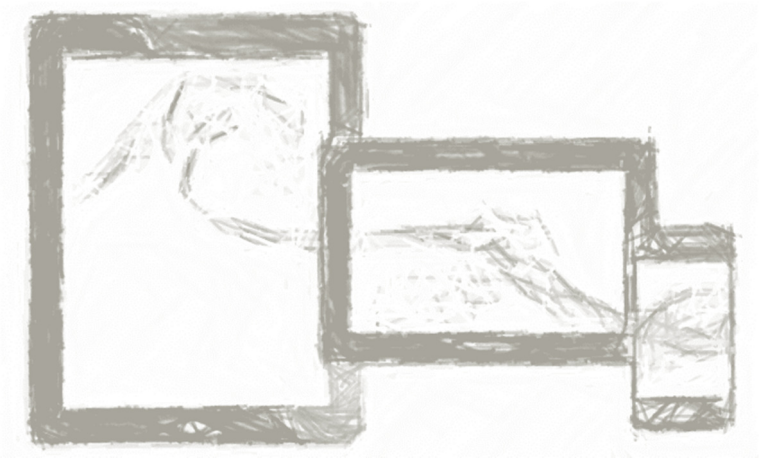
Journées Internationales de Recherche
Médias sociaux et corpus de communication médiée par les réseaux (CMR). Annotation, analyse, données libres

International Research Days
Social Media and CMC Corpora for the eHumanities

23-24 octobre/October 2015, Rennes (France)

Livret de programme
Programme Book



Comité d'organisation / Programme Committee :

Georges ANTONIADIS (U. Grenoble, France)
Valérie BEAUDOUIN (Télécom, ParisTech, France)
Michael BEISSWENGER (U. Dortmund, Germany)
Thierry CHANIER (U. Blaise Pascal, France)
Isabella CHIARI (U. Sapienza, Italy)
Linda HRIBA (U. Orléans, France)
Gudrun LEDEGEN (U. Rennes 2, France - responsable)
Julien LONGHI (U. Cergy-Pontoise, France)
Jean-Philippe MAGUE (ENS Lyon, France)
Amanda POTTS (Lancaster University, United Kingdom)
Céline POUDAT (U. Nice, France)
Ciara R. WIGHAM (U. Lyon2, France - responsable)
Torsten ZESCH (Duisburg-Essen University, Germany)

Table des matières/of contents

Vendredi/Friday 23/10

9:00-9:30	**Accueil / Welcome en B332 – Bâtiment B (plan au dos)**			
9:30-10:30	**Chercheur invité / Key note**: The DiDi Project: Collecting, Annotating, and Analysing South Tyrolean Data of Computer-mediated Communication**.** *Egon W. Stemle* **Salle E323-324 – Bâtiment E**			
	Salle B223		**Salle B332**	
10:30-11:10	**SMS**	Un corpus longitudinal de SMS d'adolescents : de la constitution du corpus à l'analyse de l'écriture SMS. *Volckaert-Legrier Olga Goumi Antonine, Bert-Erboul Alain & Bernicot Josie*	**Telecollaboration & L2**	The development of a databank in institutional-integrated teletandem: a long road towards research in Applied Linguistics. *Aranha Solange*
11:10-11:50		Méthode hybride de normalisation lexico-syntaxique des SMS. *Kogkitsidou Eleni*		Developing and sharing teletandem data. *Leone Paola*
11:50-12:30		De la constitution d'un corpus de SMS : Comment gérer un flux de données personnelles. *Ghliss Yosra & André Frédéric*		Researching identity-in-interaction in multilingual online chat: Critical multimodal discourse analysis of CMC corpora. *Klimanova Liudmila*
12:30-14:00	*Lunch*			
14:00-15:00	**Chercheure invitée / Key not**e**:** Wikipedia as a corpus resource for linguistic research. *Angelika Storer* **– Salle J. Léonard – Bâtiment A**			
	Salle B223		**Salle B332**	
15:00-15:40	**Wikipedia**	The French CoMeRe Wikiconflits subcorpus. *Poudat Céline &Grabar Natalia*	**Multimodalité/y**	Construction d'un large corpus libre de conversations écrites en ligne synchrones et asynchrones en français à partir de Ubuntu-fr *Hernandez Nicolas & Soufia Salim*
15:40-16:20		Les discussions Wikipedia : un corpus pour caractériser le genre « discussion ». *Ho-Dac Lydia-Mai, Laippala Véronika*		
16:20-16:40	*Break*			
	Salle B223		**Salle B332**	
16:40-17:20	**Grands Corpus / Large Corpora**	Towards an encoding standard for social media and CMC: Experiences from German and French corpus projects using TEI. *Beisswenger Michael, Chanier Thierry et al.*	**Evaluation de médecins / Physician Rating**	What's up doc? A corpus of Physician Rating Portals. *Johannes Knopp & Tobias Betzin*
17:20-18:00				Analyse linguistique et contrastive du genre CMC « commentaires d'évaluation des médecins en ligne ». *Coline Baechler*

Samedi/Saturday 24/10

<table>
<tr><td></td><td colspan="2">Salle B223</td><td colspan="2">Salle B332</td></tr>
<tr><td>9:00-
9:40</td><td rowspan="3">Réseaux sociaux / Social Networks</td><td>Pratiques langagières en langue bretonne sur les réseaux socionumériques : méthode d'une étude de cas.
Blanchard Jean-François</td><td rowspan="3">Annotation / POS tagging</td><td>The JANES corpus of Slovene user generated content: construction and annotation.
Fiser Darja, LjubesisNikola & Erjavec Toma</td></tr>
<tr><td>9:40-
10:20</td><td>Analyse sociolinguistique d'un Réseau Social d'Entreprise (RSE) – Données linguistiques, représentations et pratiques langagières.
Petersen Jette Milberg</td><td>sms4science.ch: A multi-lingual challenge for Part-of-Speech tagging.
Simone Ueberwasser</td></tr>
<tr><td>10:20-
11:00</td><td>LinkedIn, le média social de promotion de l'identité professionnelle quelles stratégies discursives pour la création de liens interpersonnels ?
Meyer Jeanne</td><td>An extended tag set for annotating parts of speech in CMC corpora.
Beißwenger Michael, Bartz Thomas, Ehrhardt Eric & Storrer Angelika</td></tr>
<tr><td>11:00-
11:20</td><td colspan="4">Break</td></tr>
<tr><td>11:20-
12:20</td><td colspan="4">Chercheur invité / Key note: Annotation des corpus plurilingues — l’expérience CLAPOTY. Pascal Vaillant – Salle Amphi E1 (Bâtiment E)</td></tr>
<tr><td>12:20-
13:30</td><td colspan="4">Lunch</td></tr>
<tr><td></td><td colspan="2">Salle B223</td><td colspan="2">Salle B332</td></tr>
<tr><td>13:30-
14:10</td><td rowspan="3">Twitter</td><td>Le corpus Polititweets: enjeux de constitution d'un corpus de tweets et propositions d'analyses.
Longhi Julien</td><td rowspan="3">Apprentissage-enseignement - L2 / L2-Learning-teaching</td><td>The Affordances and Disadvantages of WordReference Forums as a Space for Intercultural Exchange.
Mayne Elizabeth</td></tr>
<tr><td>14:10-
14:50</td><td>Outils linguistiques et informatiques pour l'analyse des controverses.
Jackiewicz Agata</td><td>Genre analysis of expert and learner corpora of news-based computer-mediated communication.
Marchand Tim & Akutsu Sumi</td></tr>
<tr><td>14:50-
15:30</td><td>L'influence des discours d'accompagnement sur le partage social : identifier et analyser les discours d'escorte sur Twitter.
Brigitte Sebbah, Bigey Magali, Dario Compagno, Mercier Arnaud & Pignard-Cheynel Nathalie</td><td></td></tr>
<tr><td>15:30-
15:40</td><td colspan="4">Conclusion / Closing</td></tr>
</table>

The DiDi Project: Collecting, Annotating, and Analysing South Tyrolean Data of Computer-mediated Communication.

Egon W. Stemle (Invited speaker)

Following a sociolinguistic user-based perspective on language data, the project DiDi investigated the linguistic strategies employed by South Tyrolean users on Facebook. South Tyrol is a multilingual region (Italian, German, and Ladin are official languages) where the South Tyrolean dialect of German is frequently used in different communicative contexts. Thus, regional and social codes are often also used in written communication and in computer mediated communication. With a research focus on users with L1 German living in South Tyrol, the main research question was whether people of different age use language in a similar way or in an age-specific manner. The project lasted 2 years (June 2013 - May 2015).

We created a corpus of Facebook communication that can be linked to other user-based data such as age, web experience and communication habits. We gathered socio-demographic information through an online questionnaire and collected the language data of the entire range of social interactions, i.e. publicly accessible data as well as non-public conversations (status updates and comments, private messages, and chat conversations) written and published just for friends or a limited audience. The data acquisition comprised about 150 users interacting with the app, offering access to their language data and answering the questionnaire.

In this talk, I will present the project, its data acquisition app and text annotation processes (automatic, semi-automatic, and manual), discuss their strengths and limitations, and present results from our data analyses.

Wikipedia as a corpus resource for linguistic research

Angelika Storer (Invited speaker)

Wikipedia is already known to be a valuable resource for many research fields. Until now, most linguistic studies have focused on article pages and on the content encoded in the written language. In my talk, I will demonstrate with examples how linguistics can profit from three additional perspectives on Wikipedia as a corpus resource, namely:

(1) Wikipedia as a social media corpus: in this perspective not only are the article pages relevant, but also the interaction between Wikipedia authors on talk pages and other communication channels.

(2) Wikipedia as a multimodal corpus: in this perspective not only written text is the object of analysis, but also to the integration of media objects in the article pages.

(3) Wikipedia as a multilingual corpus: Wikipedia articles of different language versions are interconnected through interlanguage links opening up innovative options for contrastive and cross-lingual research.

I will report on studies that used Wikipedia article and talk pages in order to test hypotheses about language style and register variation. On this basis I want to discuss (a) which linguistic and interactional features are most relevant for investigating wikipedia as a social media corpus, and (b) how these features may be annotated in accordance with the TEI Special Interest Group on Computer-Mediated Communication.

Annotation des corpus plurilingues — l'expérience CLAPOTY

Pascal Vaillant (Invited speaker)

Methods in corpus processing have until recently been more focused on multilingual corpora (texts in different languages about the same domain) than on plurilingual corpora (corpora with an internal linguistic heterogeneity). This may be due to the fact that they have emerged in natural language processing contexts, mostly in practical applications to written texts, and not in the field of applied linguistics, where the focus is rather on spontaneous, genuine utterances of non-standard speech, and where phenomena of combined use of different languages are not rare.

However, observing -and understanding- language contact phenomena has a growing appeal not only to linguistic specialists, but also to all those who have an interest in mining corpora of spoken language, or non-standard written language.

Within the frame of the ANR CLAPOTY project, a team of linguists and computer scientists has worked on the representation and encoding of oral transcripts, displaying different situations of language contact (with a total of 40 languages from different linguistic areas and various typological profiles).

The choice that was made, in order to allow automatic mining of the corpora without losing the complexity of real-world linguistic phenomena, was to precisely annotate all the linguistic data on the observed units, without classifying them a priori in descriptive categories, the exact definition of which is still often debatable (e.g. borrowing, calque, code switching).

To this purpose, the CLAPOTY team has developed an annotation schema in compliance with the latest standards with respect to transcription (Unicode) and markup (XML). This schema follows the inspiration of the TEI (Text Encoding Initiative), extending it where needed (namely, for the annotation of language plurality). In this model, linguistic units (at all levels) may be described as pertaining to one language or another, and even to many languages at the same time. The model is able to represent the richness and versatility of spontaneous linguistic utterances, where speakers actually often "float" between two languages.

RESUMES / ABSTRACTS
(En ordre alphabétique par le premier auteur / In alphabetical order by first author)

The development of a databank in institutional-integrated teletandem: a long road towards research in Applied Linguistics

Solange Aranha

The purpose of this presentation is to describe the process of collecting, organizing and storing data generated by learners´ online oral interactions to make them available for researchers. Teletandem (TELLES, 2006) context implies that a pair of proficient speakers of different languages are in contact via web tools (Skype, mainly) with the purpose of learning the language of the other and teaching his/her own. Teletandem activities are oriented by three theoretical principles: autonomy, reciprocity and language separation, which means that each one is responsible for establishing goals for his/her learning; both participants are expected to contribute to the learning process of the other, respecting their strategies and preferences; and the same amount of time should be devoted to each of the languages. Within this micro context of teaching and learning in each partnership, activities related to speaking/listening – when the pair is interacting via Skype – and reading/writing – tasks to be shared with the partner- occur. Within the macro context, other activities, as tutorials and questionnaires, are also part of teletandem environment.

Institutional integrated teletandem (ARANHA and CAVALARI, 2014) practice implies that a group of students from one university interacts with a group from a foreign university during the same amount of time, on the same day and time, and follows a number of procedures established by their teachers/tutors/professors. It also implies that students will be graded for the whole process and that teletandem activities are integrated into the foreign language classroom in each university. Each group has approximately 15 students at each end and they usually interact for eight weeks, during one hour per week. The groups that have been part of this data collection are English and Portuguese learners.

The amount of data generated by the groups of students involved in the activity (about five groups a year) is enormous and valuable for researches interested in studying a wide array of aspects of teaching-learning process, grammar errors, peer-correction, skills acquisition, culture, mediation, among others. The ontology (ARANHA, LUVIZARI-MURAD and MORENO, ongoing) is based on a system – institutional integrated teletandem – in which different texts are produced, divided and classified according to their nature: oral conversations through Skype, chats used during these conversations, reflexive diaries, questionnaires, written production (in three moments: first draft, revised version and final text and in two different languages. All the stored data have been used by local researchers once it is stored in one specific computer at the university, but the importance of the bank for applied researches is undeniable. Questions about on how to make the data available online and implications on ethical issues should arise with this presentation. Our group is especially concerned with exposing students´ faces, personalities and attitudes and sharing their points of view that might be aggressive and politically incorrect sometimes. Participants sign a letter of consent by which they authorize the use of personal data for specific research purposes.

