An Introduction to Automatic Text Simplification

Horacio Saggion
Universitat Pompeu Fabra
Barcelona
http://www.dtic.upf.edu/~hsaggion

TUTORIAL @ RANLP 2015
Hissar, Bulgaria
6th September 2015

NOTE

- Tutorial notes will be at:

http://www.dtic.upf.edu/~hsaggion/workshops_tutorials.html

Tutorial Presenter

- Horacio Saggion
  - http://www.dtic.upf.edu/~hsaggion
  - https://www.linkedin.com/pub/horacio-saggion/16/9b9/174
  - https://twitter.com/h_saggion
  - http://www.taln.upf.edu
- TALN - Tractament Automàtic del Llenguatge Natural http://www.taln.upf.edu
  - ~20 members: profs/lecturers, researchers, PhDs, Masters, etc.
  - We are at Universitat Pompeu Fabra, Campus de la Comunicación, Barcelona

Outline

- Introduction
  - Manual simplification
  - First approaches to automatic text simplification
- Measuring text readability / "complexity"
- Lexical Simplification
- Syntactic Simplification (and more)
- Projects and Systems
- Summary

Text Simplification

- The process of transforming a text into an equivalent which is more readable and/or understandable by a target audience
- During simplification, complex sentences are split into simple ones and uncommon vocabulary is replaced by more common expressions
- Started to attract the attention of natural language processing some years ago (1996) mainly as a pre-processing step

This is human simplification

Original Text

Amnesty International accused the U.S. authorities to provide an "inhuman" treatment to Bradley Manning, a soldier accused of leaking "wires" of American diplomacy to the website Wikileaks.

Adapted Text (by trained editor)

United States treats very bad a soldier in prison.

The soldier is called Bradley Manning.

Bradley Manning is in prison for giving information about the Government of the United States to Wikileaks.

Wikileaks is a website which provides information on matters of public interest.

Why text simplification?

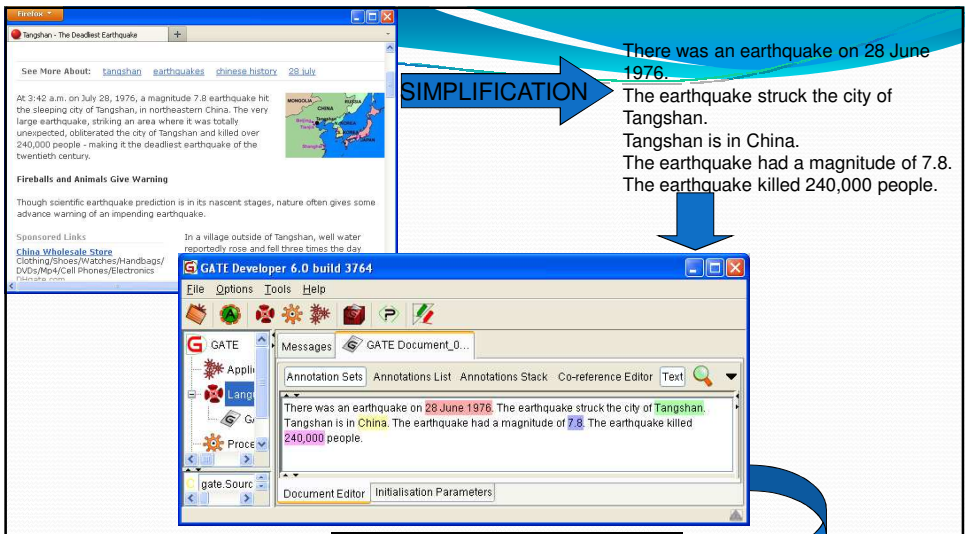
- It is an interesting research problem for the NLP community
  - Identify and measure sources of complexity / difficulty
  - Create a "paraphrase" which is easy to read
  - Several NLP expertises involved: summarization, natural language generation, sentence compression, word sense disambiguation, machine translation, etc....
- It is socially relevant
  - Unprecedented democratization of information (e.g. Web)
  - Information is not equally accessible to everyone
  - UN Enable: make information and information services accessible to different groups of persons with disability

Simplification users

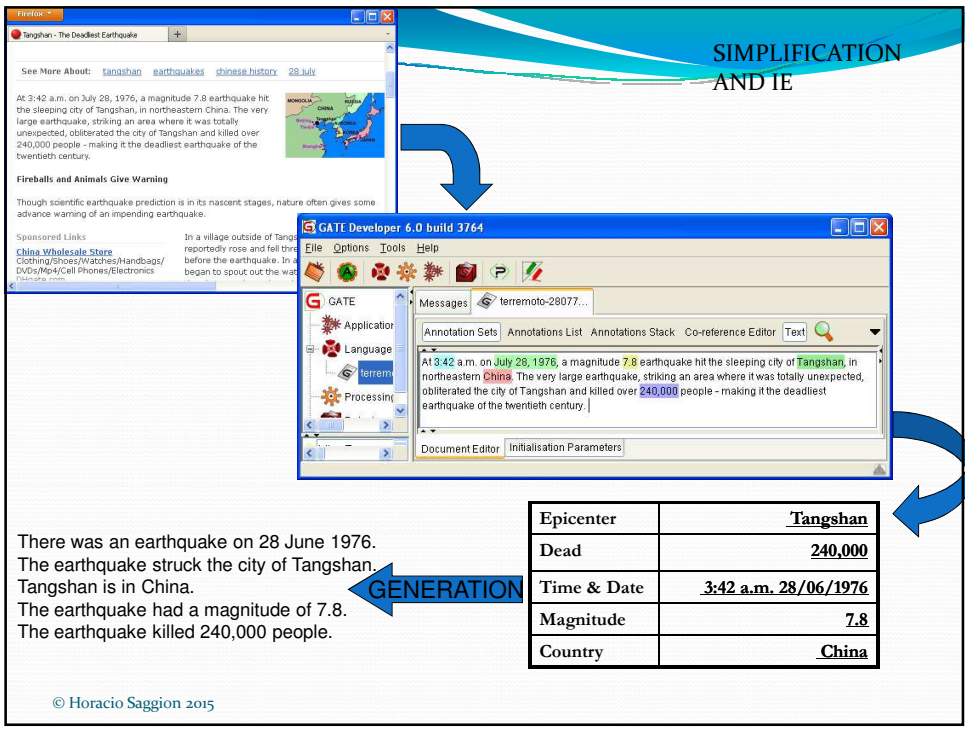
- Deaf people (Inui & al., 2003; Chung et al., 2013)
- Blind people (Grefenstette, 1998)
- People with low-literacy (Williams & Reiter, 2008; Aluísio & al., 2008)
- People with autism (Mitkov, 2012; Barbu et al., 2013; Orasan et al, 2013; Dornescu et al., 2013)
- Second language learners (Petersen and Ostendorf, 2007; Burstein et al., 2013; Eskenazi et al. 2013)
- Dyslexic people (Matausch & Peböck, 2010, Rello et al., 2013)
- People with aphasia (Carroll et al., 1999)

NLP as a simplification user

- Dealing with complex sentences
  - Initial simplification application (Chandresakar et al., 1996)
  - Improve results in IE (Jonnalagadda & Gonzalez, 2011; Evans, 2011; Minard et al., 2012)
  - Assist in question generation (Bernhard et al, 2012)
  - Text summarization (Grefenstette, 1998 , Siddharthan et al, 2004)



Epicenter	Tangshan
Dead	240,000
Time	28/06/1976
Magnitude	7.8
Country	China



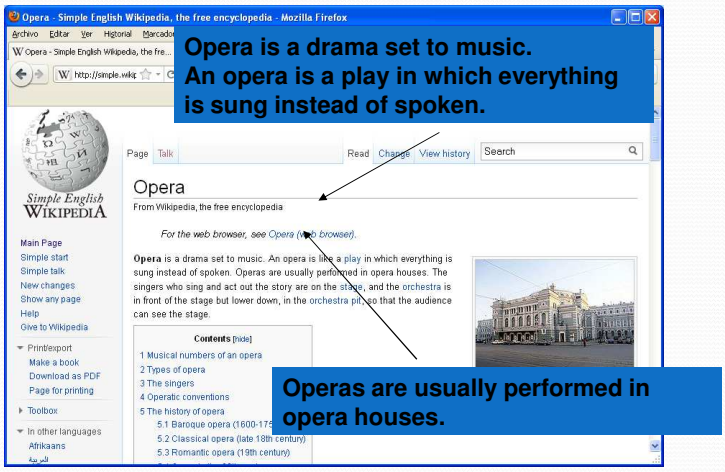
Languages

- *English* (Chandrasekar et al., 1996; Siddharthan, 2002; Carroll et al. 1998, Bouayad-Agha et al., 2009; Zhu et al., 2010; Coster & Kauchak, 2011; Yatskar et al., 2011), *French* (Seratan, 2012;François & Fairon, 2012), *Portuguese* (Aluísio et al., 2008; Specia, 2010), *Japanese* (Inui et al.,2003), *Arabic* (Al-Subaihin and Al-Khalifa, 2011), *Danish* (Klerke & Søgaard, 2012), *Swedish* (Smith et al., 2010; Keskisärkkä, 2012), *Spanish* (Saggion et al. 2011; Bautista et al., 2012; Rello et al, 2013; Mosquera et al., 2013), *Italian* (Dell'Orletta et al. 2011, Tonelli et al. 2012), *Basque* (Aranzabe et al, 2012), *Korean* (Chung et al, 2013)

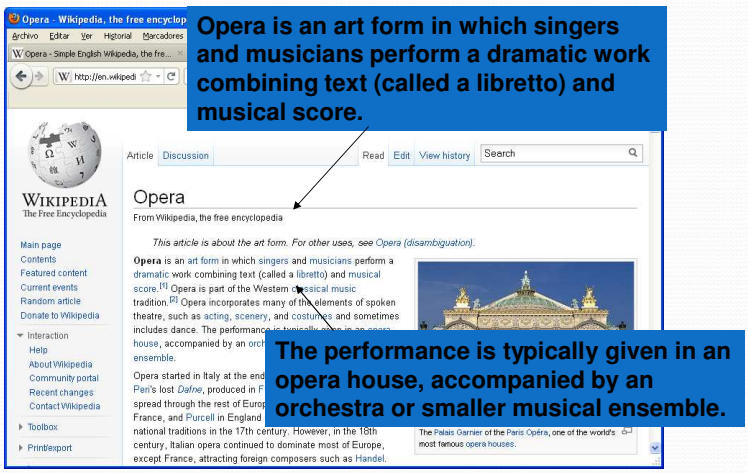
Simple language initiatives

- Easy to read initiatives
  - Plain English / Basic English (Ogden, 1930);
  - French Rationale;
  - Easy-to-Read network (Petz andTronbacke, 2008);
  - Fácil Lectura (http://www.lecturafacil.net);
  - European Association Inclusion Europe
- Guidelines
  - Simple and direct language;
  - one idea per sentence;
  - avoid jargon and technical and abbreviations;
  - one word per concept;
  - personalization;
  - use of active voice

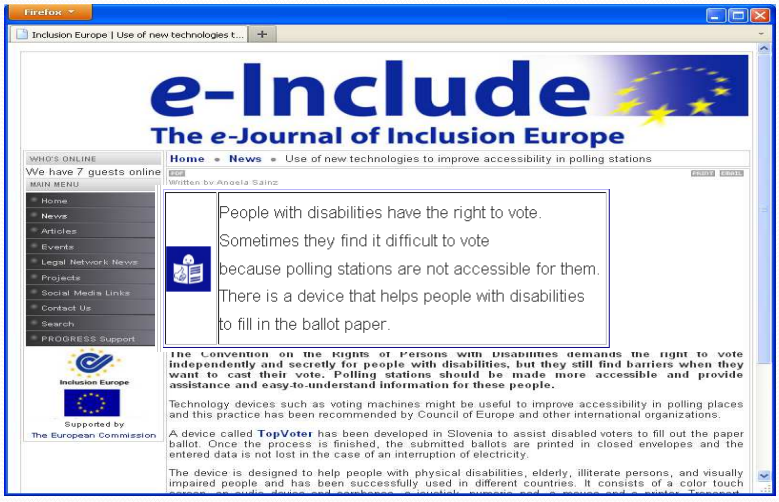
Where to find simple texts?



Where to find simple texts?



Where to find simple texts?



Where to find simple texts?

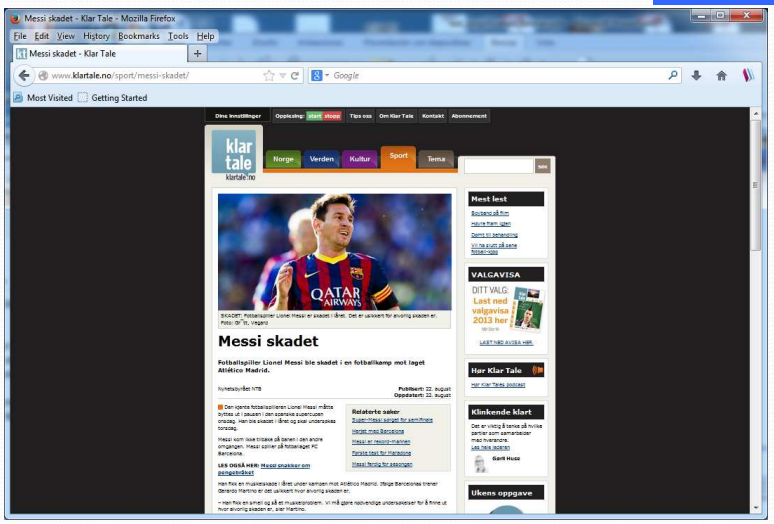
EASY NEWS

Where to find simple texts?

8 Pages News Paper (Swedish)

Where to find simple texts?

Klartale News Paper (Norwegian)



Where to find simple texts?

