I can’t believe they invented it: Simplext, the Text Simplification System for Spanish

Stefan Bott       Horacio Saggion

Universitat Pompeu Fabra, Barcelona
{stefan.bott,horacio.saggion}@upf.edu

October 17th 2012
Outline

1. Introduction
   - Text simplification
   - Related work

2. Syntactic simplification
   - The simplification grammar
   - Results

3. Applications

4. Lexical Simplification
   - Introduction and stuff
   - Setup
   - Evaluation

5. Conclusions and outlook
Text simplification is the process of transforming a text into an equivalent which is more understandable for a target user:

- Elderly people
- Language learners
- People with cognitive disabilities

The direct target user group of the Simplext project are people with Down Syndrome.
Text simplification is the process of transforming a text into an equivalent which is more understandable for a target user:

- Elderly people
- Language learners
- People with cognitive disabilities

The direct target user group of the Simplext project are people with Down Syndrome.
Text simplification is the process of transforming a text into an equivalent which is more understandable for a target user:

- Elderly people
- Language learners
- People with cognitive disabilities

The direct target user group of the Simplext project are people with Down Syndrome.
Text simplification is the process of transforming a text into an equivalent which is more understandable for a target user:

- Elderly people
- Language learners
- People with cognitive disabilities

The direct target user group of the Simplext project are people with Down Syndrome.
Text simplification is the process of transforming a text into an equivalent which is more understandable for a target user:

- Elderly people
- Language learners
- People with cognitive disabilities

The direct target user group of the Simplext project are people with Down Syndrome.
Characteristics of simplified text

Characteristics

- Simple style
- Simple vocabulary
- Unnecessary and confusing details are omitted
- One main idea per sentence
- Unfamiliar concepts explained
Characteristics of simplified text

Characteristics

- Simple style
- **Simple vocabulary**
  - Unnecessary and confusing details are omitted
  - One main idea per sentence
  - Unfamiliar concepts explained
Characteristics of simplified text

Characteristics

- Simple style
- **Simple vocabulary**
- Unnecessary and confusing details are omitted
- One main idea per sentence
- Unfamiliar concepts explained
Characteristics of simplified text

Characteristics

- Simple style
- **Simple vocabulary**
- Unnecessary and confusing details are omitted
- **One main idea per sentence**
- Unfamiliar concepts explained
Characteristics of simplified text

Characteristics

- Simple style
- **Simple vocabulary**
- Unnecessary and confusing details are omitted
- **One main idea per sentence**
- Unfamiliar concepts explained
There is a recognized need for text simplification in order to let people with comprehension problems participate in the information society.

- Human text simplification is work intense and **expensive**.
- Human text simplification is **slow** (affects online news).
The need for automatic text simplification systems

- There is a recognized need for text simplification in order to let people with comprehension problems participate in the information society.
- Human text simplification is work intense and **expensive**.
- Human text simplification is **slow** (affects online news).
There is a recognized need for text simplification in order to let people with comprehension problems participate in the information society.

- Human text simplification is work intense and **expensive**
- Human text simplification is **slow** (affects online news)
The Simplext project

- The first attempt to build a text simplification system for **Spanish**.
- A project that brings together people from different areas:
  - Theory of Simple Text
  - Natural Language Processing (NLP) / Computational Linguistics
  - Application development
  - A foundation dedicated to social inclusion of people with cognitive disabilities (PRODIS)
The Simplext project

- The first attempt to build a text simplification system for Spanish.
- A project that brings together people from different areas:
  - Theory of Simple Text
  - Natural Language Processing (NLP) / Computational Linguistics
  - Application development
  - A foundation dedicated to social inclusion of people with cognitive disabilities (PRODIS)
The Simplext project

- The first attempt to build a text simplification system for Spanish.
- A project that brings together people from different areas:
  - Theory of Simple Text
  - Natural Language Processing (NLP) / Computational Linguistics
  - Application development
  - A foundation dedicated to social inclusion of people with cognitive disabilities (PRODIS)
Data and resources

