Autonomous collocation error correction with a data-driven approach

Orsolya Vincze
Margarita Alonso Ramos

Universidade da Coruña (Spain)

Learner Corpus Research Conference
Bergen, September 27-29, 2013
What is a collocation? phraseological unit $W_1W_2$

$W_1$ = base selected according to its meaning

$W_2$ = collocate whose selection is determined by the base

pouring rain, dense fog, fierce wind

??? dense rain, fierce fog, pouring wind

Collocation learning and corpora

a) learner corpus $\rightarrow$ collocations are problematic

b) native corpus $\rightarrow$ collocation learning/teaching resource
Introduction: Corpora and error correction

**Writing aids and NLP:**
- learner writings can be checked against corpora

**Benefits for L2/FL learning:**
- DDL=data-driven learning
- authentic L2 input
- encourages inductive and autonomous learning
Introduction: The study

**Context:**
Development of an active collocation learning environment including a writing aid tool for learners of Spanish

**Preliminaries:**
Learner corpus derived typology of collocation errors (Alonso Ramos et al. 2010)

**Aim:**
Can learners autonomously correct collocation errors with the help of concordance lines? To what extent?
1. The study
   2.1. Collocation error types
   2.2. Research questions
   2.3. Methodology

2. Results
   3.1. General findings
   3.2. Correction of specific error types
   3.3. Evaluation of concordance lines as feedback
   3.4. Enhancing concordance line feedback

3. Conclusions and future work
1. The study
1.1. Collocation error types

- Typology of collocation errors based on CEDEL2 (Lozano 2009)

1) **Lexical collocation errors**, e.g.:

   Incorrect collocate: *capturar la atención* instead of e.g. *captar la atención* ‘catch sb’s attention

   Synthesis: *misinterpretaciones* instead of e.g. *malas interpretaciones* ‘wrong interpretations’

2) **Grammatical collocation errors**, e.g.:

   Governed preposition: *montar una bicicleta* instead of *montar en una bicicleta* ‘ride a bike’

   Number: *dimos bienvenidas* lit. ‘we gave welcomes’ instead of *dimos la bienvenida* ‘we gave a welcome’
1.2. Research questions

1) Can learners autonomously correct collocation errors with the help of concordance lines?

2) The correction of what error types poses more difficulty for the students when presented with the concordance lines?

3) What problems can learners have when dealing with concordance feedback?

4) How can concordance line feedback be improved in order to better assist students in the revision of collocation errors?
1.3. Methodology

**Questionnaire**
- 20 sentences from CEDEL2 (Lozano 2009) containing a collocation error
- Concordance lines: a) full sentences from esTenTen (Kilgarriff et al. 2004); b) Google Books n-grams

**Tasks**
1) propose a correction without any aid
2) propose a correction with the help of concordance lines

**Participants**
18 Spanish as a second language students working or studying in Spain at the time of the test
1.3. Methodology

Sample questionnaire item with full-sentence concordances

11. La hija está tratando de **captor la atención** de su madre, pero es muy difícil porque la madre siempre trabaja.

La ingesta de líquidos es a lo que habitualmente **prestamos** menos **atención**. Los niños sienten celos y los expresan a través de necesidades a las que los padres deben **prestar atención**. Esta siempre ha sido la enfermedad que más me ha asombrado y que me ha **llamado la atención** de una forma especial.

Aquí os vamos a describir algunas opciones que nos han **llamado la atención** por sus cualidades. Connelly **atrae la atención y no deja que decaiga en ningún momento**. El otro día estaba comprándome libros en la librería y este me **atrajo la atención**.

Desarrollar una cierta habilidad para el coqueteo te puede ayudar a aprender a **captor la atención** e interés de las personas que te agradan.

Los niños suelen portarse mal para **captor la atención** de los padres.

---

Sample questionnaire item with n-gram concordances

11. La hija está tratando de **captor la atención** de su madre, pero es muy difícil porque la madre siempre trabaja.

*capten la atención de captar la atención y la para captar la atención de concentrado la atención de los concentra la atención de de concentrar la atención de capten la atención de los concitan la atención de los concitaban la atención de los capten la atención de los*
2. Results
2.1. General findings

– with concordance lines higher number of correct suggestions, while the number of incorrect suggestions, as well as questionnaire items left blank was lower
– more positive and positive/negative changes and less negative and neutral or irrelevant changes with concordance lines

Total number of correct and incorrect suggestions, no answers provided or repeated answers (n=360)

Number of positive, positive/negative, negative and neutral changes
2.2. Correction of specific error types

- Participants were more successful in correcting lexical collocation errors than grammatical errors.
- More difficulty in noticing grammatical features.