Analyse linguistique et contrastive du genre CMC « commentaires d'évaluation des médecins en ligne »

Coline Baechler

Les dernières décennies ont été marquées par l'avènement de nouvelles formes de communication électronique scripturale autorisant divers degrés d'interaction synchrone ou asynchrone (forums, courriels, clavardage, blogues etc.). En parallèle, la diffusion d'Internet et des technologies de l'information ont permis aux internautes de multiplier et confronter leurs sources d'information, aussi bien dans le cadre professionnel que privé (encyclopédies en ligne, presse en ligne, e-commerce etc.). Ainsi, dans une situation d'achat par exemple, le consommateur potentiel peut s'informer sur le produit envisagé en consultant des commentaires d'internautes ayant utilisé le produit et évalué sa qualité. De même, les utilisateurs d'Internet ont la possibilité d'évaluer la qualité de services et de prestataires de services. Il existe donc des sites Internet d'évaluation des enseignants, des avocats ou des médecins, sur lesquels il est possible d'attribuer des notes et de publier des commentaires personnels.

Notre travail de recherche s'intéresse aux sites Internet de notation des médecins et plus particulièrement aux « commentaires d'évaluation des médecins en ligne » en tant que genre CMC en France et en Allemagne. Si les sites d'évaluation des médecins sont largement établis en Allemagne (Jameda, Docinsider, etc.), ils font encore face à de vives réactions et problèmes d'acceptation en France (Notetondoc, Mytoubib, etc.).

Cette étude s'inscrit dans le cadre d'un projet de recherche interdisciplinaire, à l'interface entre la linguistique, l'informatique et les sciences de la communication et des médias, s'interrogeant sur les facteurs qui contribuent à instaurer crédibilité et confiance dans la communication électronique anonyme.

L'apport de l'informatique se situe à deux niveaux différents. Premièrement, en ce qui concerne la formation des corpus d'analyse, le travail des informaticiens (Data and Web Science Group, University of Mannheim) permet une collecte semi-automatisée d'une grande quantité de données issues du web ainsi que leur représentation sous forme tabulaire « prêtes-à-analyser ». Deuxièmement, les différentes solutions informatiques permettent une analyse linguistique quantitative et qualitative des données du corpus.

A partir de notre approche linguistique, nous nous intéressons aux structures langagières des commentaires d'évaluation des médecins. Nous formulons l'hypothèse que nous sommes en présence d'un genre de texte en cours de développement et d'établissement. En effet, celui-ci se caractérise par une grande hétérogénéité lexicale, syntaxique et pragmatique, contredisant peut-être en partie des recherches linguistiques menées sur les types de textes ; ces dernières décrivent des caractéristiques typiques aux niveaux macro-textuels et micro-textuels, plus ou moins homogènes au sein d'une famille de textes.

Pour cela, nous analysons un corpus composé de commentaires issus de trois sites d'évaluation français et trois sites d'évaluation allemands. Le travail s'effectue à plusieurs niveaux du langage et combine des méthodes linguistiques quantitatives sur la totalité du corpus (par exemple nombre de mots, fréquence de mots, collocations) ainsi que des méthodes qualitatives sur un corpus réduit (par exemple analyse du contenu, du lexique, de la syntaxe et des actes de langage). Comme il n'existe encore que peu d'analyses linguistiques portant sur les évaluations des médecins en ligne, le schéma d'analyse a été conçu de manière résolument vaste, afin de prendre en compte le plus d'éléments possible, qui pourront être éventuellement approfondis dans des études ultérieures.

Au niveau interculturel, la confrontation du genre CMC « site d'évaluation des médecins » dans les deux aires linguistiques et culturelles permet de mettre en avant des différences au niveau de l'utilisation du langage et des stratégies évaluatives mises en place. En outre, les résultats de l'analyse mettront en exergue les caractéristiques culturelles qui influencent le jugement de la qualité du service médical, en ouvrant des perspectives vers l'anthropologie de la médecine.

Ma contribution présentera plus en détails les bases théoriques et méthodologiques de la recherche en cours, et décrira les premiers résultats de l'analyse contrastive des textes français et allemands. Ceci me donnera l'opportunité de mettre en lumière quelques particularités langagières et culturelles du genre CMC « commentaires d'évaluation des médecins en ligne ».

An extended tag set for annotating parts of speech in CMC corpora

Michael Beißwenger, Thomas Bartz, Eric Ehrhardt, Angelika Storrer

Automated part of speech (PoS) tagging is a challenging issue in building social media and CMC corpora (cf. Baldwin et al. 2013, Bartz et al. 2013, Eisenstein 2013, Giesbrecht & Evert 2009, Glaznieks & Stemle 2014, Horbach et al. 2015). Two kinds of problems have to be tackled: on the one hand, taggers trained on newspaper corpora perform poorly on parts of speech which are typical for spontaneous, spoken language. On the other hand, PoS tag sets such as the STTS ("Stuttgart Tübingen Tagset") for German do not provide tags for types of tokens that are typical for CMC discourse such as emoticons, addressing terms, hashtags or URLs.

In our presentation, we will discuss the STTS 2.0 tag set for German CMC (Beißwenger et al. 2015) as a suggestion to solve the aforementioned issues. STTS 2.0 has been created in the context of the scientific network Empirikom (http://www.empirikom.net) and builds on the categories of the canonical version of STTS (Schiller et al. 1999) which has to be regarded as a de-facto standard for annotating parts of speech in written German data. STTS (1999) focuses mainly on parts of speech in genres of edited text (e.g. newspaper articles, novels). STTS 2.0 introduces two types of new tags: (1) tags for phenomena which are specific for CMC / social media discourse, and (2) tags for phenomena which are typical for spontaneous spoken language in colloquial registers. These extensions are useful for corpus-based research of both CMC and spoken conversation. A common tag set for phenomena of type (2) will also facilitate the comparison of written CMC with transcripts of spoken conversation.

Tab. 1 provides an overview of the changes between STTS (1999) and STTS 2.0. The categories defined for CMC-specific items as well as the extensions for frequent types of colloquial contractions are true extensions to STTS (1999). The categories defined for phenomena which are typical for spontaneous spoken language restructure parts of the categories of STTS (1999). Nevertheless, all modifications and extensions defined in STTS 2.0 result in a category set which is still downwardly compatible with STTS (1999) and therefore allows for interoperability with corpora that have been tagged with STTS (1999) (e.g. the DWDS, the "Digital Dictionary of the German Language": http://www.dwds.de).

STTS 2.0 is currently being tested and evaluated in various research projects and it is used as the gold standard in a project preparing a community shared task for automatic linguistic annotation of German CMC (EmpiriST2015, http://empirikom.net/Themen/SharedTask). First results from training a tagger for applying the STTS 2.0 categories have been reported by Horbach et al. (2015) as part of the results of the BMBF project "Tools and Analyses for the Orthographical Monitoring of Current German Writing Practice" (http://www.schreibgebrauch.de/). STTS 2.0 is also used for tagging parts of speech in the project ChatCorpus2CLARIN, a curation project of the CLARIN-D infrastructure initiative (http://de.clarin.eu) led by Michael Beißwenger (TU Dortmund) and Angelika Storrer (University of Mannheim) in cooperation with researchers from the Institute for the German Language (IDS) Mannheim and the Berlin-Brandenburg Academy of Sciences (BBAW). The goal of the project is to adapt the Dortmund Chat Corpus, an existing, annotated 1MWord collection of German chat data (http://www.chatkorpus.tu-dortmund.de/, Beißwenger 2013), to the CLARIN standards and to represent the data according to the schemas developed by the TEI-SIG "Computer-mediated communication" (http://www.tei-c.org/Activities/SIG/CMC/). By using STTS 2.0 for the PoS annotations the project aims to create CMC resources that allow for a systematic corpus-based comparison of CMC genres with transcriptions of spoken dialogues available in the corpora of spoken language at the IDS Mannheim. For this purpose, the STTS 2.0 tags are defined in a way that is compatible with the extensions to STTS (1999) used at the IDS Mannheim for the PoS annotation of FOLK, the Mannheim "Research and Teaching Corpus of Spoken German" (http://agd.ids-mannheim.de/folk.shtml) (Schmidt 2014, Westpfahl & Schmidt 2013, Westpfahl 2014).

Our presentation will focus on the annotation of PoS in the ChatCorpus2CLARIN project. With the help of examples, we will introduce the CMC-specific part of the STTS 2.0 tag set and report about our experiences. As an outlook, we will discuss how the categories of STTS 2.0 may be related to PoS categories used for CMC corpora in other languages.

Towards an encoding standard for social media and CMC: Experiences from German and French corpus projects using TEI

Michael Beißwenger, Thierry Chanier (Mini panel presentation)

The internet and social media have given rise to a broad range of new communicative genres such as chats, forums, text messaging, interactions on wiki talk pages and in blog comments, via Twitter, on social network sites, and in multimodal 3D environments. A standard for the representation of these genres which is compliant with existing standards for text and speech corpora in the field of Digital Humanities would foster interoperability between language resources as well as the analysis and automatic exploitation of resources of that kind. Since 2013, corpus projects from several European countries and for different languages have been cooperating in a special interest group (SIG) of the Text Encoding Initiative (TEI, http://tei-c.org) on the creation of a standard for encoding CMC and social media genres. The TEI encoding framework (TEI P5) does not currently include any models for the representation of the linguistic and structural peculiarities of social media and CMC genres. The SIG "computer-mediated communication" is designing such models and testing them in corpus projects and with data from a broad range of genres (cf. Beißwenger et al. 2012, Chanier et al. 2014, Margaretha/Lüngen 2014) - the goal being to propose an addition to the TEI standard. The two papers of this mini panel will outline intermediate results and modeling suggestions (CMC-specific models and schemas, annotation experiences) from the work of the SIG on the example of German and French CMC and social media corpora.

Paper 1: Schemas and experiences from modeling German CMC corpora in TEI

Adrien Barbaresi, Michael Beißwenger, Eric Ehrhardt, Alexander Geyken, Axel Herold, Marc Kupietz, Lothar Lemnitzer, Harald Lüngen, Angelika Storrer, Andreas Witt

This paper reports about TEI schemas and modeling experiences from several German corpus projects: ChatCorpus2CLARIN, German Wikipedia and Usenet news corpora in DeReKo, Deutsches Referenzkorpus zur internetbasierten Kommunikation (DeRiK), German WhatsApp Corpus and the DWDS Blog corpus.

The paper will outline the encoding architecture and basic features of the TEI schemas that have been designed for these projects and give an overview of experiences and practices in using these schemas for corpus annotation. From a linguistic point of view, a fundamental challenge in adopting the TEI for the representation of social media and CMC lies in the hybrid nature of the respective genres: Written CMC shares characteristics both with written text (= written language, use of text structuring and formatting) and spoken conversation (= dialogic, sequential organization) while at the same time it differs from both genres. Even though a range of models from the current TEI P5 standard can be adopted for the representation of structural features in CMC, written CMC cannot completely be represented using the TEI modules "text structure" and "transcriptions of speech" (cf. Beißwenger et al. 2012). By the help of examples, we will discuss for which structural and linguistic features of CMC we need new, CMC-specific models, and how these models could be implemented in the TEI encoding framework. In the current version of the schemas presented in the paper, these models are defined as customizations of the TEI standard (cf. http://www.tei-c.org/Guidelines/Customization/). In view of the increasing interest in building and analyzing CMC data in the humanities and computer sciences, a future version of the TEI guidelines should implement models for CMC as part of the standard.

Paper 2: The CoMeRe French CMC corpora and their modeling in TEI

Thierry Chanier, Céline Poudat, Ciara R. Wigham

CoMeRe is a national project involving researchers from 8 different research units to develop a repository of CMC all modeled within the same extension of the TEI (Chanier et al. 2014). The project was carried out from 2013 to 2015 with the support of Corpus-Ecrits (http://corpusecrits.huma-num.fr/, a national research consortium on written corpora) and Ortolang (http://www.ortolang.fr, a national infrastructure for tools and corpora on French language). Three key principles underlie CoMeRe: variety, openness and standards.

"Variety" is one of our keywords since we have assembled interactions stemming from networks such as the Internet or telecommunications (mobile phones), as well as mono- and multimodal, and synchronous and asynchronous communications. The genres covered within CoMeRe include text or oral chats, email, discussion forums, blogs, tweets, audio-graphic conferencing systems (conference systems with text, audio, and iconic signs for communication), or even collaborative working/learning environments with verbal and nonverbal communication. The corpus also offers a variety of discourse situations: public or private conversations, as well as informal, learning, and professional situations.

"Openness" is our second keyword. The first set of 11 corpora has been released (http://hdl.handle.net/11403/comere) as open data on Ortolang. This openness is mainly driven by the future inclusion of CoMeRe within the forthcoming French National Corpus - the latter is expected to become a reference for studies in French linguistics. On the other hand, our wish to release CoMeRe corpora as open data stems from the fact that, although studies on new CMC communication genres draw much attention, there is currently no existing dataset with significant coverage to form the basis for systematic research.

"Standards" refers to two different aspects. Firstly, corpora have been structured and referred to in a uniform way. The TEI-IS is the model developed as an extension of the TEI in order to encompass the Interaction Space (IS) of CMC multimodal discourse. "Standards" also refers to the uniform basic level of automatic annotations, related to segmentation and part of speech (POS) tagging which is underway.