- Stylistic recommendations for the creation of simple text, specific for the target group
- Adapted by the DILES Research Group from Universidad Autónoma de Madrid
- Texts are automatically sentence aligned and the resulting alignments are corrected by hand
- Presently only a subset of the simplified part of the corpus is available for development purposes
Stylistic recommendations for the creation of simple text, specific for the target group


Adapted by the DILES Research Group from Universidad Autónoma de Madrid

Texts are automatically sentence aligned and the resulting alignments are corrected by hand

Presently only a subset of the simplified part of the corpus is available for development purposes
Data and resources

- Stylistic recommendations for the creation of simple text, specific for the target group
  - Adapted by the DILES Research Group from Universidad Autónoma de Madrid
  - Texts are automatically sentence aligned and the resulting alignments are corrected by hand
  - Presently only a subset of the simplified part of the corpus is available for development purposes
Data and resources

- Stylistic recommendations for the creation of simple text, specific for the target group
- Adapted by the DILES Research Group from Universidad Autónoma de Madrid
- Texts are automatically sentence aligned and the resulting alignments are corrected by hand
- Presently only a subset of the simplified part of the corpus is available for development purposes
Data and resources

- Stylistic recommendations for the creation of simple text, specific for the target group
- Adapted by the DILES Research Group from Universidad Autónoma de Madrid
- Texts are automatically sentence aligned and the resulting alignments are corrected by hand
- Presently only a subset of the simplified part of the corpus is available for development purposes
Data and resources

- Stylistic recommendations for the creation of simple text, specific for the target group
- Adapted by the DILES Research Group from Universidad Autónoma de Madrid
- Texts are automatically sentence aligned and the resulting alignments are corrected by hand
- Presently only a subset of the simplified part of the corpus is available for development purposes
The starting point: a corpus study

<table>
<thead>
<tr>
<th>operation</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>change</td>
<td>39.02 %</td>
</tr>
<tr>
<td>delete</td>
<td>24.80 %</td>
</tr>
<tr>
<td>insert</td>
<td>12.60 %</td>
</tr>
<tr>
<td>split</td>
<td>12.20 %</td>
</tr>
<tr>
<td>proximization</td>
<td>6.91 %</td>
</tr>
<tr>
<td>reorder</td>
<td>2.85 %</td>
</tr>
<tr>
<td>select</td>
<td>0.81 %</td>
</tr>
<tr>
<td>join</td>
<td>0.81 %</td>
</tr>
</tbody>
</table>

Frequencies of different editing operations
### Details of the corpus study

<table>
<thead>
<tr>
<th><strong>operation</strong></th>
<th><strong>%</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>change: lexical</td>
<td>17.48</td>
</tr>
<tr>
<td>delete: adverbial or adjectival</td>
<td>7.32</td>
</tr>
<tr>
<td>proximization</td>
<td>6.91</td>
</tr>
<tr>
<td>delete: clausal</td>
<td>4.07</td>
</tr>
<tr>
<td>insert: unrestricted</td>
<td>3.66</td>
</tr>
<tr>
<td>change: syntax</td>
<td>2.85</td>
</tr>
<tr>
<td>split: coordination</td>
<td>2.44</td>
</tr>
<tr>
<td>split: relative clause</td>
<td>2.03</td>
</tr>
<tr>
<td>delete: adverbial</td>
<td>1.22</td>
</tr>
<tr>
<td>split: participle construction</td>
<td>1.22</td>
</tr>
<tr>
<td>reorder: direct speech</td>
<td>0.81</td>
</tr>
</tbody>
</table>

**Frequencies of specific editing operations**
Some examples

(1) orig: La muestra ofrece al público la oportunidad de acercarse a la fauna, la botánica y la cultura de esta inmensa región selvática americana (...).

simp: La exposición nos muestra la cultura de esta gran selva americana. También nos muestra sus animales y plantas (...).
Some examples from the corpus

(2) orig: 5.000 metros cuadrados (…) en los cuales se han plantado 241 árboles y 4.377 arbustos.
simp: El parque tiene más de 5000 metros cuadrados de zona verde.
Hemos plantado 241 árboles y más de 4000 plantas.
Some examples from the corpus

(3) orig: 5.000 metros cuadrados situados entre las calles Doctor Fleming, Martín Chirino y Paseo de Europa en los cuales se han plantado 241 árboles y 4.377 arbustos.

