

# Data-Driven Text Simplification

Sanja Štajner and Horacio Saggion

sanja.stajner@symanto.net

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

Symanto Research

Horacio.saggion@upf.edu

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

University Pompeu Fabra, Spain

#TextSimplification2019



# Presenters

---

- Sanja Stajner
  - <http://web.informatik.uni-mannheim.de/sstajner>
  - <https://www.linkedin.com/in/sanja-stajner-a6904738>
- Symanto Research
  - <https://www.symanto.net/>
- Horacio Saggion
  - <http://www.dtic.upf.edu/~hsaggion>
  - <https://www.linkedin.com/pub/horacio-saggion/16/9b9/174>
  - [https://twitter.com/h\\_saggion](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)
  - COLING 2018 (S. Štajner & H. Saggion)
- 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](https://www.morganclaypool.com/doi/abs/10.2200/S00700ED1V01Y201602_HLT032)

# Outline

---

- Motivation for ATS
- Automatic text simplification
- TS projects
- TS resources
- Neural text simplification
- Fully-fledged ATS systems

# **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

# Text Simplification

---

- Texts are sometimes too complicated: long sentences, uncommon vocabulary
  - **Wikipedia:** Opera is an art form in which singers and musicians perform a dramatic work combining text (*libretto*) and *musical score*, usually in a *theatrical setting*.

# Text Simplification

---

- Texts are sometimes too complicated: long sentences, uncommon vocabulary
  - **Wikipedia:** Opera is an art form in which singers and musicians perform a dramatic work combining text (*libretto*) and *musical score*, usually in a *theatrical setting*.
- Easy-to-read texts usually contain shorter sentences and common vocabulary





# Text Simplification

---

- Texts are sometimes too complicated: long sentences, uncommon vocabulary
  - **Wikipedia:** Opera is an art form in which singers and musicians perform a dramatic work combining text (*libretto*) and *musical score*, usually in a *theatrical setting*.
- Easy-to-read texts usually contain shorter sentences and common vocabulary
  - **Simple Wikipedia:** Opera is a drama set to music. An opera is like a play in which everything is sung instead of spoken.



# Text Simplification

---

Article 7

Los sindicatos de trabajadores y las asociaciones empresariales contribuyen a la defensa y promoción de los intereses económicos y sociales que les son propios.



# Text Simplification

---

## Article 7

Los sindicatos de trabajadores y las asociaciones empresariales contribuyen a la defensa y promoción de los intereses económicos y sociales que les son propios.

Los trabajadores defienden sus intereses a través de los sindicatos. Las empresas defienden sus intereses a través de sus asociaciones



# Text Simplification

---



TIME

French authorities began clearing out the sprawling Calais 'Jungle' camp on Monday, with the first government buses shuttling migrants to accommodation centers leaving the camp in the morning.



# Text Simplification

---



TIME



French authorities began clearing out the sprawling Calais 'Jungle' camp on Monday, with the first government buses shuttling migrants to accommodation centers leaving the camp in the morning.



United  
Response  
support that changes with you

A camp for migrants and refugees in France has been closed.  
The camp was near a town called Calais.



# Text Simplification

---



The Parliament of the United Kingdom, commonly known as the UK Parliament or British Parliament, is the supreme legislative body in the United Kingdom, British Crown dependencies and British overseas territories.

# Text Simplification

---



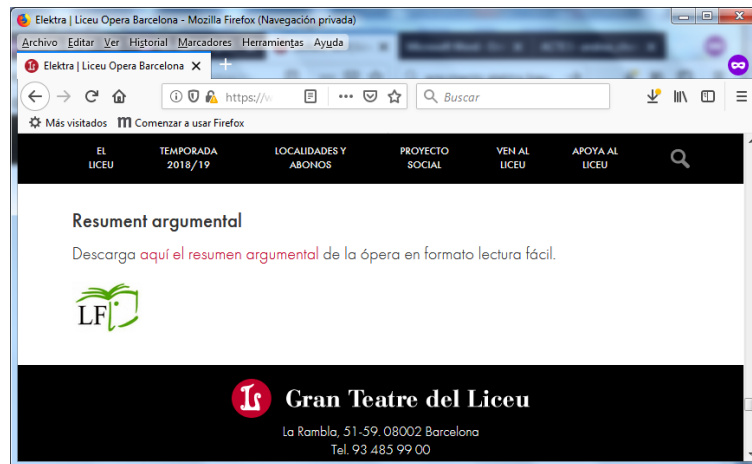
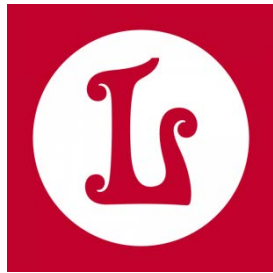
The Parliament of the United Kingdom, commonly known as the UK Parliament or British Parliament, is the supreme legislative body in the United Kingdom, British Crown dependencies and British overseas territories.

Parliament is a group of people who make laws and check what the Government is doing.



