

Data-Driven Text Simplification

Sanja Štajner and Horacio Saggion

Sanja@informatik.uni-mannheim.de

<http://web.informatik.uni-mannheim.de/sstajner/index.html>

University of Mannheim, Germany

Horacio.saggion@upf.edu

<https://www.upf.edu/web/horacio-saggion>

University Pompeu Fabra, Spain

#TextSimplification2018



Presenters

- Sanja Stajner
 - <http://web.informatik.uni-mannheim.de/sstajner>
 - <https://www.linkedin.com/in/sanja-stajner-a6904738>
- Data and Web Science Group
 - <https://dws.informatik.uni-mannheim.de/en/home/>
- University of Mannheim, Germany
- Horacio Saggion
 - <http://www.dtic.upf.edu/~hsaggion>
 - <https://www.linkedin.com/pub/horacio-saggion/16/9b9/174>
 - https://twitter.com/h_saggion
- Large Scale Text Understanding Systems Lab / TALN group
 - <http://www.taln.upf.edu>
- Department of Information & Communication Technologies
- Universitat Pompeu Fabra, Barcelona, Spain

Tutorial antecedents

- Previous tutorials on the topic given at:
 - IJCNLP 2013 and RANLP 2015 (H. Saggion)
 - RANLP 2017 (S. Štajner)
- Automatic Text Simplification. H. Saggion. 2017. Morgan & Claypool. Synthesis Lectures on Human Language Technologies Series.

<https://www.morganclaypool.com/doi/abs/10.2200/S00700ED1V01Y201602 HLT032>

Outline

- Motivation for ATS
- Automatic text simplification
- TS projects
- TS resources
- Neural text simplification

PART 1

Motivation for Text Simplification

Text Simplification (TS)

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
- TS is a complex task which encompasses a number of operations applied at different linguistic levels:
 - Lexical
 - Syntactic
 - Discourse
- 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 of providing 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 a soldier in prison very badly.

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)

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; Keskiä, 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)

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)
 - Improve results in MT (Štajner and Popović, 2016)

Where to find simple texts?

Opera is a drama set to music. An opera is a play in which everything is sung instead of spoken.

Operas are usually performed in opera houses.

The screenshot shows the browser window with the URL <http://simple.wikipedia.org/wiki/Opera>. The page title is "Opera" and it is identified as being from Wikipedia. The main text of the article reads: "Opera is a drama set to music. An opera is like a play in which everything is sung instead of spoken. Operas are usually performed in opera houses. The singers who sing and act out the story are on the stage, and the orchestra is in front of the stage but lower down, in the orchestra pit, so that the audience can see the stage." There is also a photograph of the Mariinsky Theatre with the caption "Mariinsky Theatre is a world-famous". A table of contents is visible on the left side of the article text.

Where to find simple texts?


Opera is an art form in which singers and musicians perform a dramatic work combining text (called a libretto) and musical score.

Opera

From Wikipedia, the free encyclopedia

This article is about the art form. For other uses, see [Opera \(disambiguation\)](#).

Opera is an art form in which **singers** and **musicians** perform a dramatic work combining text (called a **libretto**) and **musical score**.^[1] Opera is part of the Western **classical music** tradition.^[2] Opera incorporates many of the elements of spoken theatre, such as **acting**, **scenery**, and **costumes** and sometimes includes dance. The performance is typically given in an **opera house**, accompanied by an **orchestra** or smaller **musical ensemble**.



The performance is typically given in an opera house, accompanied by an orchestra or smaller musical ensemble.

Where to find simple texts?

Firefox

Inclusion Europe | Use of new technologies to improve accessibility in polling stations

e-Include

The e-Journal of Inclusion Europe

Home • News • Use of new technologies to improve accessibility in polling stations

WHO'S ONLINE
We have 7 guests online

MAIN MENU

- Home
- News
- Articles
- Events
- Legal Network News
- Projects
- Social Media Links
- Contact Us
- Search
- PROGRESS Support

Inclusion Europe
Supported by
The European Commission

People with disabilities have the right to vote.
Sometimes they find it difficult to vote
because polling stations are not accessible for them.
There is a device that helps people with disabilities
to fill in the ballot paper.

The Convention on the Rights of Persons with Disabilities demands the right to vote independently and secretly for people with disabilities, but they still find barriers when they want to cast their vote. Polling stations should be made more accessible and provide assistance and easy to understand information for these people.

technology devices such as voting machines might be useful to improve accessibility in polling places and this practice has been recommended by Council of Europe and other international organizations.

A device called **TopVoter** has been developed in Slovenia to assist disabled voters to fill out the paper ballot. Once the process is finished, the submitted ballots are printed in closed envelopes and the entered data is not lost in the case of an interruption of electricity.

The device is designed to help people with physical disabilities, elderly, illiterate persons, and visually impaired people and has been successfully used in different countries. It consists of a color touch screen, an audio device and keyboard, a tactile, musicalized, language and a printer. Transport

PRINT EMAIL

Where to find simple texts?

Firefox

Noticias Fácil.es, Lady Gaga, disco de oro

Noticias Fácil.es

Estás en Inicio Noticias Ocio y cultura Lady Gaga, disco de oro

Menú navegación

- Inicio
- ¿Qué es Lectura Fácil?
- Noticias
 - Buscador de noticias
 - Nacional
 - Internacional
 - Ocio y cultura
 - Deportes
 - El Tiempo
 - Economía
 - Otros
 - Tus noticias

Noticias - Ocio y cultura

Lady Gaga, disco de oro

Fecha: 15/06/2011 Fuente: discapnet Clasificación: fácil

La cantante Lady Gaga ya tiene un Disco de Oro.

El Disco de Oro se lo dan a aquellos cantantes que venden muchos discos. En total, Lady Gaga ha vendido más de 30.000 copias de su disco en España.

A la gente le gusta mucho la música de Lady Gaga. También les gustan sus vídeos musicales. Puedes verlos en Youtube (página en Internet con vídeos).



EASY NEWS

Where to find simple texts?

8 Pages News Paper (Swedish)

Firefox

8 SIDOR - Senaste numret av 8 SIDOR i bilder +

- Nyheter i bild
- Bildspel
- Om 8 SIDOR
- Ladda ned inläsningar av papperstidningen
- Kontakta oss
- Lärare
- Bibliotek
- Läsombud

Centrum för lättläst

8 SIDOR är en del av Centrum för lättläst

Centrum för lättläst

Anjas blogg

Anja Kretschmann bloggar om att vara ry i Sverige

Nyheter i bild

Lady Gaga

8 SIDOR

NUMMER 24 - 5 JUNI 2011

Lättlästa nyheter alla vardagar: www.8sidor.se

På flykt undan våldet

En kvinna flyttar in i ett av de många vita tält som sätts upp i Åkersberga för att skydda flyktingar från våld.

Dödande skott i Åkersberga

Ten svenskar sårade och dödade i Åkersberga, utanför Stockholm, i fredags kväll. Påståttligen blev en man på väg till jobbet skadad och två andra skadade. En av dem är allvarligt skadad och väntas gå utifrån.

Mycket mygg i sommar

Välkommen till sommaren! Det är dags för den årliga myggen. Det betyder att det är dags för den årliga myggen. Det betyder att det är dags för den årliga myggen.

Kroatien på väg mot EU

Kroatien ska bli medlem i Europeiska unionen. Detta är ett stort steg för Kroatien. Detta är ett stort steg för Kroatien.

Politiker oroas av gån

Gån är ett stort problem för Sverige. Detta är ett stort problem för Sverige. Detta är ett stort problem för Sverige.

Senast uppdaterat på

Klicka på bilden så kan du bläddra i nummer 24 av 8 SIDOR.

Prenumerera på 8 SIDOR

Du får åtta sidor med lättlästa nyheter i brevlådan varje vecka och du får tillgång till prenumerantwebben

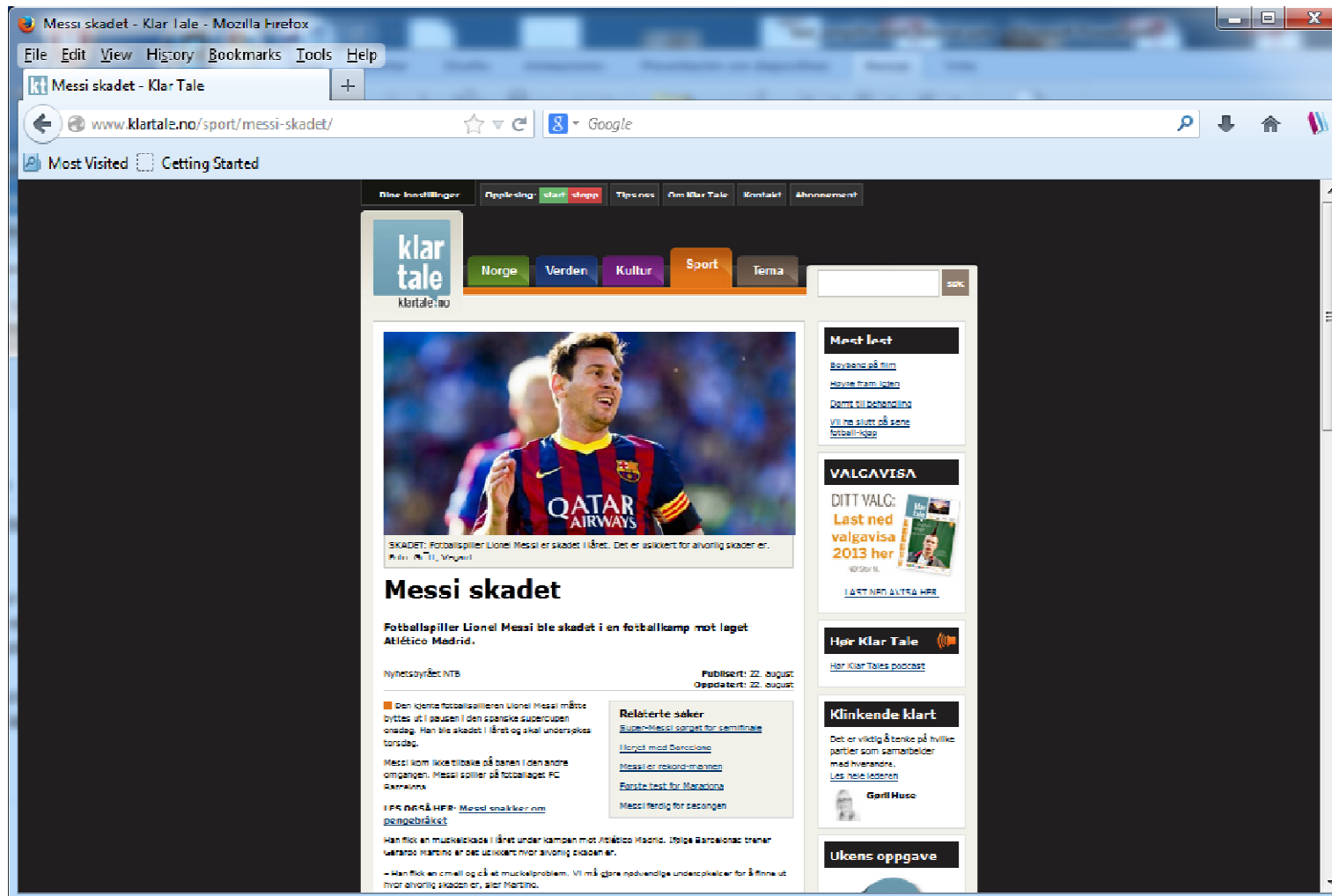
[Läs mer och beställ en prenumeration!](#)

Prenumerantwebben

Du som prenumererar på 8 SIDOR har tillgång till vår Prenumerantwebb. Där finns fakta om alla världens länder, Sverigefakta, aktuella nyhetsbilder och mycket mer. Lösenordet står längst ned på sista sidan i 8 SIDOR. Du skriver in lösenordet högst upp på den här sidan

Where to find simple texts?

Klartale News Paper
(Norwegian)



Where to find simple texts?

L'Essentiel News Paper
(French)



Essentiel - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Essentiel

www.journal-essentiel.be

Most Visited Getting Started

L'ESSENTIEL
L'information simple comme bonjour

Contact Flux RSS Se connecter

Accueil Qui sommes-nous? Articles Cahiers Forums Infographies E-learning Le Cahier des Cahiers

Cahier
Philippe, le 7^e roi

Depuis le 21 juillet, Philippe est roi de Belgique. Son père Albert II lui a laissé la couronne. Depuis plusieurs mois, on disait qu'Albert II se trouvait trop âgé et trop fatigué pour continuer à être roi. Finalement, le 3 juillet, il a annoncé la date de son abdication : le 21 juillet, jour de fête nationale. La Belgique est indépendante depuis 1830. Philippe est son septième roi. Retour sur cette histoire royale et sur l'évolution de l'état Belgique.


Voir le cahier

A la loupe

Sondage
Un couple d'homosexuels a le droit de se marier...

Where to find simple texts?

Dueparole News Paper
(Italian)



2p - Mozilla Firefox

www.dueparole.it/articolo_nav6.asp?anno=2006&numeri


dueparole

mensile di facile lettura

Pagina iniziale anno V - numero 1 - maggio 2006 Sommario

Prima pagina | Spettacoli | Vita in casa | **Notizie dall'Italia** | Notizie dall'Estero | Sport | Cultura

Il Parlamento italiano



Franco Marini e Fausto Bertinotti

- ♦ Il 10 e l'11 aprile 2006, in Italia ci sono state le elezioni politiche, cioè le elezioni per il Senato e per la Camera dei deputati. Senato e Camera dei deputati formano il Parlamento italiano.
- ♦ I cittadini italiani maggiorenni, cioè che hanno più di 18 anni, hanno votato per eleggere i deputati della Camera. I cittadini italiani con più di 25 anni hanno votato per eleggere anche i senatori.
- ♦ I deputati e i senatori italiani hanno il compito di preparare, di discutere e di approvare le leggi.

Where to find simple texts?

LiteracyWorks Web site
(English)

The screenshot shows a Mozilla Firefox browser window displaying the LiteracyWorks website. The address bar shows the URL www.literacyworks.org/learningresources/1_superfoods_. The page features two columns of content for a news story titled "Superfoods: Protect Your Body by Eating Right".