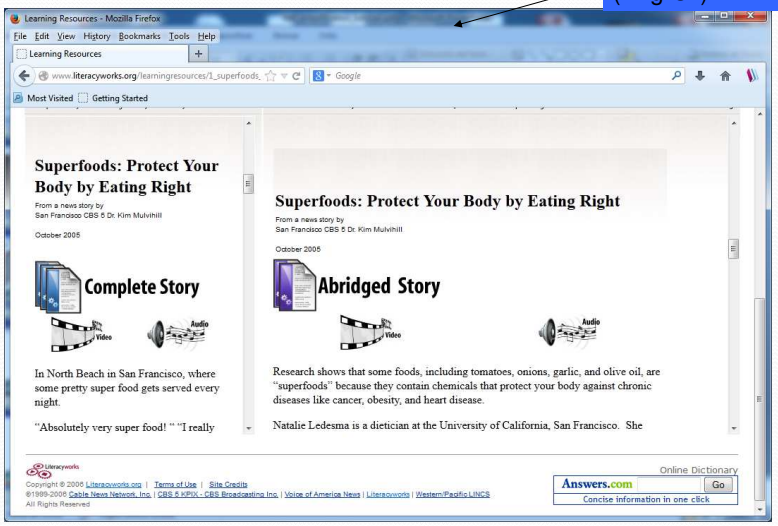
L'Essentiel News Paper (French)

Where to find simple texts?

Dueparole News Paper (Italian)

Where to find simple texts?

LiteracyWorks Web site (English)



Exercise

- Take one article from one (or more) of the web sites given, translate it into English to "verify" that it is a "simple" text
  - http://www.noticiasfacil.es/ES (Spanish)
  - http://www.dueparole.it/sommario_.asp (Italian)
  - http://www.8sidor.se/ (Swedish)
  - http://www.klartale.no/ (Norwegian)

Opportunities for NLP

- There is a heavy load associated to manual simplification of textual content
- It is impossible to wait to create texts that are adaptable
- Part of the work could be done automatically
- Machines could also help professionals
- There is also the idea that simple texts could be better dealt with by NLP tools …

Natural Language Processing for Text Simplification

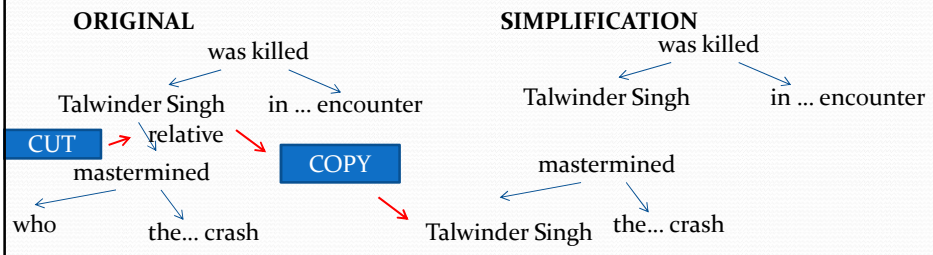
- Induced or hand-crafted rules for reducing syntactic complexity (Chandrasekar et al. 1996; Devlin & Tait, 1998; Siddharthan, 2002; Aluísio & al, 2008; Bouayad-Agha et al. 2009)
  - Simplification of relative clauses, treatment of passive constructions; substitution of pronouns by antecedents.
- Lexical substitution procedures (Caroll & al., 1998; De Belder & al, 2011; Biran & al., 2011)
  - Replacement of infrequent words by their most frequent synonym or replacement of words by a shorter synonym
- Combined approach treating the problem as MT (Zhu et al., 2010)
  - Using English Wikipedia and Simple English Wikipedia as a parallel corpus and learning from it word replacements, "drop", "copy", "split", and "reordering" operations.

First Steps: manual rules

- Rules over syntactic representations (Chandrasekar et al. 1996; Siddharthan, 2002)
  - Superficial analysis (ckunking) to identify noun and verb groups
  - Rules: W X:NP, RELPRO Y, Z. => W X:NP Z. X:NP Y. (manually developed)
  - Hu Jintao, who is the current Paramount Leader of the People's Republic of China, was visiting Bulgaria
    - W = ∅
    - X: =Hu Jintao
    - RELPRO:=who
    - Y = is the current Paramount Leader of the People's Republic of China
    - Z= was visiting Bulgaria
  - ➔ Hu Jintao was visiting Bulgaria. Hu Jintao is the current Paramount Leader of the People's Republic of China.

First Steps: rule learning

- Learning to transform from "complex" to "simple" (Chandrasekar & Srinivas, 1996)
  - (O) Talwinder Singh, who masterminded the 1984 Kanishka crash, was killed in a fierce two-hour encounter.
  - (S) Talwinder Singh was killed in a fierce two-hour encounter. Talwinder Singh masterminded the 1984 Kanishka crash.



First Steps: user's concerns

- Project PSET: "Practical Simplification of English Texts" (Devlin & Tait, 1998)
- Targeted simplification for people with aphasia
  - Transformation from passive to active voice
    - "A bid to build an incinerator on local wasteland was today accepted by the council." => The council today accepted a bid to build an incinerator on local wasteland. ✓
    - "Official documents were left on the underground by mistake." => Mistake left official document on the underground. (**note that this is a made up example !!!**) ✗
  - Resolution/replacement of anaphoric expressions
    - Standard anaphora resolution system + replacement of pronouns for antecedents/referents
  - Vocabulary simplification
    - Replacement of infrequent or little frequent words by their most frequent antecedent (use of a psycholinguistic database + WordNet)

What makes text difficult to read/understand?

- There are a number of formulas that intend to associate a readability level of **index** to a given texts
- Some formulas are extremely simple
  - they are based on concepts such as "word complexity", frequency, sentence length, etc.
- Experiments are generally carried out to show if these indices correlate with human associated indices of readability

Formulas for English

- Flesch (1949) – Flesch Reading Ease
  - S=206.835 – (1.015*ASL) –(84.6*ASW)
  - S is an index from 0 to 100, an index of 30 means very difficult, 70 means reasonable, 100 means easy
  - ASL = average sentence length
  - ASW = average number of syllables per word
- FOG index (Gunning, 1952)
  - S= 3.0680 + (0.877 * ASL) + (0.984 * PofP)
  - PofP is the percentage of polysyllables (over the total number of words)
  - (GL= 0.4 (ASL + # hard words))

Formulas for English

- Flesch-Kincaid (Kincaid et al, 1986)
  - Text Level = (.39 x ASL) + (11.8 x ASW) - 15.59.
    - ASL = average number of words per sentence (# words/# sentences).
    - ASW = average number of syllables per word(# syllables /# words).
  - a reasonable readability index will be between 6 & 10 (grade level).

Language models for predicting text complexity

- Word length and sentence length seem to be poor surrogates for text readability
- Language models are very popular in NLP and can be used to measure text complexity (Si & Callan, 2001)
  - text complexity prediction can be cast as a classic text classification problem
  - the model seeks to estimate the degree of difficulty (d) of a document (d) using a probability distribution p(g|d)
  - as always this probability is decomposed by Bayes into:

$$p(g \mid d) = \frac{p(g)\,p(d \mid g)}{p(d)}$$

Language models for predicting text complexity

- The different components of the model are then estimated as follows:
  - p(d|g) is an unigram model (the document is just a set of words)
  - p(g) is the *a priori* distribution of a difficulty degree g
  - p(d) is dropped from the equation because it does not affect the result
- The probabilistic model is combined with a sentence length model which assumes normal distribution of lengths per difficulty level
- The combined model has more predictive power than the classic Flesch-Kincaid test

Text Quality (Pitler & Nenkova, 2008)

- Study the relative importance of different factors affecting "text quality"
  - words, syntax, discourse (e.g. coreference), cohesion, etc.
- Dataset: Discourse Penn Treebank annotated with explicit and implicit discourse relations (dis.rel)
- Human assessment of a set of 30 articles
  - 3 assessors for each text
  - rating "how well this text is written?"
  - scores averaged per text
- Correlations are computed between values of features and "perceived" text quality

Text Quality (Pitler & Nenkova, 2008)

- Features in the study
  - baseline: char. per word, words per sentence, max number of words per sent, **article length**.
  - vocabulary (various Log Likelihood – LogL - unigram language models): **WSJ, AP News, and variants that take into account article length**
  - syntactic features: height of parse tree, np per sentence, **vp per sentence**, subordinate per sentence
  - cohesion: proper nouns, definite articles, various similarity features adj. sentences
  - entity coherence: 34 features
  - discourse relations: # discourse relations, **LogL of article based on disc. rel., LogL based on disc. rel. and number of disc. rel**, explicit disc. rel, implicit disc. rel.

Readability and Linguistic Features (Stajner et al. 2012)

- Study the correlation between commonly used readability formulas in English and a number of features which arise from linguistic analysis of the text
- Corpora: Simple Wikipedia, News, Health, Fiction
- Readability measures: Flesh Reading Ease Score, Flesh-Kincaid, Fog Index, SMOG grading, char. / wrds, syl / wrds, wrds / sentence
- Features (average per sentence):
  - Structural: Nouns, Adjectives, Determiners, Adverbs, Verbs, Infinitive, Coordinating Conj., Subordinating Conj., Prepositions
  - Ambiguity: Pronouns, defNPs, Word Senses

Readability and Linguistic Features (Stajner et al. 2012)

- Correlation between readability formulae in the 4 datasets is very high (over .95)
- Correlation between Flesh Reading Ease score and POS features
  - The strength of correlation depends on the dataset
  - For most of the features (V, N, Prep, Det, Adv, A, CS, CC, INF, ch/w, w/s, defNP) , the smaller the value the more readable the text
  - For word senses, the greater the value the more readable the text
- Unexpected finding: fiction texts simpler than Simple Wikipedia if we accept readability formulae as predictors of text complexity

Additional Studies

- Portuguese (Aluisio et al. 2010) and Italian (Tonelli et al, 2012)
  - Coh-Metric system (Graesser et al., 2004)
  - Portuguese: Coh-Metrix-PORT tool to compute 59 features to perform 3-way classification incorporated into the tool FACILITA
  - Italian: Coease system based on Coh-Metrix features (46) re-implemented for Italian used in 3-way classification
- Italian (Dell'Orletta et al, 2011)
  - Study readability assessment as text classification
  - 2 corpora: La Repubblica (normal) and Due Parole (simple)
  - raw textual features, lexical features, **morpho-syntactic features**, syntactic features
  - Study sentence classification
- French (François, 2011)
  - AI Readability formula – 406 features (**lexical, syntactic**, semantic, FFL) – classification into 6 readability levels

Additional Studies

- Swedish:
  - LIX formula (Björnsson, 1968) which combines average number of words per sentence with average number of "complex" words per word (other readability indexes are "nominal ratio" and "word variation index")
  - Falkenjack et al (2013) apply a machine learning approach to classify texts into two categories (simple/normal), features - shallow, morpho, syntactic, lexical – are borrowed from previous research
- Arabic (Al-Khalifa et al., 2010)
  - use single and combination of features for readability assessment (3-levels): **perplexity**, **sent. length**, word length, avg. syllables, **frequency**
- Japanese (Sato et al., 2008)
  - Character (3 alphabets) unigram LM for 13 different grades (LM1… LM13). Grade of text is the one that makes a text "more likely"

Additional Studies

- Spaulding's formula for Spanish (1956)
  - Found two factors correlated with text difficulty
  - These two factors are not correlated
    - Length of sentence (average) = ASL
    - Density of use of the vocabulary (excluding words from a pre-defined list of 1500 Spanish words considered easy) =Density
    - Difficulty= 1.609 * ASL + 331.8 * Density+22
    - a (to), abajo (down), abandonar (quit), aborrecer (hate), abrazar (hug), abrir (open),….bailar (dance), bajar (descend), bajo (short),….
  - The method comes with a graphic to map the obtained indices: 40-60 very easy; 61-80 easy;81-100 moderately difficult; 101-120 difficult: > 121 exceptionally difficult
  - único (unique),….

Additional Studies

- Two stage approach to avoid underestimation in readability assessment (Sheehan et al, 2013)
  - Classification into genre
  - Estimate difficulty within genre
  - Obtain better estimators
- New measure: lexical tightness (Flor et al., 2013)
  - Measures to what extent words used in a text are associated in the language
    - if a text uses words that are strongly associated then it is easier to read
  - It is derived from a modification of Pointwise Mutual Information
  - Better text complexity estimators

Exercise

- Compute the complexity of these two texts using Flesch Reading Ease and indicate which one is simpler (note that you have to count syllables)
  - (I) Spain is a country in Southern Europe. It is in the Iberian Peninsula near Portugal and Gibraltar. France and the little country of Andorra are on its northeast side, where the Pyrenees mountains are.
  - (II) Spain, officially the Kingdom of Spain, is a sovereign state and a member of the European Union located in southwestern Europe on the Iberian Peninsula. Its mainland is bordered to the south and east by the Mediterranean Sea except for a small land boundary with the British Overseas Territory of Gibraltar; to the north and north east by France, Andorra, and the Bay of Biscay; and to the northwest and west by the Atlantic Ocean and Portugal.