# Text Simplification

---

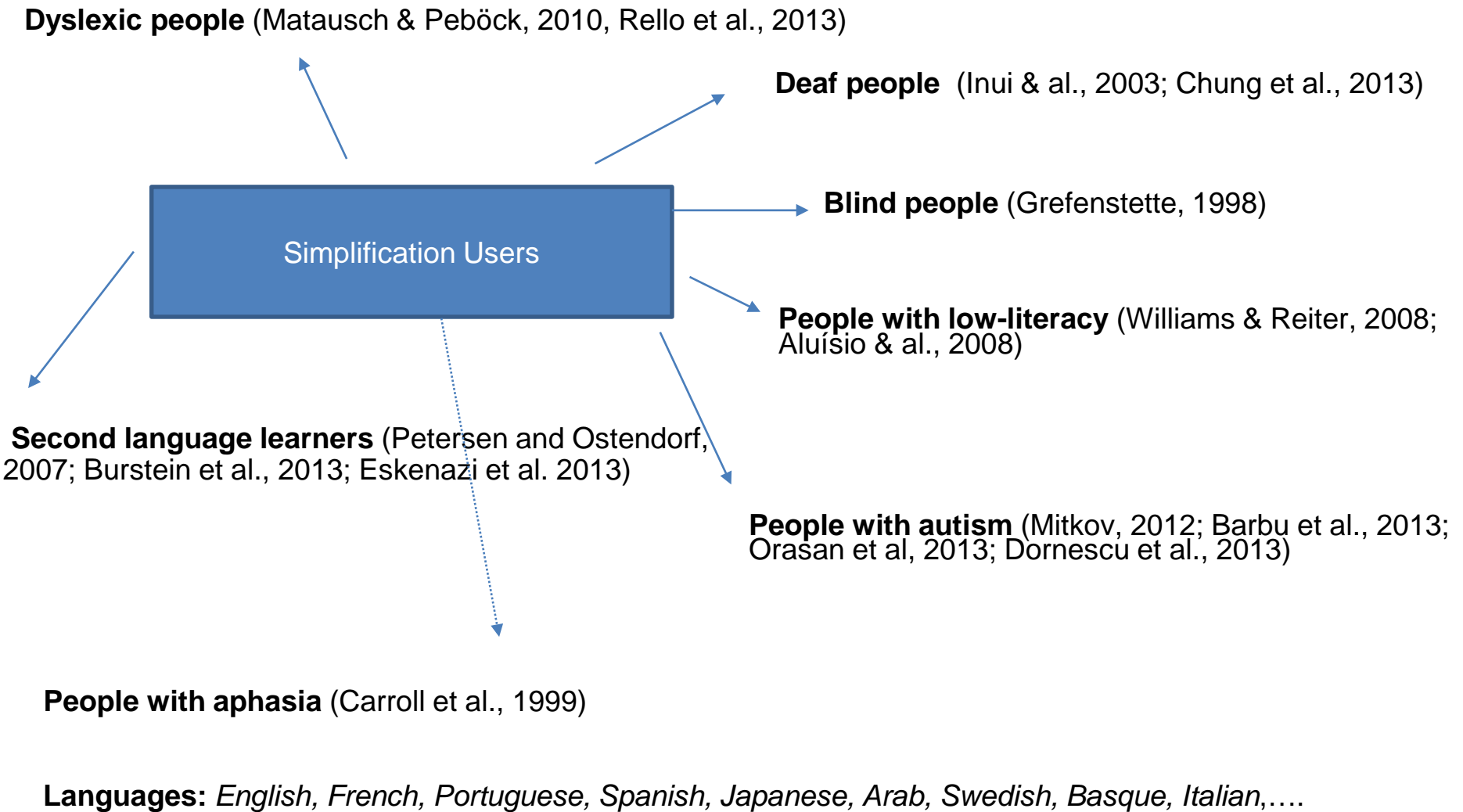




# Text Simplification

---

- Adapting the content of a text can be beneficial for different users
  - *language impairment*
  - *learning second language*
  - *non specialists in a domain*
- Simplification can also be used as pre-processing step for different NLP applications
  - Information Extraction
  - Summarization
  - Question Answering
- Manual adaptation is costly and time consuming



# What is Involved?

---

- Lexical Simplification:
  - *The play was **magnificent**.* => *The play was **great**.*
- Two main types:
  - All-in-one (Horn et al., 2014; Glavaš and Štajner, 2015)
  - Modular approach (Paetzold and Specia, 2016)
    1. Inventory of (quasi) synonyms or suitable substitutes (e.g. dictionary or word vector space)
    2. Process to decide which words need replacement (e.g. frequency, length, age of acquisition, machine learning) – Complex Word Identification
    3. Method to choose the most appropriate substitute (ranking candidates)
    4. Method to generate the substitute in context

# What is Involved?

---

- Syntactic Simplification:
  - *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.***
- 1. Inventory of syntactic phenomena which are an obstacle to comprehension / readability
- 2. Method to identify the above on sentences
- 3. Procedure to re-write/order sentences into simpler ones

# What is Involved?

---

- Discourse phenomena should also be taken into consideration
  - Sentence ordering, connectives, coreference phenomena, etc.
- Additional processes
  - Information enrichment
  - Summarization
  - Sentence compressions
  - Semantic paraphrasing
  - AAC, e.g. pictographic rendering
  - Information rendering

# The Basis for Building Simplification Systems

---

- For humans:
  - Psycholinguistic theories
  - Learning from parallel data
  - Eye-tracking
  - Human-informants (e.g. crowd sourcing)
- For NLP applications:
  - Ideally from systems' mistakes (e.g. IE system fails on certain types of sentences)

# Recap

---

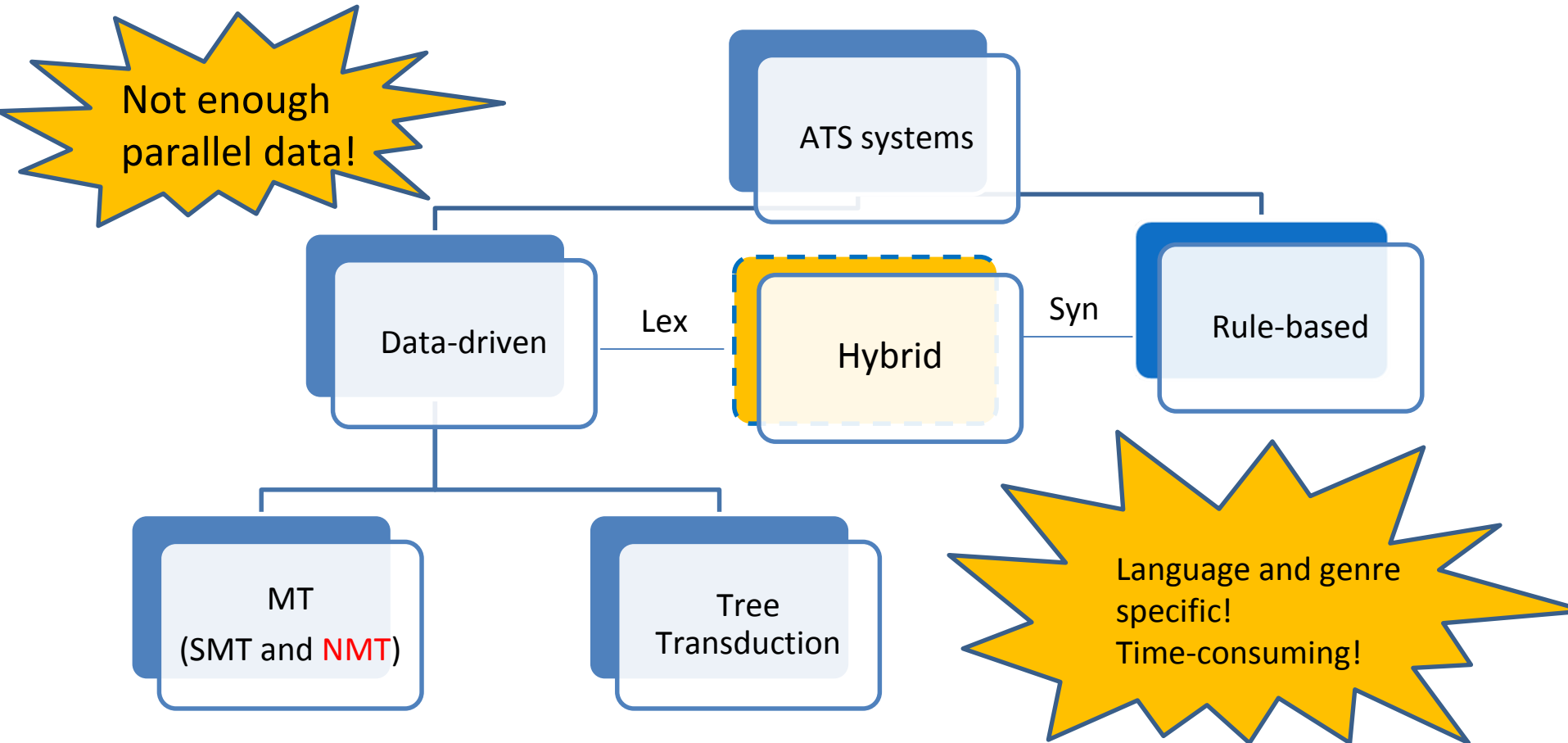
- Access to information has never been so easy, however the way in which information is transmitted can be a barrier to many people
- Simplification is a real world application which benefits a considerable number of people
- Simplification should consider different linguistics levels from words to sentences to discourse