Left Column (Complete Story):

- Title: **Superfoods: Protect Your Body by Eating Right**
- Source: From a news story by San Francisco CBS 6 Dr. Kim Mulvihill
- Date: October 2005
- Options: Complete Story, Video, Audio
- Text: "In North Beach in San Francisco, where some pretty super food gets served every night."
- Quote: "Absolutely very super food!" "I really

Right Column (Abridged Story):

- Title: **Superfoods: Protect Your Body by Eating Right**
- Source: From a news story by San Francisco CBS 6 Dr. Kim Mulvihill
- Date: October 2005
- Options: Abridged Story, Video, Audio
- Text: Research shows that some foods, including tomatoes, onions, garlic, and olive oil, are "superfoods" because they contain chemicals that protect your body against chronic diseases like cancer, obesity, and heart disease.
- Text: Natalie Ledesma is a dietician at the University of California, San Francisco. She

At the bottom of the page, there is a footer with copyright information and a search bar for Answers.com.

Simplification tasks

- Lexical simplification
 - Replace complicated words and expressions by easier to read/understand substitutes (e.g. synonyms)
 - Explain complicated words expressions by providing definitions/explanations
- Syntactic simplification
 - Transform long and complex sentences into syntactic equivalents which could be easier to read/understand
- Could be addressed independently or jointly

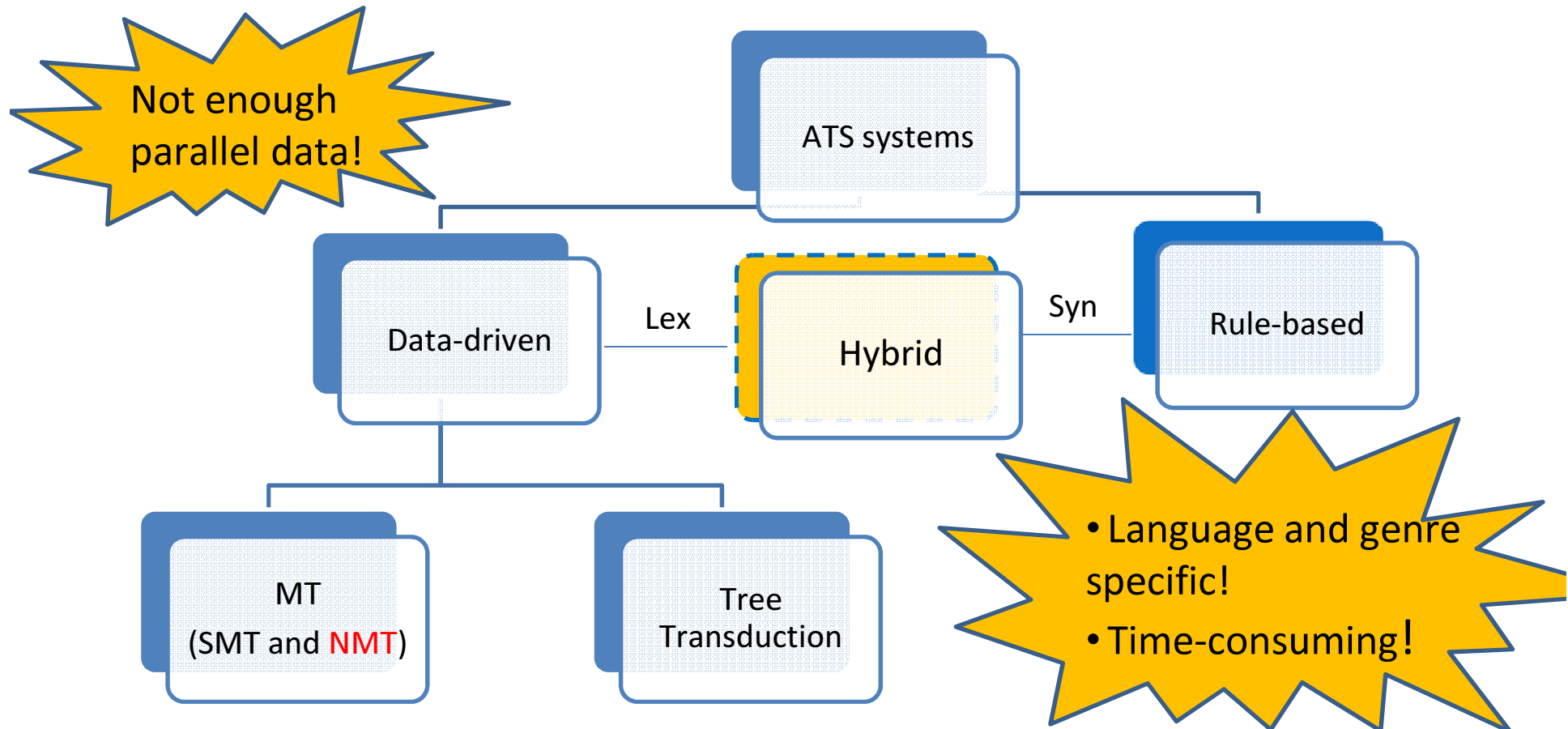
Simplification tasks

- *The play was **magnificent**. => The play was **brilliant**.*
- *The boy had **tuberculosis**. => The boy had tuberculosis, **a disease of the lungs**.*
- *The festival was held in New Orleans, **which was recovering from the hurricane**. => The festival was held in New Orleans. **New Orleans was recovering from the hurricane**.*
- *The city was destroyed **by the hurricane**. => The hurricane destroyed the city.*

PART 2

Automatic text simplification

Approaches to ATS



Stages of ATS

- Stage 1: Detection of necessary transformations
- Stage 2: Building ATS systems
- Stage 3: Evaluation of ATS systems

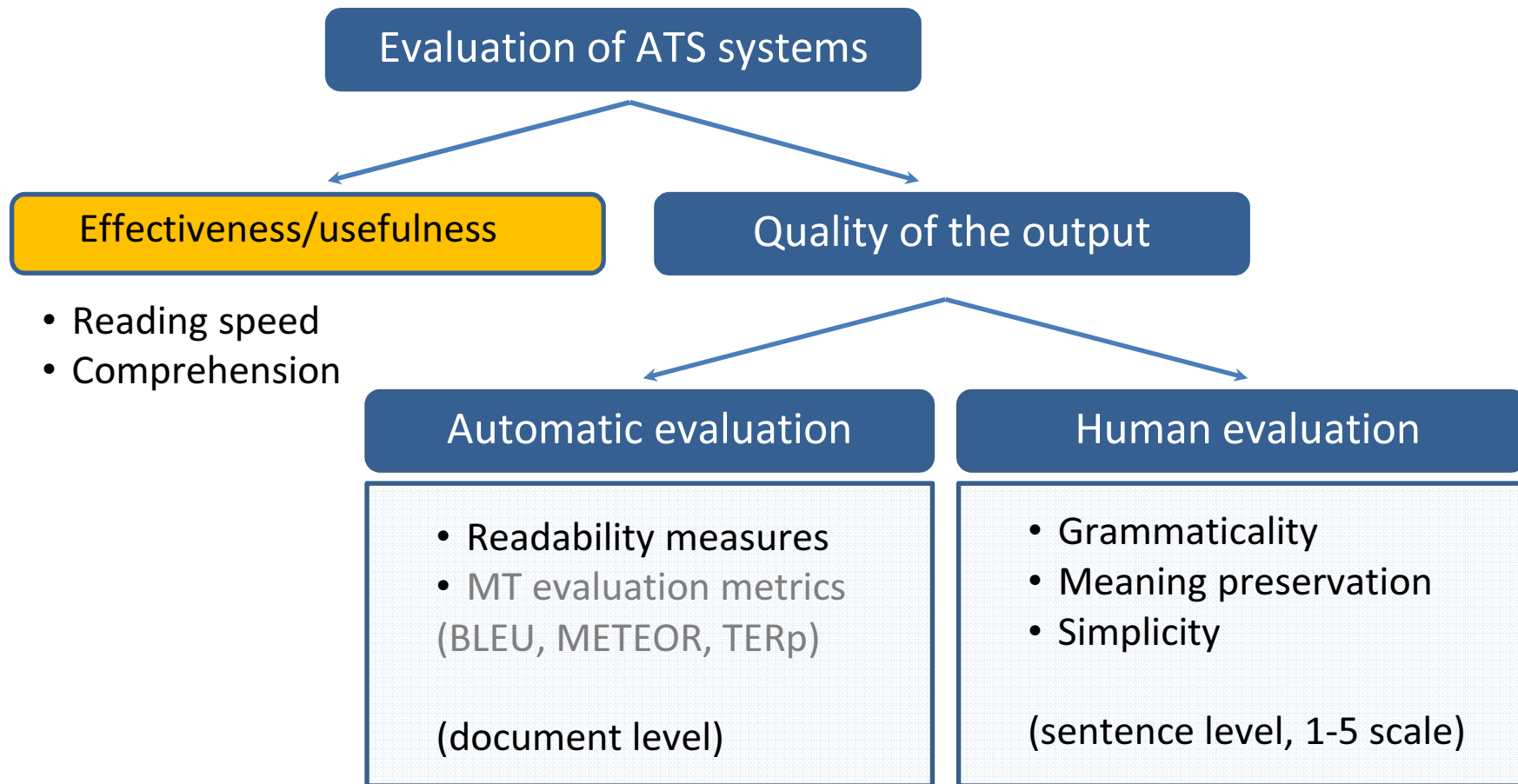
Detection of Necessary Transformations

- For humans:
 - Psycholinguistic theories (rule-based systems)
 - Learning from parallel data (data-driven systems)
 - Eye-tracking (Štajner et al., 2017, BEA)
 - Crowdsourcing (Yiman et al., 2017, RANLP)
 - **Treat everything as potentially complex (Glavaš and Štajner, 2015, ACL)**
- For NLP applications:
 - Ideally from systems' mistakes
 - **In practice: this step is skipped**

Building ATS Systems

- Modular Systems:
 - Lexical simplifier:
 - Modular (CWI module, candidate generation, ranking, substitution)
 - All-in-one
 - Syntactic simplifier
 - Adding information (e.g. definitions)
 - Elimination (content reduction)
- End-to-end systems:
 - MT-based systems (including the NMT-based system)
 - Tree transduction systems

Evaluation of ATS Systems



Quality of the Output (Human Evaluation)

Sentence	G	M	S
Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was installed on the throne.	5	/	4
Madrid was occupied by French his soldiers during the Napoleonic Wars, and Napoleon's brother Joseph was installed on the throne.	4	4	4
Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was put on the throne.	5	5	5
Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was -RRB- installed on them on the throne.	3	3	3

QATS Shared Task (LREC 2016)

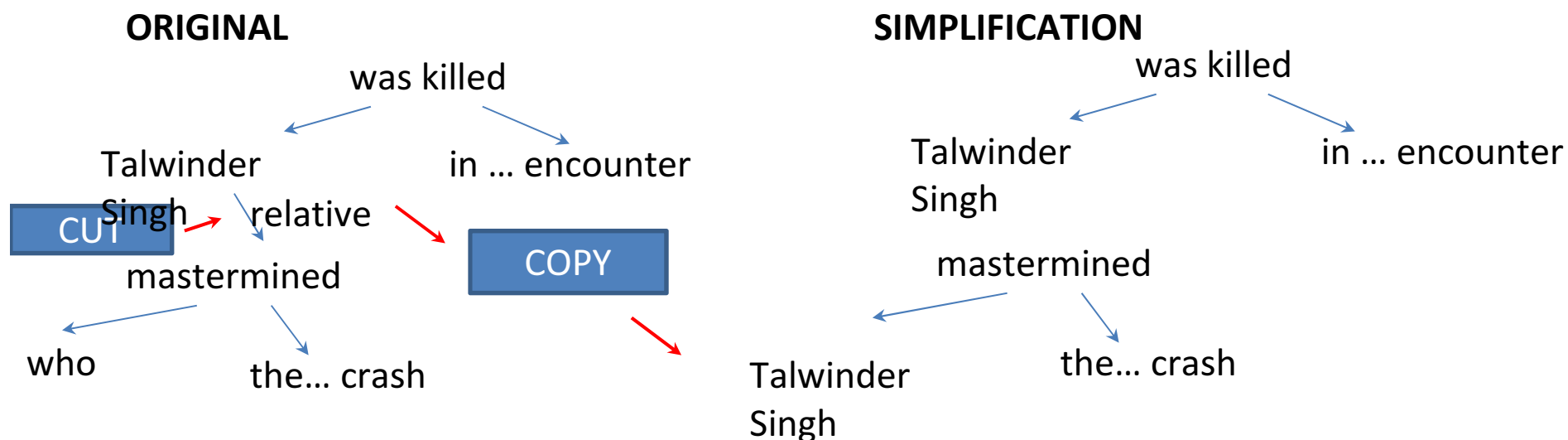
- 20 classification systems
- 12 raw metrics
- The best systems (Štajner, Popovic and Béchara, 2016) combine MT evaluation metrics (BLEU, METEOR, ...) and 17 baseline MT quality estimation features
- The highest weighted F-measure:
 - Grammaticality: 71.84 (majority class – 65.89)
 - Meaning preservation: 68.07 (majority class – 42.51)
 - Simplicity: 56.42 (majority class – 39.68)

First steps: manual rules

- Rules over syntactic representations (Chandrasekar et al. 1996)
 - Superficial analysis (ckunking) to identify noun and verb groups
 - Rules: $W X:NP, RELPRO Y, Z. \Rightarrow W X:NP Z. X:NP Y.$ (manually developed)
 - Xi Jinping, who is the current Paramount Leader of the People's Republic of China, was visiting the USA.
 - $W = \emptyset$
 - $X =$ Xi Jinping
 - $RELPRO =$ who
 - $Y =$ is the current Paramount Leader of the People's Republic of China
 - $Z =$ was visiting the USA
 - → Xi Jinping was visiting the USA. Xi Jinping 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.



Lexical simplification

- Lexical simplification is 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
- What is needed?
 - Procedure to identify which words should be “simplified”
 - Procedure to identify appropriate synonyms of the difficult words
 - Procedure to choose the simplest and more appropriate in the context
 - Procedure to adjust context to comply with the changes
- First ever lexical simplifier was implemented in the PSET project
 - WordNet was used as a source of synonyms
 - No word sense disambiguation carried out
 - Kucera-Francis frequency list used to chose simpler replacement

Lexical simplification

(De Belder & al. 2010) lexicon & language model combined

- 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.