Lexical simplification as a NLP task

- Lexical Simplification is a subtask of Text Simplification (Siddharthan, 2006) concerned with replacing words or short phrases by simpler variants in a context aware fashion (generally synonyms), which can be understood by a wider range of readers
- SemEval-2012 featured a lexical simplification task (Specia et al. 2012)
  - Based on a previous challenge on lexical substitution (McCarthy & Navigli, 2007)

Lexical simplification as a NLP task

- 3 aspects were considered
  - complexity analysis
  - search for word substitutes
  - ranking of substitutes based on context /simpler
- a dataset is created based on aggregation of information by several annotators
  - Example of Lex Subs.: "… a <u>bright</u> boy…"; bright = intelligent (3); clever (3); smart (1)
  - Lex Simp. dataset uses average ranks provided by annotators
- the task: replace a given word for its most appropriate substitute so that the sentence is easier to read
- 10 systems participated in the evaluation, a baseline system based on word frequency works fine and it is only beaten by one system

Lexical Simplification

- Combination of two sources: lexicon & language model (De Belder & al. 2010)
  - Given a word in a text, two lists of words are generated
    - L1: list of synonyms from the lexical database (authoritative source)
    - L2: list of alternative words obtained from a latent words language model (learning from non annotated data)
  - A probabilistic model estimates the probability of replacement of one word (original) by another word
    - P1(w|w_original)=P2(w|w_original,context)*P3(easy|w)
  - P2 is a language model that w fits in the given context
  - P3 can be modelled in different ways: frequency, morphosyntactic properties, complexity based on database, etc.
- (De Belder & Moens, 2012) create a dataset for evaluation of TS in English – similar to work done in SemEval 2012
  - Use data from the lexical substitution task proposed by (McCarthy & Navigli, 2007)
  - Filtering of very difficult words
  - Alternative words were ranked by difficulty by annotators
  - Rankings grouped to obtain a single model

Extracting Lexical Simplifications from Wikipedia

- Hypothesis: changes in Simple Wikipedia correspond to simplifications the author is making….(not always!) (Yatskar et al. 2010)
- A mechanism is needed to model when the change of one word by another is due to a simplification operation
- Various models are possible:
  - One model computes the "probability" that the change of a word "A" by word "a" is due to: correction, simplification, etc.
  - It is assumed that in the normal Wikipedia simplification changes are negligible
  - It is also assumed that the proportion of corrections in Simple Wikipedia is equal to those in normal Wikipedia
  - The probability of changing "A" by "a" p(a|A) is approximated by frequencies
  - A model is obtained for the most probable replacement for "A"
- Another method looks at comments left by editors identifying which association of "A" and "a" is stronger using PMI (point-wise mutual information)

Extracting Lexical Simplifications from Wikipedia

- Data extraction: pairs of words "A" and "a"
  - Sentences are aligned and words which have been substituted are identified: A => a
- Two baseline methods proposed:
  - Frequency: use the most frequent substitution
  - Random: chose a random valid substitution
- Compare with a list created automatically
  - Human > Language Model > PMI > FREQ >= RANDOM

Learning simplification rules

- Biran et al. (2011) also use English Wikipedia (EW) and Simple English Wikipedia (ESW)
- EW is used to extract context vectors for each word (co-occurrences)
- A similarity measure can be used to identify which words can be replaced by which words
  - The cosine between vector representations is used in this work
  - Some filtering applied using WordNet

Learning simplification rules

- Implementing the simplicity of a word: example "canine" and "dog"
  - check occurrences of both words in EW and SEW
  - "canine" appears 9620 times in EW
  - "canine" appears 62 times in SEW
  - "dog" appears 171000 times in EW
  - "dog" appears 1360 times in SEW
  - complexity("canine") = 9620/62 = 155
  - complexity("dog") = 171000/1360 = 125

Learning simplification rules

- The length of the word is also taken as a measure of complexity
  - len("canine")=6, len("dog")=3
  - final_complexity=complexity*len
  - fc("canine")=155*6=930
  - fc("dog")=125*3=375
- canine "is more difficult than" dog
- So "canine" can be simplifies by "dog", but "dog" can not be simplified with "canine"

Learning simplification rules

- Grammaticality: generate all equivalent pairs, if word in past tense then its simplification in past tense, etc.
- Chose as simplification of the target word w a replacement x that fits in the context
- Baseline: replace a word by its more frequent synonym
- Evaluation is a non-realistic scenario in the sense that only one word is simplified in the sentence
  - chose a sentence where only one word has been replaced by the method
  - three variables evaluated: simplification(yes/no), grammaticality (bad/ok/good), sense (yes/no)
  - the proposed method is better than the "baseline"

Lexical simplification (Bott et al., 2012)

- LexSiS: Lexical simplification for Spanish
  - Word sense disambiguation (WSD) of a target word to identify an appropriate list of synonyms
  - Simplicity computation to chose an appropriate synonym for a target word
  - Morphology generation of the simpler synonym
- Example
  - **Original:** Descubren en Valencia una nueva **ESPECIE** de pez prehistórico / A new **SPECIES** of prehistoric fish is discovered in Valencia.
  - **Simplified:** Descubren en Valencia un nuevo **TIPO** de pez prehistórico / A new **TYPE** of prehistoric fish is discovered in Valencia.

Word Sense Disambiguation

- A freely available lexical resource (Spanish Open Thesaurus - OT) can be used to select synonyms
  - mono (4 senses) / *monkey*
    - gorila, simio, antropoide / *primate*
    - simio, chimpancé, mandril, mico, macaco / *monkey*
    - overol, traje de faena / *garment*
    - llamativo, vistoso, atractivo,.... / *attractive*
- Using an 8M word corpus, a vector representations for each entry in OT can be created
  - the vector models words in context frequencies (4 words to the left, 4 words to the right)

Word Sense Disambiguation

- The final representation of a word-sense w is an aggregated vector of all synonym terms for w
  - mono (4 senses)
    - gorila, simio, antropoide => mono_vector_1
    - simio, chimpancé, mandril, mico, macaco => mono_vector_2
    - overol, traje de faena => mono_vector_3
    - llamativo, vistoso, atractivo,.... => mono_vector_4
- Given a target word w to be simplified, a vector is computed based on w *current* context
  - Two methods used: one-sense-per-discourse vs many-senses-per-discourse
- The similarity between this vector and all vectors for word w in the thesaurus are computed using *cosine*
- Pick the list of synonyms whose vector is more similar to the target vector for w

Simplicity computation

- Corpus analysis findings
  - Word length
    - short words are more abundant in simple texts than in non-simplified texts
  - Frequency (based on a word frequency list from RAE)
    - low frequency words or words not in our reference lexicon are 50% more common in non-simplified texts than in simplified texts
- This has led to the implementation of a simplicity score for words that combines frequency and length
  - Frequency (also used in previous art)
  - Length (also used in previous art)
- The target word is replaced only if a number of conditions are met

Simplicity computation

- Example
  - Descubren en Valencia una nueva **ESPECIE** de pez prehistórico (A new **SPECIES** of prehistoric fish is discovered in Valencia.)
  - Open Thesaurus lists 5 different synonym lists for ESPECIE (SPECIES)
  - Frequency-based lexical simplification procedure selects the word GRUPO (GROUP) as replacement for ESPECIE out of all available synonyms
  - LexSiS chooses one of the correct synonym lists and from the list the word TIPO (TYPE) as an appropriate replacement because the length factor makes TYPE "simpler" than GROUP

LexSiS: Evaluation

- Users presented with pairs of sentences and asked to indicate how the target word in the first sentence was when compared to the target word in the second sentence:
  - (A) El visitante puede contemplar los óleos y esculturas que se exponen en la PINACOTECA. (The visitor can appreciate the paintings and sculptures showed in the ART GALLERY.)
  - (B) El visitante puede contemplar los óleos y esculturas que se exponen en el MUSEO. (The visitor can appreciate the paintings and sculptures showed in the MUSEUM.)

simpler, more complex, equally complex/simple, not same meaning, don't understand

LexSiS: Evaluation

- Three versions compared gold standard, random, frequency, and LexSiS
- Overall LexSiS shows **higher meaning preservation** and proposes **more simpler synonyms** (Bott et al. 2012)
- Problems: replacements of words which are not difficult, replacement of words in collocations, out-of-vocabulary problem, ...

Dealing with Numbers

- Simplification of numerical expressions in text (Bautista et al. 2012) and adaptation to Spanish (Bautista & Saggion, 2014)
  - studies the problem of how to make numbers and numerical expressions simpler by the use of rounding and addition/change of modifiers

Original	Simplification
Cerca de 1,9 millones de personas asistieron al concierto (About 1.9 million people attended the concert)	Casi 2 millones de personas asistieron al concierto (Nearly 2 million people attended the concert)
Sólo se ha vendido un cuarto de las entradas (Only a quarter of the tickets have been sold)	Sólo se ha vendido ¼ de las entradas (Only ¼ of the tickets have been sold)
Uno de cada cuatro niños hablan chino (One in four children speak Chinese)	1 de cada 4 niños hablan chino (1 in 4 children speak Chinese)
Asistieron un 57% de la clase (57% of the class attended....)	Asistieron mas de la mitad de la clase (More than half of the class attended...)
Aprobaron el 98% (98% passed....)	Aprobaron casi todos (Almost everyone passed...)

Dealing with Numbers

Original expression	Rounded quantity	Case	Modifier	Simplified expression
Alrededor de 5689 millones (around 5689 million)	6000 millones (6000 million)	There is a modifier in the original expression	Original modifier is kept	*Alrededor de 6000 millones*
27.3 %	25 %	No modifier in original expression, and original > rounded	Add modifier "más de" (*more than*)	*Más de 25 % (more than 25 %)*
476	500	No modifier in original expression, and original < rounded	Add modifier "casi" (*almost*)	*Casi 500 (almost 500)*
3000	3000	No modifier in original expression, and original = rounded	No modifier	3000

- NE simplification procedure
  - Text processing using FreeLing
  - Transformation of FreeLing output into XML representation
  - Application of grammars for numerical expression recognition
  - Simplification of target numerical expression
  - Sentence rewriting

Simplifying Numerical Expressions

- Recognition grammars
  - Rules implemented to expand the Z annotations to cover different types of pre-modification
    - recognizing modifiers, quantity, mathematical representation (percentage, decimal, exact, ...), units, kind of expression
- Re-writing programs
  - Implementation of simplification operations observed in corpora or identified through a survey
- Questionnaire evaluation (15 pairs of sentences, 34 cases of NumExps)
  - Meaning preservation (M), grammaticality (G), correct simplification (S)
  - 42 experts compared original and simplified sentences
  - 81% of respondents positive about M
  - 79% of respondents positive about G
  - 72% of respondents positive about S
- Common errors
  - Comparatives: "The price of bread went up from almost 1 euro to almost 1 euro per kilo" !!!

Exercise

- Try the method with the following word pairs indicating which word is easier
  - cat, feline
  - automobile, car
  - inferno, fire

Simplification and Cohesion

- Siddharthan (2006) was concerned with generation issues during text simplification
  - sentence order, word choice, generation of referring expressions
  - *(1) Mr. Anthony, who runs an employment agency, decries program trading, but he isn't sure it should be strictly regulated.*
  - *? (2a) Mr. Anthony decries program trading. (2b) Mr. Anthony runs an employment agency. (2c) But he isn't sure it should be strictly regulated.*
- Tree stage approach: analysis, transformation, regeneration
  - analysis: text chunking
  - transformation: set of hand crafted rules
  - regeneration: sentence ordering, anaphora, conjunctive cohesion (choice of connectives)
- More recently (Siddharthan, 2011) argues for the use of dependency relations in text simplification allowing him to better model and learn lexical transformations (Siddhartan & Angrosh 2014)

Learning "all" simplification automatically

- This work considers that text simplification is a kind of machine translation (Coster & Kauchak, 2011) (but this is not totally true!)
  - normal language TO simple language
- Can be applied in contexts were original and simplification examples abound.... (original and simplification are not very different)

Learning "all" simplification automatically (PWKP aligned dataset)

Original sentence (Wikipedia)	Simplified sentence (Simple Wikipedia)
Greene agreed that she could earn more by breaking away from 20th Century Fox.	Greene agreed that she could earn more by leaving 20th Century Fox.
In 1962, Steinbeck received the Nobel Prize for Literature.	Steinbeck won the Nobel Prize in Literature in 1962.
They established themselves here and called that port Menestheus's port.	They called the port Menestheus's port.
April is the fourth month of the year in the Gregorian Calendar, and one of four months with a length of 30 days.	April is the fourth month of the year with 30 days.
August is the eighth month of the year in the Gregorian Calendar and one of seven Gregorian months with the length of 31 days.	August is the eighth month of the year. It has 31 days.
France has a close association with the Modern Olympic Games; it was a French aristocrat, Baron Pierre de Coubertin, who suggested the Games' revival, at the end of the 19th century.	France is closely associated with the Modern Olympic Games. At the end of the 19th century the Baron Pierre de Coubertin suggested to have Olympic Games again.

Learning "all" simplification automatically

- It is applied in the more specific context of statistical machine translation and in particular phrase-based translation models
- An statistical translation model is based on the following formula to model the "generation" of a translation from a source sentence
  - P(e|f) is the probability that a text e is a transation of text f
  - this formula is appoximated by P(e)*P(f|e)
    - P(e) is a language model
    - p(f|e) is a translation model
  - given a text f we look for a text e that maximizes the formula
  - finding the "best" e requires examination of all possibilities which is impractical. An heusristic search is required to find a solution.
- The probabilities as stated can not be modelled directly, one has to use text and sentence components for modelling the problem (e.g. an approximation)

Learning "all" simplification automatically

- Creation of corpus for training and text
  - Align articles of SEW and EW
  - Align paragraphs using a similarity metric
  - Align sentences using dynamic programming
- Simplification model is based on MOSES (Koehn et al, 2007)
  - each s_i is a phrase in a simple sentence and each n_i is a phrase in the original sentence

$$p(simple \mid normal) = \prod_{i=1}^{m} p(s_i \mid n_i)$$

- Phrases can be computed using the computer program GIZA++ (Och & Ney, 2000)
- In a translation model it is generally true that "null" phrases do not exist, however in simplification they are necessary

Learning "all" simplification automatically

- In the evaluation various systems are compared
  - Normal translation model
  - Translation model with delete operation
  - Do nothing system (baseline)
  - Other text simplification approaches (Cohen & Lapata, 2009; Knight & Marcu, 2002)
- Evaluation (comparison with gold standard)
  - BLEU (Papineni et al, 2002) used in machine translation
  - Simple string accuracy (Clarke & Lapata, 2006)
  - F-score over words
- For all tested measures
  - Translation system with delete operation is better than translation system without delete operation
  - The baseline is better than the sophisticated systems

Learning "Simple" Portuguese

- Specia (2010) – First to cast text simplification as a kind of translation problem
- The MOSES standard phrase-based Statistical Machine Translation system used for training a model
- PorSimples Corpus
  - 3383 sentence pairs
  - 500 sentence pairs for additional tuning
  - 500 sentence pairs for testing
- Results in terms of BLUE metric **0.60**
- System is very cautious when performing simplification resulting in an output too similar to the input

Why SMT-based Simplification doesn't Work for me?

