Xavier-Laurent Salvador, ***Solution de représentation et d'accès au sens à travers les ressources numériques***

Résumé français
La réalisation de dictionnaires spécialisés nécessite la mise en commun de compétences multiples. C'est pourquoi nous avons architecturé une méthode de saisie de données lexicographiques permettant la consultation, la saisie, mais aussi leur validation: Isilex (http://www.isilex.fr). L'ensemble constitue une application web apparentée à un système de gestion de contenu tourné vers la lexicographie et l'édition de corpus, exportable et adaptable à différents besoins. Considérant que le dictionnaire lui-même est un corpus, la conséquence de la conception innovante d'une application XML/XQuery orientée telle qu'Isilex est d'ouvrir la voie vers différentes méthodes déductives de représentation des connexions sémantiques. Nous présupposons donc qu'il est pertinent d'établir deux niveaux d'analyse des relations entretenues par les mots : la connexion et la caractérisation du degré de similarité. Le premier cas relève de l'analyse de la synonymie. Le deuxième est une entreprise de cartographie du sens d'un vocabulaire donné basé sur les calculs de similarités au sein du corpus lexicographique.

Abstract
Creation of specialized dictionaries requires a complex of multiple skills. For this reason, we have developed Isilex (http://www.isilex.fr), a method of lexicographic data management enabling the data consultation, input and validation. Altogether, Isilex is a web application similar to a content management system, but oriented towards lexicography and corpus editing, exportable and adaptable to different needs. Considering the dictionary itself as a corpus, the consequence of the innovative design of XML / XQuery oriented application, such as Isilex, is to open the way to different deductive methods of representation of semantic connections. We therefore assume that it is pertinent to establish two levels of analysis for the relations between words: the semantic connection and the degree of similarity. The first case concerns the analysis of synonymy. The second is a mapping of the meaning of a given vocabulary based on calculations of similarities within some lexicographic corpus.

Firas Hmida, ***KRCTool : un concordancier bilingue pour l'aide à la révision***

Résumé français
En traduction spécialisée, une phase de révision est nécessaire afin de valider les traductions initialement attestées par le traducteur. Cette phase, qui veille à la cohérence du document produit, nécessite la mobilisation d'informations terminologiques accessibles à travers des glossaires et des outils de gestion dédiés.
Nous proposons un prototype de concordancier bilingue qui permet de saisir un terme et sa traduction, et fournit des Contextes Riches en Connaissances (CRC) alignés en corpus comparables spécialisés. Les évaluations manuelles et expérimentales menées avec des réviseurs montrent que les CRC bilingues proposés peuvent être perçus comme utiles en complément d'autres ressources d'aide à la traduction, malgré la difficulté de l'exercice.

Abstract
In specialized translation process, a revision phase is necessary to validate the initial translation proposed by the translator. This phase, which ensures the consistency of the document produced, requires the preparation of terminological information accessible through glossaries and dedicated management tools.
We propose a first prototype of bilingual concordancer that takes as input a term and its translation, and provides not parallel but aligned Knowledge-Rich Contexts (KRC) from specialized comparable corpora. Both the manual evaluation and a real experiment with student revisers show that our concordancer can assist revisers as a complement to their habitual ressources, despite the difficulty of the task.

Boniface Gnaguenon, ***Extraction de morphèmes polysémiques intra et interlangue. Les langues fon du Bénin et le yoruba du Nigeria***

Résumé français
La Fouille de Textes (FT), est un domaine spécialisé de la fouille de données. Elle est introduite dans les années 1995 par Feldman et Degan sous le terme "Knowledge Discovery in Textual Databases" (KDT), (Feldman, 1995). C'est en un ensemble de techniques de traitements informatiques qui permet d'extraire des connaissances selon des critères bien définis.
Notre objectif est d'utiliser des méthodes de Fouille de textes, pour extraire des morphèmes polysémiques tant en intralangue qu'en interlangue (fon et yoruba). Faire de la prédiction, déterminer si des mots porteurs de sens en intralangue, peuvent-ils servir de base à un apprentissage plus aisé des langues concernées ?

Abstract
Text Mining is a specialized field of data mining. It has been introduced around 1995 by Feldman and Degan under the term "Knowledge Discovery in Textual Databases" (KDTD) [Feldman, 1995]. The aim of the field is to find

computer processing techniques that makes it possible to extract knowledge according to well-defined criteria.
In our work, we use text mining methods to extract polysemic morphemes both monolingual and bilingual (Fon and Yoruba). We investigate the influence of these techniques on learning of the concerned languages.