# **PART 2**

## Approaches



# Approaches to ATS



# Evaluation of ATS Systems

## Evaluation of ATS systems

### Effectiveness/usefulness

- Reading speed
- Comprehension

### Quality of the output

#### Automatic evaluation

- Readability measures
- MT evaluation metrics (BLEU, METEOR, TERp)

(document level)

#### Human evaluation

- Grammaticality
- Meaning preservation
- Simplicity

(sentence level, 1-5 scale)

SOME NOT  
APPROPRIATE  
FOR SIMPLIFICATION

FKBLEU  
SARI

# Quality of the Output (Human Evaluation)

---

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

# Lexical Simplification 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/>

# Lexical Simplification

---

- Rule-based systems:
  - Manual: replace **A** by **a** in certain linguistic contexts
  - Data-driven: learn rules from data
    - Use traces of edit operations such as Simple Wikipedia edit histories (Yatskar et al. 2010)
      - $X_1 X_2 X_3 \mathbf{A} X_4 X_5 X_6 \rightarrow X_1 X_2 X_3 \mathbf{a} X_4 X_5 X_6$
      - Extract the most probable rules:  $\mathbf{A} \rightarrow \mathbf{a}$
    - Use “comparable data” such as Simple Wikipedia and English Wikipedia (Biran et al. (2011) )
      - Create vectors for words and compare them to detect pairs  $\langle \mathbf{A}, \mathbf{a} \rangle$  which *seem* related and thus interchangeable
      - Filter pairs by using WordNet
      - Decide simple/complex using frequencies and length
      - Use context to **assess fitness** of replacement

# Syntactic Simplification

---

- Rule-based systems
  - Set of rules to identify specific syntactic patterns
  - Set of procedures to re-generate text
- Data-driven systems
  - Learning transformations (statistically) using “comparable” parsing trees
  - Neural Machine Translation

# First steps: manual rules

---

- Rules over syntactic representations (Chandrasekar et al. 1996)
  - Superficial analysis (chunking) 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.
    - $W = \emptyset$
    - $X = \text{Xi Jinping}$
    - $RELPRO = \text{who}$
    - $Y = \text{is the current Paramount Leader of the People's Republic of China}$
    - $Z = \text{was visiting the USA}$
  - $\rightarrow$  Xi Jinping was visiting the USA. Xi Jinping is the current Paramount Leader of the People's Republic of China.

# 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 & Angrosch 2014)

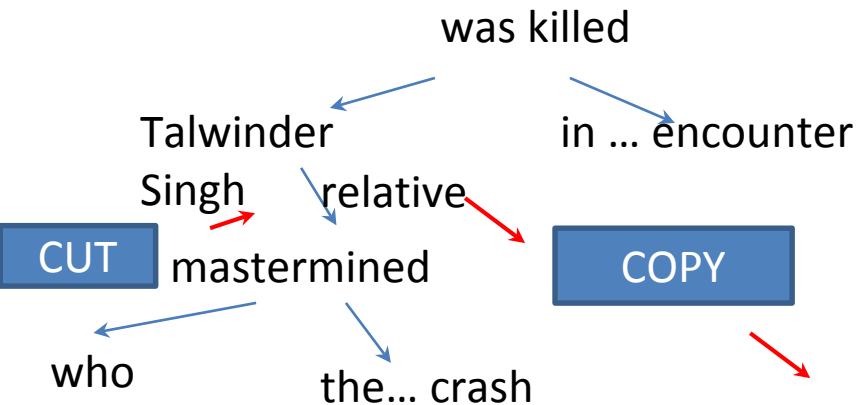


# First steps: rule learning

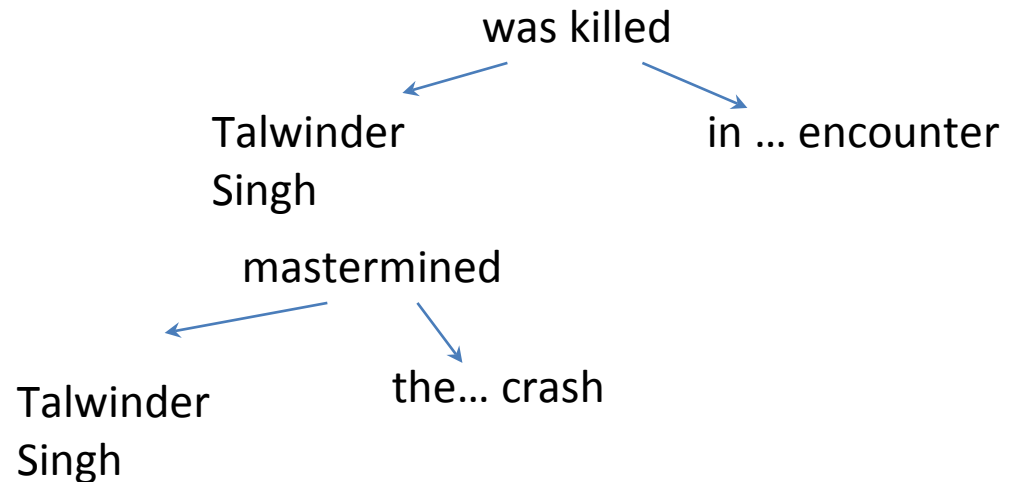
---

- 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.