Lexical Simplification

Extracting lexical simplifications using Wikipedia & Simple Wikipedia (Yatskar et al. 2010) through edit history / versions

- Hypothesis: changes in Simple Wikipedia correspond to simplifications the author is making....(not always!)
- Objective: extract rules such as $\underline{A} \rightarrow \underline{a}$, where \underline{A} and \underline{a} are synonyms and \underline{a} is easier than \underline{A}
- A mechanism is needed to model when the change of one word by another is due to a simplification operation
- Language model:
 - 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
 - The model outputs the most probable replacement for “A”
- Point-wise Information Model:
 - Search for replacements corresponding to “simplify” (editor explicitly saying so!) “A” and “a” is stronger using PMI (point-wise mutual information)

Lexical simplification

- Two baseline methods proposed:
 - Frequency model: use the most frequent substitution
 - Random model: chose a random valid substitution
- Compare with a list created automatically
 - Human list of replacements > Language Model > PMI > FREQ >= RANDOM

Lexical simplification

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

Lexical simplification

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
- $\text{complexity}(\text{“canine”}) = 9620/62 = 155$
- $\text{complexity}(\text{“dog”}) = 171000/1360 = 125$

Lexical simplification

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

Lexical simplification

- 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”

Dealing with numbers

- Simplification of numerical expressions in text (Bautista et al. 2012)
- (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 for Spanish

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...)

Syntactic Simplification

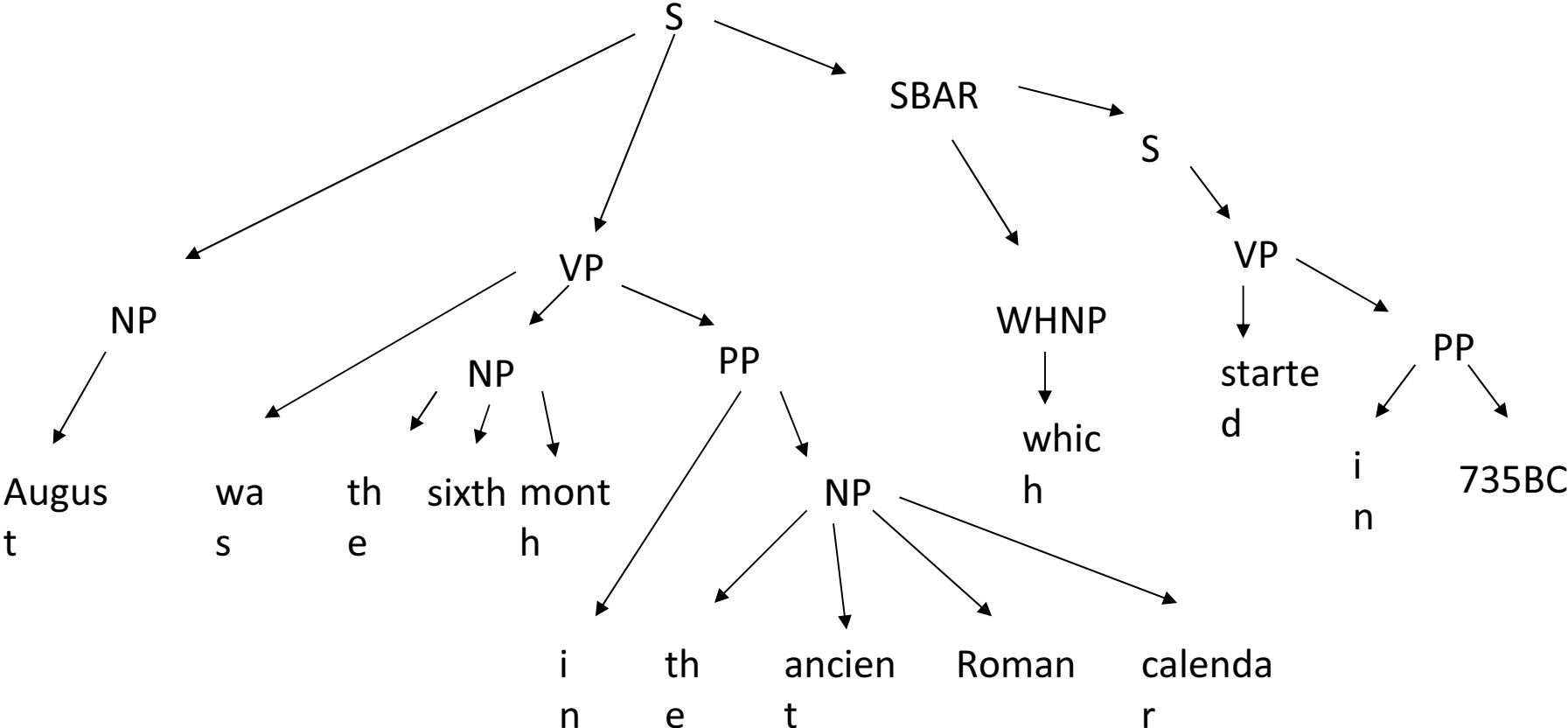
- Siddharthan (2006) was concerned with generation issues during text simplification
 - sentence order, word choice, generation of referring expressions
 - *(1) Mr. Anthony, who runs an employment agency, decries program trading, but he isn't sure it should be strictly regulated.*
 - *? (2a) Mr. Anthony 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 (Siddharthan & Angrosh 2014)

Learning simplification from parsing trees

- Based on a corpus of comparable documents $\langle C, S \rangle$ 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 (PWKP dataset)
- 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”)

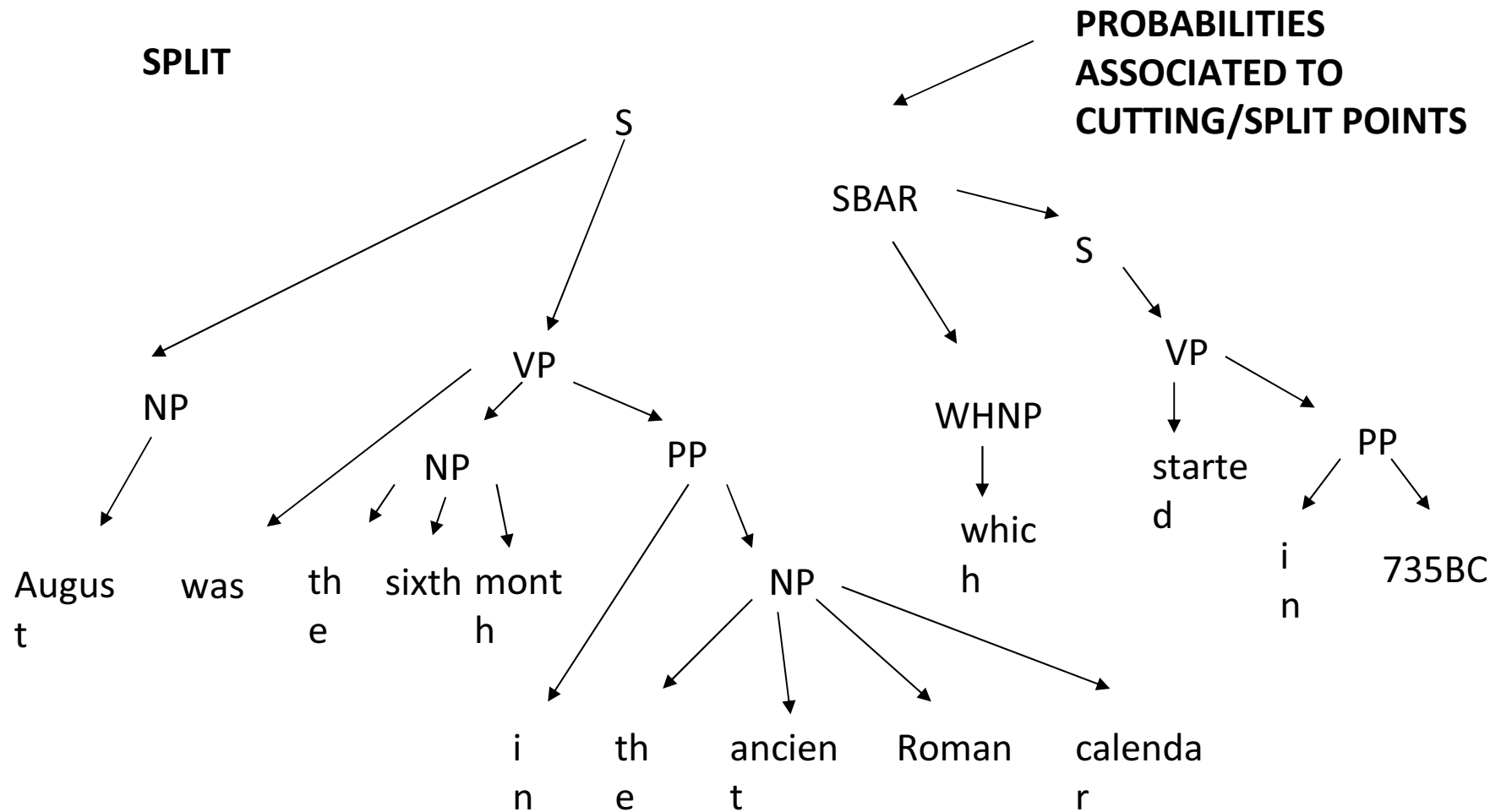
Learning simplification from parsing trees

PHRASE STRUCTURE OF COMPLEXT SENTENCE

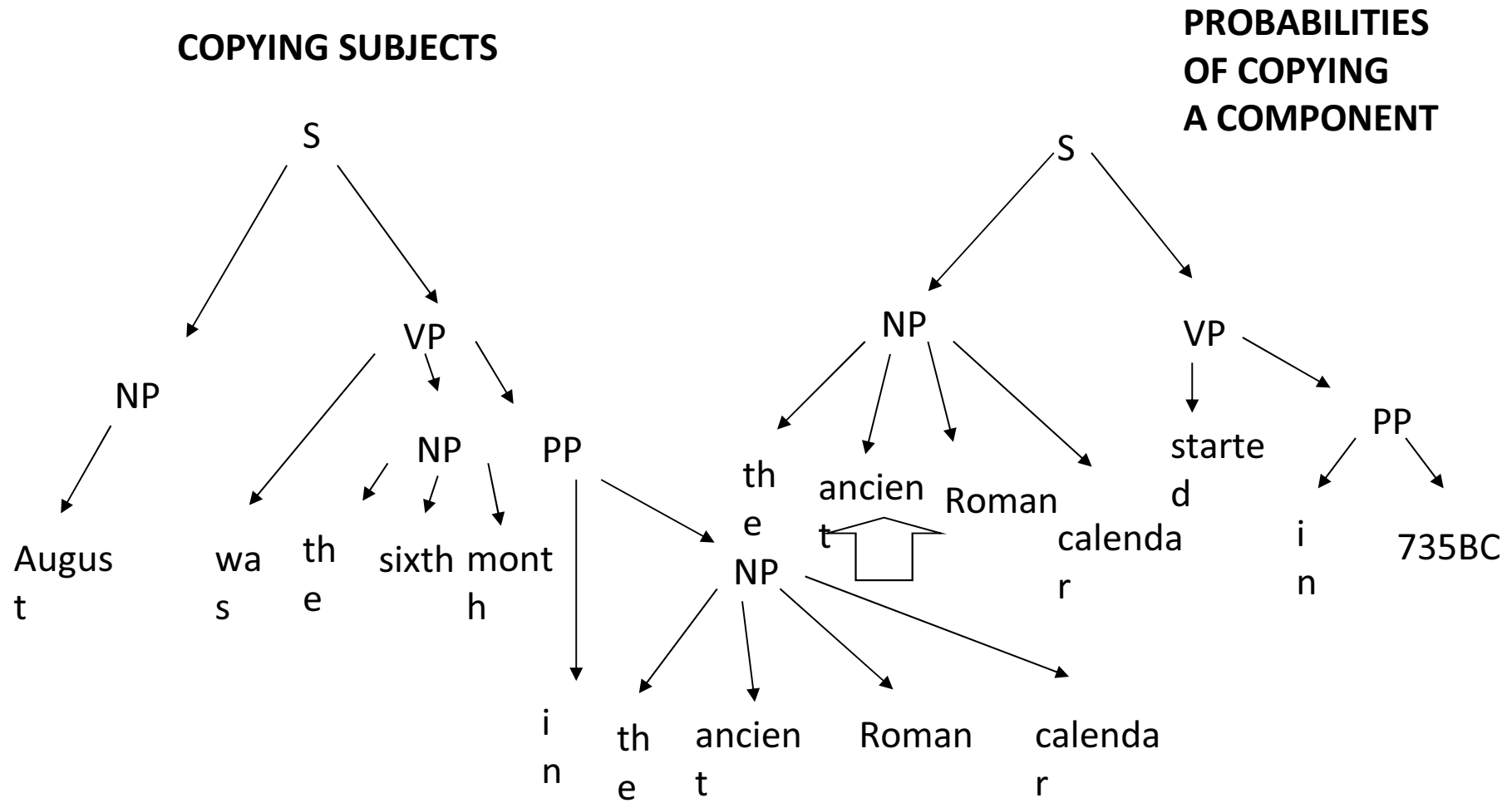


August was the sixth month in the ancient Roman calendar which started in 735BC.

