

## ASSESSING SEMANTIC ACCURACY OF MT

In HimL, we are interested in improving the semantic accuracy of machine translation. For the medical domain, it is important to avoid blatant semantic errors such as reversing negation, and mixing up semantic roles (i.e. who did what to whom). To help to measure and guide this improvement, we are developing both human and automatic mechanisms for measuring the semantic accuracy of MT.

We are experimenting with a new human evaluation measure based on a semantic annotation formalism called [UCCA \(Universal Conceptual Cognitive Annotation\)](#), developed by Omri Abdend and Ari Rappoport at the Hebrew University of Jerusalem . The idea is that we annotate the source texts using UCCA, and then map these UCCA annotations across to the MT output to measure how much of the semantics is preserved. The aim is to systematically highlight the types of semantic errors that MT systems make.

## PROBLEMS WITH MORPHOLOGY IN MT

Due to the HimL project's focus on languages of Central and Eastern Europe, we need to address an important research problem; **translation to morphologically rich languages**. While statistical techniques for machine translation have made significant progress in the last 20 years, results for translating to morphologically rich languages are still mixed.

The languages we address in the HimL project require the modeling of grammatical features such as gender and case that are rarely realized through morphology in English. The challenge is to model these features in making decisions about translating English words into our languages. For example German language:

- Grammatical features of nouns, such as number, gender and case.
- Number (plural and singular) is often easy to determine in English.
- Gender is an innate property of German nouns (and is sometimes hard to understand, such as "das Mädchen", the girl, which is neuter, rather than feminine).
- German case (loosely, the idea of subject, object and indirect object, as well as genitives) can also be quite different than in English.

## CHALLENGES FOR TRANSLATION IN THE MEDICAL DOMAIN

### Newsletter 3/2015

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402.



The public health domain is very diverse and has a number of challenges which are not unique to this domain. However, as people are using this information to guide their understanding of issues related to their health or their patients' health, providing inaccurate translations could have a serious impact.

In order to learn more about the texts we wish to translate in HimL, we have been analyzing the test sets that we created for the project. In these test sets, there are many examples which do not form proper sentences.

These can make up multi-line lists or they can be titles or subtitles. For example:

Lists:

- One unit is the equivalent of: a small glass of wine one measure of spirit half a pint of normal strength beer, lager or cider.

Titles:

- More about calcium - Amphotericin B lipid soluble formulations versus amphotericin B in cancer patients with neutropenia.

Also it is notable that even when there are full sentences available, sentence structure is quite different to news stories. There are for example many imperatives, e.g.:

- Eat a healthy balanced diet rich in calcium - Spend time outside to build up your vitamin D levels - Aim for 2 to 3 servings a day.

Finally, and most problematically, there are very many technical terms and numeric scientific results:

- Lipid-based amphotericin B was not more effective than conventional amphotericin B on mortality (relative risk (RR) 0.5; 95% confidence interval (CI) 0.64 to 1.14) - For patients on HAART, when choosing from different chemotherapy regimens, there was no observed difference between liposomal doxorubicin, liposomal daunorubicin and paclitaxel.

**All these difficulties are being addressed by the research partners in order to provide the most accurate automatic translations possible for this challenging domain.**

Newsletter 3/2015

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402.