## ORIGINAL



## SIMPLIFICATION



# Learning simplification from parsing trees

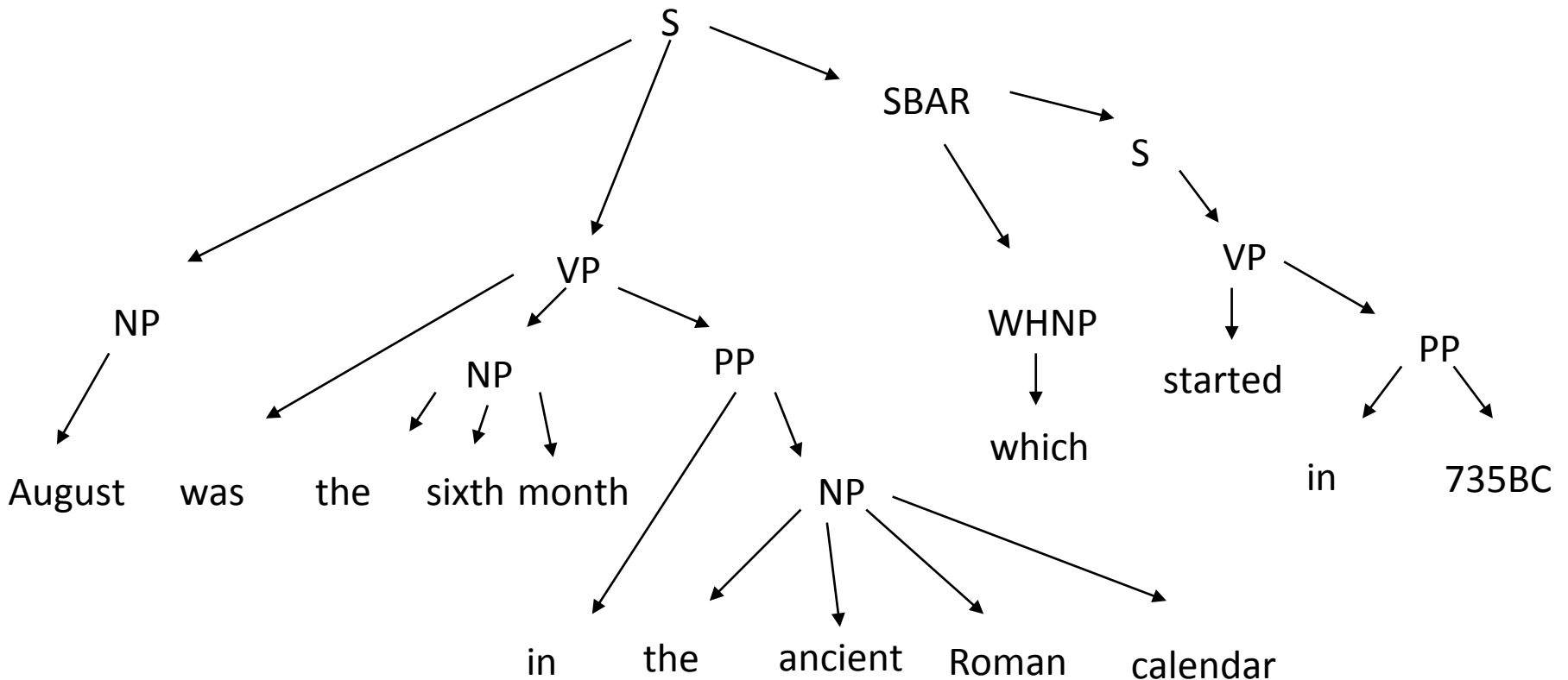
---

- 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

---

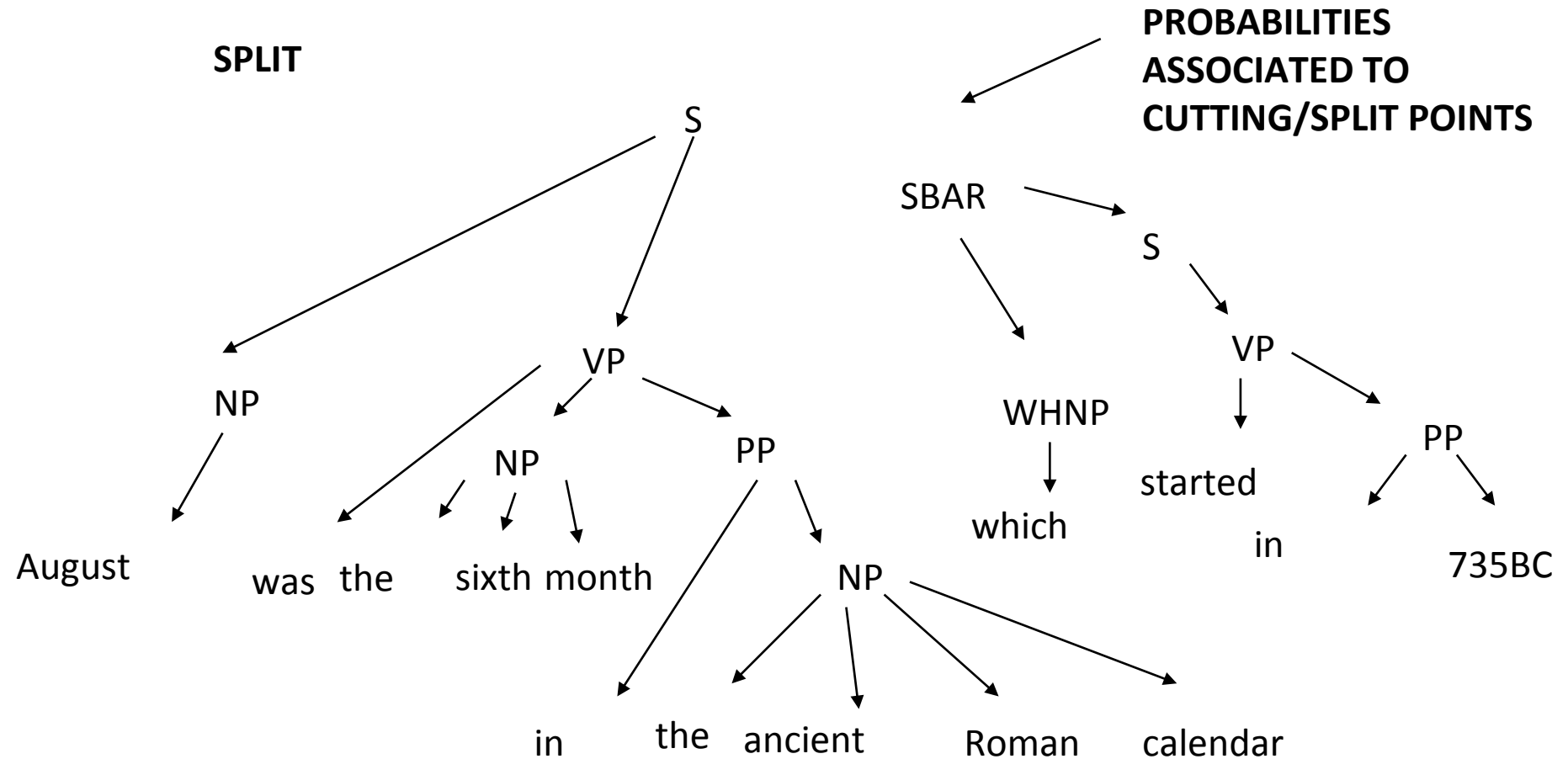
## PHRASE STRUCTURE OF COMPLEXT SENTENCE



