Learner Corpus Functionality in the Sketch Engine

Vojtěch Kovář

Lexical Computing Ltd.
vojtech.kovar@sketchengine.co.uk

SKEW 3
March 21, 2012

Outline

Learner Corpus

- Corpus with annotated errors
  - text of corrections included
  - empty errors and corrections
  - can be nested
  - “to laarn” → “to learn” → “to teach”
- Problems
  - how to encode the information
  - how to represent it to the user
  - intuitive search interface

Input Format

- Extended vertical format
  - empty tokens
  - <*err*> and <*corr*> structures
  - XML-like nesting
- Error classification
  - recorded as "*err*.*type*" attribute
  - arbitrary classification
  - hierarchy of types supported

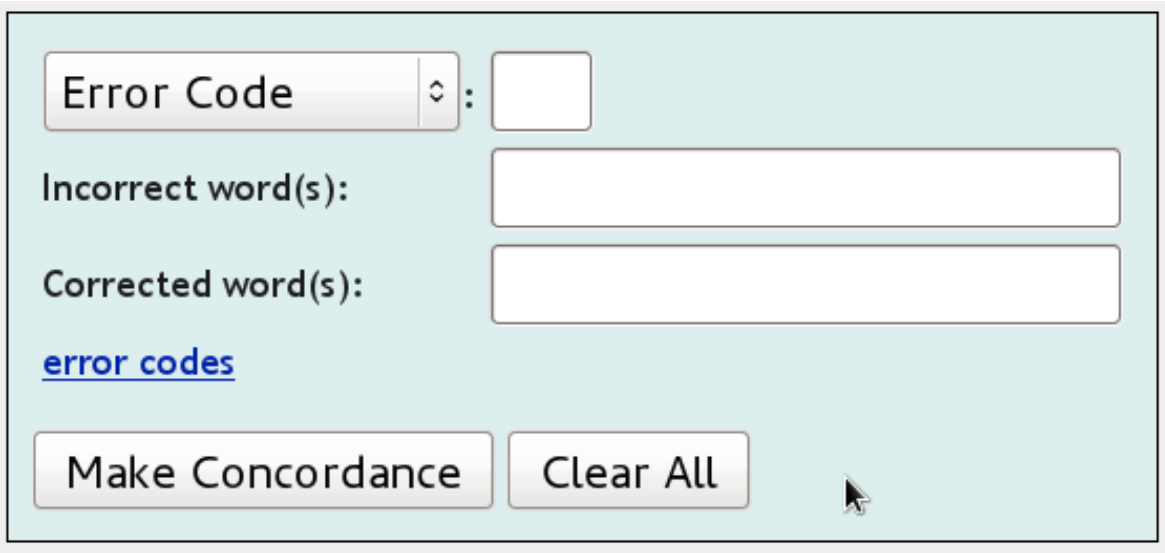
Input Format – An Example

#	word	tag	lempos
1	<doc level="B1" nationality="New Zealand">		
2	<s>		
3	The	DT	the-x
4	aim	NN	aim-n
5	of	IN	of-i
6	this	DT	this-x
7	international	JJ	international-j
8	conference	NN	conference-n
9	<err type="Repetition">		
10	<err type="Typo">		
11	cnoference	NN	cnoference-n
12	</err>		
13	<corr type="Typo">		
14	conference	NN	conference-n
15	</corr>		
16	</err>		
17	<corr type="Repetition">		
18	===NONE====	===NONE===	===NONE===
19	</corr>		
20	is	VBZ	be-v

Queries and Results

- Querying interface
  - CQL
  - standard search options
- Error search
  - by error code
  - by words within an error
  - by words within a correction
- Displaying results
  - customizable

Query and Results – Examples



Query and Results – Examples

ed <err> with </err><corr> to </corr> the <err> **teoretical** </err><corr> theoretical </corr> and applied

this international conference <err><err> **cnoference** </err><corr> conference </corr></err><corr

ects related <Prep> with # to </Prep> the <Typo> **teoretical** # theoretical </Typo> and applied aspects of

international conference <Repetition><Typo> **cnoference** # conference </Typo># </Repetition>

ll-developed projects related < with | to > the < **teoretical** | theoretical > and applied aspects of corpus

The aim of this international conference << **cnoference** | conference >| > is to bring together

Conclusion and Future Work

- Conclusion
  - interface for learner corpora
  - robust
  - intuitive interface
  - Let's use it! :)
- Usage so far
  - 2 projects (1 academic, 1 commercial)
- Future research
  - word sketches and thesaurus specific for erroneous data