The present communication will expose the IS model, and the classification of multimodal acts we developed to distinguish between verbal (as studied by corpus linguistics) and non-verbal acts related to body, groupware, and iconic systems (Ciekanski/Chanier 2008; Wigham/Chanier 2013a; Wigham/Chanier, 2013b). Examples of all these different types of multimodal interactions will be given out of samples extracted from our corpora.

Pratiques langagières en langue bretonne sur les réseaux socionumériques : méthode d'une étude de cas

Jean-François Blanchard

Tirée d'une thèse récente en sociologie, la communication proposée porte sur l'analyse d'un corpus de recherche construit autour des pratiques langagières en breton sur les réseaux socionumériques. Il s'agit principalement de l'expression en breton, ou au sujet du breton, sur l'internet. La recherche exploratoire a orienté la construction de l'objet de recherche vers une investigation portant à la fois, sur les acteurs, les usages et les représentations individuelles et collectives. Cette approche globale d'un fait social a fait naître une hypothèse dégageant trois pôles d'analyse structurés dans un modèle d'interprétation. Le modèle proposé met en présence interactive les trois éléments suivants : a) les représentations individuelles et collectives de la langue bretonne, b) les formes d'appartenance : liens et pratiques, et enfin, c) l'institutionnalisation de la langue dans l'espace numérique.

La constitution du corpus de recherche a, dans un premier temps, été guidée par le souci d'une exhaustivité en se plaçant dans une perspective synchronique — produire un compte rendu des pratiques en breton — et diachronique — reconstituer la petite histoire de l'internet en breton. Cette première étape de data mining, a été réalisée manuellement avec l'appui d'un logiciel d'exploration du web (un crawler). Le premier corpus constitué comportait environ cinq cents adresses. Cependant, l'évolution des usages, l'émergence du Web 2.0 et la montée en charge des réseaux sociaux en breton, à partir de 2011-2012, ont rendu nécessaire l'ajout de l'archivage de messages collectés sur les plateformes Facebook et Twitter sur de longues périodes. Le logiciel d'analyse qualitative Nvivo a été utilisé pour l'extraction et le codage de ces données.

Les choix méthodologiques opérés pour l'exploitation du corpus ont été guidés, naturellement, par la définition de l'objet de recherche et l'hypothèse de travail, mais aussi par les possibilités d'investigations offertes dans le corpus recueilli. En fonction des objectifs recherchés, différentes méthodes ont été mises en œuvre : quantitatives (statistiques descriptives, métriques du web, paramètres structuraux des réseaux) et qualitatives (analyse textuelle, analyse du discours, sémiotique). L'analyse du corpus a été complétée, notamment, par des entretiens avec des acteurs de l'internet en breton et une enquête auprès de jeunes du lycée Diwan. Sur le corpus initial, deux cent treize sites, ou adresses URL, ont été retenus. Les réseaux constitués par les liens hypertexte ont fait l'objet d'une analyse critique portant sur la valeur scientifique du graphe pour l'analyse des réseaux. De plus, l'intégration au corpus de l'historique complet du forum de Wikipedia en breton, depuis sa création, en 2006, a apporté des éléments sur les débats métalinguistiques (comment écrire le breton sur l'internet ?) et épilinguistiques (quelles sont les représentations du breton en débat ?). L'historique des échanges dans le groupe Facebook e brezhoneg depuis son origine ainsi que des tweet indexés par le hashtag : #bzhg ont permis d'étayer une typologie des acteurs de l'internet en breton.

Au titre des résultats, cette étude de cas, apporte, en premier lieu, des éléments permettant de montrer le rôle glottopolitique des réseaux socionumériques (RSN), dans la mesure où ils peuvent agir dans les rapports entre les langues. En deuxième lieu, la langue bretonne trouve sur les RSN des modalités de pratiques langagières structurant les formes existantes et créant de nouvelles. Enfin, en troisième lieu, la lecture politique du projet de revitalisation linguistique s'analyse comme un projet de recontextualisation de la langue. Toutefois, la présente communication porte essentiellement sur l'élaboration méthodologique du projet de recherche allant de la formulation de l'hypothèse à la présentation des résultats. Cette communication met en exergue la place centrale du corpus de communication médiée et son ajustement à l'objet de recherche. Elle souligne également l'intérêt d'un recours ponctuel aux méthodes classiques, mais aussi, et surtout, l'avantage heuristique d'une approche expérimentale en méthodologie.

The JANES corpus of Slovene user generated content: construction and annotation

Darja Fišer, Nikola Ljubešić, Tomaž Erjavec

We present an overview of the current results and on-going activities of the JANES project, http://nl.ijs.si/janes. The project, started in 2014, focuses on nonstandard Slovene, and its objectives are to build a large corpus of user-generated Slovene as found on the internet, to serve as the basis for linguistic analyses and to help improve language-technology tools for processing texts written in nonstandard Slovene.

We have compiled the first version of the JANES corpus of user-generated Slovene which contains four types of text: tweets, internet forums, comments on internet news items, and blogs, with the complete corpus containing just over 160 million tokens. Tweets were collected with the tool TweetCat (Ljubešić et al., 2014a), which was constructed specifically for compiling Twitter corpora of smaller languages. The tool, with the help of a small lexicon of language specific Slovene words, first identified users that predominantly tweet in Slovene, as well as their friends and followers. Then, in the period of almost two years, it collected their tweets, also updating the list of users. This resulted in the JANES tweet subcorpus, which contains, in addition to the actual text of each tweet also its meta-data, i.e. the author's username, date and time of the tweet, as well as the number of times the tweet was retweeted and favorited. This subcorpus currently contains 61 million tokens with most of the tweets being from 2013 -- 2014. It should be noted that the majority of these tweets turns out to be not real user-generated content but rather news feeds, adverts and similar material produced by professional authors.

For forums and news portals, from which we extracted user comments, we chose six sources which are among the more popular in Slovenia and thus contain the most texts. The forum portals selected are on motoring, health issues and science, while the news portals are of the Slovene national television and radio, and the most popular left-wing and right-wing weekly magazine. Because the structure of the crawled pages differs significantly between the sources, we wrote a per-source extractor using the Beautiful Soup (http://www.crummy.com/software/BeautifulSoup/) module. With this we were able to extract the relevant text only, avoiding the large proportion of noise, such as adverts, irrelevant links, etc. From the news portals we extracted user comments with meta-data, such as the title, identifier and URL of the article and the identifier, date, and username of the author of the comment. The same meta-data was extracted for forum posts, with additional information on the hierarchical forum thread they belong to, which makes it possible to observe only posts on particular topics, such as health/plastic-surgery. The subcorpus on forums contains 47 million tokens, while the news user comments comprise 15 million tokens. The majority of the comment texts comes, as with tweets, from 2013 - 2014, while the forums contain a longer time span, with similar portions of text coming from each of the years between 2006 and 2014.

Finally, we also made a preliminary version of the subcorpus of Slovene blogs, which we currently simply extracted from the deduplicated version of the slWaC 2.0 corpus of the Slovene Web (Erjavec and Ljubešić, 2014). From slWaC we extracted all the texts that have the string ''blog'' in their domain, resulting in a corpus of 38 million tokens. This is a temporary solution because, in contrast to forums and user comments, this subcorpus does not have a dedicated extractor, so the internal structure of blogs or their detailed metadata is not preserved. In particular, this means that the main text of the blog is not separated from any comments following it.

The subcorpora were then linguistically annotated. We first tokenised the texts, using the standard model for Slovene, a part of the ToTaLe tool (Erjavec et al., 2005). In the next stage we normalised (standardised) the word tokens, using the method from Ljubešić et al. (2014b), which uses character-based statistical machine translation. The CSMT translation model was trained on 1000 key-words from the Slovene tweet subcorpus (as contrasted with a corpus of standard Slovene), and their manually determined standard equivalents. Then, using the models for standard Slovene of the ToTaLe tool, we part-of-speech tagged and lemmatised the standardised word tokens. It should be noted that the method for word normalisation is currently only a prototype, which we plan to improve in the continuation of the project.

As mentioned, the majority of gathered tweets are in fact not user generated texts, but rather tweets posted by news agencies, adverts and similar, i.e. text in standard language that we are in fact not interested in. In order to be able to extract only texts in non-standard language, we developed a method to predict the non-standardness of text. We propose that non-standardness comes in two basic varieties, technical and linguistic, and developed a machine-learning method to discriminate between standard and non-standard text in these two dimensions. We first manually annotated a small dataset from the JANES corpus, developed the features (29 in all) used for building our regression models, and trained and evaluated the resulting system with encouraging results. For example, the model performs significantly better than the standard OOV method for predicting non-standardness, even without using an external lexicon.

As a final issue we mention our intention of making the developed corpus freely and openly available, via a concordancer and for download. We discuss issues of copyright, privacy protection and terms of service on the social media platform providers and how we plan to solve them.

De la constitution d'un corpus de SMS : Comment gérer un flux de données personnelles sensibles ?

Yosra Ghliss, Frédéric André

L'avènement de l'ère numérique a conduit, au cours de la dernière décennie, au développement de supports électroniques de médiation qui constituent de nouveaux domaines de recherches en matière de Sciences du Langage. La téléphonie mobile, participant à l'évolution de ces moyens de communications, a connu un essor sans précédent, et reste aujourd'hui un secteur d'activité marqué par une forte croissance. De de nombreux chercheurs se sont emparés de cette matière langagière afin de l'analyser de plus près, comme J. Anis, qui a par exemple commencé à définir l'ambigüité entre genre écrit et oral qui caractérise le SMS, en introduisant la notion de *parlécrit* (2005). Cougnon, pour définir cette même notion, parle quant à elle *d'écrit spontané* (2008), ou encore Panckhurst, avançant que le SMS est avant tout un genre écrit, s'intéresse à ce qu'elle nomme l'*écriture SMS* (2009).Plus récemment, J. Bernicot a également travaillé sur la pratique du SMS dans des contextes d'apprentissage de l'écrit, chez de jeunes adolescents (2011). Or, pour pouvoir étudier les discours comme les SMS, il faut avoir au préalable constitué son corpus de textos authentiques ; et c'est là où se joue le vrai défi : s'il est plus facile d'accéder et d'exploiter les types de communications comme les twitts ou les publications facebook (vu leur caractère publique), les SMS, eux, restent très problématiques dans leur collecte mais aussi leur mise en ligne.

Notre contribution propose de réfléchir sur les enjeux qui caractérisent la constitution de corpus SMS, en nous appuyant sur l'étude du cas du corpus 88milSMS. Effectuée dans le cadre du projet SMS4SCIENCE[1], la collecte de ce corpus a donné plus de 88 000SMS (d'où son intitulé 88milSMS, *Panckhurst et al.* 2014). Dans le processus de construction de ce grand corpus, nous nous intéresserons ici particulièrement à l'étape relative à l'anonymisation et l'annotation où il est question d'éthique de recherche et de droits en vue de la publication.

Tout commence alors par une collecte de SMS, rendue possible par la mise en place d'une campagne de communication intitulée « faites don de vos SMS à la science ». Elle présentait le projet au grand public notamment dans la région de Montpellier. Une fois la collecte menée à son terme, une autre étape essentielle s'ouvre: l'anonymisation. En effet, afin de transformer les SMS de l'état brut en un corpus linguistique exploitable par d'autres chercheurs, il est demandé aux linguistes de trouver un équilibre entre les contraintes juridiques liées à l'acquisition puis l'exploitation de données personnelles et la conservation de l'authenticité des SMS, afin d'assurer de futures observations empiriques et des résultats pertinents dans leur traitement.

Nous discuterons par ailleurs la posture épistémologique adoptée dans le traitement de ce corpus. Ainsi, exposerons-nous les difficultés rencontrées lors de l'étape d'anonymisation, la phase cruciale où l'enjeu était d'articuler l'anonymat des donateurs avec nos aspirations linguistiques. Puis, les différentes balises utilisées pour l'annotation. Il nous semble essentiel d'aborder ces questions afin de partager nos expériences, et ainsi trouver des solutions qui faciliteront peut-être demain la recherche concernant la communication médiée par les réseaux.

[1] Initié en 2004 par Cédrick Fairon et ses collègues à l'Université catholique de Louvain, en Belgique

Construction d'un large corpus libre de conversations écrites en ligne synchrones et asynchrones en français à partir de Ubuntu-fr

Nicolas Hernandez, Soufian Salim

Nous présentons nos efforts pour la construction d'un large corpus libre en français compilant à la fois des conversations écrites en ligne sur des modalités synchrone (chat) et asynchrones (forum et courriel) et des textes explicatifs sur une même période, et ce, autour d'un même type de situation discursive à savoir l'assistance à la résolution de problèmes.

Nos motivations scientifiques sont multiples :

- améliorer le traitement linguistique des différentes modalités en exploitant leur caractère comparable (*e.g.* en travaillant sur la portabilité d'analyseurs construits sur une modalité pour en traiter une autre)
- maîtriser l'alignement voire la "traduction" d'un contenu porté par une modalité textuelle (texte explicatif ou communication écrite en ligne) vers une autre
- mieux comprendre la structure et le fonctionnement de ce type de conversation ainsi que leurs interactions
- permettre des activités de Traitement Automatique des Langues (TAL) (e.g. construction de modèles statistiques dédiés à la reconnaissance de structures discursives)
- permettre la diffusion de nos données dans une perspective de reproductibilité et de réutilisation (*open science* (Nielsen, 2011))

Ces recherches se veulent soutenir des applications de recherche d'information inter-modalités textuelles (e.g. recherche d'une solution dans une modalité distincte de la demande), de gestion automatique de la documentation (e.g. détection de l'absence ou de l'obsolescence d'une solution) ou d'assistance à la production et au diagnostic (e.g. aide à la formulation de problème, évaluation de la complétude de messages réponses, suivi d'un sujet sur plusieurs fils de discussion et modalités).