simp: El parque tiene más de 5000 metros cuadrados de zona verde. Está entre las calles Doctor Fleming, Martín Chirino y Paseo de Europa.
Within the Simplext project, we prepare a series of simplification modules:

- Syntactic simplification (sentence splitting, reordering of constituents, etc.)
- Lexical simplification (Drndarević & Saggion, NAACL-PITR, 2012)
- Context reduction
- Simplification of numeric expressions
Within the Simplext project, we prepare a series of simplification modules:

- Syntactic simplification (sentence splitting, reordering of constituents, etc.)
- Lexical simplification (Drndarević & Saggion, NAACL-PITR, 2012)
- Context reduction
- Simplification of numeric expressions
Automatic text simplification at Simplext

Within the Simplext project, we prepare a series of simplification modules:

- Syntactic simplification (sentence splitting, reordering of constituents, etc.)
- Lexical simplification (Drndarević & Saggion, NAACL-PITR, 2012)
- Context reduction
- Simplification of numeric expressions
Within the Simplext project, we prepare a series of simplification modules:

- **Syntactic simplification** (sentence splitting, reordering of constituents, etc.)
- **Lexical simplification** (Drndarević & Saggion, NAACL-PITR, 2012)
- **Context reduction**
- **Simplification of numeric expressions**
Within the Simplext project, we prepare a series of simplification modules:

- **Syntactic simplification** (sentence splitting, reordering of constituents, etc.)
- **Lexical simplification** (Drndarević & Saggion, NAACL-PITR, 2012)
- **Context reduction**
- **Simplification of numeric expressions**
Work on automatic (structural) text simplification

- Rule based systems (Chandrasekar et al., 1996, Siddharthan, 2000)
Outline

1. Introduction
   - Text simplification
   - Related work

2. Syntactic simplification
   - The simplification grammar
   - Results

3. Applications

4. Lexical Simplification
   - Introduction and stuff
   - Setup
   - Evaluation

5. Conclusions and outlook
(4) a. Las lluvias torrenciales, que comenzaron el pasado 1 de octubre (...) y continuaron durante varios días, hicieron que los ríos y las presas se desbordaran (...)

b. Las lluvias torrenciales hicieron que los ríos y las presas (...) se desbordaran (...)

Estas lluvias comenzaron el pasado 1 de octubre. y continuaron durante varios días.
Tools

- Dependency Parser: Mate-tools (Bohnet, 2009)
- MATE: A graph transducer and grammar engineering platform (Bohnet et al., 2000)
The architecture of the syntactic simplification system

**Figure:** The architecture of the simplification system
A dependency syntax tree

Figure: An example of a dependency tree which can be simplified
A dependency syntax tree

Figure: An example of a dependency tree which can be simplified
### The simplification grammar: correctness of output

<table>
<thead>
<tr>
<th>Operation</th>
<th>Precision</th>
<th>Recall</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Clauses</td>
<td>39.34%</td>
<td>66.07%</td>
<td>20.65%</td>
</tr>
<tr>
<td>Gerundive Constructions</td>
<td>63.64%</td>
<td>20.59%</td>
<td>2.48%</td>
</tr>
<tr>
<td>Object coordination</td>
<td>42.03%</td>
<td>58.33%</td>
<td>7.79%</td>
</tr>
<tr>
<td>VP and clause coordination</td>
<td>64.81%</td>
<td>50%</td>
<td>6.09%</td>
</tr>
</tbody>
</table>

Percentage of right rule application **with correct output** and frequency of application (percentage of sentences affected) per rule type.
An experiment with statistical filtering

- Trained over hand annotated instances of grammar-detected targets (150 for simple relative clauses and 116 for complex relative clauses)
- The annotated data was taken from more online news texts (more complex than the data in our corpus)
- Two SVM classifiers (in the GATE framework implementation) were trained over this data
- Evaluation was done in isolation and in conjunction with the grammar
  - In domain: cross validation
  - Out of domain tested on data from the Simplext Corpus
An experiment with statistical filtering