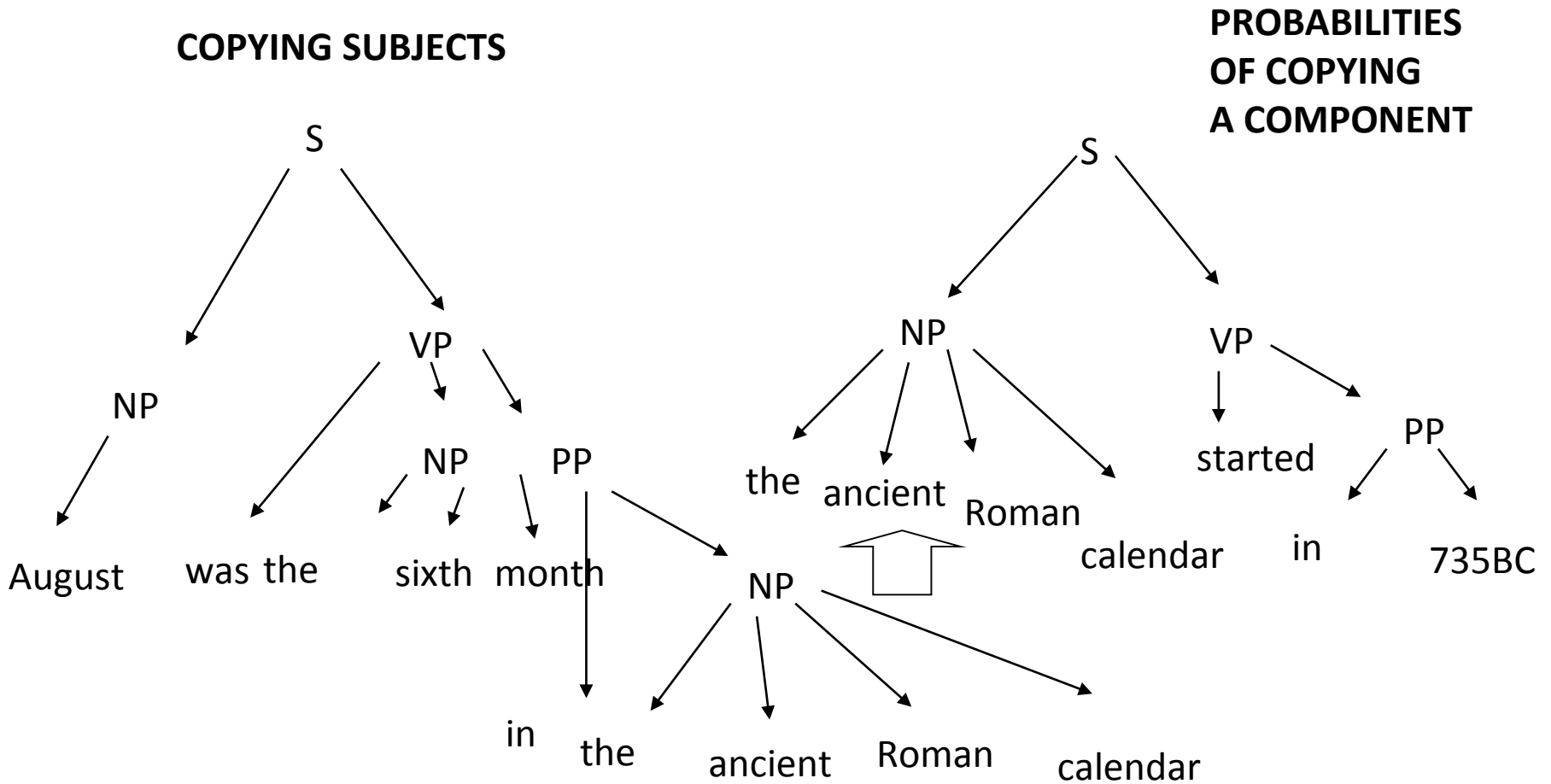
**August was the sixth month in the ancient Roman calendar which started in 735BC.**

---

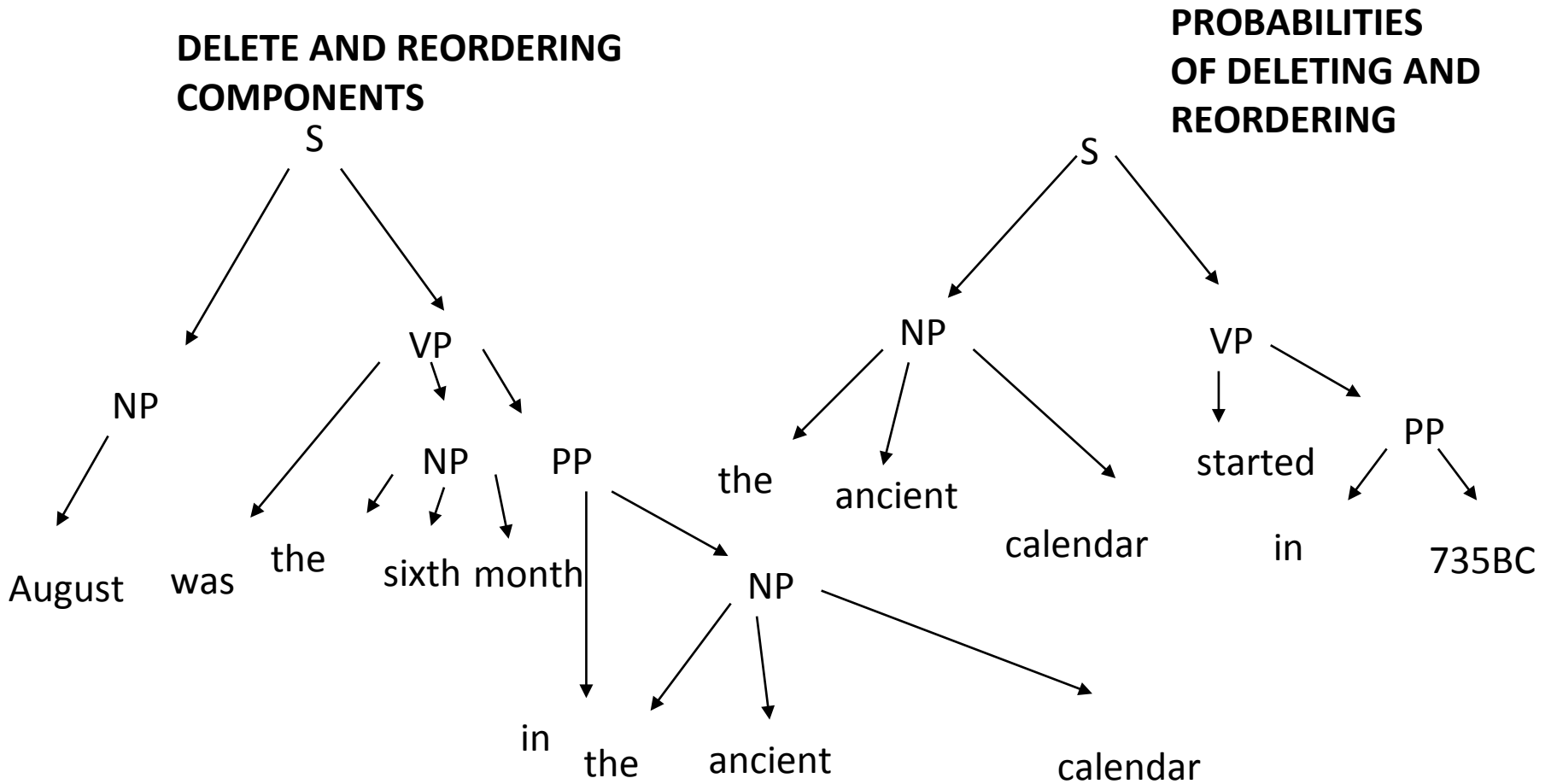
# Learning simplification from parsing trees



