we will demonstrate on the material of the Prague Dependency Treebank how the information available in a many-sided annotation of corpus may be used for a study of some of the factors concerning discourse coherence. Two case studies will be discussed in a detail, one of which uses the information on topic-focus articulation and anaphoric relations to analyze the so-called thematic progressions, and the second builds on the notion of the hierarchy of the elements of the stock of knowledge assumed by the speaker to be shared by him and the addressee and follows the dynamic development of discourse based on the changes of the activation degrees.

Eva Hajičová et **Šárka Zikánová** sont rattachées à l'Institut de linguistique formelle et appliquée de la Faculté des mathématiques et de physique de l'Université Charles de Prague. C'est à cet Institut qu'a été conçu le Prague Dependency Treebank, un corpus de tchèque contemporain, étiqueté au niveau syntaxique selon les principes de l'approche fonctionnelle générative (relations de dépendance, cadres valenciels, articulation topique-focus) qui marie la tradition fonctionnelle de la linguistique pragoise aux principes génératives inspirés des travaux de

Noam Chomsky. Grâce à son étendue et grâce à la rigueur avec laquelle il est annoté, le Prague Dependency Treebank sert de modèle de référence. Le corpus est actuellement en train d'être étiqueté au niveau discursif – relations discursives, relations anaphoriques, relations thématiques. Les fonctionnalités de ce corpus permettent de relier les niveaux syntaxique et discursif, ce qui ouvre des opportunités inédites en matière de la recherche dans le domaine du traitement automatique du langage. Pr. Eva Hajičová est l'un des concepteurs principaux du corpus ; ses travaux, co-signés souvent avec Petr Sgall et Jarmila Panevová, font référence à l'échelle internationale (articulation topique-focus, négation, présupposition). Šárka Zikánová dirige le groupe de chercheurs qui réalisent l'étiquetage du corpus au niveau du discours.

Description et modélisation des chaînes de référence : le projet Democrat

Frédéric LANDRAGIN (CNRS/ENS/Université Paris 3 Sorbonne nouvelle, Lattice)

Cette présentation fait le point des réflexions linguistiques et des avancées méthodologiques du projet ANR DEMOCRAT, « Description et modélisation des chaînes de référence : outils pour l'annotation de corpus (en diachronie et en langues comparées) et le traitement automatique des langues » (http://www.lattice.cnrs.fr/democrat/). Nous y détaillerons les étapes qui ont amené à l'obtention d'une méthode d'annotation manuelle des expressions référentielles et des chaînes de référence dans des textes écrits, et nous montrerons comment la linguistique de corpus outillée permet d'envisager des études à la fois linguistiques et statistiques de ces objets d'étude.

Références

Landragin F., 2011, « Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits », *Corpus*, 10, pp. 61-80. Accès : http://journals.openedition.org/corpus/2010.
Landragin F., Poibeau T., Victorri B., 2016, « Analyse des références et des transitions référentielles : l'apport de la linguistique outillée », pp. 123-135, *in* : Sarda L., Vigier D., Combettes B., éds, *Connexion et indexation. Ces liens qui tissent le texte*, ENS Éd., Lyon.
Poudat C., Landragin F., 2017, *Explorer un corpus textuel. Méthodes – pratiques – outils*, Louvain-la-Neuve, De Boeck Supérieur.
Schnedecker C., Landragin F., éds, 2014, « Les chaînes de référence », *Langages*, 195.
Schnedecker C., Glikman J., Landragin F., éds, 2017, « Les chaînes de référence en corpus », *Langue française*, 195.

Le Lattice (Langues, Textes, Traitements Informatiques, Cognition) est un laboratoire de linguistique sous la triple tutelle du CNRS, de l'École Normale Supérieure et de l'Université Paris 3 Sorbonne Nouvelle. Les recherches y portent sur la linguistique et le traitement automatique des langues, avec une focalisation sur la linguistique du discours et un intérêt pour la linguistique de corpus et les humanités numériques. **Frédéric Landragin** y gère l'un des trois axes de recherche, intitulé « Mécanismes de composition du discours », et participe également à l'axe « Corpus, modélisations et traitements automatiques » ainsi qu'aux études transversales sur les méthodes et outils.

Shifts in the segmentation of sentences in translation: Reasons and consequences

Olga NÁDVORNIKOVÁ (Faculty of Arts, Charles University, Prague)

Splitting and joining of sentences in translation, i.e. shifts in segmentation, may have several reasons; on the one hand, structural differences between the languages involved, on the other hand, specific features of translation (co-called *translation universals*, *cf.* Baker 1996, Blum-Kulka 1986, etc.).

