

New Learner Corpus Functionality in the Sketch Engine

Vojtěch Kovář^{1,2} and Diana McCarthy²

¹*NLP Centre
Faculty of Informatics, Masaryk University, Brno*

²*Lexical Computing Ltd., UK*

Keywords: Sketch Engine, learner corpus, concordance, error tagging

Introduction

We present an interface for efficient searching in learner corpora with annotated errors that is built into the corpus querying tool Sketch Engine (Kilgarriff et al., 2004).

We describe the data format needed for an annotated learner corpus, which allows for nested errors of different types as well as correction mark-up. Then, we show a specialized web user interface designed for searching learner error corpora.

The Sketch Engine

The Sketch Engine is a powerful corpus querying tool used by lexicographers, language teachers and language learners throughout the world. Its main functions are:

- a concordancer that allows querying in an extended CQL (corpus query language) syntax (Jakubíček et al., 2010) and provides also a simplified and intuitive querying interface for users that are not familiar with CQL
- a word list feature that allows listing of words from the corpus according to various criteria, including extracting keywords from a sub-corpus and also allows for frequency lists of meta-data (such as domain of the documents)
- a word sketch feature that provides one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour
- a statistical thesaurus based on the collocations given by the word sketches

All of these features are available as a web service at <http://www.sketchengine.co.uk>, together with preloaded corpora for more than 50 languages and a system that allows building user corpora.

Learner Corpora

Based on a customer requirement, a new functionality has been added into Sketch Engine that facilitates work with learner corpora marked for errors and corrections. Together with the newly implemented features, learner corpora can be very useful for language teachers and authors of course books to discover common types of errors and characteristic patterns of text where people tend to make errors. In the following, we will illustrate the usage of error corpora within the system.

Source Data

Usually, the Sketch Engine system requires a vertical format (word-per-line with possibly more columns for lemmas, tags etc. and the structure XML-like mark-up) as the source data format. The errors are marked up using `<err>` and `<corr>` tags for errors and their corrections respectively. The type of error can be specified with the *type* attribute of either of the tags. An example of a small source text is shown in Figure 1.

Searching possibilities

The display of the error mark-up in the concordance lines can be highly customized – the system can display the full mark-up, as in the source vertical file, or an abbreviation for readability. In Figure 2, we show what the concordance lines can look like for the data from Figure 1. Using the frequency

function, users are able to obtain the frequency distribution of e.g. error types for a given query. From this result, the concordance can be filtered for a particular error type. It is also possible to get the error type frequency distribution for the whole corpus or a sub-corpus.

We have developed a special query interface for the users that translates the commonest queries for learner corpora into CQL (words within the error, words within the correction or type of the error – see Figure 3) so that the users do not have to use the CQL themselves.

Error mark-up can be exploited in the Word Sketch or Thesaurus functions to produce word sketches specific for error mark-up. We hope to explore this in future work.

```

We          we          PR
learn       learn       VV
maths      math          NN
to         to          PP
<err type="BadWording">
<err type="Typo">
caan       caan       ??
</err>
<corr type="Typo">
can        can        VA
</corr>
</err>
<corr type="BadWording">
be         be          VB
able      able       VP
to         to          PP
</corr>
compute   compute   VV
our       our        PR
taxes     tax        NN

```

Figure 1. Example of the source learner data. **Figure 3.** Error query interface

```

Corpus: Error corpus sample
Hits: 2 (166666.7 per million)
Typo We learn maths to <BadWording><Typo> caan ||| can </Typo>||| be able to </BadWording> compute
Typo We learn maths to <BadWording><Typo> caan ||| can </Typo>||| be able to </BadWording> compute

```

Figure 2. One possible type of display of the error data in the concordance. The first line is for the “BadWording” error, the second is the “Typo” error.

Conclusion

We have introduced a powerful tool for searching learner corpora with annotated errors. We described its potential and gave examples of its use. We hope the tool will prove useful for those working with learner corpora, language teachers and authors of the learner books.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536, in the National Research Programme II project 2C06009, by the Czech Science Foundation under the project P401/10/0792 and by the EU project PRESEMT (ICT-248307). We would like to thank Cambridge University Press for their collaboration in this development.

References

- Jakubíček, M., Rychlý, P., Kilgarriff, A. and McCarthy, D. 2010. Fast syntactic searching in very large corpora for many languages. *PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. Tokyo: Waseda University, 741-747.
- Kilgarriff, A., Rychlý, P., Smrž, P. and Tugwell, D. 2004. The Sketch Engine. *Proceedings of the Eleventh EURALEX International Congress*. Oxford : Oxford University Press.