Pour pouvoir soutenir de telles applications, nous envisageons l'analyse des conversations

- en termes de leur organisation en actes de dialogue.
- et en fonction des objets conceptuels manipulés, à savoir les problèmes et leurs possibles solutions.

Nos objets d'étude sont plus spécifiques que les Communications Médiées par les Réseaux (CMR) puisque nous nous intéressons aux conversations médiées par les réseaux. Nous travaillons sur la proposition d'un modèle de conversations écrites en ligne offrant un cadre de traitement générique prenant en compte les spécificités des modalités manipulées. La définition de ce modèle passe par l'adaptation de la taxonomie des actes du dialogues et des relations discursives définies dans DIT++ par (Bunt et al., 2012) pour décrire des interactions orales.

Bien que notre initiative ne se situe pas sous l'égide du projet CoMeRe (Chanier et al., 2014), nous adhérons d'une part à la nécessité d'exploiter des standards pour décrire données et méta-données, et d'autre part à la nécessité de réfléchir aux possibilités d'exploitation (*e.g.* diffusion) d'un corpus en amont de sa construction. Notre expertise en TAL fait que nous ne retenons pas la TEI ou son extension aux CMR pour représenter nos corpus bien que nous adhérons aux modèles de données qu'ils sous-tendent.

À notre connaissance seul le corpus Simuligne du projet LETEC (Reffay et al., 2014) compile des interactions en français sur plusieurs modalités (chat, couriel et forum) s'accompagnant de ressources pédagogiques autour de la situation d'apprentissage en ligne du français langue étrangère. Notre corpus diffère dans ses objectifs et dans ses caractéristiques (e.g. objets d'étude, situation discursive, quantité de données).

Nous proposons de faire état de notre avancement dans la mise en place d'un cadre matériel nous permettant de réaliser notre recherche, en particulier : les données collectées, nos considérations pour leur diffusion, nos choix en termes de formatage des méta-données, des données et des résultats d'analyse, ainsi que les pré-traitements automatiques réalisés.

Collecte des données

Nos données ont été collectées à partir de ressources mises à disposition par les communautés du logiciel libre en particulier la communauté Ubuntu-fr (francophone).

Outre une documentation, la communauté met à disposition des moyens pour communiquer via des forums de discussion, des listes de diffusion par courriel et des canaux Internet Relay Channel (IRC). L'accès en consultation à ces contenus est public sur l'Internet. La documentation (ubuntu-fr-doc, 2015) est diffusée sous licence CC BY-SA v3.0. Les autres ressources (forum, couriel et chat) n'ont pas de licence particulière. Le droit d'auteur s'applique donc par défaut au contenu des messages. Ubuntu-fr étant l'éditeur, l'utilisateur lui délègue l'usage de ses messages dans la base de données. Si celui-ci manifeste son refus de participer à ce corpus, nous devrons supprimer ses messages de notre copie.

Contrairement aux autres ressources la documentation possède un contenu mouvant et doit être soumis au versionnage. Actuellement la documentation est seulement récupérable par aspiration du site en ligne. Nous

travaillons avec la communauté pour systématiser son archivage et versionnage. En attendant depuis novembre 2014, nous aspirons quotidiennement une copie du site. Concernant les courriels, une archive incrémentale est distribuée publiquement. Les forums évoluent aussi incrémentalement mais aucune archive publique n'est disponible. Nous les avons collectés par aspiration. Concernant le chat, nous avons mis en place une procédure de journalisation depuis novembre 2014.
Nous avons donc à disposition des conversations sur les modalités forum et courriel depuis leur création (à savoir 2004), et nous possédons environ 6 mois de données synchronisées pour toutes les modalités.
Entre autres caractéristiques, ces sous-corpus sont en croissance perpétuelle, représentatifs d'écrits récents et témoins d'une évolution des formes de communication sur une période d'une dizaine d'années. Par ailleurs, les circonstances font qu'ils pourront être complétés par des versions en langue anglaise elles-mêmes disponibles sur l'Internet (Lowe et al., 2015).

A titre indicatif, les six derniers mois de données collectées représentent pour les courriels à 80 conversations, 240 messages et 60 participants; pour les forums à 12 000 conversations, 90 000 messages et 7 000 participants, et pour le chat 60 000 messages et 1 600 participants. La documentation compte 4 631 pages HTML et 4 301 618 tokens mots.
Formatage des conversations, des méta-données et des résultats d'analyses
Nous proposons un modèle générique pour décrire les méta-données d'une conversation, de ses messages et de ses participants. Ce travail constitue un développement possible aux recommandations du projet CoMeRe. Ce modèle résulte d'une généralisation des conversations observées et tient compte des évolutions sur le format des messages Internet (http://tools.ietf.org/html/rfc6854). Il intègre des attributs communs et spécifiques permettant de catégoriser thématiquement une conversation, de comptabiliser les vues d'un message, de décrire le rôle d'un participant (e.g. ambassadeur, expert, client). Il se cristallise en schéma XML.
La question du format de sérialisation des annotations (résultats d'analyse automatique ou bien d'une activité manuelle) sur la donnée est pour nous secondaire. Les fonctions premières de ce format sont le stockage et l'échange. Ce qui nous semble primordial c'est de permettre à tout utilisateur de pouvoir éditer et projeter des annotations sur la donnée dans sa mise en forme originale sans la dénaturer. Cela implique l'adoption de certains principes (e.g. le "stand off" annotation) et l'utilisation de certains outils, tels qu'Apache UIMA (Ferrucci et Lally, 2004) ou Webanno (Eckart de Castilho et al., 2014), offrant la couche d'abstraction nécessaire à la manipulation de données annotées. En pratique, nos données annotées sont sérialisées en XML Metadata Interchange (XMI), standard pour l'échange d'informations de métadonnées UML basé sur XML. Ces considérations nous éloignent du format de la TEI mais pas nécessairement de son modèle conceptuel.
Traitements
Chaque modalité bénéficie de prétraitements particuliers (gestion de l'encodage et extraction des contenus textuels pour les courriels, segmentation textuelle qui prenne en compte le balisage HTML pour les forums, nécessité de reconstituer les conversations pour les courriels et les chats…).
En termes d'anonymisation nous discutons plusieurs alternatives : aucune anonymisation et suppression de messages à la demande, anonymisation restreinte aux méta-données, et aucune diffusion de données, seulement des outils de collecte et de traitement des corpus.
Sur la base d'une segmentation en pseudo-phrases porteuses d'actes du dialogue qui tient compte des spécifités des CMR, nous avons initié un travail d'annotation manuelle en actes du dialogue adaptés aux CMR d'après nos observations.
Conclusion
Ce travail, qui s'inscrit dans le cadre du projet ODISAE (www.odisae.com), a bénéficié du soutien du fond unique interministériel (FUI) 17. Une version du corpus sera livrée pour juillet 2015 (avec outillage et document technique pour décrire et manipuler les conversations). La dissémination par le réseau Ortolang est une option envisagée.

Les discussions Wikipedia : un corpus pour caractériser le genre «discussion»

Lydia-Mai Ho-Dac, Véronika Laippala

Cette présentation propose une description des caractéristiques intra-linguistiques des discussions Wikipedia, forum de discussion associé à chaque article de l'encyclopédie Wikipedia. Après un exposé des propriétés qui font de ces textes un objet d'étude particulièrement intéressant pour les linguistiques de corpus, nous présenterons la procédure de constitution du corpus de discussion et une première description quantitative du corpus constitué. Nous finirons sur une présentation rapide d'un ensemble d'études linguistiques envisagées sur ce corpus.

Wikipédia est une encyclopédie libre et coopérative à laquelle tout internaute peut contribuer en modifiant ou créant un article ou encore en postant un message dans une page de discussion portant sur la structure, la pertinence, le contenu de l'article. Les contributeurs peuvent également participer à des forums portant sur la totalité du projet de Wikipedia, parmi lesquels les « cafés et bistrots »[2] (e.g. « Forum des Nouveaux » pour accueillir les nouveaux, « Le salon de médiation » pour « résoudre dans un cadre serein des conflits », etc.) ; ou encore des discussions autour des choix d'édition[3], des questions légales[4]...

Cette communauté fonctionne par le travail des internautes actifs qui sont ainsi amenés à acquérir un statut dans la communauté. Ils peuvent ainsi devenir « patrouilleurs » et avoir le droit et le devoir de « marquer une modification comme n'étant pas un vandalisme », ou encore « administrateur » dont le rôle est de « protéger et maintenir la qualité des éditions du projet »[5]. Tous ces rôles participent à la modération de la Wikipédia. En effet, tout ajout ou modification (que ce soit dans un article ou une discussion) est soumis à une phase de contrôle qui décide de sa publication.

Un corpus constitué de discussions Wikipedia représente un nombre important de caractéristiques avantageuses pour les linguistiques de corpus. Premièrement, il s'agit d'un forum de discussion libre de droits (licence Creative Commons by-sa) qui existe depuis 2001 et dans lequel les contributeurs interagissent autour d'une thématique explicite et détaillée soit dans l'article associé soit dans le type de « café » (e.g. « Le salon de médiation »). Le contexte de production de ces discussions est ainsi beaucoup plus accessible que pour tout autre forum de discussion.

Autre point important, notamment pour l'application de techniques en TAL (traitement automatique des langues), de premiers travaux montrent que les discussions Wikipedia présentent relativement peu de déviance par rapport à la norme langagière. Les messages sont écrits de manière plutôt rigoureuse par rapport aux forums de discussion plus traditionnels : relativement peu de fautes d'orthographes et de grammaires, peu de recours à des modes de rédaction particuliers (lettres capitales PLUS, répétées ASSSSEEEZ, suite de ponctuation répétées !!!!). Nous présenterons quelques éléments de comparaison pour évaluer ce degré de déviance.

Troisièmement, les textes sont systématiquement associés à un nombre important de méta-données portant à la fois sur la thématique (portail thématique, article associé), le caractère subjectif de la discussion (caractère polémique, etc.) et le statut du locuteur (informations sur sa participation à la Wikipédia et sur son statut dans la communauté). Ces statuts ont fait l'objet d'un certain nombre d'études sur la corrélation entre le statut et le style langagier utilisé (Danescu-Niculescu-Mizli et al. 2012, 2013, Burke and Kraut 2008 inter alia).

Enfin, la base de données Wikipedia représente une masse de données imposante. Selon la page http://fr.wikipedia.org/wiki/Wikipédia:Statistiques (consulté le 12 mai 2015) : “Wikipédia en français compte 16192 contributeurs (Wikipédiens) ayant fait au moins une modification ces 30 derniers jours (hors utilisateurs sous IP). Parmi ceux-ci près de 5 000 contributeurs ont fait au moins 5 modifications et près de 800 ont fait au moins 100 modifications sur la même période.” La principale activité reste l'édition d'articles (1 622 066 articles au 11 mai 2015), mais la participation aux discussions est également très importante. Notre corpus est ainsi constitué de 366 326 discussions, 1 024 351 sections de discussion (topics internes à une discussion), 2 255 959 messages et 159 578 279 mots.

Afin de constituer notre corpus de discussions, plusieurs procédures automatiques ont été mises en place pour extraire et formater les discussions. L'extraction consiste à traiter la sauvegarde globale des pages courantes de la Wikipédia française (archive frwiki-20140331-pages-meta-current#.xml.bz2 diffusée librement sur la page http://dumps.wikimedia.org/frwiki/20140331/), d'y repérer les discussions et de les transformer en fichiers XML normés selon la TEI-P5.

75 % des discussions ont été exclues du corpus (1 130 227 sur les 1 496 553 contenues dans le Dump). Les critères d'exclusion sont les suivants :

---

2 La liste des cafés et bistrots est donnée dans l'article http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Avenue_des_caf%C3%A9s_et_bistros

3 http://fr.wikisource.org/wiki/Aide:Choix_%C3%A9ditoriaux, consulté le 12 mai 2015

4 http://fr.wikisource.org/wiki/Wikisource:Questions_l%C3%A9gales, consulté le 12 mai 2015

5 Une liste détaillée des statuts est donnée dans l'article http://fr.wikipedia.org/wiki/Aide:Statuts_des_utilisateurs

- La discussion porte sur un utilisateur de la Wikipedia[6]
- Indication explicite d'une redirection vers une autre discussion :

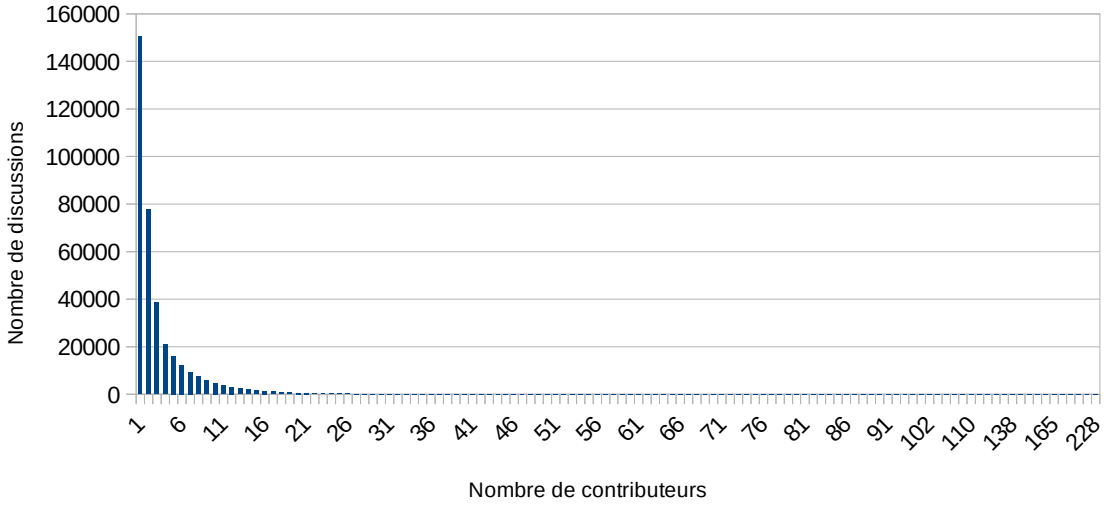
  <text xml:space="preserve">#REDIRECT [[Discuter:Iolo Morganwg]]</text>

  Exemple d'indication en tête de la discussion Discussion:Yolo Morganwg déplacée vers la page Discussion:Iolo Morganwg suite à une erreur sur l'initiale du prénom.