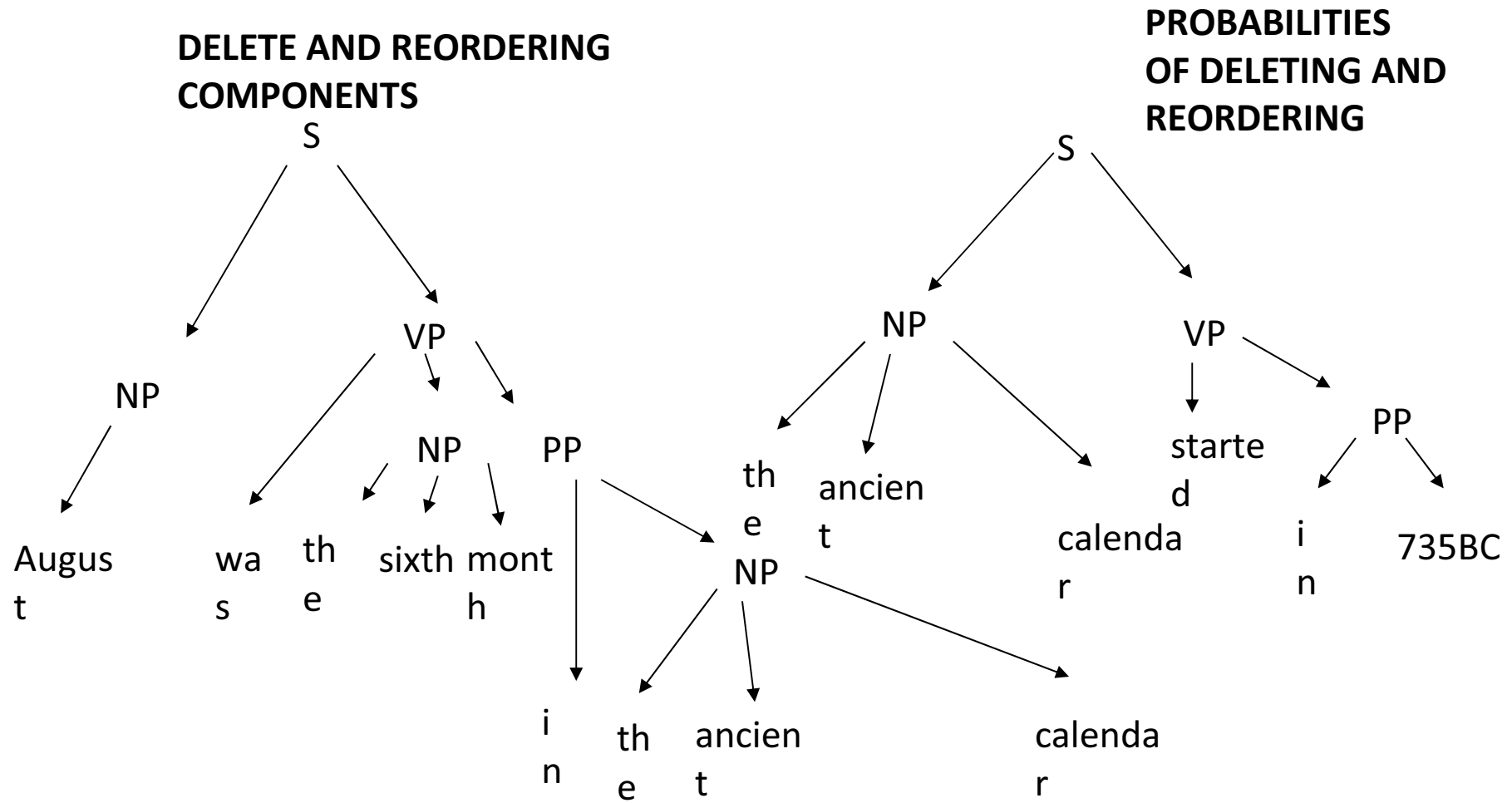
Learning simplification from parsing trees



Learning simplification from parsing trees

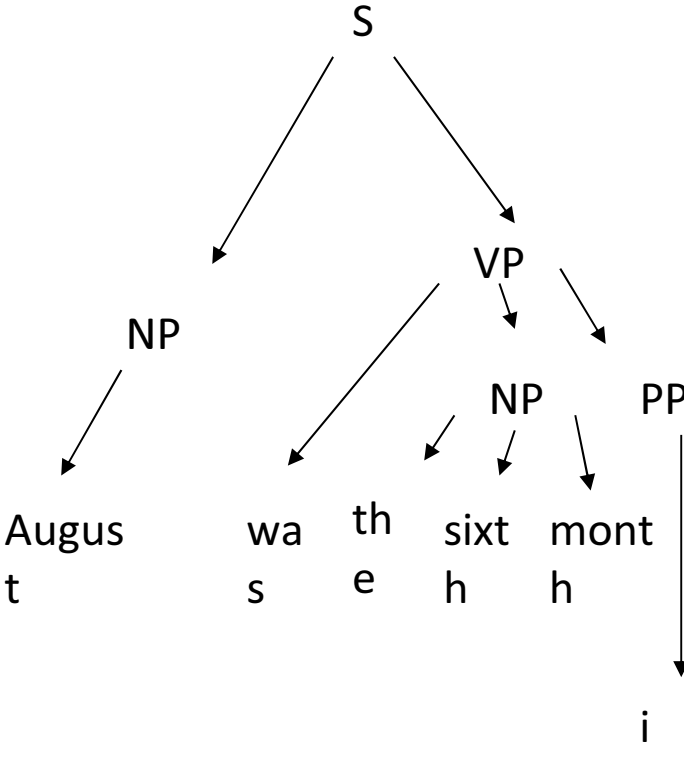


Learning simplification from parsing trees

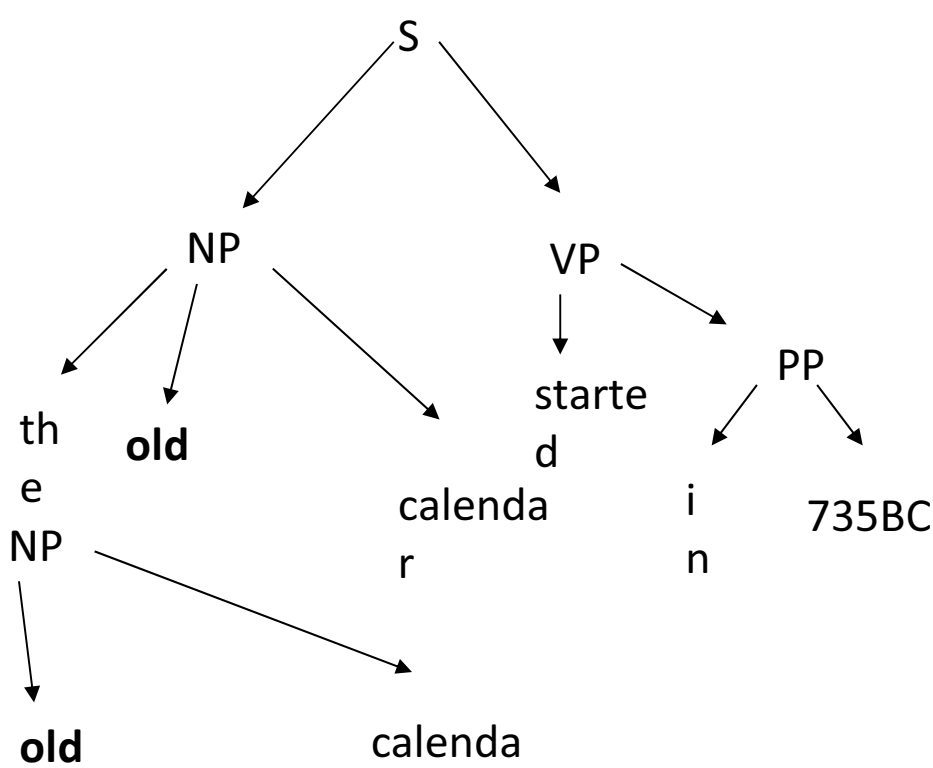


Learning simplification from parsing trees

WORD SUBSTITUTION



PROBABILITY OF REPLACING A WORD



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

Syntactic Simplification by Optimization

- 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)

Event-Based ATS System (EventSimplify)

- The core idea:
 - Events constitute relevant information in news
 - Descriptive (parts of) sentences not denoting events are informationally less relevant
- Semantic content reduction based on information relevance (opposed to traditional lexical and syntactic simplification)
- Two event-based simplification schemes:
 - Sentence-wise
 - Event-wise
- Evaluation:
 - readability (automated),
 - grammaticality and information relevance (human)

Example

Original

***Baset al-Megrahi**, the Libyan intelligence officer who was **convicted in the 1988 Lockerbie bombing** has **died at his home in Tripoli**, nearly **three years after he** was **released from a Scottish prison**. There were complications from prostate cancer and his funeral would take place on Monday.*



Simplification

***Baset al-Megrahi**, **convicted in the 1998 Lockerbie bombing** has **died at his home in Tripoli**. **Three years earlier he** was **released from a Scottish prison**.*

EventSimplify

- Build upon state-of-the-art event extraction system (Glavaš and Šnajder, 2013)
- Extract only factual events
 - Non-factual events (negated, uncertain) generally contain less important information
- Two step process:
 - Supervised extraction of factual event mentions
 - Application of event-centred simplification schemes (two different schemes)

Simplification Example

Original

*“**Baset al-Megrahi**, the Libyan intelligence officer who was **convicted in the 1988 Lockerbie bombing** has **died at his home in Tripoli**, nearly three years after **he** was **released from a Scottish prison.**”*

Sentence-wise simplification

*“**Baset al-Megrahi** was **convicted in the 1988 Lockerbie bombing** has **died at his home** after **he** was **released from a Scottish prison.**”*

Event-wise simplification

*“**Baset al-Megrahi** was **convicted in the 1988 Lockerbie bombing**. **Baset al Megrahi** has **died at his home**. **He** was **released from a Scottish prison.**”*

Event-wise with pron. anaphora resolution

*“**Baset al-Megrahi** was **convicted in the 1988 Lockerbie bombing**. **Baset al-Megrahi** has **died at his home**. **Baset al-Megrahi** was **released from a Scottish prison.**”*

Evaluation of EventSimplify

- Readability (automatically)
- Grammaticality (human)
- Information relevance (human)
- Evaluated on text snippets (280 in total)
- Baseline: retains only the main clause of a sentence and discards all subordinate clauses

Human Evaluation

Aspect	Weighted kappa	Pearson	MAE
Grammaticality	0.68	0.77	0.18
Meaning	0.53	0.67	0.37
Simplicity	0.54	0.60	0.28

IAA

Scheme	Grammaticality (1 – 3)	Relevance (1 – 3)
Baseline	2.57 ± 0.79	1.90 ± 0.64
Sentence-wise	1.98 ± 0.80	2.12 ± 0.61
Event-wise	2.70 ± 0.52	2.30 ± 0.54
Pronominal anaphora	2.68 ± 0.56	2.39 ± 0.57

Relevance = harmonic mean of Meaning and Simplicity

PART 3

TS Projects

PSET Project

- First ever project to address needs of people with language impairment
 - Targeted simplification for people with aphasia
 - Aphasia: language impairment following brain injury
 - People with aphasia may benefit from Augmentative, Alternative Communication (AAC) technology
- Complex language in newspapers
 - “Twenty-five-year-old blonde-haired mother-of-two Jane Smith....”
 - Pronouns may be difficult to interpret
 - Passive-voice constructions may be difficult to understand

PSET Project

- In PSET sentence for aphasic people should (Devlin and Tait, 1998):
 - Follow the Subject-Verb-Object (SVO) pattern
 - Be in active voice
 - Be as short as possible
 - Contain only one adjective per noun
 - Be chronologically ordered in the text
 - Be semantically non-reversible
- PSET system components
 - Syntactic simplifier
 - Anaphora resolver/substitution component
 - Lexical simplifier

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

- A text simplification system in Spanish for people with cognitive disabilities (e.g. Down syndrome)
- Lexical and syntactic simplification components implemented
- System developed based on creation/analysis of a parallel corpus of original (short) news and their simplified versions: the Simplext corpus

ORIGINAL & SIMPLIFIED SENTENCES IN SIMPLEXT

Ex.	Original	Simplified
1	<i>Abre en Madrid su primera sucursal el mayor banco de China y del Mundo.</i> (Opens in Madrid its first branch the biggest bank of China and the World.)	<i>El banco más importante de China y del mundo abre una oficina en Madrid.</i> (The most important bank of China and the world opens an office in Madrid.)
2	<i>El ICBC ha abierto ya 203 sucursales en un total de 28 países de todo el mundo, también en España desde este lunes.</i> (The ICBC has opened 203 branches in a total of 28 countries around the world, also in Spain since this Monday.)	<i>El Banco de China tiene oficinas en muchos países del mundo. Ahora, también tiene una oficina en España.</i> (The Bank of China has offices in many countries around the world. Now it also has an office in Spain.)
3	<i>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.</i> (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	<i>Arranca la liga masculina de Goiball, el único deporte específico para ciegos.</i> (Starts the men's league of Goiball, the only specific sport for the blind.)	<i>Comienza la liga masculina de Goiball. El Goiball es el único deporte específico para ciegos.</i> (Begins the men's league of Goiball. Goiball is the only specific sport for the blind.)
5	<i>La ONU prevé el fin de muertos por malaria para 2015.</i> (The UN expects the end of dead by malaria for 2015.)	<i>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 una enfermedad que se transmite gracias a un mosquito.</i> (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.)



**INFORMATIO
N
PROVIDER**

**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**

**SIMPLIFICATIO
N
EXPERT**



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 ...

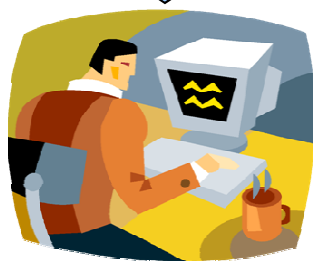
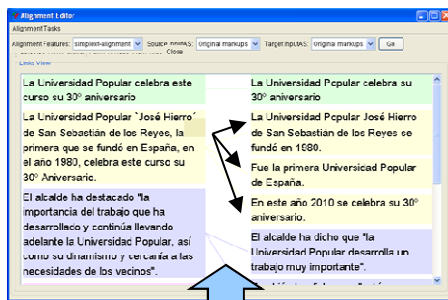
ALIGNMENT PROGRAM

