Time in translation: the kick-off workshop

The Utrecht based **Time in Translation** project (Henriëtte de Swart, Bert Le Bruyn, Martijn van der Klis) is happy to announce its kick-off workshop on Friday June 23rd in Utrecht. We're bringing together linguists from different kinds of backgrounds and with different aims and ask them to reflect on corpus methodology in their work: Stephan Th. Gries, Eva Vanmassenhove, Martijn van der Klis, Antonio Toral, Tommaso Caselli, Jet Hoek & Nicholas Asher.

Time

June 23rd, 10.00 till 17.00

Venue

Utrecht, Janskerkhof 15A, room 101

The venue is within walking distance from Utrecht Central Station. If you prefer to take a bus, you could consider the following lines that all stop at Janskerkhof:

7 (to Voordorp)
8 (to Wilhelminapark)
28 (to De Uithof P+R)
50 (to Doorn via Zeist)
51 (to Driebergen-Zeist)
52 (to Amersfoort via Vollenhove)
55 (to Maartensdijk via Tuindorp)
74 (to Zeist)
77 (to Bilthoven)
251 (to Zeist Handelsc. via Zeist-Cent)

Programme (for abstracts, see pages 3 to 5)

10.00	10.15	Coffee/Tea
10.15	10.25	Welcome
10.25	11.15	Stefan Th.Gries (University of California at Santa Barbara)
		Some "insights into[my] research": a Griesian view on quantitative corpus linguistics
11.15	11.30	Coffee/Tea
11.30	12.00	Eva Vanmassenhove (Dublin City University)

		What Do NMT and SMT KnowAbout 'Aspect', and How Does this Translate?
		Joint work with Jinhua Du and Andy Way
12.00	12.30	Martijn van der Klis (Utrecht University)
		Using multidimensional scaling to map the Perfect
12.30	14.00	Lunch (in Luden)
14.00	14.40	Antonio Toral (University of Groningen)
		Parallel Corpora in (Machine) Translation: goals, issues and methodologies
14.40	15.20	Tommaso Caselli (Free University of Amsterdam)
		Time in Context: How Theory Has Met Practice
15.20	15.40	Coffee/Tea
15.40	16.10	Jet Hoek (Utrecht University)
		Coherence Relations in (Machine) Translation
16.10	17.00	Nicholas Asher (CNRS, IRIT, Université Paul Sabatier)
		The many phases of annotating discourse and dialogue rhetorical structure
17.00		Drinks (in Luden)

Participation

Participation is for free but it would be great if you could send a short email to Bert Le Bruyn (b.s.w.lebruyn@uu.nl) if you're intending to join for lunch.

Further information and questions

If you have any further questions, please contact Bert Le Bruyn (b.s.w.lebruyn@uu.nl)

Abstracts (in order of presentation)

Stefan Th. Gries (University of California at Santa Barbara)

Some "insights into [my] research": a Griesian view on quantitative corpus linguistics

To try and answer the basic questions we were asked in the workshop overview, in this talk I will present aspects of my view on and my work in quantitative corpus linguistics as well as the goals I have for that field. Following the outline provided by the questions, I will first briefly discuss the main kinds of corpus methods I am using and exemplify them with a recent case studies from the kind of alternation research that I have long been involved in. In a second part of the talk, I will discuss a few aspects of corpus-linguistic work that are concerned with my goal of trying to help the discipline evolve (faster). Specifically, I will discuss ways in which current corpus-linguistic methods can be improved and how their statistical analysis can be improved; I will relate these ways to recent work in psycholinguistics and cognitive-linguistic/usage-based theories.

Eva Vanmassenhove (Dublin City University)

What Do NMT and SMT KnowAbout 'Aspect', and How Does this Translate?