  <text xml:space="preserve">Doublon de [[Son (physique)]].</text>

  Exemple d'indication en tête de la discussion Discussion:Onde acoustique supprimée car en doublon de celle associée à la page Son (physique).
- Aucun message dans le corps de la discussion
- Moins de 2 mots dans le corpus de la discussion

Les discussions mono-contributeurs où un locuteur lance un sujet auquel personne ne répond ont été conservées. Elles sont au nombre de 150 603 (soit un peu plus de 40 % des discussions retenues).
La délimitation des différents messages et contributeurs s'appuie sur un ensemble de règles, notamment la présence nécessaire de la date de publication du message. L'évaluation manuelle de 7 discussions comptabilisant 413 messages et 47 284 mots montre une précision de 0,92 (3 messages vides ; 5 messages scindés en 2 ; 25 messages fusionnant 2 ou 3 messages) et un rappel de 0,95 (23 messages absents).
Le modèle utilisé pour représenter les différents sujets de discussion et messages s'inspire de la norme dédiée à la fois aux dialogues oraux et aux pièces de théâtre. Ainsi, chaque message est balisé correspondant aux messages avec pour attribut "who" le nom de l'utilisateur, "when" la date de publication du message et "interactionalLevel" son niveau d'interaction (réponse au précédent message, réponse à une réponse, etc.) Chaque message est ensuite découpé en paragraphe (<p>) ou des éléments de liste (<item>).
Les discussions et les messages présentent de grandes variations de taille : entre 1 et 1 103 messages par discussion et entre 1 et 3 428 mots par message. En moyenne, une discussion implique 5 contributeurs identifiés différents, avec là aussi, de fortes variations, allant de 1 à 228 contributeurs différents pour une même discussion. Le graphique ci-dessous indique le nombre de contributeurs impliqués dans les discussions. Parmi ce décompte des contributeurs, tous les anonymes non inscrits sont regroupés, ce qui représente autour de 4,5 % des contributeurs, chiffre relativement stable.



Concernant les caractéristiques intra-linguistiques, nous proposerons un premier inventaire assez large du contenu de ce corpus : n-grams (morpho-)syntaxiques typiques (cf. Laippala et al. 2015), expression de la subjectivité (projection du lexique FEEL, Abdaoui et al. 2014), formule d'ouverture et de fermeture des messages, expression de l'accord et du désaccord.
Enfin un cas d'étude linguistique plus ciblée sera présenté autour du phénomène des noms sous-spécifiés (Schmid 2000), marqueurs de l'organisation discursive dont le rôle est susceptible de fortement varier selon les types de texte. Leur analyse contrastive permettra de tester leur pertinence en tant qu'indice distinctif entre discussions et articles.

6 Exemple : http://fr.wikipedia.org/wiki/Discussion_utilisateur:Hashar

Outils linguistiques et informatiques pour l'analyse des controverses

Agata Jackiewicz, Aymen Elkhlifi

Les réseaux sociaux numériques constituent aujourd'hui une arène privilégiée d'échanges sur les grandes questions ou controverses d'intérêt public, qu'elles soient d'ordre politique, sociétal, religieux ou économique. Collecter et analyser les traces verbales de ces controverses constitue un enjeu et une opportunité scientifique intéressante non seulement pour les sciences sociales et politiques (traçabilité du social) et l'analyse du discours (rhétorique des controverses et des polémiques), mais aussi et de plus en plus pour le traitement automatique des langues et les linguistiques de corpus (veille et fouille d'opinions). Une telle finalité nécessite des apports convergents, permettant de tenir compte (i) des caractéristiques spécifiques aux univers discursifs à support numérique, (ii) des propriétés linguistiques des discours à caractère polémique, sans oublier (iii) la dynamique proprement sociétale des processus conflictuels qui engendrent de tels discours.

Notre projet général vise la construction d'un modèle conceptuel assorti d'outils linguistiques et informatiques permettant d'interroger les traces des controverses telles qu'elles se manifestent sur la plateforme de microblogage Twitter (Jackiewicz *à par*). La communication présentera certaines phases de ce travail, en décrivant les analyses et les expérimentations menées sur plusieurs corpus de tweets, en vue de l'élaboration du modèle et de la stabilisation du répertoire des procédés linguistiques impliqués dans le marquage des tweets de nature polémique.

Ramené à ses lignes essentielles, le modèle renvoie aux entités ciblées dans le débat, à l'identité, aux postures et aux motivations de ses protagonistes, ainsi qu'aux modes de validation critique des positionnements adoptés. Sans viser une hypothétique complétude ou une forme de vérité ultime sur la réalité d'une controverse, l'idée est de pouvoir faciliter la compréhension des processus éristiques dans leur complexité, grâce au profilage des énoncés et à leur analyse linguistique. Nous défendons l'idée que le profilage notionnel des productions langagières collectées (foisonnantes et bruitées) – par lequel on cherche à ordonner analytiquement le matériau discursif associé à une controverse - facilite également l'identification des modes d'expression caractéristiques. Nous faisons l'hypothèse que la connaissance des mécanismes inhérents aux processus polémiques ainsi que des formes d'expression dédiées permet d'accéder de manière ciblée à des contenus qui en dévoilent des éléments d'intelligibilité.

Notre corpus d'étude comprend deux ensembles de tweets collectés automatiquement, concernant respectivement : (i) l'ouverture du mariage aux couples de personnes de même sexe, (ii) la question de la filiation des enfants nés d'une GPA ou d'une PMA. Les tweets ont été collectés à des périodes précises pour permettre d'adosser l'analyse aux événements sociaux qui ont constitué les temps forts de ces controverses (débats publics, auditions à l'Assemblée Nationale, manifestations nationales…).

Pour constituer nos corpus, nous avons mis en place un collecteur de tweets en Java "TCollector", qui fonctionne en temps réel, en mode "Streaming". Il se base sur des listes de mots clés thématiques. Les textes des tweets ainsi que plusieurs métadonnées sont stockés dans une base de données. Afin de pouvoir profiler le corpus et l'analyser selon différentes facettes des controverses, deux approches sont envisagées. Une première approche linguistique consiste à reconnaître des patterns. *A minima*, chaque notion du modèle est caractérisée par un champ lexical caractéristique (sa signature lexicale). Une deuxième approche d'apprentissage automatique consiste à construire des modèles de classification automatique, à partir des ensembles de tweets annotés manuellement selon les différentes dimensions de la controverse.

D'un point de vue linguistique, la perspective que nous adoptons est celle de l'analyse du discours, que nous conjuguons avec des techniques d'analyse plus formelle (détection de n-grams, définition de patrons linguistiques…). Au chapitre des controverses, notre démarche s'appuie sur des travaux de référence en sociologie et en sciences politiques (Latour 2007 ; Lemieux 2007 ; Chateauraynaud 2011 ; Julliard et Cervulle 2013), ainsi qu'en analyse de discours (Kerbrat-Orecchioni 80 ; Plantin 2003 ; Dascal and Chang 2007 ; Amossy et Burger 2011 ; Amossy 2014). Concernant les pratiques discursives propres aux réseaux sociaux, nous exploitons entre autres les travaux de Boyd *et al*. 2010, Marwick and Boyd (2010), Stenger et Coutant 2011, Jackiewicz et Vidak (2014) et Vidak et Jackiewicz (à paraître).

Researching identity-in-interaction in multilingual online chat: Critical multimodal discourse analysis of CMC corpora

Liudmila Klimanova

How is today's use of communication technology in language teaching different from 10 years ago? The previous uses of CMC in language teaching were grounded mostly in the tenets of the interactionist theory (e.g., Chapelle, 2001). The idea was that language students were to get access to input (foreign language), some of which was to become intake that was assumed to modify the learner's interlanguage system. The CMC environment would encourage the learner to produce output, negotiate meaning and form with other participants of the exchange. Over the last decade, however, we have noted a shift in our understanding of Internet mediation and its role in language instruction. This shift in conceptualizing CMC for language learning has become possible as a result of rapidly expanding social technologies. As Thorne (2008) puts it, the Internet has become less of a technological fact than a social fact. For today's language learners, Internet-mediated communication is no longer a practice environment for language learning, but the medium through which they perform social roles and engage in interpersonal activities. Internet mediation has been and will be part of the everyday life, activities, and routines. The role of internet mediation has shifted from being professionally situated, as in job-related interactions, classroom-based interactions, to more personal and inter-personal, informal ways of self-identification and self-expression. The enormous popularity of social-networking sites and Web 2.0 technologies, for example, has brought about a bulk of research on how language learners socialize into various online communities, more research on identity building, and manifestations of identity work in Internet-based communities.

My research focuses on the exploration of identity building strategies by participants of Russian-American intercultural telecollaborative exchanges in social networking spaces where textual modality of communication (i.e., synchronous and asynchronous chats) is enhanced by various tools of multimodal self-expression. The dataset consists of a large corpus of computer-mediated multimodal interactions between learners of English and Russian coded and analysed in the qualitative research program, NVIVO. The interactions were collected during five years of collaborative exchanges between college-level language classes and compiled in a large dataset for subsequent discourse analyses. Analyses of online chats were grounded in the socio-linguistic principles of identity enactment in interaction (Bucholtz & Hall, 2005) and Gee's (2005, 2011) seven tasks of language-in-use. The interaction data were triangulated by mapping micro-analyses of interactional turns onto students' local and global identity characteristics revealed in other datasets (participants' project diaries, stimulated recall sessions, and my netnographic field notes). Discourse analysis in this study refers to broad views on the types of activities, ideologies, relationships, and acts of assigning significance to certain identity tokens, which reveal participants' self-positioning strategies as L2 learners or L2 users, or which are indicative of participants' downplaying their non-native-speakerness and their striving to bring to the fore or soft-pedal certain social and linguistic identities in online communication. To categorize the types of activities and actions in which L2 learners of Russian engaged by means of language and interaction, I analyzed seven areas of reality—seven building tasks of language—that act simultaneously in every piece of language-in-use (Gee, 2005; 2007). These tasks served as coding categories and analytical blocks in my analyses. Special attention was given to the presuppositions and implicatures that carried ideological functions in discourse (Fairclough's common-sense ideologies in the service of power, Fairclough, 2001; 2003), along with explicit statements that revealed higher-level (“naturalized”) ideologies concerning non-native language use. In this dataset, textual CMC chats were often enhanced by participants' dynamic use of visuals and multimedia for self-expression that were carefully incorporated in ongoing interactions as a way to compensate for the lack of linguistic ability or to enhance certain identity representations via non-textual communication channels. This multimodality of online talk added multi-dimensional complexity and created Interstitial spaces of meaning that were to be discovered only when textual and multimodal data were treated as inseparable components of digital communication.

In my presentation I will share my data with the audience and discuss analytical approaches to CMC data researchers can use to account for multiple channels of information sharing in a single multimodal discourse analysis. I will identify difficulties that accompany such analyses and propose possible solutions.

What's up doc? A corpus of Physician Rating Portals

Johannes Knopp, Tobias Betzin

We will present a multilingual corpus (German and French) built from six different physician rating portals providing information about users, their comments, and their ratings of physicians.

The question that motivates the collection of these data is “How is trust established in computer-mediated communication (CMC)?”. In the setting of physician rating portals (PRP) patients who do not know each other share their experiences and publish recommendations based on these experiences in the sensitive field of health related issues. Other users make their decisions based on these recommendations, but how do they decide who to trust? In order to study the process of decision making, we need a data set where this problem occurs regularly and in consequence we investigated the domain of PRPs. In our talk, we will describe the gathered data in detail, the decisions made in the process of creating the dataset along with a description of the tools used to create it while the analysis with regard to the research question is addressed by the submission with the title Analyse linguistique et contrastives du genre CMC « commentaires d'évaluation des médecins en ligne by Coline Baechler.

We have identified six popular PRPs for our research – three German websites (jameda.de, docinsider.de, arzt-auskunft.de) and three French ones (notetondoc.com, mytoubib.com, les-bons-choix-sante.fr). In our current state of research a first iteration of data acquisition for all three German and one French website is completed. Each PRP lists doctor's profiles along with their comments and ratings by users in different categories presented by the website. We can extract the physicists' profiles along with more detailed information like name, address, specialization and store this information in a database. The ratings and user comments are extracted and put in the database as well, allowing a structured access to the data, e.g. we can inspect all comments that gave a bad/good rating or group the ratings by a physicist or region. As the websites are not uniform, the found information for each platform differs in details like rating schemes but usually agree on the basic facts that are provided.

In order to extract the information from the websites an extraction tool is built and adapted for each one. Usually this is a three step process: First, an initial list of all profile pages is generated for the investigated platform. Second, the relevant elements of the profile pages (physician's name, ratings, comments etc.) are identified and an extraction pattern is prepared and finally in the third step a web crawler takes the initial list along with the extraction pattern, downloads the pages and writes the found contents in the database. The implementation was carried out using a service called kimono (kimonolabs.com) which is tailored for web crawling tasks like this.

Due to access restrictions, we collected not all of the available profile data in the initial run but extracted a sample for the mentioned PRPs. Table 1 shows some statistics of the first iteration for Jameda and Docinsider. The percentage of profiles that have at least one rating differs by a large margin between Jameda and Docinsider but both platforms provide many comments for those profiles that are rated. We plan a more detailed analysis and comparison between the user behaviour on the different PRPs once the data is complete.