ALIGNMENT TABLES

BI-TEXT EDITOR

GATE bi-text editor

CORPUS for RESEARCH

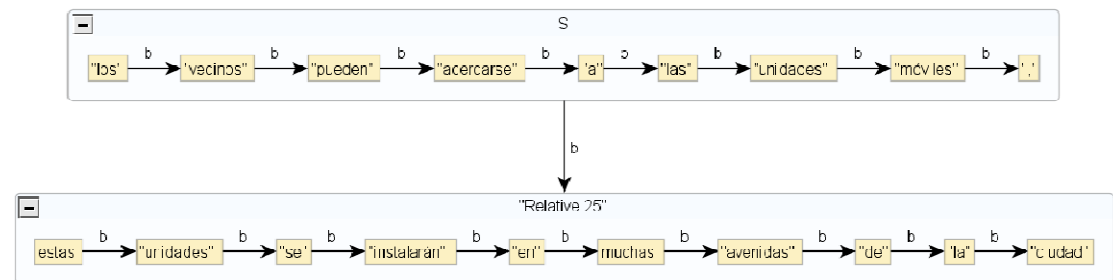
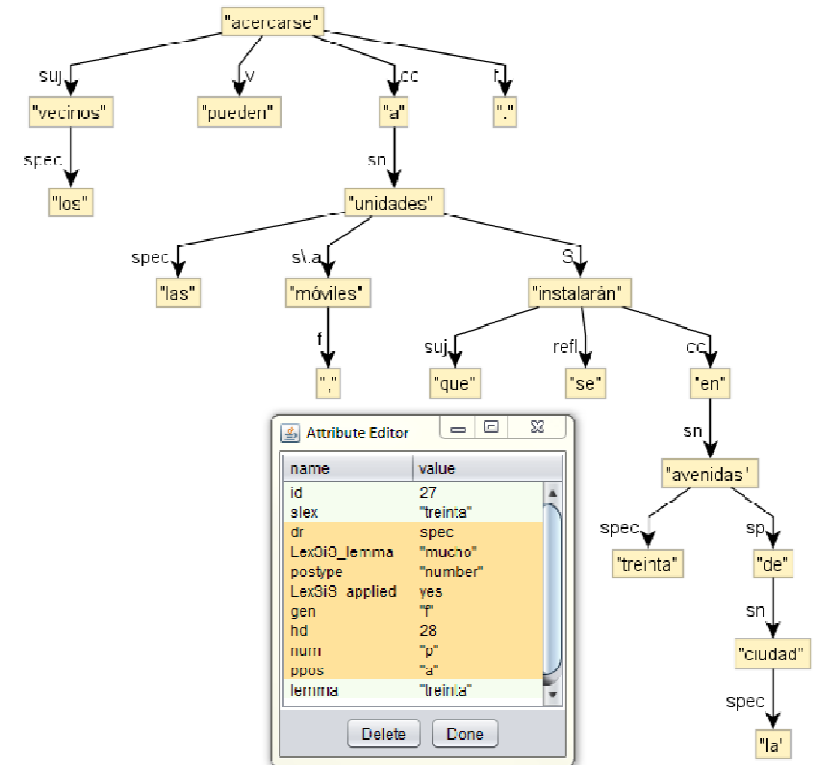
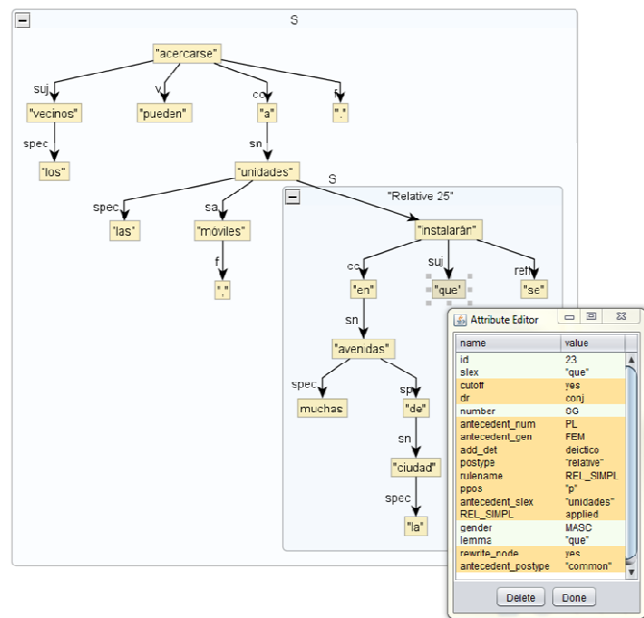


Syntactic simplification approach (Bott et al. 2012)

- Rule-based approach that transforms *dependency graphs* into “new” sentences
 - 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
<pre> leftside = [?Y1 { S->?Y1 { ?Z1->Z1 } }] </pre>	<pre> rightside = [rc:Z1 { <=>Z1 mark_relative_simple=applied }] </pre>	<pre> conditions = [?Y1.ppos="v"; ?Y1.mood="indicative" ?Z1.slex="que" or ?Z1.slex="quién" or ?Z1.slex="cuál" or ?Z1.slex="donde"; ?Z1.id<?Y1.id;] </pre>



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 simplification portal

The screenshot shows a web browser window displaying the Simplext website. The page has a dark blue header with the Simplext logo and navigation links: Inicio, Presentación, Dispositivos, Foros, and Contáctanos. A search bar is also present. The main content area features a title 'El sistema automático de simplificación de textos' and a descriptive paragraph. Below this, there are two columns of text, each with four paragraphs (T1-T4). The left column is labeled 'ORIGINAL' and the right column is labeled 'SIMPLIFICADO'. A central button labeled 'Simplificar texto' is positioned between the two columns. The text in the 'SIMPLIFICADO' column is a simplified version of the text in the 'ORIGINAL' column.

Firefox

Simplext - El sistema automático de sim... +

Iniciar sesión | Regístrate, ¡es gratis!

simplext

El sistema automático de simplificación de textos

Simplext favorece la inclusión tecnológica de personas con capacidades cognitivas limitadas a través de la simplificación automática de contenidos utilizando el paradigma de la fácil lectura.

Inicio | Presentación | Dispositivos | Foros | Contáctanos

Buscar

T1: Se registraron en España un total de 451 agresiones a facultativos.

T2: El actual estado geológico de Cataluña puede comenzar a describirse desde los primeros grandes cambios del Paleozoico. Inicialmente el territorio formaba parte de una cuenca oceánica en la que, por reposo orogénico, se depositaban materiales sedimentarios finos y arcillosos.

T3: Alrededor de 390.000 personas han regresado a sus casas desde que vieran obligadas a desplazarse por las inundaciones causadas por las lluvias monzónicas del pasado verano en Pakistán.

T4: La situación de ingobernabilidad en la que ha quedado Italia tras las elecciones ha sacudido los mercados de deuda, donde los inversores ven con preocupación que los candidatos que rechazan la austeridad impuesta por Bruselas hayan obtenido más de la mitad de los votos.

T1: Se registraron en España un total de 451 agresiones a médicos.

T2: El actual estado geológico de Cataluña puede comenzar a describir se desde los primeros grandes cambios del Paleozoico. Inicialmente el territorio formaba parte de una cuenca oceánica. En esta cuenca, por reposo orogénico, se depositaban materiales sedimentarios finos y arcillosos.

T3: Alrededor de 390.000 personas han regresado a sus casas desde que vieran obligadas a desplazarse por las inundaciones. Las inundaciones estan causadas por las lluvias monzónicas del pasado verano en Pakistán.

T4: La situación de ingobernabilidad en la que ha quedado Italia tras las elecciones ha sacudido los mercados de deuda. En estos mercados los inversores ven con preocupación que los candidatos que rechazan la austeridad. La austeridad está impuesta por Bruselas.

ORIGINAL

Simplificar texto

SIMPLIFICADO

Simplext full evaluation

- Evaluation of syntactic simplification module, rule-based simplification, and whole system (Drndarevic et al., 2013; Saggion et al. 2015)
 - 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 (PUNCT)	$N(punct)$

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	$-24.58\% \pm 3.02\%$	$-20.55\% \pm 1.96\%$	$-1.33\% \pm 4.58\%$
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%

Able to Include



- To bring for people with intellectual disabilities application which integrates:
 - Text Simplification
 - Text to Pictogram / Pictogram to Text Translation
 - Speech Synthesis
- Text Simplification
 - Spanish: technology from Simplext
 - English: new software based on Simplext

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 throughout 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

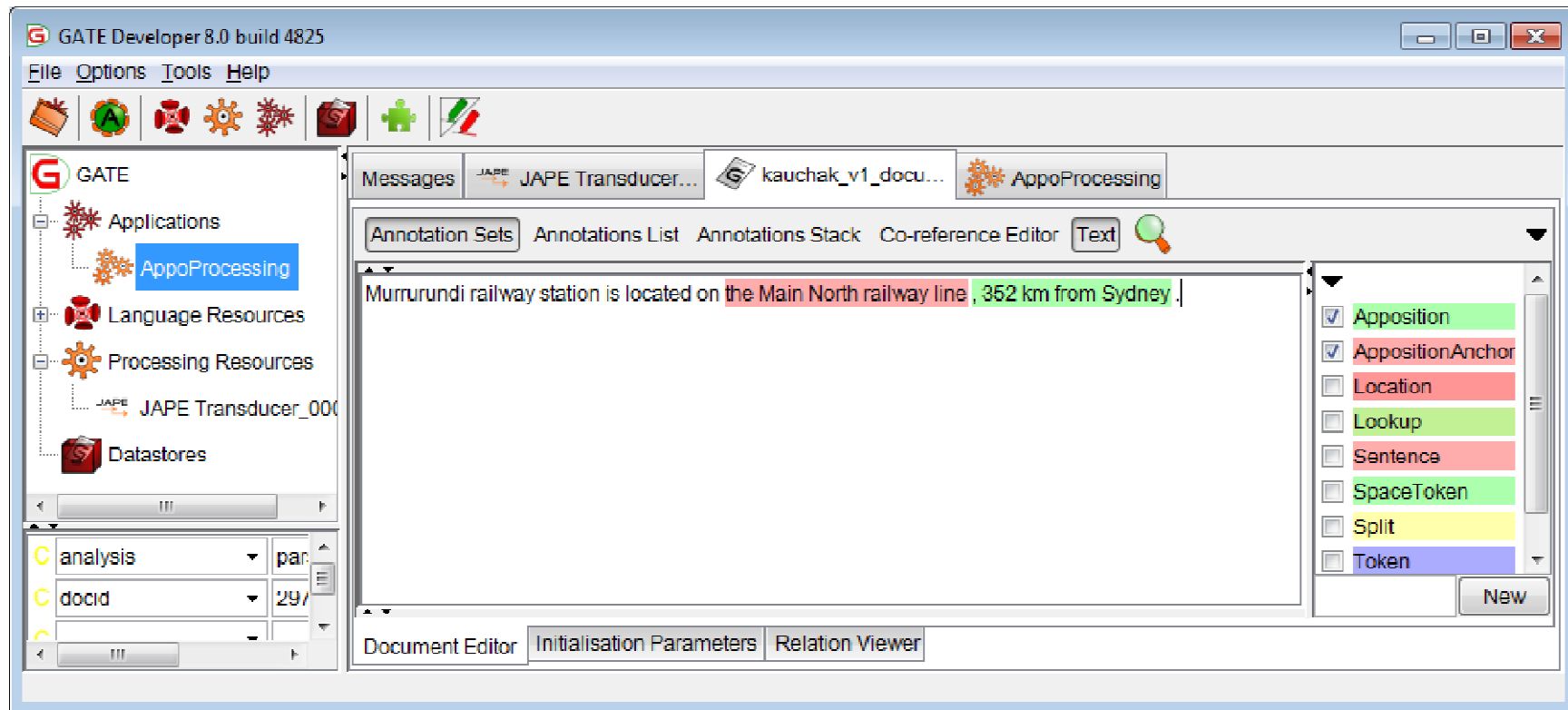
The screenshot shows the GATE Developer 8.0 interface. On the left, a tree view shows the project structure with 'AppoProcessing' selected. The main window displays a Jape rule named 'apposition'. The rule is defined as follows:

```
// Rule for appositive phrases
Phase: AppositivePhrases
Input: Token
Options: control - appait

Rule: Apposition
({Token.func != "P"}*
 {Token}):anchor_end
):anchor
(
 {Token.string == ","} | {Token.string == "."} | {Token.string == "-"}):apposition_start
 {Token.func != "APPO"}{0,6}
 {Token.category == "NNP", Token.func == "APPO"} | {Token.category == "NNPS", Token.func == "APPO"}
 | {Token.category == "NN", Token.func == "APPO"} | {Token.category == "NNS", Token.func == "APPO"}
 | {Token.category == "CD", Token.func == "APPO"}
):apposition_hd
):apposition
->
```

A blue arrow points to the regular expression part of the rule, with the text "Left Hand Right Side (regular pattern)" overlaid on it.

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	ERRORS	Parser	Rules
Apposition	79%	21%	0%	Apposition	19	2
Relative Clause	79%	14%	7%	Relative Clause	14	0
Coordination	56%	6%	38%	Coordination	4	2
Subordination	72%	25%	3%	Subordination	7	18
Passive	85%	6%	9%	Passive	5	1
Total	74%	14%	11%	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/identification
 - 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 – identify 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