- Trained over hand annotated instances of grammar-detected targets (150 for simple relative clauses and 116 for complex relative clauses)
- The annotated data was taken from more online news texts (more complex than the data in our corpus)
- Two SVM classifiers (in the GATE framework implementation) were trained over this data
- Evaluation was done in isolation and in conjunction with the grammar
  - In domain: cross validation
  - Out of domain tested on data from the Simplext Corpus
An experiment with statistical filtering

- Trained over hand annotated instances of grammar-detected targets (150 for simple relative clauses and 116 for complex relative clauses)
- The annotated data was taken from more online news texts (more complex than the data in our corpus)
- Two SVM classifiers (in the GATE framework implementation) were trained over this data
- Evaluation was done in isolation and in conjunction with the grammar
  - In domain: cross validation
  - Out of domain tested on data from the Simplext Corpus
An experiment with statistical filtering

- Trained over hand annotated instances of grammar-detected targets (150 for simple relative clauses and 116 for complex relative clauses)
- The annotated data was taken from more online news texts (more complex than the data in our corpus)
- Two SVM classifiers (in the GATE framework implementation) were trained over this data
- Evaluation was done in isolation and in conjunction with the grammar
  - In domain: cross validation
  - Out of domain tested on data from the Simplext Corpus
The simplification grammar and the statistic filter: correct identification of target structures

<table>
<thead>
<tr>
<th>Operation</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Relative Clauses (in domain)</td>
<td>85.41%</td>
<td>86.77%</td>
<td>86.06%</td>
</tr>
<tr>
<td>Complex Relative Clauses (in domain)</td>
<td>70.88%</td>
<td>71.33%</td>
<td>71.10%</td>
</tr>
<tr>
<td>Simple Relative Clauses (out of domain)</td>
<td>76.35%</td>
<td>76.35%</td>
<td>76.35%</td>
</tr>
<tr>
<td>Complex Relative Clauses (out of domain)</td>
<td>90.48%</td>
<td>85.71%</td>
<td>88.10%</td>
</tr>
</tbody>
</table>

The performance of the statistical filter in isolation

<table>
<thead>
<tr>
<th>Operation</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Relative Clauses (Grammar)</td>
<td>47.61%</td>
<td>95.24%</td>
<td>71.43%</td>
</tr>
<tr>
<td>Complex Relative Clauses (Grammar)</td>
<td>62.50%</td>
<td>55.56%</td>
<td>59.02%</td>
</tr>
<tr>
<td>Simple Relative Clauses (Grammar + Filter)</td>
<td>59.57%</td>
<td>66.67%</td>
<td>63.12%</td>
</tr>
<tr>
<td>Complex Relative Clauses (Grammar + Filter)</td>
<td>100%</td>
<td>55.56%</td>
<td>77.78%</td>
</tr>
</tbody>
</table>

The performance of grammar and the statistical filter together (grammatical correctness and parsing errors are ignored)
Outline

1 Introduction
   • Text simplification
   • Related work

2 Syntactic simplification
   • The simplification grammar
   • Results

3 Applications

4 Lexical Simplification
   • Introduction and stuff
   • Setup
   • Evaluation

5 Conclusions and outlook
Mobile and web applications

Figure: A simplified news text produced by the Simplext service on a tablet computer running Android
This work was carried out with Horacio Saggion, Biljana Drndarevic (corpus study on lexical simplicity) and Luz Rello (Evaluation)
Background

The task of substitute lexical items with a synonym which is easier to read and to understand.

- Lexical substitution for English using WordNet (Carroll et al., 1998; Lal and Ruger, 2002; Burstein et al., 2007)
- Lexical substitution using synonym dictionaries (Bautista et al., 2011)
- Lexical substitution in combination with word sense disambiguation (De Belder et al., 2010)
- SemEval-2012 Task 1: English Lexical Simplification
- No work on Spanish, so far...