After finishing the data collection for each mentioned PRP there are many ways to analyse the provided information. The general statistics about the platforms can help to form new hypotheses about individual behaviour based on language and nationality in CMC in the context of health issues and investigate them in more detail using the data base. Statistical natural language processing methods could be used to cluster similar comments and users with similar behaviour in order to find out typical behaviour and prototypical comments revealing the implicit rules that are induced by a PRP's community. Another idea would be to adapt a work by Mukherjee et. al [1] which predicted rare or unknown side-effects of medical drugs based on user comments and included the trustworthiness and credibility as factors in their probabilistic graphical model. We could translate their model on the PRP data and try to identify trustworthy comments automatically which could help to investigate our initial research question about establishing trust in CMC.

Méthode hybride de normalisation lexico-syntaxique des SMS

Eleni Kogkitsidou, Georges Antoniadis

L'étude de la communication par SMS, et par conséquent du langage SMS qui la supporte, se focalise, pour une grande partie, sur les particularités de ce langage. Le langage SMS comporte plusieurs particularités ; il s'agit, notamment, d'un code écrit qui combine plusieurs procédés pour raccourcir les phrases et les mots (Stark E. 2011). Il est proche de l'oral tout en étant une forme écrite. L'étude de ces caractéristiques a suscité l'intérêt de nombreux chercheurs (Antoniadis et al., 2011).

L'observation et l'étude de ces caractéristiques nécessite l'utilisation de données authentiques dans le but d'obtenir un point de vue plus objectif (Fairon and Paumier, 2006). Notre étude a comme point de départ le corpus de SMS, collectés et traités semi automatiquement, dans le cadre du projet alpes4science[1]. Le projet alpes4science fait partie du projet international sms4science[2] qui a pour objectif l'étude de la communication par SMS. La collecte s'est déroulée du 1er octobre 2010 au 31 janvier 2011 dans les Hautes-Alpes et l'Isère. Des informations variées qui sont liées à des questions démographiques et comportementales associées à l'utilisation de SMS sont disponibles sur la base de données alpes4science. Au total, 22 054 SMS authentiques ont été recueillis, envoyés par 359 personnes ; dont, 240 parmi eux (auteurs de 96,7 % des SMS collectés) ont rempli le questionnaire d'informations démographiques et comportementales associées. La base de données issue de la collecte contient les SMS collectés (anonymisés, transcrits en langue standard et alignés avec les SMS bruts), le lexique (mots SMS avec leurs traductions en langue standard et leurs fréquences), ainsi que des informations variées sur les expéditeurs (âge, sexe, niveau d'études, langue maternelle, mode d'utilisation des SMS, etc.).

La plupart de ces messages courts présentent des différences significatives en comparaison avec le langage standard ; leurs auteurs utilisent diverses formes courtes pour abréger les mots dans l'objectif de gagner du temps (parfois aussi de l'espace) et faire, sans doute, un effort moindre.

Un des obstacles auquel nous devons faire face avec les systèmes de traitement automatique du langage standard est la morphologie particulaire des mots SMS (fusion de mots, formes abrégées imprévisibles, suppression de caractères, manque de ponctuation, etc.). Pour mieux observer ces particularités dans notre corpus de travail, nous avons appliqué un traitement dans le but de mettre en évidence les phénomènes liés aux problèmes du traitement automatique du langage SMS, en particulier concernant son étiquetage morphosyntaxique. L'étiquetage morphosyntaxique constitue une étape fondamentale afin de pouvoir traiter davantage des données textuelles, comme, par exemple, pour la reconnaissance des entités nommées, la traduction automatique, l'extraction d'information etc.

Les logiciels actuels pour l'étiquetage morphosyntaxique des textes standards donnent, en général, des résultats bien satisfaisants. Cependant, leurs résultats sont quasi inacceptables lorsque l'étiquetage concerne des textes dégradés (comme les SMS) et courts. L'objectif de notre proposition est de présenter une méthode hybride de normalisation lexico-syntaxique des SMS.

Comme Sproat et al. (2001) le signalent, il est essentiel d'appliquer un processus de normalisation avant tout autre « traitement basique » de TAL. En ce qui concerne les SMS, notre processus de normalisation a comme objectif de convertir un texte informel dans un texte grammaticalement correct.

Le processus de normalisation proposé est basé sur une approche divisée en plusieurs modules, chacun étant responsable d'une sous-tâche spécialisée. Le premier groupe de modules est en charge du processus de normalisation structurelle : normalisation des séparateurs, traitement des symboles de ponctuation, détection des émoticons, découpage en phrases, etc. ; pour ceci nous faisons appel à l'utilisation des heuristiques développées à partir d'un corpus d'apprentissage ainsi que par la découverte automatique des règles issues du calcul des statistiques durant le déroulement d'analyses. Un deuxième groupe de modules est responsable du processus de normalisation linguistique : expansion de formes non ambiguës, résolution des abréviations, découpage en unités lexicales, traduction des mots SMS en langue standard. Pour cette partie nous utilisons deux approches : symbolique et probabiliste. Dans la première démarche, nous concevons des Grammaires Locales (Réseaux des Transduction Récursives) utilisées en combinaison avec des dictionnaires linguistiques et une base de connaissance regroupant des informations spécifiques aux mots SMS. Pour pallier quelques limitations associées à une faible tolérance aux problèmes morphosyntaxiques de notre démarche symbolique, nous nous inspirons des approches de la traduction automatique pour construire une paire de langues (FR_SMS, FR_STANDARTD) et ensuite appliquer des règles de transfert superficielles/lexicales à l'aide d'un moteur de traduction libre. Comme résultat nous obtenons la transcription normalisée vers la langue standard pour chaque message SMS. Nous considérons que les résultats de cette approche hybride nous permettrons d'appliquer notre méthode à d'autres types de textes contenant de bruit (chats, e-mails, forums de discussion, tweets, etc.).

[1] http://www.alpes4science.org/ [2] http://www.sms4science.org/

Developing and sharing teletandem data

Paola Leone

The current paper illustrates and discusses the building and publication of a corpus of video-recorded teletandem oral interactions, called RETI (Repository of Teletandem Interactions). Teletandem is a form of telecollaboration by which two students, proficient in two different languages, interact orally via Voice Over Internet Protocol (VOIP) technology in order to socialize and develop their L2 competence. During Teletandem sessions, the students talk about different topics, their learning experience, needs and objectives. These communicative exchanges have a dual focus on language and contents. However, as Leone (2014) showed, a strong role in Teletandem sessions is also played by the social dimension, i.e. the interlocutors' desires to establish a solid interpersonal relationship and to talk about their learning experiences. Teletandem is the on-line version of tandem language practice/learning (e.g. Brammerts & Kleppin, 2001), a practice which, starting from the '80s, has used dialogue between paired speakers of different native languages as a framework for the development of L2 language competence. Just like tandem, Teletandem is based on the principle of reciprocity, which implies that each party must have identical opportunities to develop his/her L2 language competence. This means that with an English-Italian pair of students, for instance, the first half of the virtual meeting will be in English, and the second half in Italian. This kind of learning scenario is called "alternate monolingual teletandem" (Leone, 2014). At the University of Salento, Teletandem has been used to develop the inter-comprehension skills (see ***A Framework of Reference for Pluralistic Approaches to Languages and Cultures***; FREPA, 2012, p. 7) of speakers of related languages, such as Italian and Portuguese. Thus, students were made to "work on two or more languages of the same linguistic family […] being one of these languages the learner's mother tongue". In each Teletandem session, each speaker talked exclusively in his/her native language. I will call this learning scenario "teletandem intercomprehension".

RETI is a databank of spontaneous authentic conversations between L2 learners. When transcribed and annotated, these data can become a learner corpus. A learner corpus can help linguists to investigate participants' communicative strategies and the intercultural dimension of dialogue. Furthermore, it can be used to identify "what is difficult for the learner as it is revealed by errors (misuse) but also, most interestingly, by overuse, underuse and avoidance of specific language choice with reference to a selected norm" (Prat-Zagrebelsky, 2004: 43). Therefore, the RETI database contains relevant information on how the students' communicative competence in L2 progresses and on how students deal with L2 culture in spontaneous oral computer-mediated interactions.

The present paper addresses the process of building the corpus, focussing on metadata and annotation. Metadata will be added, in order to provide information about situation (e.g., type of learning scenario; objectives; materials; teacher's role; language for communication), participants (e.g., numbers or pseudonymous for identifying each participant; gender; first language and other known languages; vehicular L2 language level; language experience), and discourse (e.g., discourse type: free conversation, topic based discussion, task completion; task type and conversational topics) (see also Chanier et al. 2014).

Transcriptions of the computer mediated oral interactions will be annotated during different stages and tagging will be useful for two different research perspectives: discourse analysis and the intercultural dimension of the Teletandem exchange.

As for other corpora such as CoMeRe (Chanier et al., 2014) 'openess' and the possibility of sharing data with other colleagues will be key issues for corpus implementation (OpenData, 2013). Therefore, questions of ethics and rights raised by publishing oral video corpora as open data will be addressed.

Le corpus Polititweets: enjeux de constitution d'un corpus de tweets et propositions d'analyses

Julien Longhi

L'analyse du discours politique connaît un renouvellement important, dû notamment aux nouveaux supports et formats d'expression, comme les réseaux sociaux numériques (RSN). Or, ces lieux de production d'écrits sont le plus souvent saisis par des disciplines qui les traitent comme des données sociales, plutôt que comme des discours. C'est diversement le cas pour l'opinion-mining, le data-mining, la fouille de données, etc. Les données textuelles y sont considérées comme des « sacs de mots » (Rastier 2011) et les critères sémiotiques et discursifs sont mis au second plan au détriment de caractérisations qui s'appuient sur des ontologies, des thésaurus, ou des associations sémantiques.

La communication proposée vise à décrire les enjeux philologiques, herméneutiques, et également institutionnels et interdisciplinaires, de la constitution d'un corpus de tweets politiques. Le corpus Polititweets (Longhi et al. 2014 : 34273 messages, 205 utilisateurs) a été élaboré selon le format TEI (avec des pistes d'extension aux formats CMC proposées par un groupe européen qui s'est constitué autour de cette question) dans le cadre du projet CoMeRe, afin de tenir compte des éléments spatio-temporels, contextuels, technologiques, interactionnels, thématiques, dialogiques, etc. des messages produits. Il s'agira donc dans un premier temps de montrer la pertinence du corpus face aux données pour construire une sémiotique des discours politiques numériques.

Dans un second temps, nous nous pencherons sur la dimension analytique du corpus :

- la détection et la caractérisation d'idéologies (selon des critères définis dans Sarfati 2014), à partir de la constitution de « règles » linguistiques pour repérer les spécificités formelles de l'idéologie ;

- l'analyse du corpus selon la caractérisation de sous-corpus spécifiques (tweets idéologiques, tweets efficaces, tweets négatifs, tweets dialogiques) avec le logiciel d'analyse de données textuelles Iramuteq.

Dans une conception englobante de la discursivité (qui intègre les paliers du genre et du texte), nous décrirons plus largement, pour des applications concrètes de sémantique du discours outillée (en écho à des travaux en sémantique des textes outillée, par exemple Valette 2004), l'efficacité de notre perspective vis-à-vis des traitements concurrents issus de l'informatique ou des sciences cognitives. Nous ouvrirons notamment sur la persceptive de constitution d'autres corpus de tweets relatifs à des objets d'analyse spécifiques, notamment en lien avec des controverses ou événements (corpus #intermittent, #mariagepourtous, etc.).

Genre analysis of expert and learner corpora of news-based computer-mediated communication

Tim Marchand, Sumie Akutsu

This paper investigates the nature of one particular type of computer-mediated communication (CMC), the comments section following online news stories. The study compares the CMC of two such corpora: an expert, or native-speaker, corpus created from messages posted on the BBC's Have Your Say website (BBC 2001-2015), and a learner corpus formed through the collection of comments on a blog used as part of a year-long EFL course for university students in Japan. Each week, an article about a recent news item, together with supporting class materials, is posted online for the learners to access. The students read the news story and, after a classroom session, write their reactions to the story on the class blog, thereby replicating the actions of the contributors to the BBC website. After several years of running the course, the learner corpus is now approaching 300,000 tokens in size, while the reference corpus has over 1.5 million tokens.

Previous research of native English-speaker CMC has suggested that its grammatical and lexical features differ significantly from both written and spoken registers of English, although overall its features may be considered as an intermediary register between the two (Yates, 1996; Murray, 2000; Marchand, 2013). This paper seeks to advance that research by outlining a dual approach to text classification, and exploring the extent to which the learner texts are convergent with those in the expert corpus from a genre classification viewpoint.

The first approach follows the research undertaken by Biber et al. (1999) into lexical bundles. Biber et al. compared the distribution of lexical bundles across typical written discourse (academic writing) with typical spoken discourse (conversation) and found there to be a marked contrast in the form and function of the most predominant chunks of language in these two registers. This study uses a similar methodology to determine the extent that each kind of discourse more closely matches the CMC corpora.

The second approach also builds on the work of Biber, this time his seminal study in text classification using multidimensional factor analysis (Biber, 1988). Both corpora were analysed using the Multidimensional Analysis Tagger (MAT), which offers an approximate replication of the tagger used in Biber's original study (Nini, 2014). The MAT program generates grammatically annotated versions of the corpora in addition to the statistics required for a text-type or genre analysis, plotting the input texts along Biber's six Dimensions of register variation. Under this analysis and following Biber's (1989) list of text types, both corpora were determined as being closest to the text type Involved Persuasion. However, the results will show that there were considerable differences within some of the Dimension variables. For example, the learner corpus produced much higher scores for Dimension 1, characterized by Biber (1988) as contrasting Involved vs Informational texts, predicated by a greater use of personal pronouns and clausal expressions of stance.