- Stajner (2015) studied SMT-based simplification for the Spanish language
- Simplext corpus
  - 200 pairs or <original, simplified> documents
  - 925 "sentence" pairs - very limited training data
  - simplifications are for people with intellectual disabilities
- Replication of Specia and Coster & Kauchak approaches
- BLEU score for Spanish very low - why???
  - Portuguese corpus – 4483 sentence pairs
  - English corpus – 137000 sentence pairs
  - Spanish corpus – may be size is a factor???

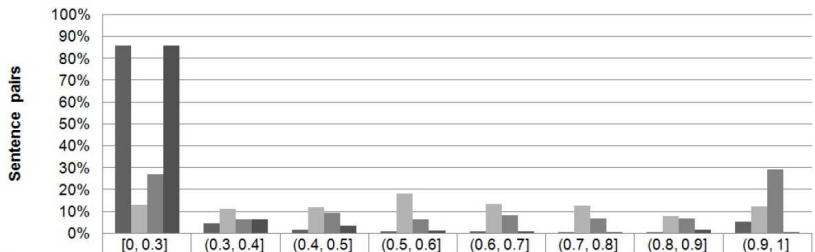
Why SMT-based Simplification doesn't Work for me?

- Compare: Spanish, English, Portuguese
- WikiPedia TS corpus (English), Simplext (Spanish), PorSimples (Portuguese)
- Same size of training, development, and test
- "similar" type of language model in the 3 languages

Corpora	Training	Dev.	Test	BLEU
Simplext	741	94	90	10.05
PorSimple	741	94	90	48.06
WikiPedia	741	94	90	51.47

Quality vs Quantity of Simplification Data

- Original vs Simplified sentences for 4 different datasets (+ Encyclopaedia Britannica)
  - Similarity metric S-BLEU



	[0, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1]
Simplext	85.96%	4.45%	1.62%	0.94%	0.94%	0.27%	0.40%	5.40%
Porsimples	12.96%	11.20%	11.74%	18.08%	13.23%	12.82%	7.83%	12.28%
Wikipedia	26.86%	6.48%	9.31%	6.34%	8.37%	6.88%	6.75%	29.15%
EncBrit	85.86%	6.49%	3.49%	1.16%	0.83%	0.50%	1.50%	0.17%

Quality vs Quantity of Simplification Data

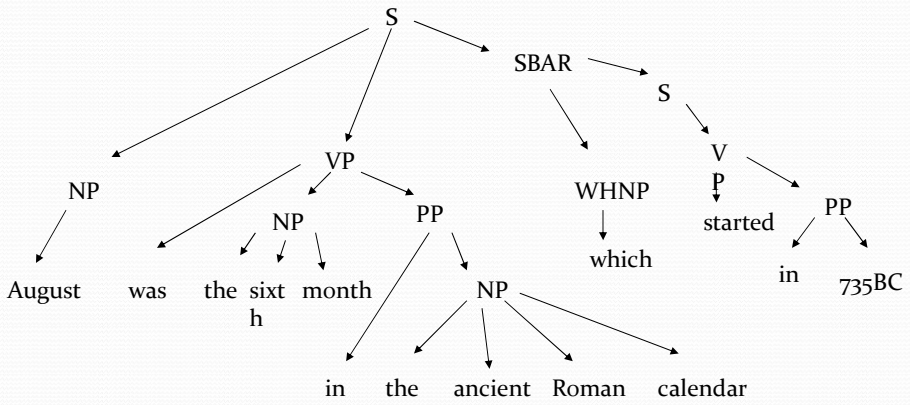
- Simplext and EncBritannica datasets have a higher number of substitutions and insertions than in PorSimples and Wikipedia
- PorSimples corpus used in the experiments contain "natural" simplifications
- Wikipedia TS corpus contains sentence pairs which already have a cosine similarity over 0.50
- It seems that when sentence pairs in training have moderate differences in terms of similarity, the system is able to produce sentences that are grammatical and meaning preserving but not simpler

Learning simplification from trees

- Based on a corpus of comparable documents <C,S> of complex and simplified versions (Zhu et al. 2010)
  - English Wikipedia/Simple English Wikipedia
  - Align EW & SEW using a TF*IDF method and allow 1 to n alignments
- This work models the following aspects:
  - replacement of words and phrases
  - syntactic simplification seen as composition of the following operations on a tree
    - "Split", "Drop", "Copying", "Reordering"

Learning simplification from trees

PHRASE STRUCTURE OF COMPLEXT SENTENCE

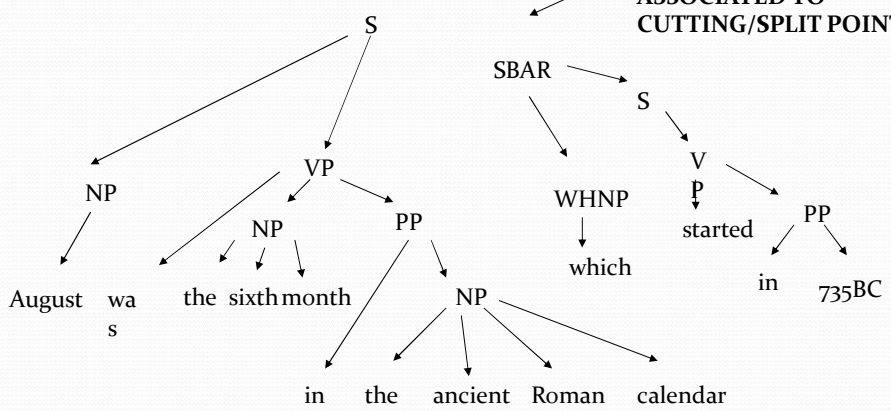


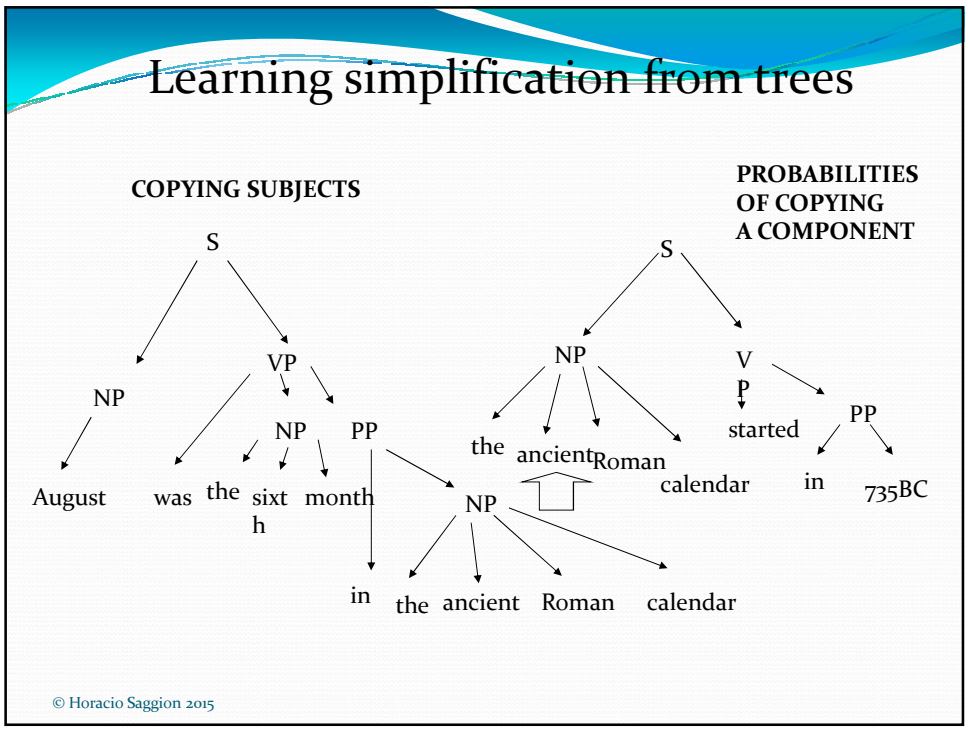
August was the sixth month in the ancient Roman calendar which started in 735BC.

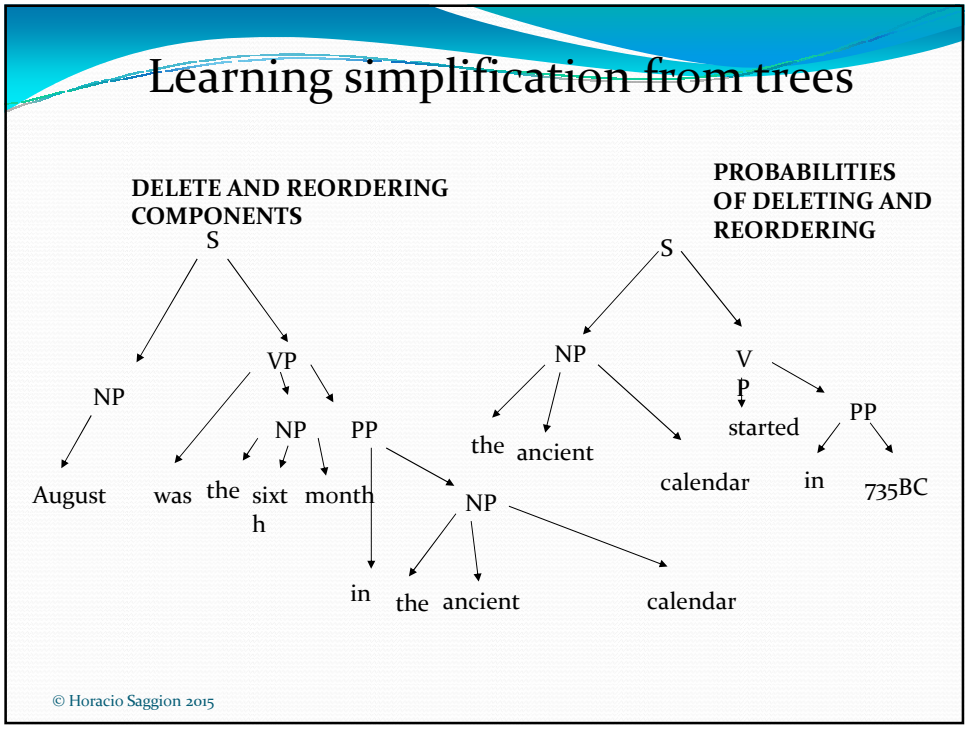
Learning simplification from trees

SPLIT

PROBABILITIES ASSOCIATED TO CUTTING/SPLIT POINTS



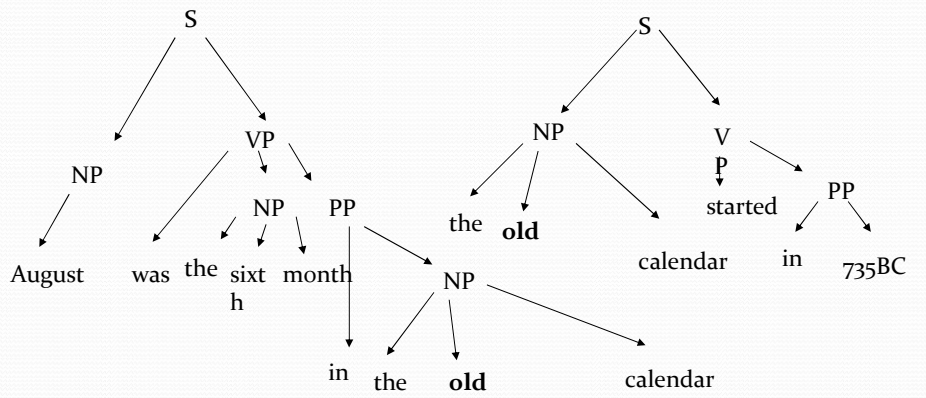

Learning simplification from trees
COPYING SUBJECTS
PROBABILITIES OF COPYING A COMPONENT
S
NP
VP
NP
PP
August
was
the
sixth
month
in
NP
the
ancient
Roman
calendar
S
NP
VP
started
PP
the
ancient
Roman
calendar
in
735BC
© Horacio Saggion 2015


Learning simplification from trees
DELETE AND REORDERING COMPONENTS
PROBABILITIES OF DELETING AND REORDERING
S
NP
VP
NP
PP
August
was
the
sixth
month
in
NP
the
ancient
calendar
S
NP
VP
started
PP
the
ancient
calendar
in
735BC
© Horacio Saggion 2015

Learning simplification from trees

WORD SUBSTITUTION

PROBABILITY OF REPLACING A WORD



August was the sixth month in the old calendar. The old calendar started in 735BC.