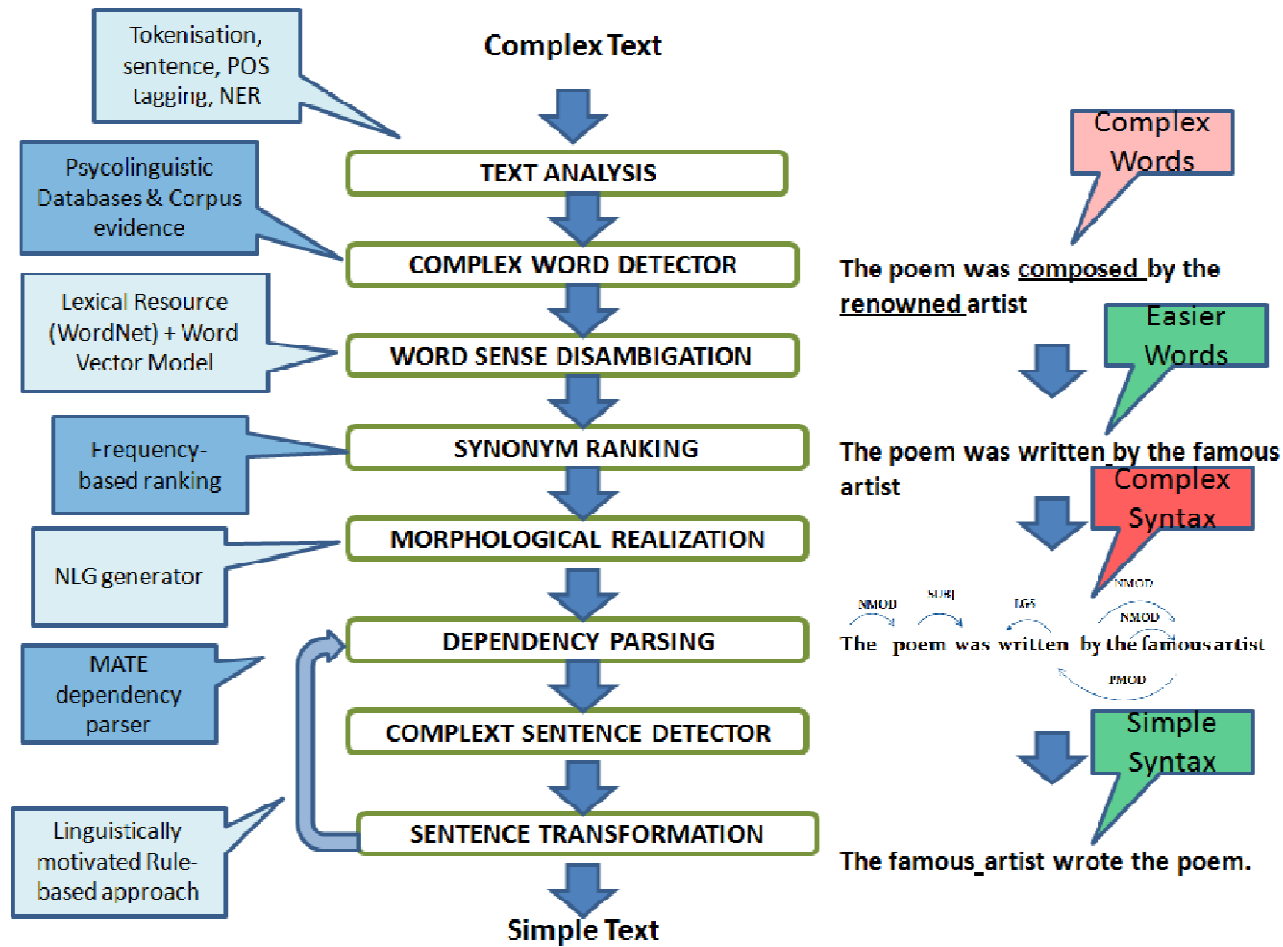
Synonym Ranking

- 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

Overall architecture

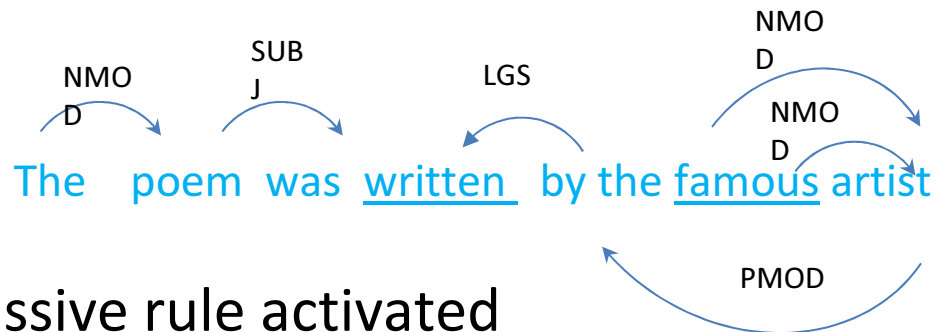


Example

- “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.*

Demonstrations

- Simplext / Able to Include (interface)
- Simplification embedded in e-mail application (interface)
- Multilingual lexical simplifier (console)

PorSimples project

- A text simplification system Brazilian Portuguese aimed at people with low-literacy (11% of Brazilian population)
- Mainly syntactic simplification developed
- Corpus developed for the project containing newspaper articles and two different simple versions (natural and strong)

PorSimples

- Simplifications in the PorSimples corpus
 - Appositions; relative clauses; subordinate clauses; coordinate clauses; sentences w/non-finite verbs; passive voice
- Rule-based syntactic simplification
 - Hand-made developed procedures applied in cascade
 - Passive voice, apposition, subordination, non-restrictive clauses, restrictive relatives, and coordination
 - Iterative process: sentence is re-parsed after each rule application
 - Treatment of relative clauses:
 - Input: “The book, which John gave me, belongs to Paul”
 1. Find relative pronoun and check the relative is non restrictive
 2. Find where the relative ends
 3. Generate a sentence with the relative
 4. Generate a sentence with the main clause.
 5. Reorder
 - Output: “The book belongs to Paul. John gave me the book.”

PorSimples

- Learning components in PorSimples
 - Sentence split algorithm using Support Vector Machine classifier (features from basic text analysis + rhetorical relation inspired features)
 - Simplification as machine translation using the corpus
 - 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
 - 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
- SIMPLIFICA is a project related to PorSimples whose aim is to provide an authoring tool for the production of adapted texts
 - Incorporates readability-assessment prediction
 - Proposes simplifications to the user who can correct them to achieve good text quality

Simplification for people with autism

- The FIRST project “Flexible Interactive Reading Support Tool” (Mitkov, 2012; Martín-Valdivia et al., 2014)
 - Reading obstacles faced by people with autism:
 - complex sentences, ambiguity, figurative language, rare and specialized terms, etc.
 - Open Book a multilingual (Spanish, Bulgarian, English) tool simplifies and summarizes input text, replace or explains difficult vocabulary, provides navigation mechanisms, add pictograms for better understanding....
 - Simplification is not only for the final user but also for their carers who can adapt input text using the tool
- (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

More on simplifying for end users

- Adapting texts for people with dyslexia (Rello et al. 2013)
 - People with dyslexia have problems with word recognition
 - Word frequency and length are factors which influence readability and comprehensibility in people with dyslexia
- Adapting texts to poor readers (Williams and Reiter, 2005)
 - Natural Language Generation techniques addressing word selection, sentence generation, and discourse
 - Easy to understand words, short sentences, simple syntactic structures, easy to understand discourse connectives, etc.

Simplifying for NLP

- To improve parsing results (Jonnalagadda et al. 2009)
 - Concerned with problems when parsing biomedical texts with parsers developed for newspaper articles
 - Simplifying named entities (e.g. gene names into placeholders), rudimentary segmentation of sentences based on punctuation
- To improve summarization output (Lal and Ruger, 2002; Siddarthan et al. 2004)
 - Lexical simplification of the output summary
 - Simplification before content selection
- To improve information extraction (Evans, 2011)
 - Classification algorithm to detect and segment different types of coordination
 - Simple IE patterns over simplified input work better than complex IE patterns over non-simplified input
- To improve comprehension of patent documents (Bouyad-Aga et al, 2009)
 - Segmentation of long claim sentences and re-generation as simpler sentences

TS Helps Machine Translation

Focus: English to Serbian MT (Štajner and Popović, 2016)

Two different automated TS systems:

1. **TS-C:** lexico-syntactic TS system (Angrosh and Siddharthan, 2014) with no content reduction (completely preserving the original meaning)
1. **TS-A:** LightLS + EventSimplify which performs syntactic simplification and a significant content reduction by only preserving the most relevant parts of the sentence (Štajner and Glavaš, 2017)

Fluency and Adequacy (Example 1)

Original (A = 2, F = 3):

”As we emerge from a decade of conflict abroad and economic crisis at home, it’s time to renew America,” Obama said, speaking against a backdrop of armored vehicles and a U.S. flag.

TS-C (A = 4, F = 4):

Speaking against a backdrop of armored vehicles and a U.S. flag, Obama said it’s time to renew America as we emerge from a decade of conflict abroad and economic crisis at home.

Fluency and Adequacy (Example 2)

Original (A = 2, F = 3):

Several Israeli security delegations have visited Egypt during the past two months to **decide on** a new embassy location.

TS-C (A = 4, F = 4):

Several Israeli security delegations have visited Egypt during the past two months to **choose** a new embassy location.

Fluency and Adequacy (Example 3)

Original (A = 4, F = 3):

A Florida mother shot her four children early Tuesday morning **before turning** the gun on herself at her home in Port St. John, police said.

TS-A (A = 5, F = 5):

A Florida mother shot her four children early Tuesday morning. **After that, the Florida mother turned** the gun on herself at her home.

PART 4

Existing resources for text simplification

Text Simplification Resources

- Lexical Resources for Simplification
 - Synonym inventories in several languages: Word Nets / Multilingual Central Repository; various Open Thesaurus (Spanish, English, Catalan, etc.)
 - Compiled lists of frequencies: Kucera-Francis (Kucera & Francis, 1967), Age of Acquisition (Kuperman et al., 2012)
 - Lists of familiar words (Dale & Chall, 1948)
- Corpora
 - Comparable corpus: Wikipedia \leftrightarrow Simple Wikipedia; Edit histories
 - Parallel corpora: Newsela (Xu et al, 2015), Simplext (Saggion et al. 2015) , PorSimples (Aluisio et al. 2008), FIRST (Stajner et al, 2014)

Lexical Simplification Resources

- Based on the SemEval Lexical Substitution the English Lexical Simplification dataset is created (Specia et al., 2012)
- Based on the lexical substitution dataset (McCarthy and Navigli, 2009)
- 201 words in 10 different contexts

Original sentence: During the siege, George Robertson had appointed Shuja-ul-Mulk, who was a *bright* boy only 12 years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral.

Set of possible substitutes: intelligent; bright; clever; smart

Simplicity Gold Rankings (average of human annotators): intelligent
clever smart bright

Lexical Simplification Resources

- Also based on the lexical substitution dataset, (De Belder and Moens, 2012) created a similar dataset. Difficulty based on grades provided by informants.

1161	acquire.v	Thus , the analyst <u>acquires</u> knowledge about the nature of the patient through an awareness of something going on in him.
1165	acquire.v	How many times have I caught up with those people several years later , to discover that they have <u>acquired</u> a lifestyle , a car and a mortgage to match their salary , and that their initial ideals have faded to the haziest of memories , which they now dismiss as a post adolescent fantasy?
986	liberal.a	Municipal housing schemes with <u>liberal</u> aid from the central government will be encouraged for those who do not wish to establish their own houses .
987	liberal.a	We're both in our early thirties , both grew up in the suburbs of east coast US cities , raised by <u>liberal</u> parents who pushed us towards soccer , the progressive , globalized , nonviolent sport of choice for seventies and eighties US parents .
1575	scene.n	Every <u>scene</u> seems totally natural like it could have really happened , and yet the movie is not a dull slice-of-life diorama either .
1577	scene.n	On the plus side , the immediate mode offers the possibility of exploring dynamic <u>scenes</u> .

1161	acquire.v	[[gain, gather, collect], [acquire], [amass]]
1165	acquire.v	[[get], [obtain, achieve, gain], [acquire, procure]]
986	liberal.a	[[generous], [abundant, plentiful, liberal, social]]
987	liberal.a	[[open minded, free thinking], [broad minded], [tolerant], [progressive, liberal]]
1575	scene.n	[[part, act], [setting, scene], [sequence]]
1577	scene.n	[[picture, area], [scene, setting], [sight], [sequence]]

Lexical Simplification Resources

- Horn et al. (2014) created a 500 sentences, crowd-sourced lexical substitution resource sampled from alignments between English Wikipedia and Simple English Wikipedia

occurrences	A haunted house is defined as a house that is believed to be a center for supernatural occurrences or paranormal phenomena.	events (24); happenings (12); activities (2); things (2); accidents (1); activity (1); acts (1); beings (1); event (1); happening (1); instances (1); times (1); situations (1)
acquired	Dodd simply retained his athletic director position, which he had acquired in 1950.	gotten (13); gained (11); got (7); received (7); obtained (5); achieved (3); amassed (1); inherited (1); taken (1); started (1)

Lexical Resources

- CASSA (Baeza-Yates et al., 2015) is a lexical database created automatically from the Spanish Open Thesaurus and the 5-gram Google Books Ngram Corpus

Frequency	Target	Context ($w_1, w_2, ?, w_3, w_4$)	Substitutes	Lemma
60285	ámbitos	todos los ? de la	[campo,ambiente,terreno]	ámbito
59886	ocurre	lo que ? es que	[pasar,suceder,acontecer]	ocurrir
58326	tercio	el primer ? del siglo	[doblar,desplazar,inclinar]	terciar
58026	facultades	de las ? que le	[poder,licencia,autorización]	facultad
57511	mitad	a la ? de la	[parte,fracción,porción]	mitad

Lexical Resources

- French lexicon annotated with degrees of difficulties (Gala et al, 2013)
 - Words extracted from spoken transcriptions of Parkinson's disease affected people – considered a sample of “simple” text productions
 - Words graded based on two existing lexical resources in French: Manulex: lexicon with grades extracted from educational resources, and JeuxDeMots: semantic network with synonymic and other lexical relations
 - A classifier is trained to associate grades to JdM entries not present in Manulex
- Brooke et al (2012) presents a method to create a readability lexicon with three levels of difficulty from a small (15k) set of seed words.
 - A regression model to associate a score and a classification algorithm to discriminate two given words are used to expand the lexicon

Simple English Wikipedia Dataset

- Called PWPK dataset, it has been compiled by Zhu et al. (2010)
- 65K articles from SEW aligned to EW
- Sentences aligned using tf*idf + cosine similarity
- Final dataset contains 108K sentence pairs

Ex.	English Wikipedia	Simple English Wikipedia
1	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.
2	This month was originally named Sextilis in Latin, because it was the sixth month in the ancient Roman calendar, which started in March about 735 BC under Romulus.	This month was first called Sextilis in Latin, because it was the sixth month in the old Roman calendar. The Roman calendar began in March about 735 BC with Romulus.
3	Dombasle-sur-Meurthe is a commune in the Meurthe-et-Moselle department in northeastern France.	Dombasle-sur-Meurthe is a town in France.
4	Konkani is the official language in the Indian state of Goa and is also one of the Official languages of India.	It is the official language of Goa, a state in India.
5	The male fertilises the eggs externally by releasing his sperm onto them, and will then guard them for at least three months, until they hatch.	After the fertilization of the eggs, the male will guard them for at least six months.
6	Transport Marske is served by Longbeck and Marske railway stations, which connect to Darlington mainline station.	The Longbeck railway station and Marske railway station, which connect to Darlington mainline station, are the only means of transport there.

Newsela corpus (English + Spanish)

- Xu et al. (2015) heavily criticizes PWKP since it has alignment errors and contains inadequate simplifications
- 50% of pairs in PWKP are not simplifications
- Newsela is controlled for quality
- 1,130 news articles re-written 4 times for children at different grade levels
- Freely available for research purposes upon request at:
<https://newsela.com/data/>

Newsela corpus (English + Spanish)

Ex.	Text Fragments of Four Simplified Versions of the Same Original Text
Original	CHICAGO - On a recent afternoon at Chicago's Dewey Elementary Academy of Fine Arts, Ladon Brumfield asked a group of 9- and 10-year-old African-American girls to define beauty. The nearly 20 girls unanimously agreed that if a woman had short, kinky hair, she was not beautiful. But when Brumfield, the director of a project empowering young girls, passed around a photograph of Lupita Nyong'o, the dark-brown-skinned actress who sports an extra-short natural, the girls were silent for a moment. Then, once again, their answer was unanimous: They agreed Nyong'o was beautiful.
Simp. 1	CHICAGO - On a recent afternoon at a Chicago elementary school, Ladon Brumfield asked a group of 9- and 10-year-old African-American girls to define beauty. The nearly 20 girls unanimously agreed that if a woman had short, kinky hair, she was not beautiful. But then Brumfield, the director of a project empowering young girls, passed around a photograph of Lupita Nyong'o, the dark-brown-skinned actress who wears an extra-short Afro. The girls, who attend Dewey Elementary Academy of Fine Arts, were silent for a moment. Then, once again, their answer was unanimous: They agreed Nyong'o was beautiful.
Simp. 2	CHICAGO - On a recent afternoon, Ladon Brumfield asked a group of 9- and 10-year-old African-American girls to define beauty. The nearly 20 girls unanimously agreed that if a woman had short, kinky hair, she was not beautiful. They thought women with smooth, straight hair were more beautiful. But then Brumfield passed around a picture of Lupita Nyong'o, the dark-skinned actress who wears her hair extra-short. The girls, who attend Dewey Elementary Academy of Fine Arts, were silent for a moment. Then, once again, their answer was unanimous: They agreed Nyong'o was beautiful.