Yuliya Korenchuk, ***Présentation des résultats de l'extraction terminologique sous forme de familles morphologiques***

Résumé français
Les résultats de l'extraction terminologique mono ou multilingue sont généralement présentés sous forme de listes. Cette présentation permet de les trier en fonction des valeurs utilisées pour leur identification. Toutefois, elle ne permet pas de visualiser les termes proches si leurs valeurs ne le sont pas. Par exemple, si le terme « géométrie » est beaucoup plus fréquent que « géométrique », il est possible qu'ils soient éloignés sur la liste de résultats. Cela entraîne une perte de temps pour un spécialiste humain qui se charge, par exemple, de les insérer dans un dictionnaire spécialisé.
La présentation des termes extraits sous forme de familles morphologiques mono ou multilingues permet de pallier au problème présenté ci-dessus. Les familles morphologiques sont des regroupements de mots partageant un ou plusieurs éléments qui peuvent être, en fonction de l'approche, des morphèmes ou des séquences de caractères. Dans le cadre de notre projet, les familles sont construites grâce aux n-grammes de caractères se substituant aux préfixes et aux racines. Nous pouvons imaginer comme exemple une famille de mots partageant le quadri-gramme de caractères correspondant au préfixe « nano » : « nanoparticule, nanostructure, nanotube », etc. Dans notre projet, les familles sont construites par rapports aux termes du domaine ou par rapport aux candidats extrait à partir du corpus.
Sur la base des exemples issus de notre projet de recherche, je présenterai les avantages et les limites de ces familles pour le regroupement de termes (français, anglais et allemands) et de leurs traductions, ainsi que l'utilité d'une telle organisation pour suivre les évolutions de la terminologie d'un domaine dans le temps.

Abstract
The results of mono or multilingual terminology extraction are usually presented in the form of lists. This presentation allows to sort them according to the values used for their identification. However, it does not allow visualization of close terms if their values are quite different. For example, if the term "geometry" is much more frequent than "geometric", it is possible that they would be distant on the results list, which leads to a loss of time for a human specialist willing, for example, to insert them into a specialized dictionary.
Presenting extracted terms in the form of mono or multilingual morphological families makes it possible to overcome the problem presented above. Morphological families gather words sharing one or more elements that may be, depending on the approach, morphemes or sequences of characters. In our project, the families are built using characters n-grams which replace prefixes and roots. We can imagine as an example a family of words sharing the character quadri-gram corresponding to the prefix "nano": "nanoparticle, nanostructure, nanotube", etc. In our project, the families are built for domain terms or for candidates extracted from a corpus.
On the basis of the examples from our research project, I will present the advantages and inconvenients of these families for presentation of terms (in French, English and German) and their translations, as well as the benefits of such structure for handling diachronical evolutions of terminology of some domain.

Gaël Lejeune, ***Découverte automatique multilingue de néologismes dans la presse en ligne***

Résumé français
Dans cette communication nous présentons une nouvelle méthode de détection automatique des néologismes dans la presse en ligne. Pour mettre au point cette méthode, nous avons exploité des résultats issus de la plateforme Neoveille. Cette plateforme a permis de collecter ces derniers mois plusieurs milliers de néologismes en associant une phase de pré-filtrage automatique de candidats néologismes et une phase d'expertise linguistique visant à classer ces candidats. Nous utilisons les données qui ont été annotées pour le français afin de construire un système remplissant deux objectifs : (I) diminuer la quantité de travail des experts linguistes en opérant une pré-sélection plus efficace des candidats et (II) être en mesure de traiter d'autres langues pour lesquelles nous n'aurions pas ou peu de ressources linguistiques et d'experts disponibles.

Abstract
We present a method for automatic detection of neologisms in online newspapers. We rely on the Neoveille platform which combines state-of-the-art processes to track linguistic changes for detecting neologisms candidates and a web platform for linguists to classify these candidates. We use human annotated data for french collected through this platform to build an automated neologism detection system with a two fold objective : (I) lessen the amount of work for linguists by improving the quality of the candidates (discarding noise) and (II) get results for languages for which we do not have experts or computational resources.