Optimization Approaches

- Woodsend & Lapata (2011) propose two components to "learn" to simplify English
  - Learn from corpora simplification transformations
  - Optimize rule application
    - Given a sentence produce "all" possible simplifications licensed by the grammar
    - select the "simplest" one using a number of constraints
- Quasi-synchronous grammars allow them to model non-isomorphic transformations
  - lexical rules and splitting rules are extracted from aligned corpus
- Integer Linear programming (ILP) is used to select an optimal simplification
  - cost function: grammaticality + readability
- ILP is also applied to English by De Belder (2014) and to French by Brouwers et al. (2014)

Semantic Representations

- Narayan & Gardent (2014) argue that previous approaches fail because they only look into syntax, disregarding semantics
  - Choice of pronouns are not treated appropriately
  - Current MT systems can not handle syntactic transformations, only reordering and word substitution
- They propose that splitting and deletion operations need semantic information
- They model the text using Kamp's Discourse Representation Structure having separate models for syntactic simplification and word substitution and ordering
- Their method outperforms all previous approaches

Simplification as NLP "facilitator"

- Jonnalagadda et al (2009) use simplification to improve parsing results of biomedical texts
  - Two-step method: domain dependent text transformations & sentence splitting
  - "Simplified" sentences obtain better parsing results than non-simplified sentences
- Evans (2011) studies complex coordinations and subordinations (aboundant in medical texts)
  - Instead of having complex extraction patterns, sentences are simplified
    - complex coordination: *Examination shows jaundice, hypothermia, hypotonia, large anterior and posterior fontanels, and a hoarse cry*
  - Simplification helps a process of information extraction

Simplification applications

- Text Summarization (Pal & Ruger, 2002)
  - Complexity of the sentence analyzed using a psycholinguistic database
  - Replace complex words by simple synonyms using WordNet
  - Also Siddharthan et al (2004) use simplification in summarization
- Subtitling : reducing the number of characters given physical constrains of the medium (Daeleman et al, 2004)
  - One important operation is deletion of material
  - dataset: alignments between transcriptions and subtitles
  - machine learning algorithms (using parallel corpus)
  - knowledge based system (linguistic rules for deletion)
  - machine learning approach performed badly
  - rule-based approach worked better

Simplification applications

- Simplification of patent claim "sentences" (Bouayad-Agha et al. 2009) - PatExpert

An optical disk drive comprising: a laser light source for emitting a laser beam; an optical system for conversing the laser beam from the laser light source on a signal plane of optical disk on which signal marks are formed and for transmitting the light reflected from the signal plane; one or more optical components, arranged in the optical path between the laser light source and the optical disk, for making the distribution of the laser beam converged by the conversing means located on a ring belt just after the passage of an aperture plane of the optical system; a detection means for detecting the light reflected from the optical disk; and a signal processing circuit for generating a secondary differential signal by differentiating the signals detected by the detection means and for detecting the edge positions of the signal marks by comparing the secondary differential signal with a detection level.

An optical disk drive comprises a laser light source, an optical system, a detection means, and a signal processing circuit.
The laser light source emits a laser beam.
The optical system converses the laser beam from the laser light source on a signal plane of optical disk.
On the latter, signal marks are formed.
The optical system also transmits the light reflected from the signal plane.
.....
The signal processing circuit generates a secondary differential signal.
To do so, it differentiates the signals detected by the detection means.
It also detects the edge positions of the signal mark.
To do so, it compares the secondary differential signal with a detection level.

Adapt simplification to the user

- PorSimples project (Aluísio & al, 2008)
- Target population: people with low literacy
- Study of simplification cases based on corpus and specification of simplification procedures
  - Treatment of relative clauses:
    - Input: "The book, which John gave me, belongs to Paul"
    - Find relative pronoun and check the relative is non restrictive
    - Find where the relative ends
    - Generate a sentence with the relative
    - Generate a sentence with the main clause.
    - Reorder
    - Output: "The book belongs to Paul. John gave me the book."
- The are various procedures in charge of dealing with different phenomena. These are cascaded in specific order to achieve a correct simplification.

Adapt simplification to the user

- The FIRST project "Flexible Interactive Reading Support Tool" (funded by FP7) (Mitkov, 2012)
  - Simplification for people with autism: they have problems with complex sentences, ambiguity, figurative language, etc.
  - Adaptive tool will simplify text, provide navigation mechanisms, and add pictograms for better understanding....
  - Simplification is not only for the final user but also for their carers
- (Dornescu et al, 2014) describe how relative clauses and other syntactic phenomena are dealt with in FIRST
  - Rule-based + sequence learning (CRFs) to identify complex constructions

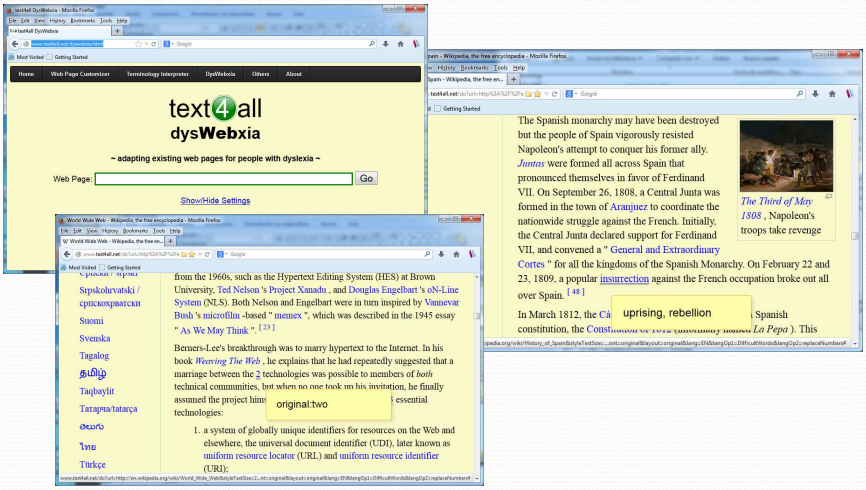
Adapt simplification to the user

- DysWebxia (Rello 2014): Simplification/adaptation for dyslexic users
- Dyslexia is a neurological reading disability that is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities
- how to make textual content more accessible to people with dyslexia
  - 10-17% of native English speakers are affected by dyslexia
  - 7.5-10% of native Spanish speakers are affected by dyslexia
- basis for the work: use of eye-tracking methodology to verify different word/text conditions (frequency, length, use of paraphrase, etc.)

Adapt simplification to the user

- Studying frequency and length factors (Rello & al. 2013)
  - Frequency and readability: significant effect
  - Length and readability: significant effect
  - Frequency and understandability: no effect
  - Length and understandability: significant effect
- Substitute or Help? (Rello & al. 2013)
  - Comparison between
    - automatically replacing a word by a synonym using LexSiS (Bott &al., 2012)
    - providing an interactive way of accessing to a synonym list (after word sense disambiguation and simplicity scoring)
    - dyslexic users prefer the helping aid to the substitution by synonyms

DysWebxia browser



The Simplext Project (Saggion et al., 2011; Saggion et al. 2015)

- A text simplification system for people with cognitive disabilities
- One of the first applications of simplification to the Spanish language
- Technological objective
  - Develop an ubiquitous simplification solution
  - Simplified content can be delivered to different devices
- Scientific objectives
  - Development of resources for simplification in Spanish
  - Development of methods and algorithms to simplify content
- Social objective: simplifying content for people with cognitive disabilities
- Spanish: spoken by over 500 million people, official language of 21 countries, 3rd language on Internet

Simplext Demo

INFORMATION PROVIDER

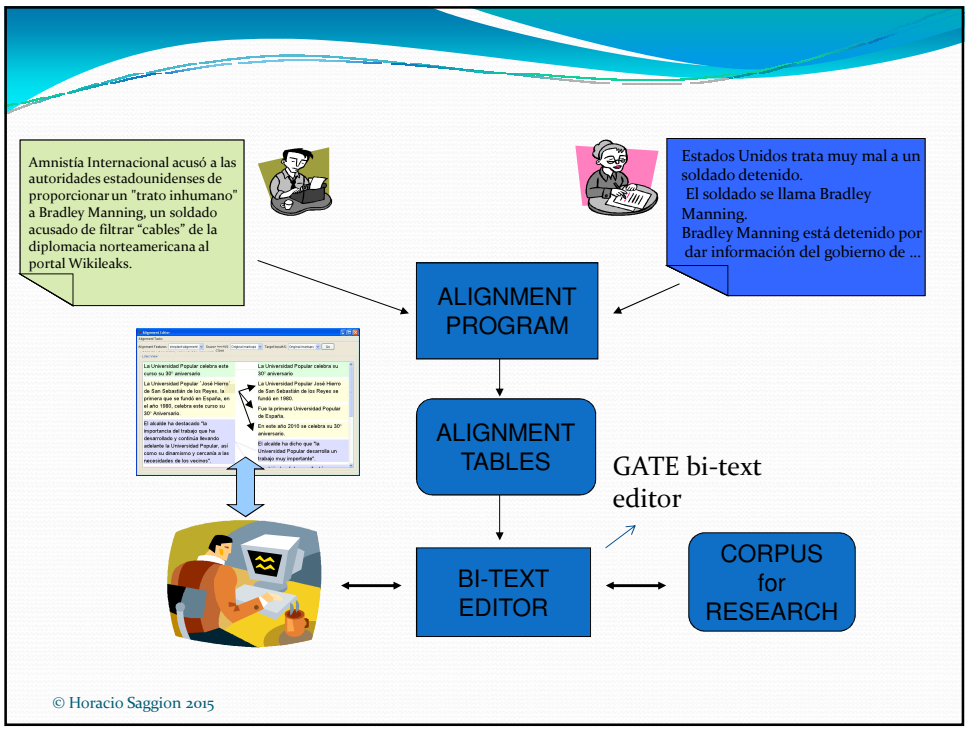
SIMPLIFICATION EXPERT

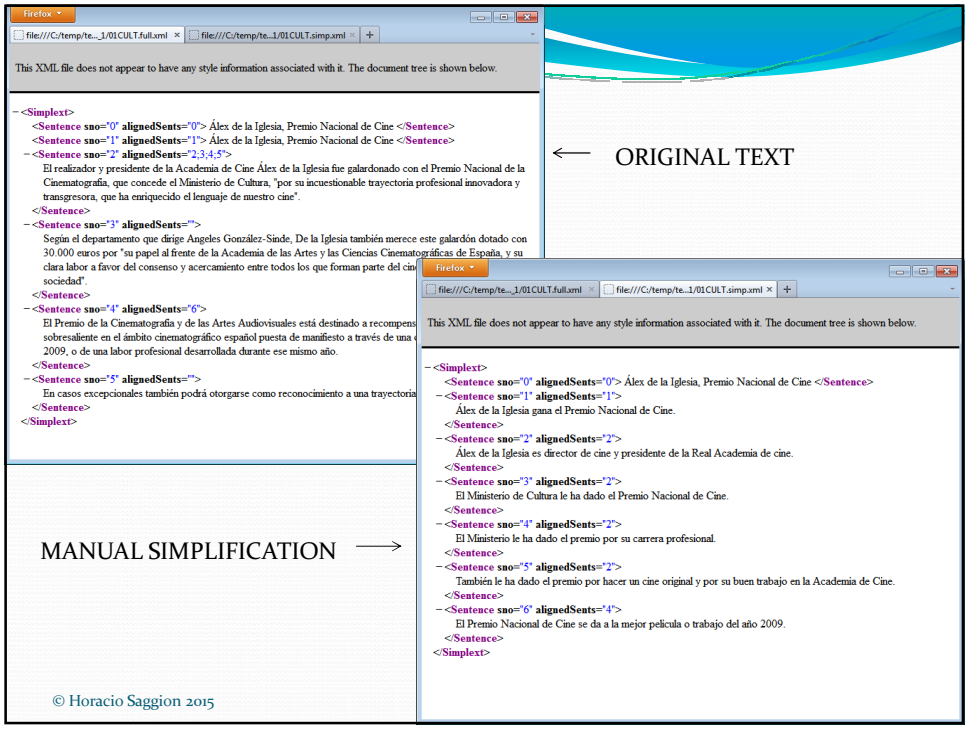
COMPLEX SENTENCE

Amnistía Internacional acusó a las autoridades estadounidenses de proporcionar un "trato inhumano" a Bradley Manning, un soldado acusado de filtrar "cables" de la diplomacia norteamericana al portal Wikileaks.

4 SIMPLER SENTENCES

Estados Unidos trata muy mal a un soldado detenido.
El soldado se llama Bradley Manning.
Bradley Manning está detenido por dar información del gobierno de Estados Unidos a Wikileaks.
Wikileaks es una página web donde se da información sobre asuntos de interés público.


Amnistía Internacional acusó a las autoridades estadounidenses de proporcionar un "trato inhumano" a Bradley Manning, un soldado acusado de filtrar "cables" de la diplomacia norteamericana al portal Wikileaks.
Estados Unidos trata muy mal a un soldado detenido.
El soldado se llama Bradley Manning.
Bradley Manning está detenido por dar información del gobierno de ...
ALIGNMENT PROGRAM
ALIGNMENT TABLES
GATE bi-text editor
BI-TEXT EDITOR
CORPUS for RESEARCH
© Horacio Saggion 2015


ORIGINAL TEXT
MANUAL SIMPLIFICATION
© Horacio Saggion 2015

Simplifications in Spanish

Ex.	Original	Simplified
1	*Abre en Madrid su primera* **sucursal** *el mayor banco de China y del Mundo.* (Opens in Madrid its first branch the biggest bank of China and the World.)	*El banco más importante de China y del mundo abre una* **oficina** *en Madrid.* (The most important bank of China and the world opens an office in Madrid.)
2	*El ICBC ha abierto ya 203* **sucursales** *en un total de 28 países de todo el mundo, también en España desde este lunes.* (The ICBC has opened 203 branches in a total of 28 countries around the world, also in Spain since this Monday.)	*El Banco de China tiene* **oficinas** *en muchos países del mundo. Ahora, también tiene una* oficina *en España.* (The Bank of China has offices in many countries around the world. Now it also has an office in Spain.)
3	*Como muestra de su envergadura, según datos de 2009, el ICBC tenía en nómina a un total de 386.723 empleados, sólo en China, en un total de 16.232 sucursales.* (As a sign of its size and according to data from 2009, the ICBC had a total of 386,723 employees in China only, in 16,232 branches.)	
4	**Arranca** *la liga masculina de Goalball, el único deporte específico para ciegos.* (Starts the men's league of Goalball, the only specific sport for the blind.)	**Comienza** *la liga masculina de Goalball. El Goalball es el único deporte específico para ciegos.* (Begins the men's league of Goalball. Goalball is the only specific sport for the blind.)
5	*La ONU prevé el fin de muertos por malaria para 2015.* (The UN expects the end of dead by malaria for 2015.)	*La ONU cree que ninguna persona morirá por malaria a partir de 2015. La ONU es la Organización de las Naciones Unidas. La malaria es ua enfermedad que se transmite gracias a un mosquito.* (The UN believes that nobody will die of malaria from 2015. The UN is the United Nations Organization. Malaria is a disease transmitted by a mosquito.)

