Research Agenda

October 2014

Lexical Computing's research interests lie at the intersection of corpus and computational linguistics, and the company is committed to an empiricist approach to the study of language in which corpora play a central role. For a very wide range of linguistic questions, if a suitable corpus is available, it will help our understanding. Its strap line is ‘corpora for all’.

As in case of any interdisciplinary research Lexical Computing faces scientific challenges in both fields: linguistics and computer science. Hereby we list the most important ones in both areas:

1. **Parallel and Distributed Processing of Very Large Text Corpora**
   As the volume of textual data to be processed is growing and often reaches dozens of terabytes, entirely new approaches need to be developed in order to achieve satisfying processing times. These approaches often use parallel and distributed processing and require redesign of the related algorithms as much of the processing cannot be trivially paralellized.

   - Miloš Jakubíček, Adam Kilgarriff, Pavel Rychlý. Effective Corpus Virtualization. In: *Challenges in the Management of Large Corpora (CMLC-2)* Workshop Programme. p. 7.

2. **Building Very Large Text Corpora from the Web**

   The web is a vast supply of textual data, for many languages and text types, but there are assorted challenges in turning that data into corpora that are useful for linguists. Lexical Computing is a leader in the field and explores new methods for new tasks.

   - Jan Pomikalek, Pavel Rychly, Adam Kilgarriff 2009. Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics. Special Issue of *Research in Computing Science* Vol 41, Mexico City.

3. **Corpus Heterogeneity and Homogeneity**

   While many people use corpora, our ability to describe them, and compare them, in a scientific and quantifiable manner, is limited, and this has long been on our research agenda.

   - Adam Kilgarriff 2001. Comparing Corpora *International Journal of Corpus Linguistics* 6 (1): 1-37.
   - Adam Kilgarriff 2012. Getting to know your corpus. In: Proc. Text, Speech, Dialogue (TSD 2012), Lecture Notes in Computer Science. Sojka, P., Horak, A., Kopecek, I., Pala, K. (eds). Springer.

4. **Corpus Evaluation**

   Which of a set of corpora is best, for general language research, lexicography and technology development? Scientists very rarely give any justification for their choice of corpus, beyond "it was available". It is not obvious how they should, and this is our challenge.

   - Adam Kilgarriff and Pavel Rychlý and Milos Jakubicek and Vojtěch Kovář and Vit Baisa and Lucia Kocincová 2014. Extrinsic Corpus Evaluation with a Collocation Dictionary Task. LREC (Language Resources and Evaluation Conference), Reykjavik, Iceland.

5. **Terminology Extraction**

   Finding the terms, in a set of texts for a domain, as input for terminologists preparing a terminology for the domain.

   - Adam Kilgarriff et al. Finding Terms in Corpora for Many Languages with the Sketch Engine. *EACL 2014*, 2014, 53.

6. **Corpora and Language Teaching**

   There are many language learners, language teachers and textbook authors who see the benefits of using corpora as banks of examples of language use. But most contain many examples of unhelpful or incomprehensible sentences, which will confuse and dismay learners, as well as many helpful ones. While manually selecting examples is the traditional method for dictionaries and coursebooks, this limits the number of examples that are available (by orders of magnitude). The challenge here is to automatically identify the useful examples, and to present them to language learners in a user-friendly way, as and when they want them.

   - Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX*, Barcelona, Spain.

7. **Language change over time**

   For students of language change (including dictionary companies wanting to include new words in a new dictionary) corpora are an enticing prospect: if there are corpora from different time points, then, all being well, the words in the newer data that were not in the older data will be new words. Efforts along these lines so far have been a little frustrating: between many pairs of corpora, even if seemingly well matched, there are differences of topic and composition which usually dominate differences due to language change. Responses to this include taking greater care over composition, and taking multiple data sets from different times. The challenge is then to find the profile of changes in frequency over time that delivers the highest-accuracy.

- Ondřej Herman and Vojtěch Kovář. Methods for Detection of Word Usage over Time. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013*. Brno: Tribun EU, 2013. p. 79-85, 7 pp. ISBN 978-80-263-0520-0.