Simp. 3	CHICAGO - On a recent afternoon, a group of 9- and 10-year-old African American girls talked about beauty. They all agreed that women with short, kinky hair were not beautiful. But then Ladon Brumfield, founder of the group Girls Rule!, passed around a photograph of Lupita Nyong'o. The dark-skinned actress wears her hair extra-short. The girls were silent for a moment. Then, once again, they all agreed: Nyong'o was beautiful.
Simp. 4	CHICAGO - Ladon Brumfield asked a group of African American girls to think about beauty. Brumfield began Girls Rule!, a girl empowerment project. The girls agreed that women with short, kinky hair were not beautiful. But then Brumfield passed around a picture of Lupita Nyong'o. She is a famous actress. She has dark skin. And Nyong'o wears her hair extra-short. The girls, who are 9 and 10 years old, were silent. Once again, they all agreed. Nyong'o was beautiful.

(Newsela, 2016)

Sentence and paragraph alignment

- English Newsela corpus manually aligned (Xu et al., 2016)
- English Newsela corpus automatically aligned (Štajner et al., 2017; Paetzold et al., 2017)
- English and Spanish Newsela corpus automatically aligned (Štajner et al., 2018)

Automatic alignment tools

- CATS (Štajner et al., 2017; Štajner et al., 2018):
 - three text similarity measures
 - two alignment strategies (preserving order or not)
 - <http://cats-demo.informatik.uni-mannheim.de/demo.jsp>
 - Freely available: <https://github.com/neosyon/SimpTextAlign>
- MASSAlign (Paetzold et al., 2017):
 - <https://github.com/ghpaetzold/massalign>

PorSimples corpus (Brazilian Portuguese)

- Aluísio and Gasperin (2010) Parallel corpus of news articles (Zero Hora) together with human simplifications
- Two simplifications: *natural* and *strong*
- Sentences: 2,116 original; 3104 natural simplifications; 3,537 strong simplifications

Original	As salas de cinema de todo o mundo exibiam uma produção do diretor Joe Dante em que um cardume de piranhas escapava de um laboratório militar e atacava participantes de um festival aquático. (<i>Movie theaters around the world exhibited a production of director Joe Dante where a school of piranhas escape from a military laboratory and attacked participants of an aquatic festival.</i>)
Natural	As salas de cinema de todo o mundo exibiam uma produção do diretor Joe Dante. Em a produção do diretor Joe Dante, um cardume de piranhas escapava de um laboratório militar e atacava participantes de um festival aquático. (<i>Movie theaters around the world exhibited a production of director Joe Dante. In production of director Joe Dante, a school of piranhas escape from a military laboratory and attacked participants of an aquatic festival.</i>)

Strong	As salas de cinema de todo o mundo exibiam um filme do diretor Joe Dante. Em o filme, um cardume de piranhas escapava de um laboratório militar. O cardume de piranhas atacava participantes de um festival aquático. (<i>Movie theaters around the world show a film of director Joe Dante. In the film, a school of piranhas escape from a military laboratory. The school of piranhas attacked participants of an aquatic festival.</i>)
--------	---

Simplext Corpus (Saggion et al. 2015)

- 200 short news articles from Spanish news agency
- Each simplified by a text adaptation expert
- Corpus aligned at sentence level (Bott & Saggion, 2011) automatically and manually corrected
- Sentences: 1,149 original; 1,808 simplified
- Most are 1-to-1 alignments with content reduction
- Splits (1-to-2 and 1-to-n) are the second most frequent alignment
- Also deletions and insertions are observed

Ex.	Original	Simplified
1	<i>Licenciada en Bellas Artes por la Universidad Politécnica de Valencia, Ana Juan es ilustradora, escritora y pintora.</i> (Bachelor of Fine Arts from the Polytechnic University of Valencia, Ana Juan is an illustrator, writer and painter.)	<i>Ana Juan es ilustradora, escritora y pintora. Estudió Bellas Artes en la universidad de Valencia.</i> (Ana Juan is an illustrator, writer and painter. She studied Fine Arts at the University of Valencia.)
2	<i>La ONU celebra en 2011 el Año Internacional de la Química para fomentar el interés de los jóvenes por esta ciencia y mostrar cómo, gracias a ella, se puede “responder a las necesidades del mundo”.</i> (The UN celebrates in 2011 the International Year of Chemistry to promote the interest of young people in this science and show how, thanks to it, we can meet the needs of the world.)	<i>En 2011 se celebra el Año Internacional de la Química.</i> (2011 marks the International Year of Chemistry.)
3	<i>2011, AÑO INTERNACIONAL DE LA QUÍMICA.</i> (2011, International Year of Chemistry.)	<i>El 2011 es el Año Internacional de la Química.</i> (2011 is the International Year of Chemistry.)
4	<i>El jugador del Fútbol Club Barcelona Andrés Iniesta colaborará de nuevo con la Federación Española de Enfermedades Raras y pondrá cara a su campaña de sensibilización de 2011.</i> (Barcelona Football Club player Andres Iniesta will collaborate again with the Spanish Federation for Rare Diseases and will give his image in the 2011 awareness campaign.)	<i>Andrés Iniesta ayudará este año a la Federación Española de Enfermedades Raras. Andrés Iniesta es jugador de fútbol en el Fútbol Club Barcelona. También prestará su imagen a la campaña de esta Federación.</i> (Andres Iniesta will help this year the Spanish Federation for Rare Diseases. Andres Iniesta is football player in the Football Club Barcelona. He will Also lend his image to the campaign of the Federation.)

PART 5

Neural Approaches

Lexical Simplification Systems

- Two main types:
 - Modular approach (Paetzold and Specia, 2016)
 - All-in-one (Horn et al., 2014; Glavaš and Štajner, 2015)
- Modular approach:
 - Complex word identification (CWI)
 - Substitution candidate generation
 - Substitution candidate ranking
 - Substitution

LS Approaches

- Devlin and Tait (1998): uses WordNet (rule-based)
- Yatskar et al. (2010): uses EW meta-data (unsupervised)
- Biran et al. (2011): uses co-occurrence statistics of SEW (unsupervised)
- Horn et al. (2014): uses sentence-aligned EW-SEW (supervised)
- Glavaš and Štajner (2015): uses word embeddings (unsupervised)
- Paetzold and Specia (2016): uses word embeddings with POS (unsupervised)
- Implementation of many LS systems:
<http://ghpaetzold.github.io/LEXenstein/>

Light-LS (Glavaš and Štajner, 2015)

- Pros:
 - No need for parallel data or simplified data
 - Better coverage than other LS systems
- Cons:
 - Simplifying only single words (no multi-word expressions)
 - Problem with antonyms (due to word embeddings)

Light-LS: Main Idea

- “Simple” words are also present in “non-simple” texts
- We need:
 - Good semantic similarity measure (to retrieve substitution candidates)
 - Good measure of word complexity (to rank substitution candidates)

Light-LS (Glavaš and Štajner, 2015)

- Simplification candidate selection:
 - Using only content words
 - Using 200-dimensional GloVe vectors pretrained on English Wikipedia and Gigaword 5
 - For each content word select 10 most similar content words (cosine similarity) excluding morphological derivations
- Ranking:
 - Context similarity (symmetric window of size 3)
 - Simplicity (frequency in a large corpora)
 - Fluency (language model)

Evaluation

- Automatic evaluation on two datasets:
 - Replacement task (Horn et al., 2014)
 - Ranking task (SemEval-2012 Task 1)
- Human evaluation on a 1 – 5 Likert scale:
 - Grammaticality
 - Meaning preservation
 - Simplicity

Replacement Task Results

- Precision: the percentage of correct simplifications (i.e. the system simplification was found in the list of manual simplifications)
- Accuracy: the percentage of correct simplifications out of all words that should have been simplified
- Changed: the percentage of target words changed by the system

Model	Precision	Accuracy	Changed
Biran et al. (2011)	71.4	3.4	5.2
Horn et al. (2014)	76.1	66.3	86.3
LIGHT-LS	71.0	68.2	96.0

Ranking Task Results

- Task: for each target word (one per sentence) and three given substitution candidates, rank the substitution candidates from simplest to most complex
- Evaluation: the official SemEval-2012 Task 1 script for calculating Cohen's kappa

Model	Cohen's kappa
Baseline-random	0.013
Baseline-frequency	0.471
Jauhar and Specia (2012)	0.496
LIGHT-LS	0.540

Results of Human Evaluation

Source	G	Smp	MP	Ch
Original	4.90	3.36	--	--
Manual	4.83	3.95	4.71	76.3%
Biran et al.	4.63	3.24	4.65	17.5%
LIGHT-LS	4.60	3.76	4.13	68.6%
Biran et al. Ch.	3.97	2.86	3.57	--
LIGHT-LS Ch.	4.57	3.55	3.75	--

Example

Source	Sentence
Original	The contrast between a high level of education and a low level of political rights was particularly great in Aarau, and the city <u>refused</u> to send troops to defend the Bernese border.
Biran et al.	The separate between a high level of education and a low level of political rights was particularly great in Aarau, and the city refused to send troops to defend the Bernese border.
LIGHT-LS	The contrast between a high level of education and a low level of political rights was especially great in Aarau, and the city <u>asked</u> to send troops to protect the Bernese border.

LS-NNS (Paetzold and Specia, 2016)

- Similar idea of using word embeddings for unsupervised LS
- Difference: context-aware word embeddings (POS tags instead of sense labels)
- Difference: modular approach
- Difference: used a corpus of subtitles

Model	Precision	Accuracy	Changed
Biran	0.121	0.121	1.000
Kauchak	0.364	0.172	0.808
Glavas	0.456	0.197	0.741
LS-NNS	0.464	0.226	0.762

(Paetzold and Specia, 2016)

Exploring Neural TS Models (Nisioi et al., 2017)

- First attempt at using sequence to sequence neural networks to model text simplification
- The model simultaneously performs lexical simplification and content reduction
- Almost perfect grammaticality and meaning preservation
- Higher level of simplification than state-of-the-art ATS systems

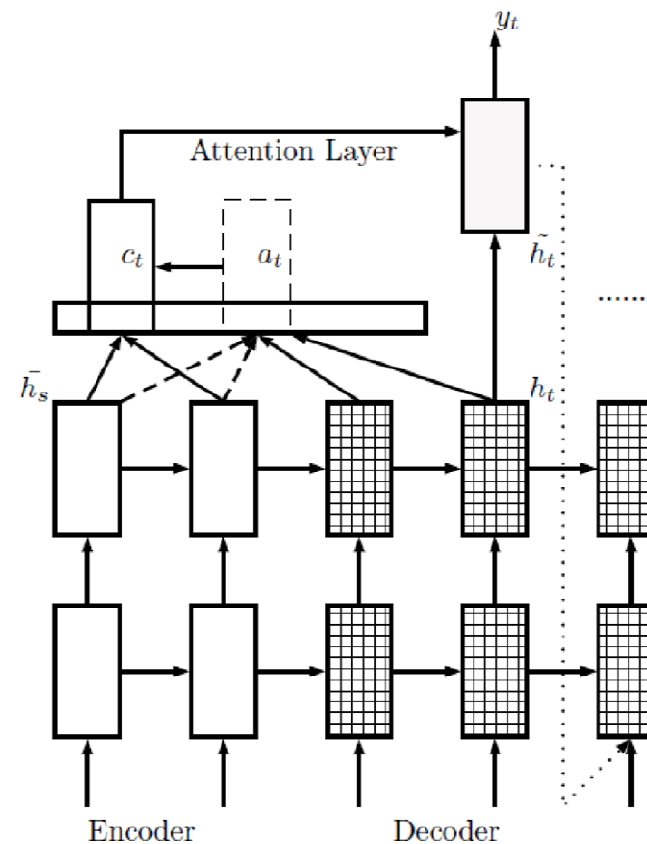
Dataset

- EW-SEW (Hwang et al., 2015): 150,000 full matches and 130,000 partial matches
- Manually created multi-reference development and test set (Xu et al., 2016): 2,000 + 359 (each with eight references)
- High number of named entities
- High lexical richness

	Original	Simplified
Locations	158,394	127,349
Persons	161,808	127,742
Organisations	130,679	101,239
Misc	95,168	71,138
Vocabulary	187,137	144,132
Tokens	7,400,499	5,634,834

NTS System

- OpenNMT framework
- Two LSTM layers
- Hidden states of size 500
- 500 hidden units
- A 0.3 dropout probability
- Vocabulary: 50,000



(Nisioi et al., 2017)

Training

- Training the model for 15 epochs with plain SGD optimiser
- After epoch 8, halve the learning rate
- At the end of every epoch save the current state of the model and predict perplexity values of the models on the dev set
- Early-stopping and selecting the model with best perplexity
- Parameters initialised over uniform distribution with support $[-0.1, 0.1]$
- Global attention in combination with input feeding for the decoder
- The architecture configuration, data, and pretrained models available at:
<https://github.com/senisioi/NeuralTextSimplification>

What's New Here?

- Word embeddings
- Kauchak (2013) showed that adding original language to the simple language in LMs improves ATS
- Encoder: original English + Google News (word2vec)
- Decoder: simplified English + Google News (word2vec)

What About the OoV Words?

- Vocabulary size: 50,000
- Those not present in the vocabulary are replaced with 'UNK' symbols during training
- At the prediction time, we replace unknown words with the highest probability score from the attention layer

How to Find the Best Hypothesis?