Making texts easy to read

- Lexical simplification
  - Chose more accessible vocabulary / apply recurrent re-writing strategies
- Reduction of structural complexity by splitting sentences
  - Reduce sentence length and use simpler syntax
- Content reduction
  - drop unnecessary material
- Clarification
  - provide explanations/definitions of difficult terms

Syntactic simplification

- Treatment of relative clauses, gerundive and participle constructions, coordination of clauses, verbs, and objects, and quotation inversion
  - "The torrential rains **that** began on October 1 …, causing the worst flooding…" -> *"The torrential rains caused the worst… These rains began on…"*
  - Amnesty International pointed out that people have been arrested before the elections **scheduled** for November 29." -> *"Amnesty International pointed out… The elections are scheduled for November 29."*
  - "For this reason the people **could not live** in their houses **and were given** shelter in camps" -> *"For this reason the people could not live in their houses. These people were given shelter in camps."*

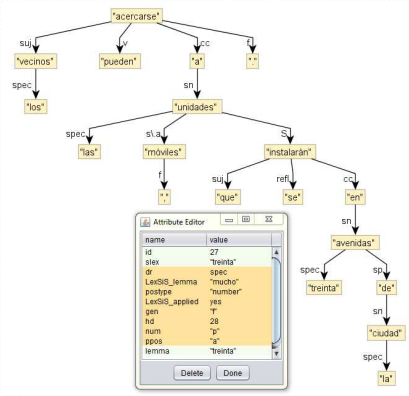
Syntactic simplification approach (Bott, Saggion, Mille; 2012)

- Rule-based approach that transforms dependency-trees / graphs
  - Performs various sentence splitting operations on subordinate and coordinate structures
  - Copies subjects and verbs (e.g. relative pronoun replaced by lexical NP)
  - Orders the various clauses
  - Produces the output from the resulting dependency graphs
- Tools: dependency parser (Bohnet, 2009) and MATE framework (Bohnet et al., 2000)

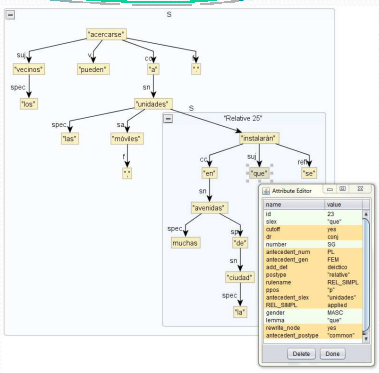
Working on Dependencies

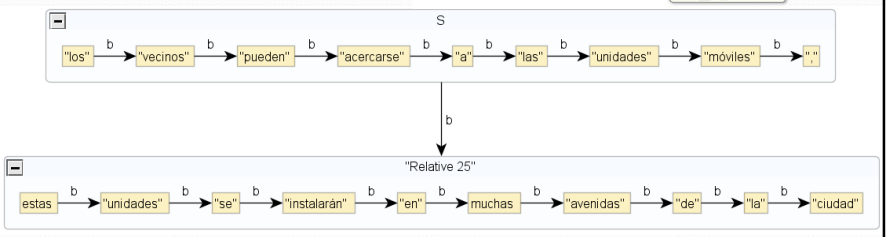
INPUT	OUTPUT	CONSTRAINTS

leftside = [
?Xl {
S-> ?Yl {
?r-> ?Zl
}
}
]
	rightside = [
	rc:?Zr {
	<=> ?Zl
	mark_relative_simple=applied
	}
	]
		conditions = [
		?Yl.ppos="v";
		?Yl.mood="indicative"
		?Zl.slex=que or ?Zl.slex=quién
		or ?Zl.slex=cuál
		or Zl.slex=donde;
		?Zl.id<?Yl.id;
		]



Working on Dependencies





Rule Evaluation

Simplification	Precision	Recall
Relative clauses	40%	66%
Gerundive	63%	20%
Object coordination	42%	58%
VP/Clause coordination	64%	50%

- Evaluation carried out over 886 sentences of the corpus
- Sources of error: parsing, restrictive clauses, idioms, etc.

Other Simplification Mechanisms

- Sentence deletion
  - Productive operation, around 84% of documents have delete operations applied
  - Baseline: delete last sentence achieves 0.73 F-score
  - A classification system for sentence deletion using features such as position, cohesion, presence of numerical expressions, presence of named entities achieves 0.81 F-score
- Definition insertion
  - Also very productive: just over 57% of simplified texts contain definitions not present in the original text
  - 70% of definitions correspond to named entities (people, organizations, places)
  - Problem: what terms should be explained? what is a simple definition? Where to insert the definition in the simplified text?

Other simplification operations (Drndarevic & Saggion, 2012)

- Transformation of nouns/adjectives of nationality
  - "Tunisian authorities" => "authorities of Tunisia"
- Transformation of numerical expressions
  - year clarification "2010" => "the year 2010"
  - deletion "end of may of 2010" => "2010"
  - small numbers in digits "seven books" => "7 books"
  - replace by definition ("2 decades" => "20 years")
- Normalization of "reporting" verbs to the form *decir* ("say")
  - "X explained that Y" => "X said that Y"
- Removal of information in parenthesis

Simplext full evaluation

- Evaluation of syntactic simplification module, rule-based simplification, and whole system (Drndarevic et al., 2013)
  - tested the degree of simplification achieved => "readability" indices
  - tested the grammaticality and meaning preservation => human evaluation
  - readability + understandability => target user evaluation

Readability Assessment

- Degree of simplification
  - a set of readability indices

Formula	Calculation
Average Sentence Length (ASL)	N(w)/N(s)
Index of Complex Sentences (ICS)	N(cs)/N(s)
Sentence Complexity Index (SCI)	(ASL+ICS)/2
Lexical Density Index (LDI)	N(dcw)/N(s)
Index of Low-Frequency Words (ILFW)	(N(lfw)/N(cw))*100
Lexical Complexity (LC)	(LDI+ILFW)/2
Spaulding Density (SD)	N(w)/N(rw)
Spaulding Spanish Readability (SSR)	1.609*ASL+331.8*SD+22.0
Average Word Length (AWL)	N(char)/N(w)
Number of NumExp (NUM)	N(NumExp)
Number of punctuation marks (PUNC)	N(punct)

Readability	Original	Autom. Simp.	Man. Simp.
LC	11.27 ± 0.26	9.35 ± 0.25	4.29 ± 0.27
SSR	179.89 ± 1.50	164.70 ± 1.50	120.90 ± 1.74
ASL	33.08 ± 0.56	25.43 ± 0.53	13.81 ± 0.16
CS	69.15 ± 1.39	55.11 ± 1.82	52.05 ± 2.04
SCI	51.11 ± 0.82	40.27 ± 1.08	32.93 ± 1.05
DEPTH	9.85 ± 0.14	8.50 ± 1.14	5.87 ± 0.06
PUNCT	17.22 ± 0.72	14.07 ± 0.59	3.40 ± 0.33

Readability Assessment

- Differences between text conditions
  - In general manual and automatic simplification "reduce" the value of the readability indexes

Readability	Original vs. Manual	Original vs. Autom. Simp.	Autom. Simp. vs. Man. Simp.
LC	$-62.92\% \pm 1.90\%$	$-17.00\% \pm 1.08\%$	$-54.64\% \pm 2.23\%$
SSR	$-32.74\% \pm 0.89\%$	$-8.39\% \pm 0.46\%$	$-25.90\% \pm 1.02\%$
ASL	$-56.92\% \pm 0.85\%$	$-22.32\% \pm 1.31\%$	$-43.40\% \pm 1.15\%$
CS	$-24.58\% \pm 3.02\%$	$-20.55\% \pm 1.96\%$	$\mathbf{-1.33\% \pm 4.58\%}$
SCI	$-34.43\% \pm 2.31\%$	$-21.16\% \pm 1.63\%$	$-14.52\% \pm 3.15\%$
DEPTH	$-39.46\% \pm 0.77\%$	$-13.12\% \pm 1.15\%$	$-29.42\% \pm 1.03\%$
PUNCT	$-77.28\% \pm 2.37\%$	$-17.37\% \pm 1.41\%$	$-72.28\% \pm 2.79\%$

Simplext full evaluation

- Grammaticality and meaning preservation
  - 25 human evaluators asked to read and assess original and simplified sentences
  - Questionnaires with 38 pairs of original (O) and simplified (S) sentences
  - Pairs O-S contained at least one lexical change and one syntactic change
  - Order was altered random to counterbalance the sequence effect

Non-target User Evaluation

- Grammaticality and meaning preservation
  - Questions:
    1. paragraph A is grammatical
    2. paragraph B is grammatical
    3. paragraphs A & B have the same meaning
  - Answers on Likert scale: 1 completely disagree – 5 completely agree

Score	Simplicity (O)	Simplicity (AS)	Gramm. (O)	Gramm. (AS)	Meaning
5 – Strongly agree	20%	30%	63%	24%	45%
4 – Agree	20%	20%	23%	21%	25%
3 – Neutral	30%	10%	7%	24%	10%
2 – Disagree	20%	20%	5%	18%	11%
1 – Strongly disagree	10%	20%	2%	13%	9%
Mean	3.20	3.20	4.40	3.25	3.86
Median	3	3.5	5	3	4
Mode	3	5	5	3 and 5	5
Positive	40%	50%	86%	45%	70%
Neutral	30%	10%	7%	24%	10%
Negative	30%	40%	7%	31%	20%

Final User Evaluation

- Extrinsic evaluation with Simplext final users
- Performed by the PRODIS foundation
  - 44 subjects with Down Syndrome
  - 3 texts in 3 different versions: original, manually simplified, automatically simplified
    - text 1 = pets
    - text 2 = soccer museum
    - text 3 = Braile Writing System
  - 4 "inferential" questions for each text (only correct/incorrect options were considered)
    - text 1 = What is this news about?
    - text 2 = What museum is this news about?
    - text 3 = What is the Braile Writing System?
  - measurements:
    - readability = reading time
    - understandability = number of correct answers

Final User Evaluation

nice generalizations

one long sentence

three sentences

Condition	Example
Original	El 90% de los españoles prefiere adoptar un perro o gato antes que comprar el animal, en tanto que el 10 por ciento restante optaría por pagar porque prefiere un animal de raza, seg'un un sondeo elaborado por eBay Anuncios sobre la percepción ante las adopciones de mascotas. (90% of Spaniards prefer to adopt a dog or cat before buying the animal, while the remaining 10 percent would opt to pay because they prefer an animal of good breed, according to a survey prepared by eBay Ads on perceptions before pet adoption.
Manual	La mayoría de los españoles prefiere adoptar un perro o un gato a comprarlo. El resto de los españoles prefiere comprar el animal. Así están seguros de que el animal es de buena raza. (Most Spaniards prefer to adopt a dog or cat to buy it. The rest of the Spaniards prefer to buy the animal. So they are sure that the animal is of good breed.
Simplext	El 90% de los españoles prefiere adoptar un perro o gato antes que comprar el animal, en tanto que el 10 por ciento restante optaría por pagar porque prefiere un animal de raza, según un sondeo sobre la percepción ante las adopciones de mascotas. El sondeo está elaborado por eBay Anuncios. (90% of Spaniards prefer to adopt a dog or cat before buying the animal, while the remaining 10 percent would opt to pay because they prefer an animal of good breed, according to a survey on perceptions before pet adoption. The survey is prepared by eBay Ads.

two sentences

only one subordination tranformed

Final User Evaluation

- Differences between the original and simplified conditions with simplified texts obtaining on average more correct answers than the original texts, however
- Differences not statistical significant
- Qualitatively:
  - Participants found very positively the existence of a tool such as Simplext that provides a simplification solution accessible through different technological channels (e.g., computer, smart-phone, tablet).
  - Participants were also able to perceive differences in the texts making them more confident in the reading task.
- Recent research (Fajardo et al., 2013) indicates that "easy-texts" have a role to play in making texts more understandable for people with intellectual disabilities

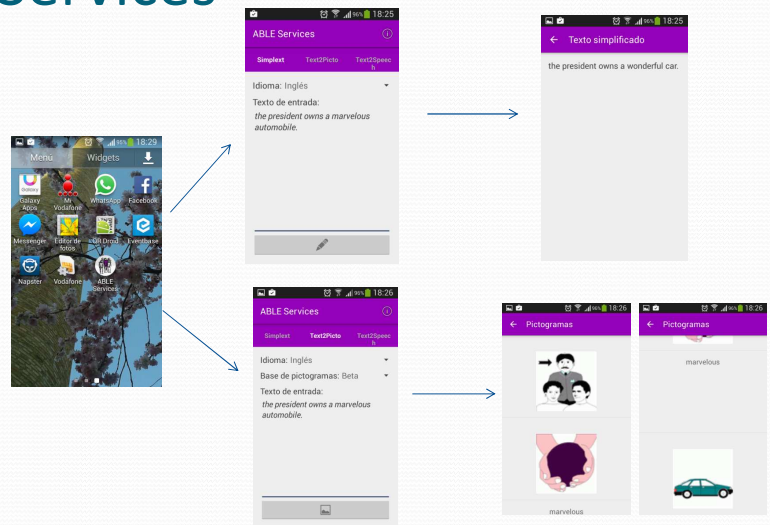
Current systems for Spanish

- Lexical Simplification: CASSA (Baeza-Yates et al. 2015)
  - Developed a context-aware simplification resource for Spanish based on Google Books n-grams
  - *una nueva ESPECIE de planta [TIPO, CLASE, GÉNERO] / a new SPECIES of plant [TYPE, CLASS, GENRE]*
  - *de una ESPECIE nueva de [GRUPO, FAMILIA, CLASE] / of a SPECIES new of [GROUP, FAMILY, CLASS]*
- Full simplification (Stajner et al. 2015)
  - Trained an MT-based simplification system based on a "light" simplification corpus of around 800 sentence pairs improving over Simplext on grammaticality & meaning preservation