The paper concludes by suggesting how the genre classification of learner texts over the course of the academic year can be used to trace student longitudinal development in terms of register awareness and convergence to native-speaker norms.

The Affordances and Disadvantages of WordReference Forums as a Space for Intercultural Exchange

Elizabeth Mayne

"The Internet has done an incredible job of bringing the world together in the last few years. One of the greatest barriers has been language". This is how Michael Kellog, the founder of WordReference (www.wordreference.com), begins the introduction to his site. One of his goals, among providing bilingual dictionaries and excellent translations for as many languages as possible, is to "provide the world's best language forums, dedicated to relatively serious discussion about the meanings and translations of words, terms and expressions in many languages".

This paper presents the affordances as well as the disadvantages of using WordReference Forums both as a second language learner and as a researcher. The data analyzed in this paper consists of a 216-post merged thread on WordReference Forums, entitled "tu/vous-tutoyer-vouvoyer-tutoiement/vouvoiement". Thus, the topic of the thread is a pragmatic variable in French, tu vs. vous. This is a topic that has garnered a lot of attention in sociolinguistic research, but continues to be a challenge for second language learners of French.

WordReference is a free, open tool, which allows anyone with access to the Internet the ability to read or post to the language forums of their interest. Anyone can start a new discussion thread, although moderators have the ability to merge threads if they feel that they are redundant. Unlike projects such as Cultura (Furstenberg et al., 2001), where instructors guide second language students' topics of interaction with native speakers, threads in WordReference Forums typically begin with a question that a learner posts about a translation or concept that he or she cannot find in the WordReference dictionary.

Using Lave and Wenger (1991)'s notion of communities of practice, I argue that the engagement of the native speaker informants as well as that of the moderators in these discussion threads is crucial to the success of WordReference Forums. Another affordance of this tool is that it offers the opportunity for intercultural exchange to a potentially worldwide audience. When the topic of discussion such as *tu* vs. *vous* is of pragmatic interest, this tends to spark discussions in which learners attempt to make comparisons with pragmatic features in their own languages, or in other languages and cultures with which they are familiar. In discussions about *tu* vs. *vous*, learners attempt to make comparisions with address pronouns in other languages, or with first name vs. last name usage in English. Native French speakers as well as learners who are living or have lived in France provide personal anecdotes in attempt to explain when exactly they believe the usage of *tu* vs. *vous* is appropriate.

One of the disadvantages for both the learner and the researcher is that you do not really know the identity of the posters to the Forums. Therefore, for the learner, the disadvantage is that someone positioning himself as an "authority" on the French language may not actually be a native speaker, or even a competent speaker of the language. For the researcher, the data provided by WordReference Forums posts often does not give the demographic information that one might want; posters are asked only to provide a username, native language, and current location, although some do not even provide the second two pieces of information. Another disadvantage to using this form of Computer-Mediated Communication is that it is asynchronous. Since posters from all over the world participate, time zone differences can mean that a question is answered by several non-native speakers while the native speakers are sleeping.

Still, WordReference Forums present a unique opportunity for intercultural exchange, particularly when a pragmatic variable is the topic of discussion.

LinkedIn, le média social de promotion de l'identité professionnelle quelles stratégies discursives pour la création de liens interpersonnels ?

Jeanne Meyer

La proposition de communication s'inscrit dans un projet de recherche en sociolinguistique portant sur les mises en discours de l'appartenance communautaire médiée sur le réseau social professionnel LinkedIn. LinkedIn, « c'est le plus grand réseau social professionnel en ligne » avec plus de 55 millions de membres (Real Del Sarte, p.308).

> « Ce n'est pas un réseau social et bien d'avantage qu'un site de recrutement. Les membres de LinkedIn utilisent le réseau pour proposer leur expertise,[...] échanger des connaissances[...] et nouer des liens avec des personnes qui ont le même état d'esprit à travers le monde » (Fanelli-isla, p.86).

Considéré comme un réseau de rencontre de profils professionnels, LinkedIn serait donc un lieu où se tissent potentiellement des liens interpersonnels entre individus à des fins de formation de communauté(s) professionnelle(s). Le réseau met ainsi à disposition une panoplie d'outils pour promouvoir l'identité professionnelle d'individus en vue de construire leur réseau. L'utilisateur peut notamment avoir recours à différentes rubriques-type à compléter pour nourrir son profil parmi lesquelles nous pouvons trouver Résumé, Compétences, Projets, Formation, etc. La rubrique Résumé est particulièrement intéressante dans la mesure où elle suppose une mise en discours *personnalisée* et promotionnelle des principaux éléments du profil et par conséquent implique (ou non) l'explicitation de l'intentionnalité interpersonnelle de l'individu quant à la constitution de son profil.

J'entame ici une recherche sur les stratégies discursives utilisées au sein de cette rubrique Résumé pour créer un ou des lien(s) communautaire(s). Il s'agit donc de voir comment est dit l'appel à une communauté professionnelle. Cette recherche s'articule en deux temps : un premier temps de pré-enquête consistant en l'analyse de la pertinence et de la faisabilité de ce projet et un deuxième de temps d'extension sur corpus afin de proposer des éléments de réponse à ces objectifs scientifiques testés sur un plus grand nombre de profils.

Ici, la communication traite de la phase de pré-enquête. Pour ce faire, un corpus de profils a été créé : les mots-clés directeurs de la recherche des profils compatibles avec mon objectif scientifique sont *chargé de recrutement +LinkedIn* – en émettant l'hypothèse que les profils dans lesquels seront présents ces deux mots-clés concernent des professionnels sensibles au recrutement 2.0 et donc à la mise en discours d'un appel à la communauté professionnelle virtuelle médiée par le réseau social. A partir de cette recherche, j'obtiens 17 profils compatibles (j'entends par compatible les profils qui comportent un résumé sur les fiches LinkedIn). Ces 17 profils me servent d'échantillon pour pré-enquête à l'analyse plus linguistique des caractéristiques de la rubrique Résumé.

Les premiers éléments de cette analyse révèlent certains particularismes concernant la forme des discours et attestant du caractère hybride de ces textes concernant les trajectoires pronominales (de l'identité au collectif mutualisé) et les marques oralisantes (avec adresse au lecteur notamment).

Analyse sociolinguistique d'un Réseau Social d'Entreprise (RSE) – Données linguistiques, représentations et pratiques langagières

Jette Milberg Petersen

Cette communication est proposée dans l'axe 1 "Le développement de corpus CMR" même si j'aborderai également la communication en situation de langues en contact. Il peut en effet exister une corrélation entre les données linguistiques que les membres d'un RSE saisissent volontairement dans leurs profils et leurs pratiques langagières.

Ma recherche-action réalisée entre mai 2012 et décembre 2013 sur le Réseau Social d'Entreprise (RSE) "SolidarNet" du groupe GDF SUEZ (aujourd'hui Engie) - en 2014 le groupe comptait 147 200 collaborateurs dans 70 pays - m'a permis de développer des schémas présentant d'une part les répertoires linguistiques auto-déclarés par les utilisateurs dans leurs profils (suite invitation des concepteur/créateurs du RSE à "indiquer les langues que vous *parlez*") et d'autre part les langues réellement utilisées dans leurs contributions. Celles-ci sont des interactions plurilingues asynchrones distantes écrites. En plus, nous avons pu observer les changements de code et des stratégies d'accommodation adoptés par les utilisateurs lors de leurs échanges écrits.

Grâce à un accès administrateur au RSE, nous avons pu recueillir manuellement les langues "maternelles" et "étrangères" dans les profils de 200 membres, sur 472 inscrits au moment de notre recueil[7]. Cela a permis de présenter les répertoires linguistiques dont disposaient en quelque sorte ces membres pour communiquer sur le RSE. Les choix linguistiques des 40 utilisateurs ayant contribué dans les huit "groupes thématiques" du RSE ont ensuite été mis en rapport avec leurs propres répertoires linguistiques. Ces pratiques langagières ont également été recueillies manuellement, par observation et relèvement (direct sur le RSE ou sur des captures d'écran). Pour de futures recherches, je souhaiterais connaître les avis des participants de nos JIR pour trouver des solutions alternatives, inspirées par exemple des outils de *Social Network Analysis (SNA)* : quant aux données linguistiques des profils, après décision du contenu exact, une solution automatique pourrait être envisageable, mais pour les pratiques langagières, d'autres outils seront nécessaires.

Les choix de langue(s) ont été étudiés dans les communications sur Facebook (J. Androutsopoulos, 2013). L'intérêt de faire apparaître et d'exploiter les données linguistiques sur les RSE se trouve dans la prise en compte des langues sous un angle sociolinguistique et non comme une contrainte technique, comme c'est souvent le cas pour la gestion du "contenu multilingue". Cela permettrait d'analyser les interactions plurilingues et d'envisager le développement de nouvelles pratiques langagières dans le but d'optimiser l'outil de communication.

D'après Lüdi (2003), en effet, les domaines dans lesquels les langues sont pratiquées joueraient un rôle dans le choix de langue. Si un domaine « peut exercer des contraintes sur le choix de langue », il joue aussi un rôle dans l'acquisition des langues. Selon Mucchielli (2000), cité par C. Batazzi-Alexis (2002), "l'interaction entre la technologie (qui détermine le fonctionnement) et l'usage (conditionné par l'organisation) qui en est fait (...) occasionne indubitablement de nouveaux savoirs".

Pour ce qui concerne l'usage des langues, et selon les résultats d'un questionnaire remis aux utilisateurs du RSE, des représentations - que j'évoquerai lors de ma présentation - inhibent et même excluent certains utilisateurs de la communication. Dans le contexte des RSE des groupes internationaux, on peut alors se poser la question si une "accommodation" (H. Giles, 1991) est réellement efficace et s'il ne faudrait pas autoriser et même encourager une non-accommodation (Bilaniuk, L., 2010). Cela consisterait à développer les compétences de compréhension des textes écrits dans une langue "étrangère" et en faire un usage complémentaire. Cette pratique, basée sur le concept de l'intercompréhension, permettrait alors de rendre davantage d'utilisateurs actifs.

Des Interactions plurilingues asynchrones distantes écrites (Ipade) ont été expérimentées avec le projet Galanet dans le cadre de l'enseignement/apprentissage des langues romanes basé sur le concept de l'Intercompréhension (Degache, C., Tea, E., 2003). Inspirés de ce projet, nous cherchons actuellement une entreprise pour expérimenter l'intercompréhension sur un RSE. L'objectif est de vérifier si dans ce contexte spécifique de travail, les utilisateurs arrivent à s'approprier des pratiques de communication similaires et à faire appel à une médiation linguistique collaborative.

Avec ma communication, j'espère illustrer l'importance du rôle de l'entreprise dans l'évolution des pratiques et des représentations des langues. A long terme, on peut ainsi espérer qu'*une prise en compte des langues sur les RSE contribuera à une meilleure acquisition et à un usage plus varié des langes* (voir aussi Berthaud, 2010)... *au bénéfice futur des entreprises elles-mêmes*.

[7] Le résultat de notre recherche juridique nous a appris qu'une utilisation de ces données pour la recherche est compatible avec les finalités initiales de la collecte des données, à condition d'être réalisée dans le respect des principes et des procédures prévus.

The French CoMeRe Wikiconflits subcorpus

Céline Poudat, Natalia Grabar, Camille Paloque-Berges

In order to improve the representativity of the CoMeRe corpus and to enrich it with subcorpus built with Wikipedia pages, a small workgroup Wikipédia - nouvelles collectes has been created. It counts N. Grabar (STL, Lille 3), C. Paloque-Berges (HT2S and DICEN, CNAM) and C. Poudat (BCL, Nice Sophia Antipolis). Further to various discussions on research interest of the involved people, we decided to start the collection of data on pages that show conflictual discussions and are related to disputes and controversies positioned in the STEM fields (Science, Technology, Engineering and Mathematics).

Selection of pages has been performed in three steps:

(1) detection of Wikipedia pages prone to contain conflictual sequences; in that respect, we carefully examined four sets of pages in October 2013: (i) the 73 disputes then handled through the French “Salon de médiation” (Dispute resolution noticeboard in English Wikipedia); as conflictual pages are also signalled in Wikipédia with specific tags, we considered pages labeled with (ii) “désaccord de neutralité” (NPOV dispute, 214 pages) and (iii) “désaccord de pertinence” (relevance or content dispute, 546 pages); finally, protected and semi-protected articles (169 pages) were examined.

(2) We obtained a pre-selection of 1,002 articles to be analyzed and assessed, enabling us to draw up a first typology of conflicts in the STEM fields;

This served as the basis for (3) selection of articles on conflictual themes. Seven articles were finally chosen, which appear to be representative of conflicts within the STEM fields: "Quotient Intellectuel" (Intelligence Quotient), "Igor et Grichka Bogdanoff", "Organisme génétiquement modifié" (Genetically modified organism), "Chiropratique" (Chiropratic), "Histoire de la Logique" (History of logic), "Eolienne" (Wind turbine), and "Psychanalyse" (Psychoanalysis).

If the number of these articles may seem to be small, it should be kept in mind that each of them must be analyzed in order to reflect on its nature, structuring and content. All of these characteristics must help us in following and studying the conflict, its genesis, apogee, and its resolution when relevant. Hence, we chose to group together pages, that are potentially relevant to conflicts, in page clusters provided by a given Wikipedia article. Such an approach allowed us to drastically reduce the number of articles selected given the large amount of data extracted for each cluster. Thus, the seven conflictual pages currently gathered within the Wikiconflits corpus include 4,456 posts published by 3,971 contributors (489,000 tokens in discussions / 330 Mo all subcorpus zip).

The Wikipedia data have been extracted and annotated automatically according to the TEI-CMC recommendations. This task has been done by Kun Jin and Thierry Chanier (LRL, Clermont-Ferrand). Because the source Wikipedia code is often not respected by the contributors, which makes proper exploitation of discussions impossible, these automatic extractions have been checked and corrected manually - except for the GMO cluster to date.