As for the first reason, according to Fabricius-Hansen (1996 and 1999) and Solfjeld (1996) languages differ in their (relative) information density. High information density languages (such as German or French) encode the discourse information in complex, hierarchical sentences, whereas low information density ones (e.g. Norwegian or Czech) prefer a more incremental, paratactic style. Thus, while translating from a high information density language to a low information density one,

splitting of sentences occurs more often than in the other direction of translation (*cf.* Nádvorníková 2017). More importantly, shifts in segmentation involve other changes: adding of connectives, moving non-finite constructions (gerunds, participial adjuncts, etc.) to the sentence level, changing textual coreference relations, etc. Nevertheless, shifts in segmentation of sentences occur also *independently* of the level of information density of the source/target languages. This may be explained by the influence of a general tendency of translations to the explicitation, simplification or normalization of the source text (so-called *translation universals*, Baker 1996, Blum-Kulka 1986).

In this paper, we will explore these reasons for and consequences of the shifts in segmentation of sentences on the basis of data from the core of the Czech-French-English part of the InterCorp parallel corpus (www.korpus.cz/intercorp, Čermák and Rosen 2012), which contains mainly narrative texts.

References

Baker M., 1996, "Corpus-based translation studies: The challenges that lie ahead", pp. 175-186, in: Somers H., ed., *Terminology, LSP and Translation: Studies in language engineering, in honour of Juan C. Sager*, Amsterdam, J. Benjamins.

Blum-Kulka S., 1986, "Shifts of Cohesion and Coherence in Translation", pp. 17-35, in: House J., Blum-Kulka S., eds, *Interlingual and Intercultural Communication: discourse and cognition in translation and second language acquisition studies*, Tübingen, Narr.

Čermák F., Rosen A., 2012, "The case of InterCorp, a multilingual parallel corpus", *International Journal of Corpus Linguistics*, 13, 3, pp. 411-427.

Fabricius-Hansen C. 1996, "Informational density: a problem for translation and translation theory", *Linguistics*, 34, pp. 521-565.

Fabricius-Hansen C., 1999, "Information packaging and translation: aspects of translational sentence splitting (German–English/Norwegian)", pp. 175-214, in: Doherty M. ed., *Sprachspezifische Aspekte der Informationsverteilung*, Berlin, Akademie Verlag.

Nádvorníková O., 2017, "Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus", pp. 445-461, in: Emonds J., Janebová M., eds, *Language Use and Linguistic Structure*, Palacký University, Olomouc.

Nádvorníková O., Šotolová J., 2016, "Za hranice věty: analýza změn v segmentaci na věty v překladových textech na základě francouzsko-českého paralelního korpusu", pp. 188-235, in: Čermáková A., Chlumská L., Malá M., eds, *Jazykové paralely*, Praha, NLN.

Solfjeld K., 1996, "Sententiality and translation strategies German-Norwegian", *Linguistics*, 34, pp. 567-590.

Olga Nádvorníková est enseignante-chercheure à la Faculté des Lettres de l'Université Charles de Prague (Institut des Études Romanes). Elle est responsable de la partie française du corpus InterCorp, un corpus synchronique parallèle couvrant un grand nombre de langues. Chaque texte de ce vaste corpus a une contrepartie tchèque. Ainsi, le tchèque est la langue pivot : pour chaque texte, il existe une version tchèque unique (originale ou traduction), alignée avec une ou plusieurs versions en langue étrangère. Le corpus représente ainsi un outil précieux pour la recherche contrastive au niveau de la langue et du discours. Olga Nádvorníková a menée plusieurs projets de recherche reliées à l'exploitation de ce corpus (structures gérondives, analyse du fonctionnement des verbes *dicendi*, etc.)

Interroger la cohérence de l'interaction dialogique par la formalisation

Maxime AMBLARD (CNRS/Université de Lorraine/Inria, Loria)

SLAM (Schizophrénie et Langage – Analyse et Modélisation) est un projet interdisciplinaire rassemblant linguistes, psychologues, informaticiens et philosophes. Dans ce cadre, les recherches s'articulent autour d'un corpus de transcription d'entretiens entre des patients diagnostiqués schizophrènes et des psychologues. Dans ces échanges, au delà de simples incongruités, des moments présentent de véritables difficultés d'interprétation sémantique et pragmatique. Nous

présenterons d'une part le corpus et les enjeux méthodologiques de son interrogation et d'autre part comment la formalisation, à partir de la SDRT, permet de rendre compte des dysfonctionnements apparaissant à cet endroit.

Références

Amblard M., Fort K., Demily C. *et al*., 2015, "Analyse lexicale outillée de la parole transcrite de patients schizophrènes", pp. 91-115, in: Sharp B., Delmonte R., eds, *Natural Language Processing and Cognition*, Berlin, De Gruyter. Repris en ligne : https://hal.univ-lorraine.fr/hal-01188677v2.

Asher N., Lascarides A., 2003, *Logics of Conversation*, Cambridge, Cambridge University Press.

Chaika E., 1974, "A linguist looks at 'schizophrenic' language", *Brain and Language*, 1, 3, pp. 257-276.

Fromkin V. A., 1975, "A linguist looks at "a linguist looks at 'schizophrenic language'"", *Brain and Language* 2, pp. 498-503.