Able to Include

- European project funded under the Competitiveness and Innovation Framework Programme of the EU
- To bring for people with intellectual disabilities
  - Text Simplification
  - Text to Pictogram / Pictogram to Text Translation
  - Speech Synthesis
- Text Simplification
  - Spanish: technology from Simplext
  - English: new software based on Simplext

Services



Able to Include Simplification System for English

- Lexical Simplification system based on LexSiS (Ferrés et al., 2015)
  - Java library + resources adapted for English
- Syntactic Simplification system (Marimon et al., 2015)
  - Method based on Simplext
  - Tools are different (mainy Java + GATE)

Syntactic Simplification System

- Linguistically motivated approach
  - Identification of phenomena known to cause reading/understanding problems
- NLP pipeline composed of the following modules:
  - Document Analysis
    - Tokenization and Sentence Splitting
    - Mate Parser (Bohnet, 2010) trained on the CoNLL dataset
    - A set of grammars implemented in the JAPE language
      - identify "complex" syntactic constructions
      - annotate the text with useful information to facilitate re-writing
  - Sentence Generation / re-writing
    - Java programs deal with the transformation of the sentence (copying, reordering, capitalization, etc.)
- Iterative process

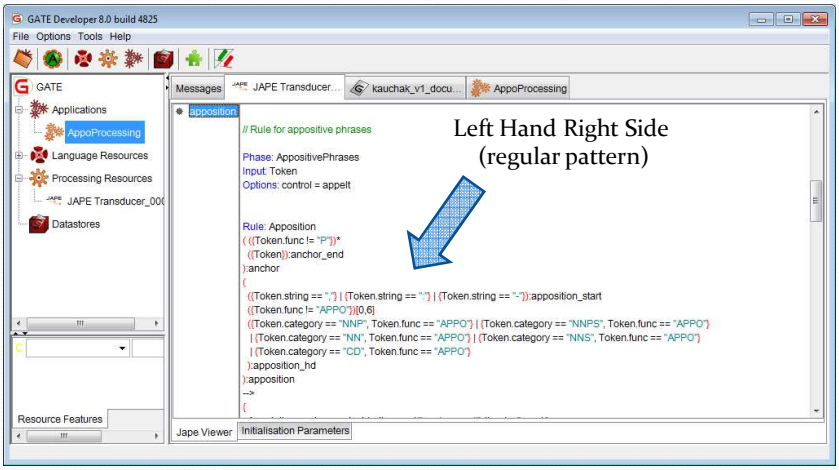
Syntactic Phenomena

- **passive constructions**
  - *a. The release **was accompanied by** a number of TV appearances, including a full hour on On the Record.*
  - *b. A number of TV appearances, including a full hour on On the Record **accompanied** the release.*
- **appositive constructions**
  - *a. The moon is named after **Portia, the heroine of William Shakespeare's play The Merchant of Venice.***
  - *b. The moon is named after Portia. **Portia is the heroine of William Shakespeare's play The Merchant of Venice.***
- **relative clauses**
  - *a. The festival was held in **New Orleans, which was recovering from Hurricane Katrina.***
  - *b. The festival was held in New Orleans. **New Orleans was recovering from Hurricane Katrina.***
- **coordinated constructions**
  - *a. Tracy **killed** 71 people, **caused** $ 837 million in damage and **destroyed** more than 70 percent of Darwin's buildings, including 80 percent of houses.*
  - *b. Tracy **killed** 71 people. Tracy **caused** $ 837 million in damage. **And** Tracy **destroyed** more than 70 percent of Darwin's buildings, including 80 percent of houses.*

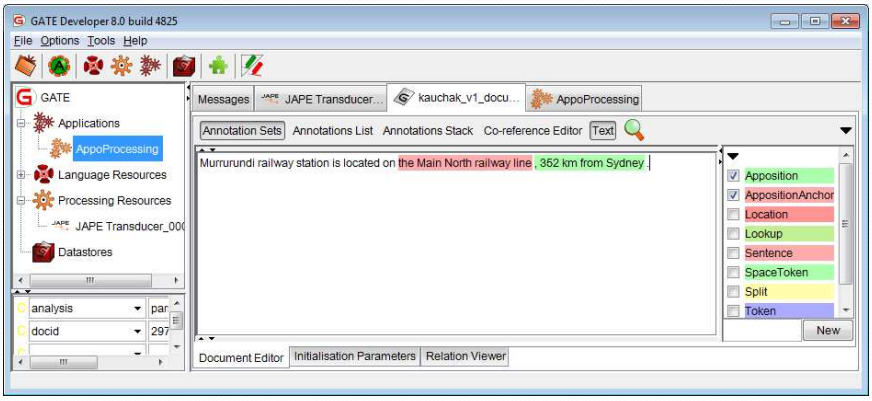
Syntactic Phenomena

- **correlated correlatives**
  - *a. A hypothesis requires more work by the researcher in order to **either** confirm **or** disprove it.*
  - *b. A hypothesis **requires** more work by the researcher in order to confirm it. **Or** a hypothesis **requires** more work by the researcher in order to disprove it.*
- **subordinate clauses**
  - *a. He is perhaps best known for his design for the Natural History Museum in London, **although** he also built a wide variety of other buildings throughout the country.*
  - *b. He also built a wide variety of other buildings thought the country. **But** he is perhaps best known for his design for the Natural History Museum in London.*
- **adverbial clauses**
  - *a. Oxfordshire is a county in the South East England region, **bordering on** Northamptonshire, Buckinghamshire, Berkshire, Wiltshire, Gloucestershire and Warwickshire.*
  - *b. Oxfordshire is a county in the South East England region. Oxfordshire **borders on** Northamptonshire, Buckinghamshire, Berkshire, Wiltshire, Gloucestershire and Warwickshire.*

Identification and Annotation



Identification and Annotation



Murrurundi railway station is located on the Main North railway line, 352 km from Sydney

Murrurundi railway station is located on the Main North railway line. The Main North railway line is 352 km from Sydney.

English Rule-based System

- Appositions: 1 rule
- Relative clauses: 17 rules
- Coordination: 10 rules
- Correlatives: 4 rules
- Subordination: 8 rules
- Adverbial: 12 rules
- Passive: 14 rules

- Rules are applied iteratively until no more simplifications are fired

Evaluation of English Rules

RULES	Right	Wrong	Ignored
Apposition	79%	21%	0%
Relative Clause	79%	14%	7%
Coordination	56%	6%	38%
Subordination	72%	25%	3%
Passive	85%	6%	9%
Total	74%	14%	11%

ERRORS	Parser	Rules
Apposition	19	2
Relative Clause	14	0
Coordination	4	2
Subordination	7	18
Passive	5	1
Total	49	23

- Evaluation Dataset (English Wikipedia Simplification Dataset)
  - 100 sentences for each phenomena
  - right: all required elements correctly identified
  - wrong: at least one element is missing
  - ignored: structure not identified

Lexical Simplification System

- System Architecture
  - Document Analysis
  - Complex Word Detection
  - Word Sense Disambiguation
  - Synonym Ranking
  - Language Realization

Document Analysis

- Objective: Linguistically analyze the input document
- Approach: Make use of available Natural Language Processing Tools
  - GATE (Cunningham et al. 2002) – well known library for Natural Language Processing
  - ANNIE pipeline from the GATE system
    - Tokenization – identify words
    - Sentence Splitting – identify sentences
    - Part-of-Speech (POS) tagging – identity the category of each word
    - Lemmatization – obtain the lemma of each word
    - Named Entity Recognition and Classification – recognize names of people, places, etc.

Complex Word Detection

- Goal: to detect which words might be complex for the target audience
- Approach: rely on available psycholinguistic data
  - Age-of-Acquisition (AoA) norms (Kuperman et al., 2012)
    - Ratings of the age at which words are learned
    - English words: nouns, verbs, and adjectives
    - 30121 rating (51715 inflected words)
  - Kucera-Francis word frequency counts (Kucera & Francis, 1967)
    - 43299 words (all POS categories)
    - Extracted from the Brown Corpus (over 1M words)

Word Sense Disambiguation

- Goal: obtaining the most appropriate sense for a given word in a given context
- Approach: Vector Space Model approach for Lexical Semantics
  - Dictionary of Target Words and Senses: OpenThesaurus (transformed)

bill {
bill, measure – > sense_1
bill, note, banknote - > sense_2
bill, invoice -> sense_3

  - We model each target word in the dictionary (and each sense) as a vector of "context words" extracted from text collections (e.g. Simple Wikipedia)
  - Given a complex word in a sentence we compare its context against the vectors in the dictionary and select the most appropriate list of synonyms

Synonym Ranking

- Goal: obtain the "simplest" and "most appropriate" synonym word for the given context
- Approach: Ranking of synonyms by **lexical simplicity**
- Measures used: word frequency (more frequent means simpler)
- Resources
  - British National Corpus (BNC) frequency list, Google Web 1T Corpus unigram frequencies, Simple English Wikipedia : frequency counts, Normal English Wikipedia: frequency counts, Kucera-Francis norms, Age-of-Acquisition rankings

Language Realization

- Goal: Generate the correct inflected forms of the final selected synonym word substitutes in context
- Approach: Use the SimpleNLG Java API
  - Default SimpleNLG Lexicon used
  - Rules that use lemmas, and Part-of-Speech tags of the simple word and context
  - Complement this module with heuristics to repair contexts

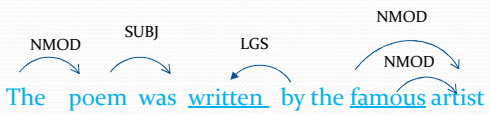
    an *automobile* -> ***a*** *car*

Example

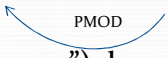
- "The poem was composed by the renowned artist"
  - *compose* and *renowned* detected as complex words because of their low frequency
    - *compose* has various meanings: {write} {compile} {pen, write, indite}....
    - *renowned* has one meaning: {celebrated, famous,....notable, noted}
  - *write* selected as substitute of *compose*
  - *famous* selected as substitute of *renowned*

Example

- "The poem was composed by the renowned artist"

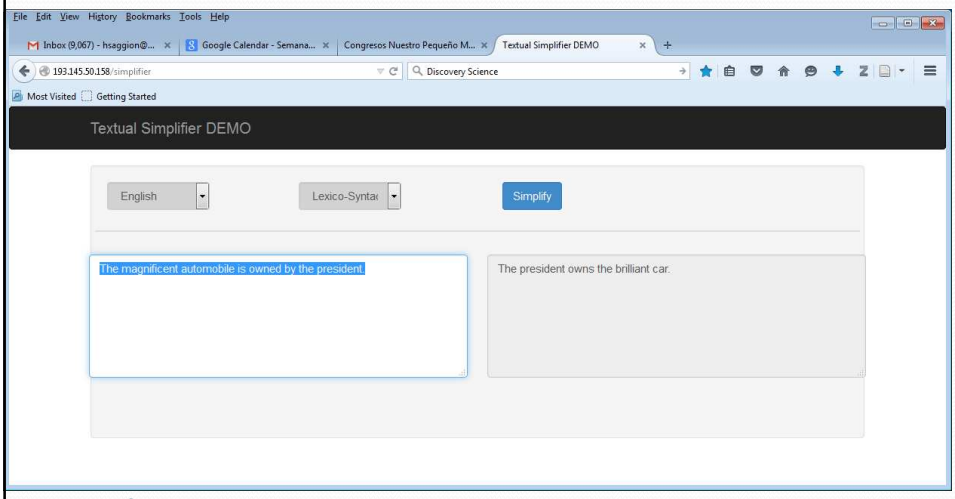


- Passive rule activated



  - passive subject ("the poem"), by-agent ("the famous artist")
  - verb is re-written: *was written* => *wrote*
  - elements re-ordered
- Result: *The famous artist wrote the poem.*

Able to Include Demo



Summary

- Objectives of text simplification
  - transform a text into a version which is easier to read and (hopefully understand) and which has more or less the same content than the original
  - contribute with resources to assist the process of simplification, resources that could contribute to tasks other than the actual simplification
- There is an increasing interest in NLP for the topic, and certainly in IR and Web Accessibility
- In coming years we will see application which will make textual access easier for different groups of people (and for many languages)

Summary

- There are 3 research axes in text simplification
  - Detecting the complexity/simplicity of a text (probably with respect to external factors)
  - Simplification of the lexicon
  - Simplification of the syntactic structure
- There are several techniques and a trend for corpus-based supervised methods, however many languages lack appropriate resources so linguistic intuitions are paramount here
- Text simplification (word, sentence) is not the only way to make text accessible
  - images, tables, pictures, animations
  - disposition of the textual elements
  - speech

Text Simplification Resources

- Corpora
  - Wikipedia + Simple Wikipedia
  - Edit histories
  - Newsela (Xu et al, 2015)
  - Simplext (Saggion et al. 2015) – upon request
  - PorSimples (Aluisio et al. 2008)
- Lexicons / Datasets
  - WordNets / Multilingual Central Repository
  - Various Open Thesaurus (Spanish, English, Catalan, etc.)
  - Kucera-Francis (Kucera & Francis, 1967)
  - Age of Acquisition ()

Text Simplification Resources

- Web sites of simplified news
  - LiteracyWorks: http://www.literacyworks.org/
- Working Systems
  - PorSimples
  - FACILITA
  - PorSimples
  - Simplext
  - Able to Include

References

Al-Ajlan, A., Hend S. Al-Khalifa, and AbdulMalik S. Al-Salman. Towards the development of an automatic readability measurements for arabic language ICDIM, page 506-511. IEEE, (2008)

Aluísio, S., Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick Galani Maziero, Renata Pontin de Mattos Fortes: Towards Brazilian Portuguese automatic text simplification systems. ACM Symposium on Document Engineering 2008: 240-248

Aluísio, S., Specia, L., Gasperin, C. and Scarton, C. (2010) Readability Assessment for Text Simplification. In Proceedings of the Fifth Workshop on Innovative Use of NLP for Building Educational Applications. Los Angeles, USA.