Criteria of *Simplicity*

- Word frequency
- Word length
Example of a Lexical Simplification

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>Descubren en Valencia una nueva ESPECIE de pez prehistórico.</td>
</tr>
<tr>
<td></td>
<td><em>A new SPECIES of prehistoric fish is discovered in Valencia.</em></td>
</tr>
<tr>
<td>Simplified</td>
<td>Descubren en Valencia un nuevo TIPO de pez prehistórico.</td>
</tr>
<tr>
<td></td>
<td><em>A new TYPE of prehistoric fish is discovered in Valencia.</em></td>
</tr>
</tbody>
</table>

**Table:** Example of a lexical simplification
A small corpus study on lexical simplicity

- 6595 words of original (O) and 3912 words of simplified text (S)
- Processed with FreeLing (Padrò et al., 2010)
- Word length:
  - No word longer than 15 characters in the S set
  - Short words (1-5 characters) is more abundant in the S set (95.54%) than in the O set (64.10%)
- Word frequency:
  - Use of a frequency lexicon extracted from the RAE corpus. Frequency indices from 0 (out of dictionary) to 6 (most frequent).
  - lower frequency words (FI 3 and FI 0) are around 50% more common in O texts than in S texts
### Word frequency distribution in original and simplified text

<table>
<thead>
<tr>
<th>Frequency index</th>
<th>Original</th>
<th>Simplified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq. 0</td>
<td>10,53%</td>
<td>4,71%</td>
</tr>
<tr>
<td>Freq. 3</td>
<td>1,36%</td>
<td>0,74%</td>
</tr>
<tr>
<td>Freq. 4</td>
<td>1,35%</td>
<td>1,00%</td>
</tr>
<tr>
<td>Freq. 5</td>
<td>6,68%</td>
<td>5,67%</td>
</tr>
<tr>
<td>Freq. 6</td>
<td>80,08%</td>
<td>87,88%</td>
</tr>
</tbody>
</table>

**Table:** The distribution of n-frequency words in original and simplified texts.
Our method

Try to find the best (most simple) substitution candidate in a given context

- Thesaurus lookup for synonyms groups (i.e. word senses)
- Determine the best word sense (with a vector space model)
- Determine the best (i.e. simplest) substitution candidate
- Output a lemma
Resources for the experiment

- A small parallel corpus (6595 words of original and 3912 words of simplified text)
- Training corpus for language models: 8M of Spanish news text (web data)
- OpenThesaurus Spanish
OpenThesaurus

(a) mono | 4
- |gorila|simio|antropoide
- |simio|chimpancé|mandril|mico|macaco
- |overol|traje de faena
- |llamativo|vistoso|atraactivo|sugerente|
provocativo|resultón|bonito
Vector Space Model

- Trained over 8M words, lemmatized with FreeLing
- Extract a word vector for each lemma with all occurrences of this lemma within a 9-lemma window
- Use of a simple cosine measure for word similarity distances
Simplicity Measure

- **Word length**
  \[ score_{wl} = \begin{cases} 
  \sqrt{wl} - 4 & \text{if } wl > 4 \\
  0 & \text{otherwise.} 
  \end{cases} \]

- **Word (lemma) frequency**
  \[ score_{freq} = \log c_w \]

- **Global simplicity score**
  \[ score_{simp} = \alpha_1 score_{wl} + \alpha_2 score_{freq} \]

- Weights were set to \( \alpha_1 = -0.39 \) and \( \alpha_2 = 1.11 \)
  We determined weights by annually selecting 100 good simplification candidates proposed by OpenThesaurus for given contexts and calculating the average difference for \( score_{wl} \) and \( score_{freq} \) between the original word lemma and the simplified lemma.
Thresholds

- Ignore words which are more frequent than 1%
- Ignore substitute which have a simplicity score difference less than 0.5
- Ignore substitutes with low cosine similarity distance to the target word. A threshold of 0.013 was found through experimentation
Evaluation

- 2 Baselines
  - Most frequent synonym
  - Random synonym
- Gold Standard (simplifications found in the corpus)
Evaluation

- Multiple choice questionnaire
- 5 possible answer:
  {simper, similarly difficult, more complex} synonym, no synonym, dunno/can’t understand)
- Up to 4 simplification possibilities for each sample (gold (if available), system, 2 baselines)
- In each instance 2 alternatives are shown (target vs simplified (system, baselines, gold))
- 3 annotators with no previous experience (native speakers of Spanish)
**Example of an evaluation item**