Kamp H., Reyle U., 1993, *From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory*, Dordrecht, Kluwer Academic.

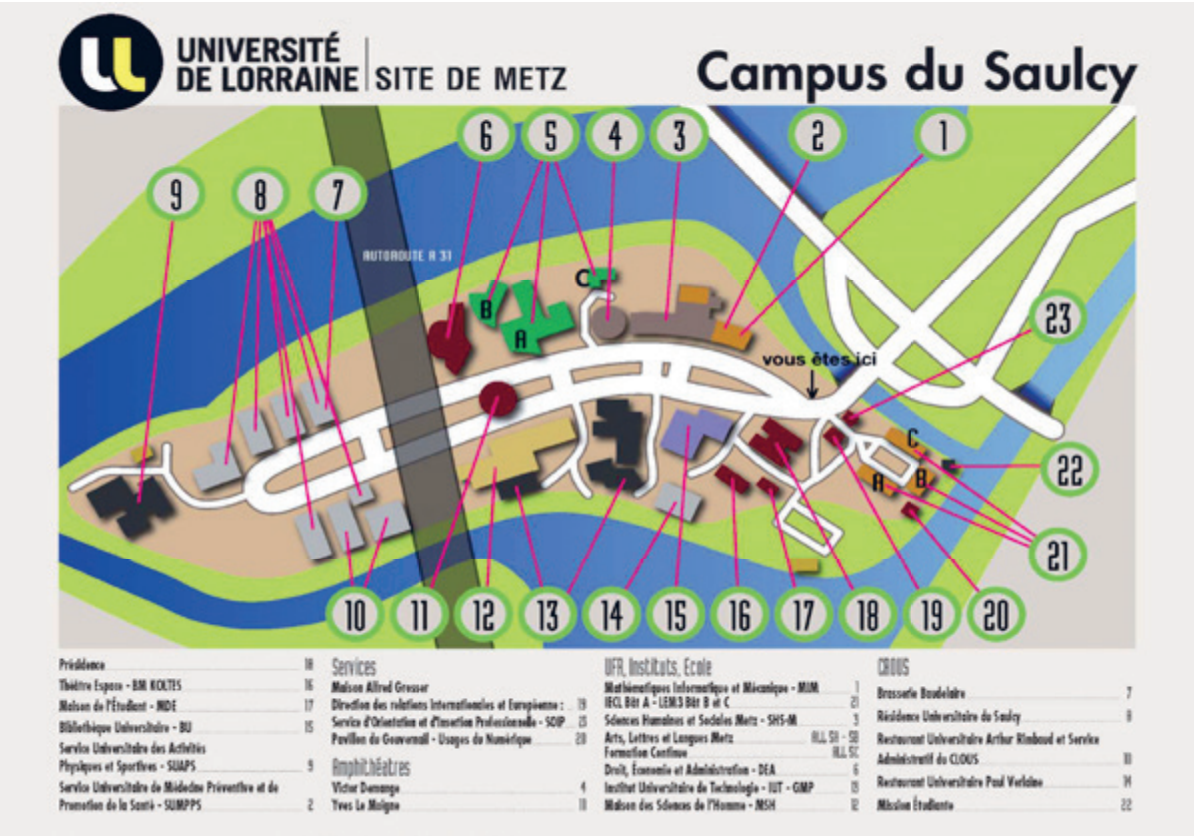
Musiol M., 2009, « Incohérence et formes psychopathologiques dans l'interaction verbale schizophrénique », pp. 219-238, in: Rozenberg J., Franck N., Hervé C., eds, *Des neurosciences à la psychopathologie: Action, Langage, Imaginaire*, Bruxelles, De Boeck.

Maxime Amblard est maître de conférences HDR en informatique à l'Université de Lorraine, rattaché au laboratoire d'informatique Loria UMR 7503 (http://www.loria.fr), équipe Sémagramme (http://semagramme.loria.fr) qui s'intéresse à la modélisation formelle de la sémantique de la langue. Ses travaux portent sur l'utilisation de la logique pour rendre compte de la sémantique, et par extension sur les aspects formels de la linguistique. Il est responsable du master en TAL, support français du master Erasmus Mundus *Language and Communication Technologies*.

Accès

La salle Ferrari correspond au nº 21 du plan, bâtiment A

L'arrêt du bus-tram *Mettis* au nº 15 (Bibliothèque Universitaire)



CREM
UNIVERSITÉ DE LORRAINE
UFR SHS-METZ — BP 60228
57045 METZ CEDEX
TÉL. : 03 72 74 83 35
CREM-CONTACT@UNIV-LORRAINE.FR

CREM.UNIV-LORRAINE.FR

PRATIQUES
JOURNALS.OPENEDITION.ORG/PRATIQUES

PUBLICTIONNAIRE
DICTIONNAIRE ENCYCLOPÉDIQUE ET CRITIQUE DES PUBLICS
PUBLICTIONNAIRE.HUMA-NUM.FR