# Learning simplification from parsing trees

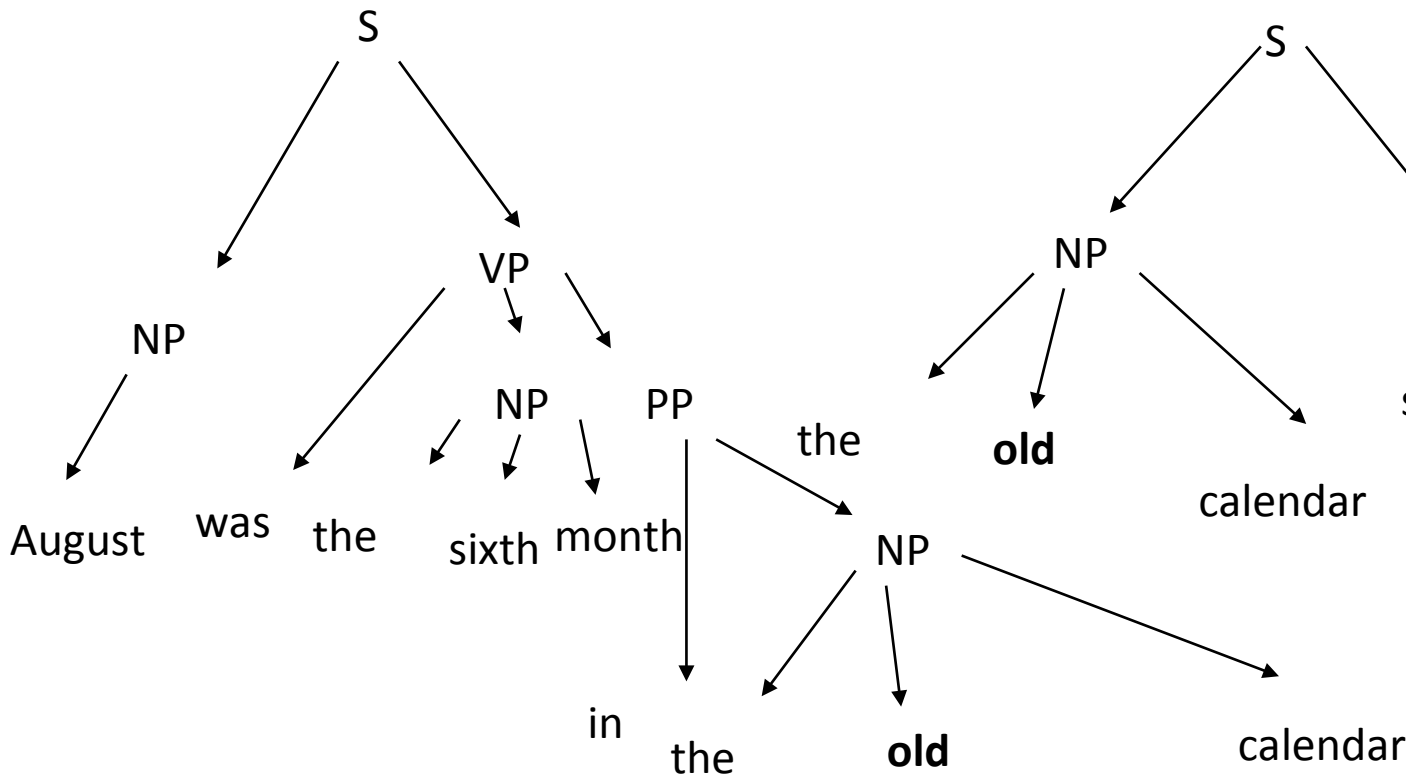


# Learning simplification from parsing trees

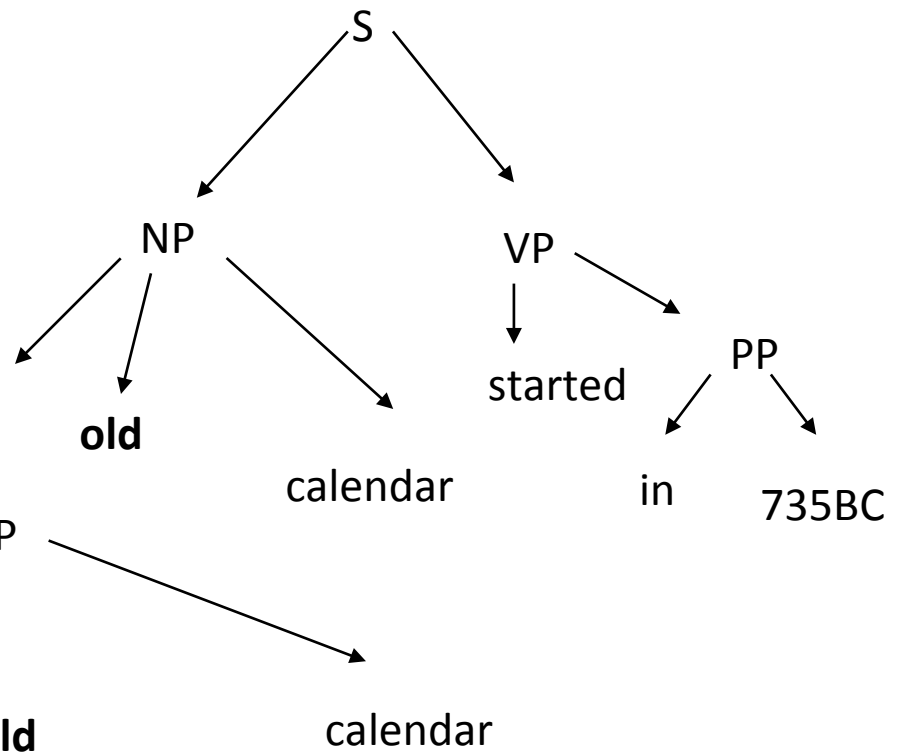


# Learning simplification from parsing trees

## WORD SUBSTITUTION



## PROBABILITY OF REPLACING A WORD



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

# 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
Grammaticality	0.68
Meaning	0.53
Simplicity	0.54

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	<b>2.70 ± 0.52</b>	<b>2.30 ± 0.54</b>
Pronominal anaphora	<b>2.68 ± 0.56</b>	<b>2.39 ± 0.57</b>