Joint work with Jinhua Du and Andy Way

One of the important dierences between English and French/Spanish grammar is related to how their verbal systems deal with aspectual information. While the English simple past tense is aspectually neutral, the French and Spanish past tenses are linked with a particular imperfective/perfective aspect. This study examines what Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) learn about `aspect' and how this is reflected in the translations they produce. We use their main `knowledge sources', phrase-tables (SMT) and encoding vectors (NMT) to examine which aspectual information they encode. Furthermore, we examine whether this encoded knowledge is actually transferred during the decoding and thus reflected in the translations they produce. Our study is based on the translations of the English simple past and present perfect tenses into French and Spanish imperfective and perfective past tenses. We examine the interaction between the lexical aspect of English simple past verbs and the grammatical aspect expressed by the tense in the French/Spanish translations. On the one hand, it results that SMT phrase-tables contain information about the basic lexical aspect of verbs. Although the lexical aspect is often closely related to the grammatical aspect expressed by the French and Spanish tenses, for some verbs more contextual information is required in order to select an appropriate tense. The SMT n-grams provide insufficient context to grasp other aspectual factors included in the sentence to consistently select the tense with the appropriate aspectual value. On the other hand, the encoding vectors produced by our NMT system do contain information about the entire sentence. An analysis based on the English NMT encoding vectors shows that a logistic regression model can obtain an accuracy of 90.95% when trying to predict the correct tense based on the encoding vectors. However, this positive results do not seem to transfer to the actual translations.

Martijn van der Klis (Utrecht University)

Using multidimensional scaling to map the Perfect

The PERFECT (which we define as the construction 'HAVE/BE' + past participle) is subject to widespread cross-linguistic variation (Lindstedt, 2000). Despite extensive literature, the goal of providing a compositional semantics of the PERFECT has not yet been reached (Ritz, 2012). We would like to use semantic maps (Haspelmath, 1991) for this purpose. However, instead of starting from prototypical examples, we would rather use multilingual parallel corpora like EuroParl (Tiedemann, 2012), as translation equivalents provide us with form variation across languages in contexts where the meaning is stable.

To create a semantic map, we first extract PERFECTs in languages under study from the corpus, and then mark the corresponding verb phrases in the translations of each extracted fragment. We then assign tenses to all fragments, and use a simple distance function to create a dissimilarity matrix. Applying multidimensional scaling then allows us to visualize the level of similarity of individual cases of the dataset. We created an innovative visualization that allows to filter and switch between tense labeling, as wells as to drill down to the raw data. Our methodology mimics that of Wälchi and Cysouw (2012), though we focus at the level of grammar rather than the lexical domain.

Haspelmath, M. (1997). *Indefinite pronouns*. Oxford: Clarendon Press.

Lindstedt, J. (2000). The perfect – aspectual, temporal and evidential. In Östen Dahl, editor, *Tense and Aspect in the languages of Europe,* pages 365-384. De Gruyter, Berlin.

Ritz, M. E. (2000). Perfect tense and aspect. In Östen Dahl, editor, *The Oxford Handbook of Tense and Aspect*, pages 881-907. Oxford University Press, Oxford.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012),* pages 2214-2218.

Wälchli, B., & Cysouw, M. (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics, 50*(3):671-710.

Antonio Toral (University of Groningen)

Parallel Corpora in (Machine) Translation: goals, issues and methodologies

Parallel corpora play a central role in current approaches to machine and computer-assisted translation and also in any corpus-based study that involves original text and its translation. This talk motivates the use of parallel data, as well as its desired properties. It then introduces practical methodologies to automatically acquire and prepare parallel data for the task at hand. Finally, it glances at the neighbouring field of Translation Studies to assert that translations can differ to a great extent depending on the strategy followed by the translator, which might lead to the translation being more or less appropriate for its use in corpus-based studies.

Tommaso Caselli (Free University of Amsterdam)

Time in Context: How Theory Has Met Practice

This talk will present a personal perspective on how theoretical framework for tense and aspect have been "translated" in practical solutions for Natural Language Processing tasks. Three different, partially related, case studies will be illustrated: Timeline Extraction, Storyline extraction, and Content Type Identification.

Jet Hoek (Utrecht University)

Coherence Relations in (Machine) Translation

Coherence relations can be made linguistically explicit by means of a connective (e.g., *because,* if) or a cue phrase (e.g., *on the other hand*, *that's why*), but this need not necessarily be the case. We investigate when and why relations are explicitly marked using directional parallel corpora extracted from the Europarl corpus (Europarl Direct; Cartoni, Zufferey, & Meyer 2013, Koehn 2005). By using multiple language pairs, we aim to go beyond describing differences between individual languages in a language pair, and make observations that hold cross-linguistically. The results of our study do not only further our understanding of discourse coherence, but are also intended to be used to improve the quality of Machine Translation systems at the discourse level.