Yves Bestgen, ***Simplification et normalisation en traduction : évaluation d'une prédiction à propos de l'emploi des collocations par l'analyse automatique d'un corpus parallèle et comparable***

Résumé français
Un des objectifs principaux de la traductologie de corpus est de décrire les différences entre les textes traduits et les textes non traduits. Parmi les traits mis en évidence, la simplification et la normalisation sont susceptibles d'affecter la présence d'expressions phraséologiques dont l'importance dans l'emploi du langage est bien établie. Durant ces quinze dernières années, plusieurs études ont été menées afin d'évaluer l'hypothèse selon laquelle les traducteurs auraient tendance à suremployer des collocations fréquentes dans la langue cible au détriment des collocations rares, souvent plus créatives. Elles ont abouti à des conclusions contradictoires, mais, surtout, elles ont été menées sur un nombre très limité de séquences phraséologiques, faisant planer un doute sur la possibilité de généraliser les conclusions. La présente recherche a pour objectif de tester cette prédiction au moyen d'une analyse totalement automatique d'un corpus parallèle et comparable. La méthode proposée consiste à extraire d'un texte traduit ou non traduit toutes les séquences de deux mots consécutifs et à les rechercher dans un vaste corpus de référence natif afin de leur attribuer deux scores d'associations, l'un privilégiant les collocations observées très fréquemment dans la langue alors que l'autre privilégie les collocations nettement plus rares. L'analyse de la section journalistique du corpus PLECI met en évidence, tant en français qu'en anglais, des différences statistiquement significatives entre les textes traduits et non traduits en accord avec l'hypothèse, mais les tailles d'effet sont faibles.

Abstract
One of the main objectives of corpus-based translation studies is to describe the differences between translated and untranslated texts. Among the features highlighted, simplification and normalization are likely to affect the presence of phraseological expressions whose importance in the use of language is well established. During the last fifteen years, several studies have been carried out to evaluate the hypothesis that translators tend to overuse frequent collocations in the target language to the detriment of rare, often more creative collocations. They yielded contradictory conclusions, but above all they were conducted on a very limited number of phraseological sequences, casting doubt on the possibility of generalizing the conclusions. The aim of this research is to test this prediction by means of a fully automatic analysis of a parallel and comparable corpus. The proposed method consists in extracting from a text all the sequences of two consecutive words and searching them in a large native reference corpus in order to attribute to them two associations scores, one privileging the collocations observed very frequently in the language while the other favors much rarer collocations. The analysis of the newspaper section of the PLECI corpus reveals statistically significant differences between translated and untranslated texts consistent with the hypothesis, both in French and in English, but the effect sizes are small.

Hanno Biber et Evelyn Breiteneder, ***Wittgenstein's Nephew's Nephews. Creating and Making Use of Parallel Text Corpora of the Literary Text*** **Wittgensteins Neffe** ***by Thomas Bernhard***

Abstract
Wittgensteins Neffe by the Austrian writer Thomas Bernhard (1931-1989) is about the narrator's relationship with his "friend Paul, the nephew of the philosopher whose Tractatus Logico-philosophicus is now known to the whole of the scholarly world, to say nothing of the "pseudoscholarly world." The narrative literary text explores the memorizing recognition of a friend, whereby the theme of a failed recognition of the genius in society is presented in form of an attempt of recognition of a friendship between the suffering artist and his pal, a mad genius: "At the very time when I was lying in the Hermann Pavilion, my friend Paul was some two hundred yards away in the Ludwig Pavilion, though this, unlike the Hermann Pavilion, did not belong to the pulmonary department, and hence to the so-called Baumgartnerhöhe, but belonged to the mental institution Am Steinhof."
New methods of creating and analysing parallel text corpora for translation studies and comparative investigations into language use are going to be presented in this paper. A parallel corpus of the text *Wittgensteins Neffe* is going to be aligned on a sentence level and in this manner compared with other language versions of the translated text. At a first step, an English version of the text will be compared to its German original. The language of the text will be made an object of investigation in electronic form, which needs to be carefully analysed by taking into account all structural elements on all levels of linguistic as well as structural literary features.
The principles of digital literary studies in the context of corpus research can be used for a detailed analysis of the parallel texts and are to be considered particularly fruitful for this purpose when combined with methods of computational philology. The methods of digital literary studies help to create new perspectives on the text in translation, which becomes accessible to the readers and the translators in various ways by making use of the parallel corpus. A literary text can therefore be studied and presented as an entity that is part of a corpus of texts which are historically and linguistically related to its source. Even a rather small narration can be used as such an exemplary object of scholarly and scientific interest to determine the potential of digital corpus research methods in literary text studies, in translation studies as well as in the field of computational philology.