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target</strong></td>
<td>Descubren en Valencia una nueva ESPECIE de pez prehistórico.</td>
</tr>
<tr>
<td></td>
<td><em>A new SPECIES of prehistoric fish is discovered in Valencia.</em></td>
</tr>
<tr>
<td><strong>System</strong></td>
<td>Descubren en Valencia un nuevo TIPO de pez prehistórico.</td>
</tr>
<tr>
<td></td>
<td><em>A new TYPE of prehistoric fish is discovered in Valencia.</em></td>
</tr>
<tr>
<td><strong>Frequency Baseline</strong></td>
<td>Descubren en Valencia un nuevo GRUPO de pez prehistórico.</td>
</tr>
<tr>
<td></td>
<td><em>A new GROUP of prehistoric fish is discovered in Valencia.</em></td>
</tr>
<tr>
<td><strong>Random Baseline</strong></td>
<td>Descubren en Valencia un nuevo LINAJE de pez prehistórico.</td>
</tr>
<tr>
<td></td>
<td><em>A new LINEAGE of prehistoric fish is discovered in Valencia.</em></td>
</tr>
</tbody>
</table>

**Table**: Example extracted from dataset (T-S/B) Target vs. LexSIS and Baselines.
Evaluation Datasets

- **(T-S/B) Target vs. System and Baselines**
  - 50 unique target words together with their synonymous generated by the system and the baselines
  - duplicates removed (e.g. system = baseline 1)

- **(T-G/S/B) Target vs. Gold, System and Baselines**
  - 12 cases out of the 26 cases of single lexical substitution in the Simplext Corpus

- **(G-T) Gold vs. Target**
  - system did not propose a simplification
  - 14 instances
Global Results

<table>
<thead>
<tr>
<th>System</th>
<th>Synonym (%)</th>
<th>Simpler Syn. (%)</th>
<th>Equal Syn. (%)</th>
<th>Complex Syn. (%)</th>
<th>Simpler Syn. Glob. (%)</th>
<th>Equal Syn. Glob. (%)</th>
<th>Complex Syn. Glob. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>65.03</td>
<td>15.13</td>
<td>24.37</td>
<td>58.82</td>
<td>9.94</td>
<td>16.02</td>
<td>38.67</td>
</tr>
<tr>
<td>Frequency</td>
<td>66.12</td>
<td>42.98</td>
<td>19.01</td>
<td>38.02</td>
<td>28.42</td>
<td>12.57</td>
<td>25.14</td>
</tr>
<tr>
<td>LexSiS</td>
<td>72.49</td>
<td>40.88</td>
<td>17.52</td>
<td>41.61</td>
<td>29.63</td>
<td>12.70</td>
<td>30.16</td>
</tr>
<tr>
<td>Gold</td>
<td>97.44</td>
<td>71.05</td>
<td>17.11</td>
<td>11.84</td>
<td>69.23</td>
<td>16.67</td>
<td>11.54</td>
</tr>
</tbody>
</table>

**Table:** Evaluation of all the datasets.
### Detailed Results

<table>
<thead>
<tr>
<th>System</th>
<th>Synonym (%)</th>
<th>Simpler Syn. (%)</th>
<th>Equal Syn. (%)</th>
<th>Complex Syn. (%)</th>
<th>Simpler Syn. Glob. (%)</th>
<th>Equal Syn. Glob. (%)</th>
<th>Complex Syn. Glob. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>100</td>
<td>78.57</td>
<td>16.67</td>
<td>4.762</td>
<td>78.57</td>
<td>16.67</td>
<td>4.762</td>
</tr>
</tbody>
</table>