We will present different steps of this work, and the criteria we adopted for the selection and structuring of the data. We will then describe our first results and the directions for future work.

L'influence des discours d'accompagnement sur le partage social : identifier et analyser les discours d'escorte sur Twitter

Bénédicte Toullec, Magali Bigey, Justine Simon

L'étude du partage des liens URL d'actualité sur un réseau socionumérique comme Twitter permet de mettre au jour les dynamiques langagières qui se jouent sur ces réseaux pour exposer ses convictions et pour encourager le partage. Une quantité importante de tweets ne se contentent en effet pas de mettre en circulation le titre de l'article et son lien URL, mais y ajoutent (voire substituent) des propos personnels, des marques linguistiques diverses qui traduisent les intentions de communication de leurs auteurs et peuvent avoir un impact sur les retweets de ces messages, sur la lecture des articles associés, etc. Il nous est donc apparu indispensable de mettre en place des méthodologies idoines pour repérer et isoler ces tweets, pour offrir une typologie des formes de discours d'accompagnement, et réfléchir à leur éventuelle efficacité en termes de partage.

L'objectif consiste d'un point de vue global à saisir les formes discursives de partage de l'information brute et des méta-discussions (Lovink 2012) sur l'information constituée de commentaires qui visent à fixer un sens et à faire circuler cette information. Au-delà des enjeux liés à la constitution (à l'accès et à l'exhaustivité) de corpus volumineux de données, les corpus de tweets posent des problèmes liés à la définition et à l'analyse des discours qui y sont présents. En partant de l'URL d'information d'actualité comme support de partage, l'analyse des textes entourant celle-ci nécessite qu'une première opération soit conduite afin d'identifier ce qui relève du texte produit automatiquement (notamment par des « bots », des boutons de partage ou encore des retweets) et ce qui relève véritablement d'un acte d'énonciation. Il s'agit ensuite d'opérer une première classification de ces tweets.C'est en procédant à un double travail de repérage empirique et itératif (par tâtonnement) des constantes et des différences que nous affinons la constitution de corpus.

Cette communication vise à explorer certaines dimensions contrastives propres à un corpus de tweets permettant par la suite de distinguer différents sous-corpus. Ces contrastes conduiront à identifier des discours d'escorte - notion proposée par Alain Rabatel (2011) pour l'analyse du discours numérique - soit tout ce qui entoure un lien URL en vue de donner des indications sur l'interprétation de celui-ci. Le discours d'escorte ne se contente pas d'accompagner le lien, il cherche à éveiller l'intérêt du lecteur à propos du contenu de l'article mis en lien et constitue un discours d'influence sur son interprétation. L'engagement énonciatif est un aspect signifiant dans les discours d'escorte étudiés, cela pouvant se traduire par une intervention graphique plus ou moins forte, du point de vue du nombre de caractères. Les contrastes choisis pour approfondir cette analyse pourront tout aussi bien être internes (liés à des typologies d'acteurs) que reposant sur des contrats différents (contrat informationnel, contrat humoristique, etc.).

Le corpus général de tweets correspond à l'enregistrement de tous les partages de liens URL de 32 médias français d'actualité, sur une période de 5 mois (15 mai-15 octobre 2014). Le corpus restreint constitué pour la présentation aux JIR de Rennes partira d'un corpus aléatoire contribuant à analyser et poser les fondements d'une analyse de ces textes et permettant d'identifier une typologie de discours.

sms4science.ch: A multi-lingual challenge for Part-of-Speech tagging

Simone Ueberwasser

The aim of this talk is to present possible solutions to the challenges written non-standard multilingual data can represent for automatic Part-of-Speech tagging. The linguistic rules as well as the server based software (cf. Ruef/Ueberwasser 2013) used to create a normalized layer as a starting point for the PoS annotation will be at the center of this presentation.

Based on a corpus of about 26'000 original text messages from Switzerland, we have found out that, using manual, interlinear glossing to create a normalized layer parallel to the original SMS layer proved to be a successful approach to Part-of-Speech tagging. This normalized layer was used as an input to Helmut Schmid's TreeTagger (cf. Schmid (1995)) and, with some training of the tagger, resulted in a precision of 96%, a rate that can be considered to be very high considering the nature of the input.

We feel that other projects affronting similar data might profit from our experience. So far, our approach has been tested against the Swiss SMS corpus (freely available for academic research at http://sms.linguistik.uzh.ch), which contains 25'947 text messages (~500'000 tokens) in all four national languages (German, French, Italian, Romansh, mainly dialectal data except for French). The data are anonymized and annotated for their main languages (which has become necessary due to an extremely high percentage of multilingual messages with massive code-switching), also for borrowings.

Text messages, especially those written before the massive advent of smartphones, are generally renowned for their specific spelling strategies (cf. e.g. Crystal, 2008), heavily deviating from the written norm. However, PoS taggers are trained on these official orthographic norms, so they are not directly applicable to SMS data. In the case of the Swiss SMS corpus, the situation is even more difficult, first because of the intensive use of code-switching, as example (1) shows:

(1) I like you saumässig and my little härzli pöpperlet toujour per te! You are mon ceur, tu sei min stärn ...
('I like you extremely[German] and my little [English] heart[SGD, diminutive] beats[SGD] always[French] for[Italian] you[Italian]! You are my[French] heart[French], you[Italian or French] are[Italian] my[SGD] star[SGD]')
[SGD = Swiss German dialect]

In text messages where the number of alternations, insertions and borrowings is as high as in this example, they are all annotated as foreign material without proper PoS tags. However, in less extreme cases, the tagger can integrate borrowings, as example (2) shows:

(2) [Hey][,][very][nice][!][Freut]]mi][für][di][!]
[ITJ][$][ADV][ADJD][$][VVFIN][PRF][APPR][PPER][!]
('hey, very nice! I am pleased for you!)

Second, as an additional challenge, the greatest part of the Swiss SMS data are dialectal (SGD, Northern-Italian from Ticino, Romansh varieties from the Grisons) and lack any standard orientation point as for their spelling (cf. Dürscheid/Stark 2011). What is more, the Swiss German dialects do not only deviate from Standard German in their spelling, but on every level of linguistic structure. Most importantly, there are morphosyntactically necessary elements in the dialects that do not have an equivalent in the Standard language and vice versa:

(3) Swiss German dialect:
[aber][guet][häschs][][em][eugen][gseit]
Normalized layer:
[aber][gut][hast][es][dem][Eugen][gesagt]
'[but][good][have][it][to the][Eugen][told]'
('how good that **you** told Eugen')
automatically generated PoS:
[ADV][ADJD][VAFIN][PPER][ART][NE][VVPP]

In this example, Standard German requests the (inverted) subject du as does English, but in SGD, the subject would add an unwanted emphasis.

As a way of overcoming all these challenges, the Swiss SMS team decided to write an application for interlinear glossing that allowed for every token to be manually converted into a standardized spelling. Because the software is server based, consistency between glossers could be maintained by presenting suggestions based on the work already performed by any glosser. Morphosyntax was left untouched wherever possible, resulting in a normalized layer that cannot considered as Standard German but represents the SGD's morphosyntax. Following these procedures allowed us to create an annotation that will allow us to keep up the flow of publications published with the Swiss SMS corpus as a basis (cf. www.sms4science.ch).

Un corpus longitudinal de SMS d'adolescents : de la constitution du corpus à l'analyse de l'écriture SMS

Olga Volckaert-Legrier, Antonine Goumi, Alain Bert-Erboul, Josie Bernicot

Les SMS ont célébré leurs vingt ans en décembre 2012. Leur caractéristique principale est une forme orthographique particulière, qu'on appelle textisme (1 pw1 sr la kestion). Un textisme est défini comme un changement dans la forme orthographique d'un mot par rapport à l'écrit traditionnel (Bernicot, Goumi, Bert-Erboul, & Volckaert-Legrier, 2014 ; Bernicot, Volckaert-Legrier, Goumi, & Bert-Erboul, 2012). Du point de vue de la recherche fondamentale, l'étude des SMS permet de répondre aux questions concernant l'acquisition et le fonctionnement du langage avec des données nouvelles. Ce matériel linguistique inédit est particulièrement intéressant dans une perspective pragmatique dont l'objet est de mettre en relation les caractéristiques des productions linguistiques avec celles des situations de communication (Austin, 1962; Josie Bernicot, Veneziano, Musiol, Bert-Erboul, 2010; Grice, 1975; Searle, 1969; Verschueren, 1999). Il s'agit de déterminer les spécificités linguistiques des SMS, définis comme un registre de la communication écrite (Ravid & Tolchinski, 2002).

L'étude des SMS doit passer par la constitution de très grands corpus. C'est la première étape nécessaire pour la description d'un phénomène à la fois nouveau et complexe. Différentes bases de données de SMS francophones existent. En 2004, le CENTAL de l'Université de Louvain La Neuve (Belgique) a coordonné un projet international « sms4Science » qui a permis de recueillir 75000 messages et de constituer une base de données informatisée de 30 000 SMS. Le but de ce projet est de contribuer à l'étude de la communication par SMS et du langage qu'elle véhicule. (Fairon, Klein, & Paumier, 2006). Pour répondre à cet objectif, des chercheurs de plusieurs pays s'associent pour constituer, dans différentes langues, de vastes corpus de SMS pour la recherche scientifique. Ce projet a été répliqué dans différents pays francophones : en France (La Réunion : 12 000 SMS ; Montpellier : 88 000 SMS ; Grenoble : 4500 SMS), en Suisse (26 000 SMS), au Canada (20 000 SMS).

Pour aller plus loin dans la compréhension du mode d'acquisition du registre SMS, on doit utiliser une méthode permettant de déterminer les caractéristiques des messages réellement produits par les scripteurs. La méthodologie de recueil longitudinal nécessaire pour comprendre l'acquisition de l'écriture SMS n'est pas utilisée dans les études, exceptés dans les recherches de Wood et al. (2009), Wood et al. (2011a) et Wood et al. (2011b). Il faut noter que dans les 2 premières études, la période de recueil est relativement courte (9 et 10 semaines) et dans la dernière, la période correspond à une année scolaire mais les SMS ont été collectés en début et en fin d'année. Le corpus de 4524 SMS que nous avons recueillis de façon longitudinale pendant une année est unique en langue française (et même en considérant les autres langues). La méthodologie de notre étude permet de recueillir des SMS produits dans des conditions de vie quotidienne. La collecte est réalisée de façon longitudinale (mois par mois) sur une longue période (12 mois) auprès de jeunes adolescents n'ayant eu aucune pratique des SMS avant le début de l'étude. Nous pouvons ainsi mettre au jour le processus d'évolution de la forme des SMS et contrôler l'ancienneté de la pratique de ce moyen de communication.

Dix-neuf jeunes adolescents scolarisés en classe de 6ème et de 5ème ont participé à l'étude : 10 filles et 9 garçons (âge moyen 11,79 ans). Les élèves n'ayant jamais possédé ou utilisé de téléphone portable ont été invités à participer à l'étude. La proposition était la suivante : être équipé gratuitement d'un téléphone mobile pendant un an et s'engager à «donner» au moins 20 SMS (rédigés par l'élève lui-même) par mois à l'équipe de recherche qui garantissait l'anonymat à toutes les étapes de l'étude. Le consentement et l'engagement écrits des parents et des enfants ont été obtenus. Tous les participants sont issus de la classe moyenne, dans l'âge scolaire légal, et de langue maternelle française. Les participants étaient équipés de téléphones portables avec clavier alphanumérique. Pour créditer les téléphones en communication, des cartes de rechargement ont été utilisées. Un logiciel permet la réception par les chercheurs des SMS «donnés» chaque mois par les participants. Chaque début de mois, les téléphones des participants sont automatiquement crédités de 30 minutes de communication voix ou 150 SMS. Une fois par mois, les téléphones des participants sont en plus crédités d'un forfait illimité de SMS pendant 5 jours. C'est pendant cette période qu'ils doivent rediriger vers l'équipe de chercheurs au moins 20 SMS parmi les SMS qu'ils ont envoyés. Le recueil des données s'est déroulé pendant l'année scolaire 2009-2010. Pour tous les participants, nous avons également le niveau orthographique mesuré par un test standardisé.

A partir de ce corpus, nous avons analysé différents indices : la longueur des messages, la densité et la nature des textismes, ainsi que le lien entre les textismes et le niveau en écrit traditionnel. Ces premières analyses ont donné lieu à une publication dans la revue Journal of Computer Assisted Learning (Bernicot, Goumi, Bert-Erboul, & Volckaert-Legrier, 2014). Cet article a fait l'objet d'un communiqué de presse du CNRS (http://www2.cnrs.fr/presse/communique/3475.htm). Celui-ci a été relayé par de nombreux média nationaux et européens (presse, radio, télévision). Les résultats ont également été présentés lors de colloques internationaux (CECI à Brest en juin 2014, Colloque ACFOS XI en novembre 2014) et des chapitres d'ouvrage (Bernicot, Volckaert-Legrier, Goumi, & Bert-Erboul, 2014 ; Volckaert-Legrier, Goumi, Bert-Erboul, & Bernicot, 2015).

L'exploitation plus approfondie du corpus va se poursuivre, en analysant des aspects structuraux du langage comme la diversité lexicale et la complexité syntaxique. Ces indices seront mis en relation avec la durée de pratique de l'écriture SMS. Il serait aussi intéressant de mettre ce corpus à la disposition de la communauté scientifique, ce qui implique un travail de formatage des données et la constitution d'une interface informatisée d'interrogation. De plus les résultats obtenus pourraient être le point de départ d'innovations pédagogiques comme le mLearning (mobile learning ; ex : l'enseignant envoie sur le mobile des élèves des mots de vocabulaire à apprendre dans une langue seconde (Lu, 2008).