Résumé français
Wittgenstein Neffe de l'écrivain autrichien Thomas Bernhard (1931-1989) parle de la relation du narrateur avec son «ami Paul, neveu du philosophe dont le Tractatus Logico-philosophicus est maintenant connu de tout le monde savant, pour ne rien dire du monde pseudoscholaire». Le texte littéraire narratif explore la reconnaissance mémorisante d'un ami, par laquelle le thème d'une reconnaissance échouée du génie dans la société se présente sous la forme d'une tentative de reconnaissance d'une amitié entre l'artiste souffrant et son copain , un génie fou: «Au moment même où je me trouvais dans le pavillon Hermann, mon ami Paul était à deux cents mètres du pavillon Ludwig, bien que, contrairement au pavillon Hermann, n'appartenait pas au service pulmonaire, donc à la Baumgartnerhöhe soi-disant, mais appartenait à l'institution mentale Am Steinhof. "
Cette article présentera de nouvelles méthodes de création et d'analyse de corpus de textes parallèles pour les études de traduction, ainsi que les enquêtes comparatives sur l'utilisation de la langue. Un corpus parallèle du texte *Wittgensteins Neffe* sera aligné sur un niveau phrastique permettant ainsi une comparaison avec d'autres versions linguistiques du texte traduit. Dans un premier temps, une version anglaise du texte sera comparée à son original en allemand. La langue du texte (sous forme électronique) fera l'objet d'une étude minutieuse, tenant compte de tous les éléments à tous les niveaux linguistiques, ainsi que des traits structurels propres à la littérature.
Les principes des études littéraires numériques s'appuyant sur des corpus, peuvent être utilisés pour une analyse détaillée des textes parallèles et doivent être considérés comme particulièrement fructueux à cet effet lorsqu'ils sont combinés avec des méthodes de philologie computationnelle. Les méthodes des études littéraires numériques aident à créer de nouveaux angles de vue sur le texte en traduction, qui deviennent accessibles aux lecteurs et aux traducteurs de diverses façons via le corpus parallèle. Un texte littéraire peut donc être étudié et présenté comme une entité faisant partie d'un corpus de textes qui sont historiquement et linguistiquement liés à sa source. Même une narration de taille modeste peut être utilisée comme un objet exemplaire d'intérêt scientifique et scientifique pour déterminer le potentiel des méthodes exploitant des corpus numériques en littérature, en traduction, ainsi que dans le domaine de la philologie informatique.

Anastasiia Andrusenko, ***Comparative Analysis of Meta Discourse in Arabic and Spanish Research Articles***

Résumé français
Pléthore d'études dans le métadiscours utilisent l'anglais comme point de référence commun, ce qui reflète l'importance de l'anglais comme lingua franca dans le monde de l'éducation et de la recherche (Markkanen et al., 1993, Mauranen, 1993, Valero-Garcés, 1996, Moreno, 1997, 2004, Mur-Dueñas, 2011). Les anthologies sur la rhétorique contrastive n'incluent d'études sur l'espagnol (Connor, 1996: 52). Cependant, des recherches approfondies sur les contrastes entre l'anglais et l'espagnol ont été menées par divers linguistes espagnols (Dafouz-Milne, 2008, Milne, 2003, 2006, Moreno, 1997, 2004, Mur-Dueñas, 2011; Valero-Garcès, 1996). L'absence de littérature comparant les conventions rhétoriques espagnoles et arabes a motivé cette étude. Sur la base d'un corpus de 90 articles recueillis dans 6 revues de linguistique, cette étude cherche à identifier les similitudes et les différences dans l'utilisation d'euphémismes et d'emphases dans des articles scientifiques écrits en langue maternelle espagnole et arabe.

Abstract
A plethora of studies in metadiscourse use English as a common point of reference. This reflects the importance of English as a lingua franca in the global education and research community (Markkanen et al., 1993; Mauranen, 1993; Valero-Garcés, 1996; Moreno, 1997, 2004; Mur-Dueñas, 2011). Anthologies on contrastive rhetoric have not included studies of Spanish (Connor, 1996: 52). However, extensive research on English-Spanish contrasts has been conducted by various Spanish linguists (Dafouz-Milne, 2008; Milne, 2003, 2006; Moreno, 1997, 2004; Mur-Dueñas,2011; Valero-Garcés, 1996). Concerning English-Arabic contrastive studies very interesting seems the study of El-Seidi (El-Seidi, 2000). The lack of literature comparing Spanish and Arabic rhetorical conventions has been the motivation for this study. Based on a corpus of 90 articles collected from 6 journals of linguistics, this study seeks to detect the similarities and differences in the use of "hedges" and "boosters" in native Spanish and native Arabic linguistics research articles.