- We generate first two candidate hypotheses from each beam size from 5 to 12
- Try to find the best beam size and hypothesis based on:
 - BLEU with NIST smoothing (Bird et al., 2009)
 - SARI (Xu et al., 2016)
- Development dataset (2,000 sentence pairs) is used for finding the best beam size and hypothesis according to BLEU and SARI

Evaluation

- First 70 sentences from the test set (Xu et al., 2016)
- Automatic evaluation (BLEU and SARI)
- Human evaluation:
 - Number of changes
 - Correctness of changes
 - Grammaticality
 - Meaning preservation
 - Relative simplicity

Comparison with the State of the Art

- SBMT system (Xu et al., 2016)
- Unsupervised s.o.t.a. LS system LightLS (Glavaš and Štajner, 2015)
- PBSMT system with output reranking (Wubben et al., 2012)
- We use original systems in all three cases

Correctness and Number of Changes

- The whole phrase counted as one change:
 - e.g. *become defunct* → *was dissolved*
- **If** grammatically correct and preserves the meaning (2 native speakers) **and if** it makes the sentence easier to understand (2 non-native speakers) → **correct**
- Where not agreed, we asked a third annotator and used the majority vote

Grammaticality and Meaning Preservation

- 1 – 5 Likert scale (1 – very bad, 5 – very good)
- 3 native English speakers
- Quadratic Cohen's kappa:
 - 0.78 for Grammaticality
 - 0.63 for Meaning preservation

Simplicity

- Evaluated by 3 non-native but fluent English speakers
- Relative simplicity (evaluating sentence pairs Original – Simplified):
 - +2 much simpler
 - +1 somewhat simpler
 - 0 equal
 - 1 somewhat more difficult
 - 2 much more difficult
- Quadratic Cohen's kappa: 0.66

Results

Approach	Training		LM		Changes		Scores		Rank
	Dataset	Size (sent)	Corpus	Size (sent)	Total	Correct	G	M	S
NTS	Default (beam 5, hypothesis 1)				36	72.2%	4.92	4.31	0.46
NTS	Best SARI (beam 5, hypothesis 2)				72	51.6%	4.19	3.62	0.38
NTS	Best BLEU (beam 12, hypothesis 1)				44	73.7%	4.77	4.15	0.92
NTS-w2v	Default (beam 5, hypothesis 1)				31	54.8%	4.79	4.17	0.21
NTS-w2v	Best SARI (beam 12, hypothesis 2)				110	68.1%	4.53	3.83	0.63
NTS-w2v	Best BLEU (beam 12, hypothesis 1)				61	76.9%	4.67	4.00	0.40
PBSMT	Wiki (Good+Partial)	284,499	Wiki	391,572	76	35.5%	4.09	3.31	0.26
PBSMT	Newsela (neighb.)+Wiki	593,947	Newsela+Wiki	766,446	81	46.9%	4.40	3.84	0.30
PBSMT	Newsela (all) + Wiki	764,571	Newsela+Wiki	766,446	87	48.3%	4.25	3.73	0.30
	PBSMT-R (Wubben et al., 2012)				171	41.0%	3.10	2.71	-0.55
s.o.t.a.	Supervised SBMT (PPDB+SARI)(Xu et al., 2016)				143	34.3%	4.28	3.57	0.03
	Unsupervised (LightLS) (Glavaš and Štajner, 2015)				132	26.6%	4.47	2.67	-0.01

NTS vs. State-of-the-Art ATS

- NTS models have **higher percentage of correct changes**
- NTS models have **more simplified output** than any other ATS system
- NTS with custom word2vec embeddings, ranked with SARI:
 - the highest total number of changes among NTS models
 - one of the highest number of correct changes
 - the second highest simplicity score
 - solid grammaticality and meaning preservation scores

Customised NTS Models

- Ranking predictions with **SARI** → the highest number of changes
- Ranking predictions with **BLEU** → the highest number of correct changes
- Customised word embeddings in combination with **SARI** seem to work best among all our NTS systems

System	Output
NTS-w2v default	Perry Saturn (with terri) defeated Eddie Guerrero (with chyna) to win the WWF European Championship (8:10); Saturn pinned Guerrero after a diving elbow drop.
NTS-w2v SARI	Perry Saturn pinned Guerrero to win the WWF European Championship.
NTS-w2v BLEU	Perry Saturn pinned Guerrero after a diving drop drop.
NTS default	He (with terri) defeated Eddie Guerrero (with chyna) to win the WWF European Championship (8:10); Saturn pinned Guerrero after a diving elbow drop .
NTS BLEU/SARI	He defeated Eddie Guerrero (with Chyna) to win the WWF European Championship (8:10); Saturn pinned Guerrero after a diving elbow drop.
LightLS (Glavaš and Štajner, 2015)	Perry Saturn (with terri) defeated Eddie Guerrero (with chyna) to win the WWF European Championship (8:10); Saturn pinned Guerrero after a swimming shoulder fall .
SBMT (Xu et al., 2016)	Perry Saturn (with terri) beat Eddie Guerrero (with chyna) to win the WWF European League (8:10); Saturn pinned Guerrero after a diving elbow drop.
PBSMT-R (Wubben et al., 2012)	Perry Saturn with terri and Eddie Guerrero , chyna , to win the European Championship then-wwf 8:10); he pinned Guerrero after a diving elbow drop.
Original	Perry Saturn (with terri) defeated Eddie Guerrero (with chyna) to win the WWF European Championship (8:10); Saturn pinned Guerrero after a diving elbow drop.

(Nisioi et al., 2017)

Reinforcement Learning for TS (Zhang and Lapata, 2017)

- Encoder-decoder model embedded in reinforcement learning framework (DRESS)
- Motivation: generic encoder-decoder models favorise copying over rewriting
- DRESS rewards: simplicity, relevance, and fluency
- Trained and evaluated on three datasets: WikiSmall (Zhu et al., 2010), WikiLarge (Kauchak, 2013), and Newsela (Xu et al., 2015)
- Output compared with: PBMT-R (Wubben et al, 2010), Hybrid (Narayan and Gardent, 2014), SBMT-SARI (Xu et al., 2016),

Output Examples



Complex	There's just one major hitch: the primary purpose of education is to develop citizens with a wide variety of skills.
Reference	The purpose of education is to develop a wide range of skills.
PBMT-R	It's just one major hitch: the purpose of education is to make people with a wide variety of skills.
Hybrid	one hitch the purpose is to develop citizens.
EncDecA	The key of education is to develop people with a wide variety of skills.
DRESS	There's just one major hitch: the main goal of education is to develop people with lots of skills.
DRESS-LS	There's just one major hitch: the main goal of education is to develop citizens with lots of skills.
Complex	"They were so burdened by the past they couldn't think about the future," said Barnett, 62, who was president of Columbia Records, the No.1 record label in the United States, before joining Capitol.
Reference	Capitol was stuck in the past. It could not think about the future, Barnett said.
PBMT-R	"They were so affected by the past they couldn't think about the future," said Barnett, 62, was president of Columbia Records, before joining Capitol building .
Hybrid	"They were so burdened by the past they couldn't think about the future," said Barnett, 62, who was Columbia Records, president of the No.1 record label in the united states, before joining Capitol.
EncDecA	"They were so burdened by the past they couldn't think about the future," said Barnett, who was president of Columbia Records, the No.1 record labels in the United States.
DRESS	"They were so sicker by the past they couldn't think about the future," said Barnett, who was president of Columbia Records.
DRESS-LS	"They were so burdened by the past they couldn't think about the future," said Barnett, who was president of Columbia Records.

(Zhang and Lapata, 2017)

Comparison of NTS Systems (Štajner and Nisioi, 2018)

- Systems evaluated for:
 - The percentage of sentences which undergone at least one change;
 - The total number of changes;
 - The percentage of correct changes;
 - Grammaticality of the simplified sentence;
 - Meaning preservation of the simplified sentence;
 - Relative simplicity of the simplified sentence in comparison to the original sentence.

Evaluation results (Štajner and Nisioi, 2018)

Domain	Train-Test	Rerank	Hypoth.	Changed sent.	Total changes	Correct	G	M	S
In	News-News	default	h1	27.1%	23	21.7%	4.52	2.31	+0.02
In	News-News	SARI	h3	90.0%	76	54.1%	4.97	3.87	+0.50
In	Wiki-Wiki	default	h1	41.4%	37	48.6%	4.59	3.41	+0.30
	In	Wiki-Wiki	SARI	87.1%	78	59.0%	4.77	4.05	+0.49
Cross	Wiki-News	default	h1	47.1%	46	28.3%	3.87	2.55	+0.23
Cross	Wiki-News	SARI	h4	85.5%	77	25.0%	4.48	3.52	+0.21
Cross	News-Wiki	default	h1	40.0%	37	18.9%	3.86	2.90	+0.04
	Cross	News-Wiki	SARI	97.1%	102	23.4%	4.34	3.19	+0.28
SBMT (SARI+PPDB) (Xu et al., 2016)				82.9%	143	34.3%	4.28	3.57	+0.03
Dress-LS (Zhang and Lapata, 2017)				67.1%	63	42.9%	4.27	3.80	+0.14

Outputs

Ex.	System	Output
1a	Original-W, NTS-h1-any	She remained in the United States until 1927 when she and her husband returned to France.
1b	NTS-SARI-any	She stayed in the United States until 1927 when she and her husband returned to France.
1c	SBMT	She <i>is still</i> in the United States until 1927 when she and her husband returned to France.
1d	Dress-LS	She stayed in the United States until 1927 when she <i>was married</i> to France.
2a	Original-N, NTS-h1-any	Both newcomers and advanced learners trained together, but those with more experience were given more challenging training.
2b	NTS-SARI-in	Both newcomers and advanced learners trained together. However, those with more experience were given more challenging training.
2c	NTS-SARI-cross	Both newcomers and advanced <i>atheists</i> trained together, but those with more experience were given more challenging training.
3a	Original-W, NTS-h1-cross	Disney received a full-size Oscar statuette and seven miniature ones, presented to him by 10-year-old child actress Shirley Temple.
3b	SBMT	Disney won a full size Oscar statue and seven <i>the mini, made</i> to him by 10 year child actress Shirley Temple.
3c	NTS-h1-in	It was presented to him by 10-year-old child actress Shirley Temple.
3d	NTS-SARI-in, Dress-LS	Disney received a full-size Oscar statuette and seven miniature ones.
3e	NTS-SARI-cross	Disney received a full-size Oscar statuette and seven miniature ones. They presented to him by 10-year-old child actress Shirley Temple.
4a	Original-W, Dress-LS	At the Voyager 2 images Ophelia appears as an elongated object, the major axis pointing towards Uranus.
4b	NTS-h1-cross, NTS-SARI-in	At the Voyager 2 images Ophelia seems as an elongated object, the main axis show up on Uranus.
4c	NTS-h1-in	At the Voyager 2 images Ophelia appears as a stretched object, the major axis pointing towards Uranus.
4d	NTS-SARI-cross	At the Voyager 2 images Ophelia, the major axis pointing towards Uranus.
5a	Original-W, NTS-h1-out, SBMT	Graham attended Wheaton College from 1939 to 1943, when he graduated with a BA in anthropology.
5b	NTS-SARI-cross	Graham attended Wheaton College from 1939 to 1943. He graduated with a BA in anthropology.
5c	NTS-SARI-in, NTS-h1-in	Graham graduated from Wheaton College from 1939 to 1943.
5d	Dress-LS	Graham went to Wheaton College from 1939 to 1943.
6a	Original W, NTS h1 any	As a result, although many mosques will not enforce violations, both men and women when attending a mosque must adhere to these guidelines.
6b	Dress-LS	As a result, although many mosques will not enforce violations, both men and women.
6c	NTS-SARI-cross	As a result, many mosques will not enforce violations, both men and women when attending a mosque must follow these guidelines.
6d	NTS-SARI-in	As a result, although many mosques will not enforce violations, both men and women when attending a mosque must stick to these guidelines.
6e	SBMT	As a result, while many mosques will not meet the breach, both men and women when go to a mosque must meet these guidelines.

(Štajner and Nisioi, 2018)

Comparison of Fully-fledged Systems: Example 1

Original, Angrosh et al. (2014), Woodsend and Lapata (2011):

They drove a **patrol** car onto the lawn **in an attempt to rescue her**.

EvLex, LexEv:

They drove a **police** car onto the lawn. ← content reduction

EventSimplify + Light LS = EvLex

LightLS + EventSimplify = LexEv

Angrosh et al. (2014) is a hybrid system

Woodsend and Lapata (2011a) is a supervised system based on EW-SEW

Comparison of Fully-fledged Systems: Example 2

Original, Woodsend and Lapata (2011):

Jonson was rushed to hospital but died from her **wounds**, Goodyear said.

Angrosh et al. (2014):

Goodyear said Jonson was rushed to hospital but died from her wounds.

EvLex, LexEv:

Jonson was rushed to hospital. **Jonson** died from her **injuries**.

← syntactic (reordering)

← syntactic (sentence splitting)

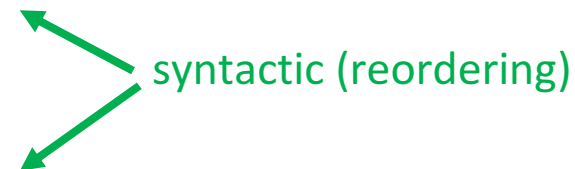
Comparison of Fully-fledged Systems: Example 3

Original:

“The ambassador’s arrival has not been announced and he flew in complete secrecy.” the official said.

Woodsend and Lapata (2011):

“The ambassador’s arrival has not been announced,” **the official said. He flew in complete secrecy.**



Angrosh et al. (2014):

The official said The ambassador’s arrival has not been announced. **And** he flew in complete secrecy.

EvLex, LexEv:

He **arrived** in complete secrecy. ← content reduction

Summary of the tutorial

- Text simplification is a complex task which requires considerable linguistic and world knowledge
- Automatic text simplification, although still imperfect, has the potential to serve a variety of users with special needs
- Text simplification has been addressed with a variety of techniques including rule-based methods, unsupervised approaches, and current/innovative data-driven techniques
- The techniques will depend on several factors such as availability of resources or what you are aiming for (e.g. just try a new approach or create a system for an end user)

Summary of the tutorial

- For the time being, and except for few works, text simplification is being approached at word and sentence, neglecting discourse issues such as cohesion and coherence
- There is much to be done to take text simplification research to the next level

Data-Driven Text Simplification

Sanja Štajner and Horacio Saggion

Many thanks for attending the tutorial !!!!

#TextSimplification2018

