Background

• Selection of examples for L2 training and Lexicography:
  - Invent – subjective and time-consuming
  - Select manually – hundreds of corpus hits, selection becomes time-consuming
  - (Semi-)automatic pre-selection – a possible alternative
• Principle: rank examples according to their appropriateness or “goodness”; the best ones come to the top
• Definition of “goodness” in linguistic parameters:
  - Optional sentence length
  - Optional word length
  - Presence of subject and finite verb etc.
• Previous tests with automatic ranking: for English (KGLARFF et al. 2008), for Slovene (Kosem et al. 2011), for German (Segler 2007, Didakowski et al. 2012)

Algorithm for Swedish (hit-ex)

• Algorithm #1 at the moment:
  - some example is scored independently of all other examples using a manually defined set of heuristic rules, each of which has an associated weight
  - parameters under consideration: sentence length; word frequency; keyword position in the sentence; presence of finite verb;
  - only “soft” parameters, i.e. if they are not met, examples are considered anyway, through punished by reducing points;
  - equal “punishment” for each parameter;

• Algorithm #2 (Borin et al. 2012): (Semi-)automatic pre-selection – a possible alternative
• Principle: examples should not only be typical but also different
• Difference is measured as a similarity metric, based on the Euclidean distance between feature vectors (words and syntactic relations);
• Vectors represent words in the context of the search terms, as well as a number of syntactic features derived from dependency trees.

Evaluation set-up 1

• Critical questions:
  - Can the two algorithms satisfactorily rank examples?
  - Which of the two performs better?
  - What other parameters might be necessary to include to improve their performance as predictors of good examples.
• Evaluators background:
  - L2 teachers/computational linguists (2)
  - lexicographers/computational linguists (2)
  - lexicographer (1)
  - all 5 have doctoral degrees
• Evaluators’ mother tongues:
  - 3 native and 2 non-native Swedish speakers
• Evaluators’ gender:
  - 2 men, 3 women

• Test items (keywords):
  - 60 test items from Kelly list, 10 per proficiency level defined in CEFR terms (Council of Europe, 2001)
  - Only lexical word classes: nouns, verbs, adjectives, adverbs
  - Number of items from each word class reflects word class distribution per proficiency level

Evaluation results. Quantitative data.

<table>
<thead>
<tr>
<th>user group</th>
<th>acc</th>
<th>unacc</th>
<th>doubtful</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alg #1</td>
<td>66,1%</td>
<td>18,6%</td>
<td>15,3%</td>
<td>100%</td>
</tr>
<tr>
<td>alg #2</td>
<td>61,1%</td>
<td>21,4%</td>
<td>17,5%</td>
<td>100%</td>
</tr>
<tr>
<td>L2 teachers, total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alg #1</td>
<td>45,6%</td>
<td>25,5%</td>
<td>27%</td>
<td>100%</td>
</tr>
<tr>
<td>alg #2</td>
<td>50,2%</td>
<td>20,4%</td>
<td>29,3%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Lexicographers more positive than L2 teachers: 63.6% vs 46.7% accepted

alg1 “won” by 5% for lexicographers and by 7% for L2 teachers

Evaluation. Quantitative data

User comments fall into 4 categories:

1. Comments/criticism of structural features, e.g. ellipsis, passive, limited context, word order, anaphora, pronouns, long phrase structure; lack of word class specific patterns
2. Lexical features: stricter word frequency filtering, proper names, acronyms and abbreviations, compounds, keyword repetition
3. Annotation: part-of-speech specific searches, e.g. exclude proper names when searching for nouns; some annotation errors
4. Heterogeneous: metaphorical use, abstract use, strange examples, etc.

Future

1. Algorithms improvement (based on results of the evaluation):
   - for alg #1 and #2 additional parameters, e.g. voice, word order, proper names, pronouns, strength of collocation with contextual words, valency for verbs, word class specific approaches, vocabulary frequency
   - for alg #2 additional techniques: word sense discrimination (Purandare and Pedersen 2004)

2. Second evaluation set-up:
   - Parameter configuration customizable in terms of strength of “punishment” per parameter
   - Larger output set for better overview (esp. for alg #2)
   - Based on polysemous words (esp. for alg #2)
   - More specific study over different user group needs (L2 teachers, lexicographers, linguists)

3. Potential results:
   - Suggest best parameter configuration per user group
   - Setup web service for example rating
   - Include web service into the present applications, e.g. Lärka, Korp, editing tool for Swedish FrameNet