**Table:** Evaluation of dataset (G-T): Gold vs. Target.
## Detailed Results

<table>
<thead>
<tr>
<th>System</th>
<th>Synonym (%)</th>
<th>Simpler Syn. (%)</th>
<th>Equal Syn. (%)</th>
<th>Complex Syn. (%)</th>
<th>Simpler Syn. Glob. (%)</th>
<th>Equal Syn. Glob. (%)</th>
<th>Complex Syn. Glob. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>57.14</td>
<td>4.76</td>
<td>42.86</td>
<td>47.62</td>
<td>2.86</td>
<td>25.71</td>
<td>28.57</td>
</tr>
<tr>
<td>Frequency</td>
<td>47.22</td>
<td>47.06</td>
<td>17.65</td>
<td>35.29</td>
<td>22.22</td>
<td>8.33</td>
<td>16.67</td>
</tr>
<tr>
<td>LexSiS</td>
<td>55.56</td>
<td><strong>65.00</strong></td>
<td>15.00</td>
<td><strong>20.00</strong></td>
<td><strong>36.11</strong></td>
<td><strong>8.33</strong></td>
<td>11.11</td>
</tr>
<tr>
<td>Gold</td>
<td>95.83</td>
<td>58.70</td>
<td>17.39</td>
<td>21.74</td>
<td>61.11</td>
<td>16.67</td>
<td>16.67</td>
</tr>
</tbody>
</table>

**Table**: Evaluation of dataset (T-G/S/B): Target vs. Gold, LexSiS and Baselines.
## Detailed Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>66.00</td>
<td>18.18</td>
<td>20.20</td>
<td>60.61</td>
<td>12.08</td>
<td>13.42</td>
<td>40.27</td>
</tr>
<tr>
<td>Frequency</td>
<td>72.11</td>
<td>43.40</td>
<td>18.87</td>
<td>37.74</td>
<td>31.29</td>
<td>13.61</td>
<td>27.21</td>
</tr>
<tr>
<td>LexSiS</td>
<td>76.00</td>
<td>35.96</td>
<td>18.42</td>
<td>45.61</td>
<td>27.33</td>
<td>14.00</td>
<td>34.67</td>
</tr>
</tbody>
</table>

**Table:** Evaluation of dataset (T-S/B): Target vs. LexSiS and Baselines.
Error Analysis

Why do we have so little overlap between gold standard and system simplifications?

- Target word is already too frequent
- Best substitution candidate is not much simpler than target word
- Target word and best substitution candidate are the same
- Vector distance between substitute and target is too high
Discussion

- Task is hard: In SemEval-2012 only one system (out of nine) scored above the frequency baseline
- Our results are roughly comparable to Biran et al. (2011)
- Evaluation is comparable to Yatskar et al. (2010) (who used simple english wikipedia edit histories)
Conclusions and outlook

- We are developing a system Spanish Text Simplification
- The syntactic simplification system is running (and subject to improvements)
- A lexical simplification system is under development
- Other modules are in an early stage
  - Content reduction
  - Clarification insertion
- A first evaluation with target users is being carried out this month
Conclusions and outlook

- We are developing a system Spanish Text Simplification
- The syntactic simplification system is running (and subject to improvements)
- A lexical simplification system is under development
- Other modules are in an early stage
  - Content reduction
  - Clarification insertion
- A first evaluation with target users is being carried out this month
Conclusions and outlook

- We are developing a system Spanish Text Simplification
- The syntactic simplification system is running (and subject to improvements)
- A lexical simplification system is under development
- Other modules are in an early stage
  - Content reduction
  - Clarification insertion
- A first evaluation with target users is being carried out this month
Conclusions and outlook

- We are developing a system Spanish Text Simplification
- The syntactic simplification system is running (and subject to improvements)
- A lexical simplification system is under development
- Other modules are in an early stage
  - Content reduction
  - Clarification insertion
- A first evaluation with target users is being carried out this month
We are developing a system Spanish Text Simplification

The syntactic simplification system is running (and subject to improvements)

A lexical simplification system is under development

Other modules are in an early stage
  - Content reduction
  - Clarification insertion

A first evaluation with target users is being carried out this month
We are developing a system Spanish Text Simplification

The syntactic simplification system is running (and subject to improvements)

A lexical simplification system is under development

Other modules are in an early stage
  - Content reduction
  - Clarification insertion

A first evaluation with target users is being carried out this month
We are developing a system Spanish Text Simplification

The syntactic simplification system is running (and subject to improvements)

A lexical simplification system is under development

Other modules are in an early stage
  - Content reduction
  - Clarification insertion

A first evaluation with target users is being carried out this month
Thanks

Thanks a lot!