Andrea Petz, Beor Tronback (2008) People with Specific Learning Difficulties: Easy to Read and HCI. ICCHP 2008: 690-692

Barbu, E.; Martín-Valdivia, M.; Alfonso, L. & Lopez, U. (2013), 'Open Book: a tool for helping ASD users' semantic comprehension', *Proceedings of the 2th Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Atlanta ,United States, 14 June, pages 11-19.

Barthe, Kathy; Juaneda, Claire; Lesegneur, Dominique; Lesquet, Jean-Claude; Morin, Claude; Escande, Jean; Vayrette, Annick. (1999) GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French. Technical Communication, Volume 46, Number 2, May 1999 , pp. 220-229(10)

Bautista, S. Raquel Hervás, Pablo Gervás, Richard Power, and Sandra Williams (2011). How to Make Numerical Information Accessible. In Part 1 Proceedings of 13th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT 2011)

Bernhard, Delphine and Louis De Viron, Véronique Moriceau, Xavier Tannier. 2012. Question Generation for French: Collating Parsers and Paraphrasing Questions. Dialogue & Discourse. Vol 3, No 2 (2012).

Björmsson, C.H. (1968). Läsbarhet. Stockholm: Liber.

Bott S, Saggion H, Mille S. (2012). Text Simplification Tools for Spanish. Proceedings of Language Resources and Evaluation Conference, 2012.

Bott S, Saggion H. (2011). An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. ACL Workshop on Monolingual Text-to-Text Generation.

Bott S, Saggion H. (2012). Automatic Simplification of Spanish Text for e-Accessibility. Proceedings of the 13th International Conference on Computers Helping People with Special Needs. :54–56.

Bouayad-Agha, N., Gerard Casamayor, Gabriela Ferraro, and Leo Wanner. (2009). Simplification of patent claim sentences for their paraphrasing and summarization. 2009 FLAIRS Conference.

Brouwers, Laetitia and Delphine Bernhard, Anne-Laure Ligozat and Thomas Francois. **Syntactic Sentence Simplification for French. The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations.**

Burstein, Jill and Sabatini, John and Shore, Jane and Moulder, Brad and Lentini, Jennifer. 2013. A User Study: Technology to Increase Teachers' Linguistic Awareness to Improve Instructional Language Support for English Language Learners. In Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility.

Carroll, J. Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. (1998). Practical Simplification of

Chandrasekar, R. Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. COLING 1996 , pp1041–1044.

Chung, Jin-Woo and Ho-Joon Lee, Jong C. Park. 2013. Enhancing readability of web documents by text augmentation for deaf people. WIMS 2013.

Coster, W. and David Kauchak. 2011. Learning to simplify sentences using wikipedia. Proceedings of Text-To-Text Generation, Portland, Oregon. Association for Computational Linguistics.

Daelemans, W. and Anja Höthker and Erik Tjong Kim Sang. (2004). Automatic Sentence Simplification for Subtitling in Dutch and English. Proceedings of the Language Resources and Evaluation Conference.

De Belder, Jan. (2014) Integer Linear Programming for Natural Language Processing , PhD thesis, Informatics Section, Department of Computer Science, Faculty of Engineering Science.

De Belder, Jan; Deschacht, Koen; Moens, Marie-Francine. (2010) Lexical simplification, 1st international conference on interdisciplinary research on technology, education and communication, Kortrijk, Belgium, 25-27 May 2010

Dell'Orletta, F., Montemagni, S., Venturi, G. (2011). READ-IT: assessing readability of Italian texts with a view to text simplificationProceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies. 73-83

Devlin, S., Tait, J. (1998) The use of a psycholinguistic database in the simplification of text for aphasic readers. In J. Nerbonne, editor, Linguistic Databases, 161–173. CSLI Publications, Stanford, California

Dornescu, I.; Evans, R., & Orasan, C. (2013), A Tagging Approach to Identify Complex Constituents for Text Simplification', *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)*, Hissar, Bulgaria, pp. 221-229.

Drndarevic B, Saggion H. (2012a). Towards Automatic Lexical Simplification in Spanish: An Empirical Study. Workshop Predicting and improving text readability for target reader populations (PITR2012). NAACL 2012.

Drndarevic B, Saggion H. (2012b). Drndarevic B, Saggion H. 2012. Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish. SEPLN Journal . 49

DuBay, W.H. (2004), The Principles of Readability. Impact Information.

Eskenazi, Maxine and Lin, Yibin and Saz, Oscar. 2013. Tools for non-native readers: the case for translation and simplification. In Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility.

Evans, R. J. (2011) Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing*. Vol. 26, Nr. 4, pp. 371-388.

References

Falkenjack, J., Katarina Heimann Mühlenbock, and Arne Jönsson . (2013) Features Indicating Readability in Swedish Text, Proceedings of NODALIDA 2013. Oslo, Norway.
Falkenjack, Johan and Katarina Heimann Mühlenbock, Arne Jönsson
Features Indicating Readability in Swedish Text. Proceedings of the 19th Nordic Conference of Computational Linguistics, NODALIDA 2013
Fellbaum, C. *WordNet*. An Electronic Lexical Database. The MIT Press.
Ferrés, Daniel and Montserrat Marimon, Horacio Saggion. 2015. A Web-Based Text Simplification System for English. In Procesamiento de Lenguaje Natural Journal 2015.
Flesch, R. (1949) The art of readable writing. Harper, new York.
Flor, Michae and Beata Beigman Klebanov and Kathleen M. Sheehan . Lexical Tightness and Text Complexity. In Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility.
François T., La lisibilité computationnelle : un renouveau pour la lisibilité du français langue première et seconde ? In International Journal of Applied Linguistics (ITL), vol. 160, 75-99.
François, T. and Fairon, C. (2012) An « AI readability » formula for French as a foreign language. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea.
Graesser, Arthur C. and Danielle S. McNamara , Max M. Louwerse, Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers. May 2004, Volume 36, Issue 2, pp 193-202.
Grefenstette, G. 2008. Producing Intelligent Telegraphic Text Reduction to provide an Audio Scanning Service for the Blind. 1998 AAAI Spring Symposium on Intelligent Text Summarization.
Gunning, R. (1952). The technique of clear writing. McGraw-Hill, New York.
Inui, K. Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In Proceedings of the second international workshop on Paraphrasing - Volume 16 (PARAPHRASE '03), Vol. 16. Association for Computational Linguistics, Stroudsburg, PA, USA, 9-16.
Jonnalagadda, Siddhartha and Graciela Gonzalez. 2009. Sentence Simplification Aids Protein-Protein Interaction Extraction. The 3rd International Symposium on Languages in Biology and Medicine, Jeju Island, South Korea, November 8-10, 2009.
Keskisärkkä, R. (2012). Automatic Text Simplification via Synonym Replacement. Master's thesis, Linköping University.
Kincaid, J.P, Fishburne, R.P., Rogers, R.L., Chisson, B.S. (1986) Derivation of new readability formulas for Navy enlisted personell.
Klerke, Sigrid; Søgaard, Anders. 2012. DSim, a Danish parallel corpus for text simplification. The 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey.
Kučera, H., Francis, W. N. (1967) Computational analysis of present-day American English. Providence, RI: Brown University Press.
Kuperman V, Stadthagen-Gonzalez H, Brysbaert M. 2012. Age-of-acquisition ratings for 30,000 English words. Behavior Research Methods. December 2012, Volume 44, Issue 4, pp 978-990.
Luz Rello, Ricardo Baeza-Yates, Horacio Saggion, Jennifer Pedler (2012). A First Approach to the Creation of a Spanish Corpus of Dyslexic Texts. Proceedings of NLP4ITA. Istambul.
Marimon, Montserrat and Horacio Saggion and Daniel Ferrés. 2015. Porting a methodology for syntactic simplification from English to Spanish. Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software at IJCAI 2015.
Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, Lillian Lee (2010) For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. HLT-NAACL 2010: 365-368
Matausch, K. and B. Pebück, EasyWeb – A Study How People with Specific Learning Difficulties Can Be Supported on Using the Internet, in: Miesenberger, K.; Klaus, J.; Zagler, W.; Karshmer, A., 12th International Conference on Computers Helping People with Special Needs (ICCHP 2010), pp. 641-648.
McCarthy Diana and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval '07). Association for Computational Linguistics, Stroudsburg, PA, USA, 48-53.
Minard, Anne-Lyse and Anne-Laure Ligozat, Brigitte Grau, Simplification de phrases pour l'extraction de relations, TALN 2012, Grenoble.
Mitkov, R. (2012) Natural Language Processing and Language Disabilities. Invited Talk. NLP4ITA LREC Workshop, Istambul. http://www.taln.upf.edu/nlp4ita/
Narayan, Shashi, Claire Gardent. 2014. Hybrid Simplification using Deep Semantics and Machine Translation. The 52nd Annual Meeting of the Association for Computational Linguistics, Jun 2014, Baltimore, United States. pp.435 – 445.
Ogden, C.K. (1930) Basic English: A General Introduction with Rules and Grammar.
Or Biran, Samuel Brody, and Noémie Elhadad. Putting it Simply: a Context-Aware Approach to Lexical Simplification. 2011. ACL, pp. 496-501. Portland, OR.
Orasan, C.; Evans, R. & Dornescu, I. (2013), 'Text Simplification for People with Autistic Spectrum Disorders', in D. Tufis; V. Rus; & C. Forascu (eds.), Towards Multilingual Europe 2020: A Romanian Perspective, Romanian Academy Publishing House, Bucharest, pp. 287-312.
P. Lal and S. Rüger (2002) Extract-based Summarization and Simplification. In Proc. Workshop on Text Summarization Document Understanding Conference (DUC) 2002
Petersen, S.E. and Ostendorf, M. (2007)Text Simplification for Language Learners: A Corpus Analysis. In Proceedings of SLaTE-2007, 69-72.
pps1353–1361, Beijing, China, Aug.

References

Pitler, E. and Nenkova, A. (2008) Revisiting Readabiliy: A Unified Framework for Predicting Text Quality. In Proceedings of EMNLP, 2008.
Rello, L., Ricardo A. Baeza-Yates, Laura Dempere-Marco, Horacio Saggion: Frequent Words Improve Readability and Short Words Improve Understandability for People with Dyslexia. INTERACT (4) 2013: 203-219.
Rello, L., Ricardo A. Baeza-Yates, Stefan Bott, Horacio Saggion: Simplify or help?: text simplification strategies for people with dyslexia. W4A 2013: 15
Rello, L., Susana Bautista, Ricardo A. Baeza-Yates, Pablo Gervás, Raquel Hervás, Horacio Saggion: One Half or 50%? An Eye-Tracking Study of Number Representation Readability. INTERACT (4) 2013: 229-245
Saggion H, Gómez-Martínez E, Etayo E, Anula A, Bourg L. (2011). Text Simplification in Simplext: Making Text More Accessible. Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural.
Saggion, Horacio andSanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. ACM Trans. Access. Comput. 6, 4, Article 14.
Sato, S., Matsuyoshi, S., Kondoh, Y. (2008) Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In Proceedings of LREC 2008.
Sato, Satoshi and Suguru Matsuyoshi and Yohsuke Kondoh .Text Readability and Word Distribution in Japanese. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2008).
Seratan, V. (2012) Acquisition of Syntactic Simplification Rules for French. In Proceedings of LREC 2012. Istanbul, Turkey.
Si, L. and Callan, J. (2001). "A statistical model for scientific readability." In Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM) ACM.
Siddharthan, A. (2002). An architecture for a text simplification system. Proceedings of LREC 2002. pp64-71
Smith, C. & A. Jönsson (2011). Automatic Summarization As Means Of Simplifying Texts, An Evaluation For Swedish. In Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011.
Spaulding, S. 1956. "A Spanish Readability Formula" Modern Language Journal 40(8) 433-441.
Specia, L (2010). Translating from Complex to Simplified Sentences. Computational Processing of the Portuguese Language. Lecture Notes in Computer Science Volume 6001, 2010, pp 30-39
Specia, L., Jauhar, S.K. and Mihalcea, R. (2012) SemEval-2012 Task 1: English Lexical Simplification. First Joint Conference on Lexical and Computational Semantics (*SEM). Montréal, Canada.
Štajner, S. , Evans, R. , Orasan, C. and Mitkov, R. (2012) What Can Readability Really Tell Us About Text Complexity? In Proceedings of NLP4ITA. Istanbul, Turkey.
Štajner, Sanja (2015) New Data-Driven Approaches to Text Simplification. PhD Thesis, University of Wolverhampton, UK.
Susana Bautista, Biljana Drndarevic, Raquel Hervás, Horacio Saggion, Pablo Gervás: Análisis de la Simplificación de Expresiones Numéricas en Español mediante un Estudio Empírico. Linguamática 4(2): 27-41 (2012)
Susana Bautista, Horacio Saggion: Can Numerical Expressions Be Simpler? Implementation and Demostration of a Numerical Simplification System for Spanish. LREC 2014: 956-962.
Tonelli, S., Ke Tran Manh, Emanuele Pianta (2012). *Making Readability Indices Readable*. In Proceedings of the NAACL Workshop on Predicting and Improving Text Readability for Target Reader Population (PITR2012), Montréal, Canada.
Vari-Cartier, P.: (1981): Development and Validation of a New Instrument to Asses the Readability of Spanish Prose, Modern Language Journal, 65, 141-148.
Williams, S. and Ehud Reiter [2008]. Generating basic skills reports for low-skilled readers. Journal of Natural Language Engineering, Vol. 14, Issue 4, pp. 495-525.
Woodsend, K. and Lapata, M. (2011) Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK.
Wubben, S., van der Bosch, A., Krahmer, E. (2012). Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea.
Xu, Wei and Chris Callison-Burch and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. In TACL 2015.
Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee.L. (2011) In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
Zhu, Z., Delphine Bernhard, and Iryna Gurevych. (2010). A monolingual tree-based translation model for sentence simplification. Proceedings of The 23rd International Conference on Computational Linguistics, pps1353–1361, Beijing, China, Aug.