Relevance = harmonic mean of Meaning and Simplicity

# Recap

---

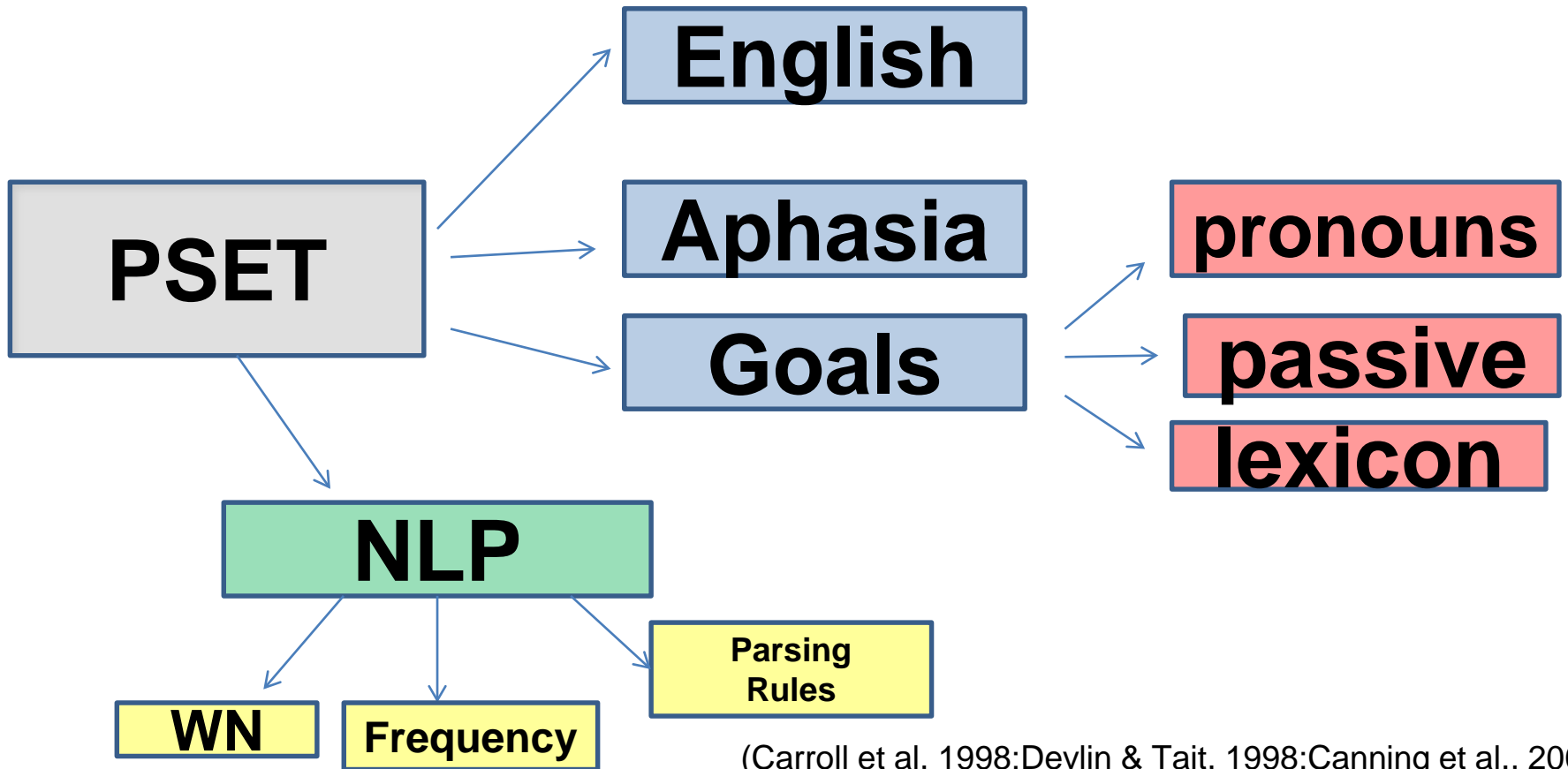
- (generally) Two main tasks addressed:
  - Lexical Simplification
  - Syntactic Simplification
- Approaches can be unsupervised and supervised
- Evaluation is a key aspect of text simplification with many angles (benchmarking, human evaluation, metrics)

# PART 3

## Simplification Projects

# Simplification Projects

---

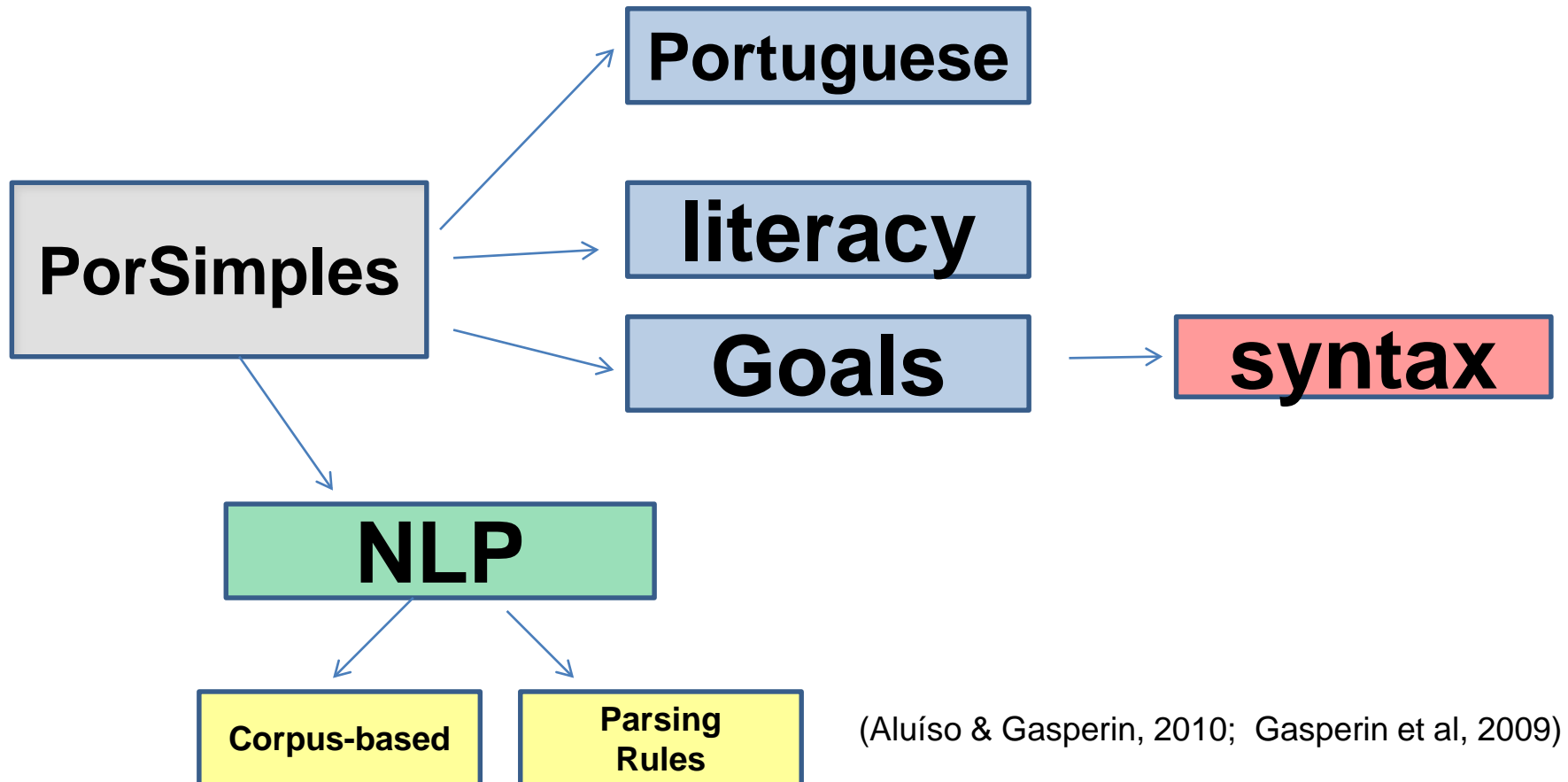


(Carroll et al, 1998;Devlin & Tait, 1998;Canning et al., 2000)