Participants’ success in correcting different collocation error types with the help of concordance lines.
2.3. Problems with concordance lines as feedback: comparing full-sentence and n-gram concordances

- full-sentence concordance lines are more effective: higher number of correct and lower number of incorrect answers

Number of correct, incorrect suggestions, no answers provided or repeated answers according to concordance type
2.3. Problems with concordance lines as feedback: Analyzing incorrect suggestions

1) New errors in participants’ answers

   a) non-concordance induced errors

<table>
<thead>
<tr>
<th>Erroneous segment in original learner sentence</th>
<th>Erroneous correction suggestion</th>
<th>Expected correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>..nos despedimos, y *gracias, y caminamos hacia el puerto..</td>
<td>..nos despedimos, y *gracias a Dios caminamos hacia el puerto..</td>
<td>..nos despedimos, y les dimos gracias, y caminamos hacia el puerto..</td>
</tr>
<tr>
<td>lit. ‘we said goodbye, thanks, and we started to walk towards the port’</td>
<td>‘we said goodbye, and thank God we started to walk towards the port’</td>
<td>‘we said goodbye, and thanked them, and started to walk towards the port’</td>
</tr>
</tbody>
</table>
2.3. Problems with concordance lines as feedback: Analyzing incorrect suggestions

1) New errors in participants’ answers

   a) non-concordance induced errors
   b) meaning-related concordance-induced errors: probably due to lack of sufficient context

<table>
<thead>
<tr>
<th>Erroneous segment in original learner sentence</th>
<th>Erroneous correction suggestion</th>
<th>Expected correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mi futuro no <em>tiene limitades.</em></td>
<td>Mi futuro no <em>tiene limitaciones.</em></td>
<td>Mi futuro no tiene limites.</td>
</tr>
<tr>
<td>‘My future has no limits.’</td>
<td>lit. ‘My future has no limitations.’</td>
<td>lit. ‘My future has no limits.’</td>
</tr>
</tbody>
</table>
2.3. Problems with concordance lines as feedback: Analyzing incorrect suggestions

1) New errors in participants’ answers

a) non-concordance induced errors
b) meaning-related concordance-induced errors: probably due to lack of sufficient context
c) concordance-induced errors involving the inappropriate application of a pattern observed in the concordances

<table>
<thead>
<tr>
<th>Erroneous segment in original learner sentence</th>
<th>Erroneous correction suggestion</th>
<th>Expected correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>*La película se trata de una mujer soltera, su hija y sus amigas…‘The film is about a single woman, her daughter and her friends..’</td>
<td>*La película, que se trata de una mujer solera, su hija y sus amigas..‘The film, which is about a single woman, her daughter and her friends..’</td>
<td>La película trata de una mujer solera, su hija y sus amigas..‘The film is about a single woman, her daughter and her friends..’</td>
</tr>
</tbody>
</table>
2.3. Problems with concordance lines as feedback: Analyzing incorrect suggestions

1) Negative changes in participants’ answers
   a) non-concordance induced errors
   b) meaning-related concordance-induced errors: probably due to lack of sufficient context
   c) concordance-induced errors involving the inappropriate application of a pattern observed in the concordances

2) Incomplete correction of learner sentences

<table>
<thead>
<tr>
<th>Erroneous segment in original learner sentence</th>
<th>Erroneous correction suggestion</th>
<th>Expected correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>...y entonces *encendió el fuego que quemó la casa y los mató.</td>
<td>...y entonces *prendió el fuego que quemó la casa...</td>
<td>...y entonces prendió fuego a la casa...</td>
</tr>
<tr>
<td>lit. ‘... and then she lit the fire that burnt the house’</td>
<td>lit. ‘... and then she set the fire that burnt the house...’</td>
<td>lit. ‘... and then she set fire to the house’</td>
</tr>
</tbody>
</table>
3.4. Enhancing concordance line feedback

- **grammatical errors are less salient in concordance lines**
  - group concordance lines in order to emphasize patterns (similar to Wu et al. 2010)

- **implicit nature of concordance feedback**
  - Pro: promotes inductive learning
  - Con: participants do not always manage to identify the errors
  - should there be an explicit indication of errors?
  - multiple-step feedback: 1) only concordances, 2) additional aid

- **concordance-induced errors due to lack of context**
  - allow users to check wider context and more corpus examples if needed

- **students might need information not provided by concordance lines**
  - integration with dictionary (meaning related errors)
  - incorporate information on verb conjugation
3. Conclusions and future work
3. Conclusions

1) Can learners autonomously correct collocation errors with the help of concordance lines?
Yes, our study shows that concordance lines do have a favorable effect on learners’ autonomous error correction.

2) The correction of what error types poses more difficulty for the students when presented with the concordance lines?
Grammatical collocation errors are less salient than lexical collocation errors.
3. Conclusions

3) What problems can learners have when dealing with concordance feedback?
- lack of context: identify/distinguish meanings
- identification/noticing of error

4) How can concordance line feedback be improved in order to better assist students in the revision of collocation errors?
- emphasize grammatical patterns in the presentation of concordances
- allow more context
- more explicit indication of error (optional)
- integration with other language learning resources
3. Future work

**HAREnEs prototype interface**
Herramienta de ayuda a la redacción en español = Spanish writing aid tool
3. Future work

HAREnEs prototype interface
Herramienta de ayuda a la redacción en español = Spanish writing aid tool
Thank you for your attention!

This research has been supported by Ministerio de Economía y Competitividad (FFI2011-30219-C02-01), and the FPU grant (AP2010-4334).