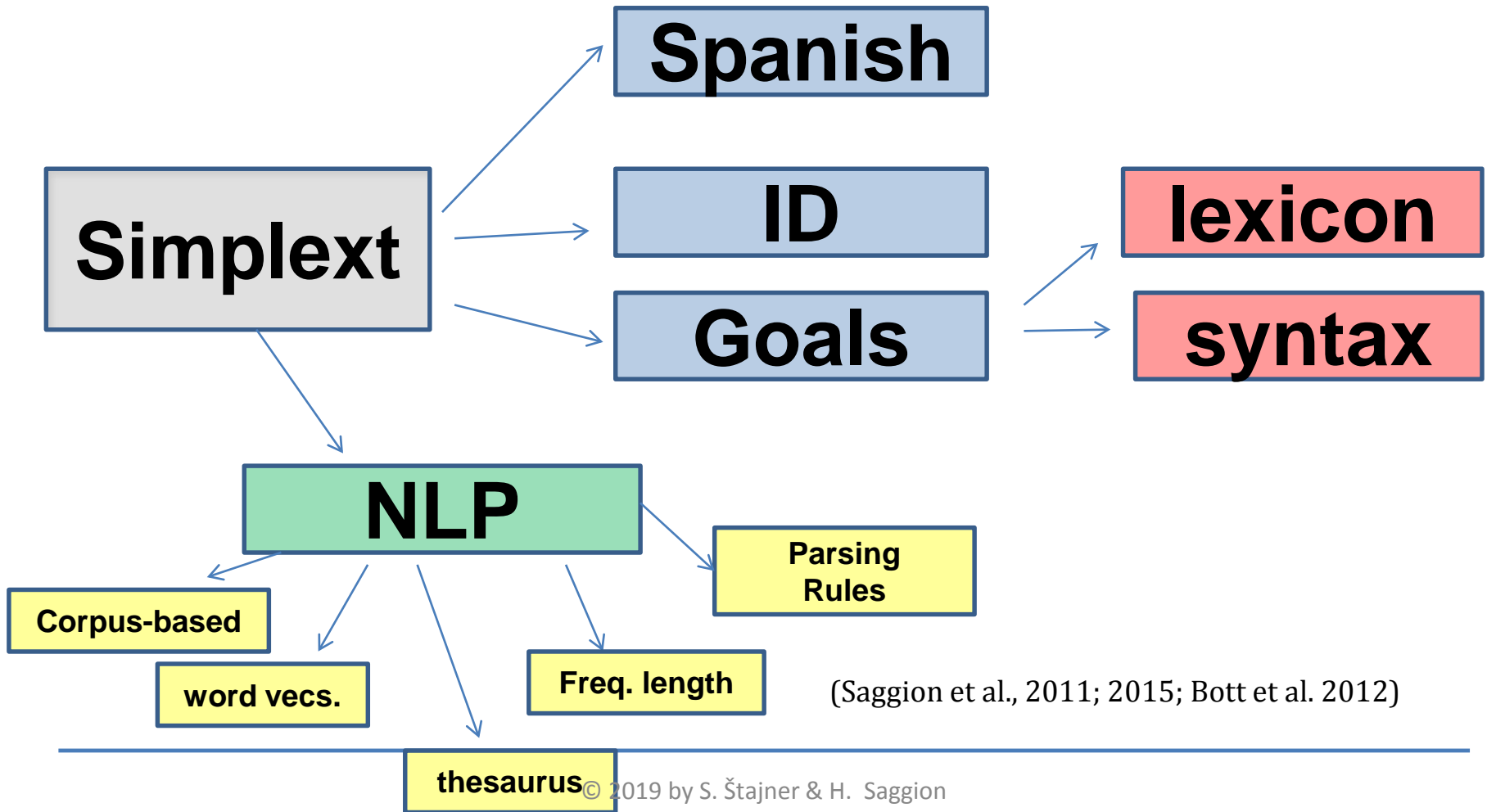
# Simplification Projects

---



# Simplification Projects

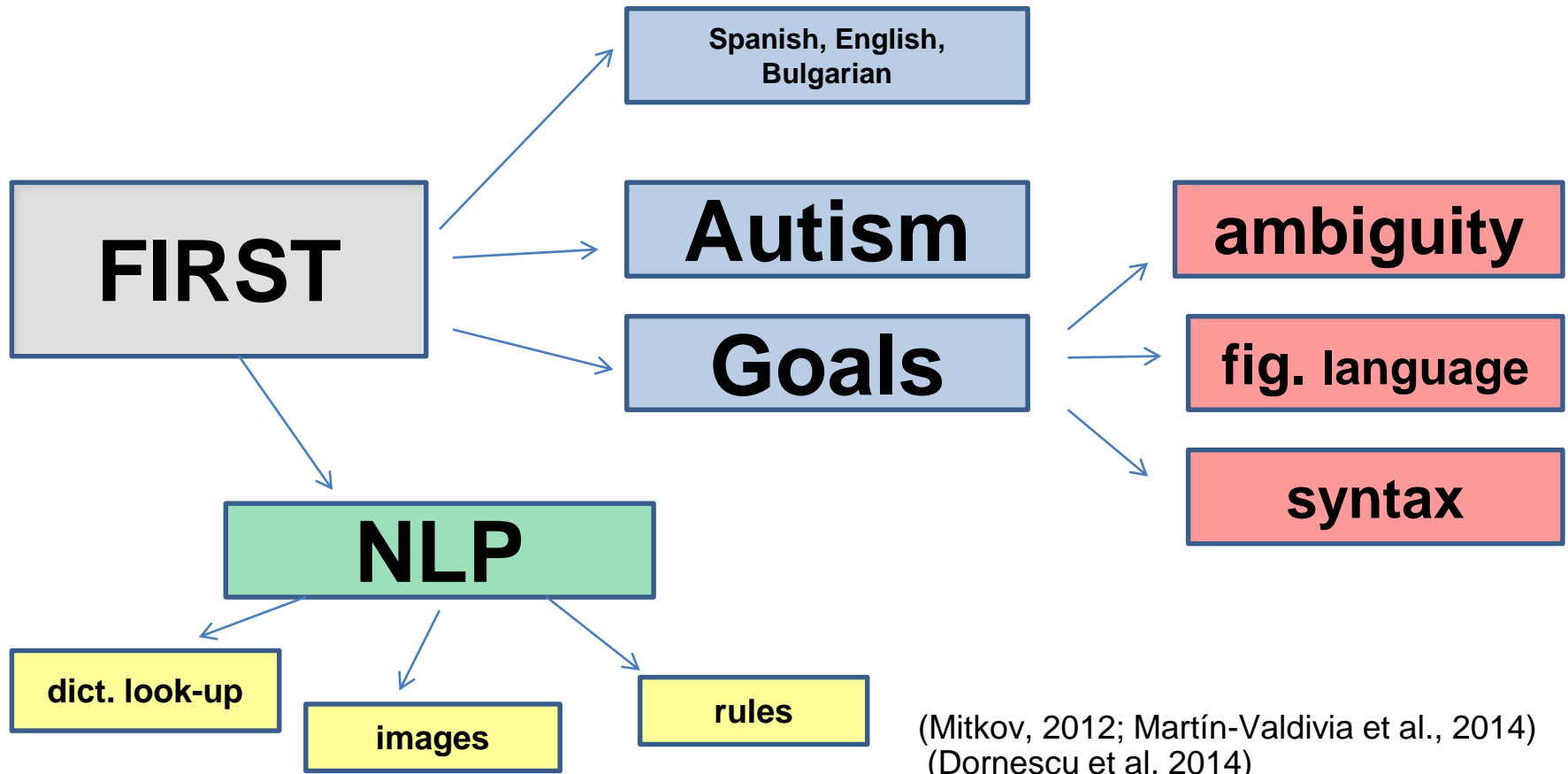
---



(Saggion et al., 2011; 2015; Bott et al. 2012)

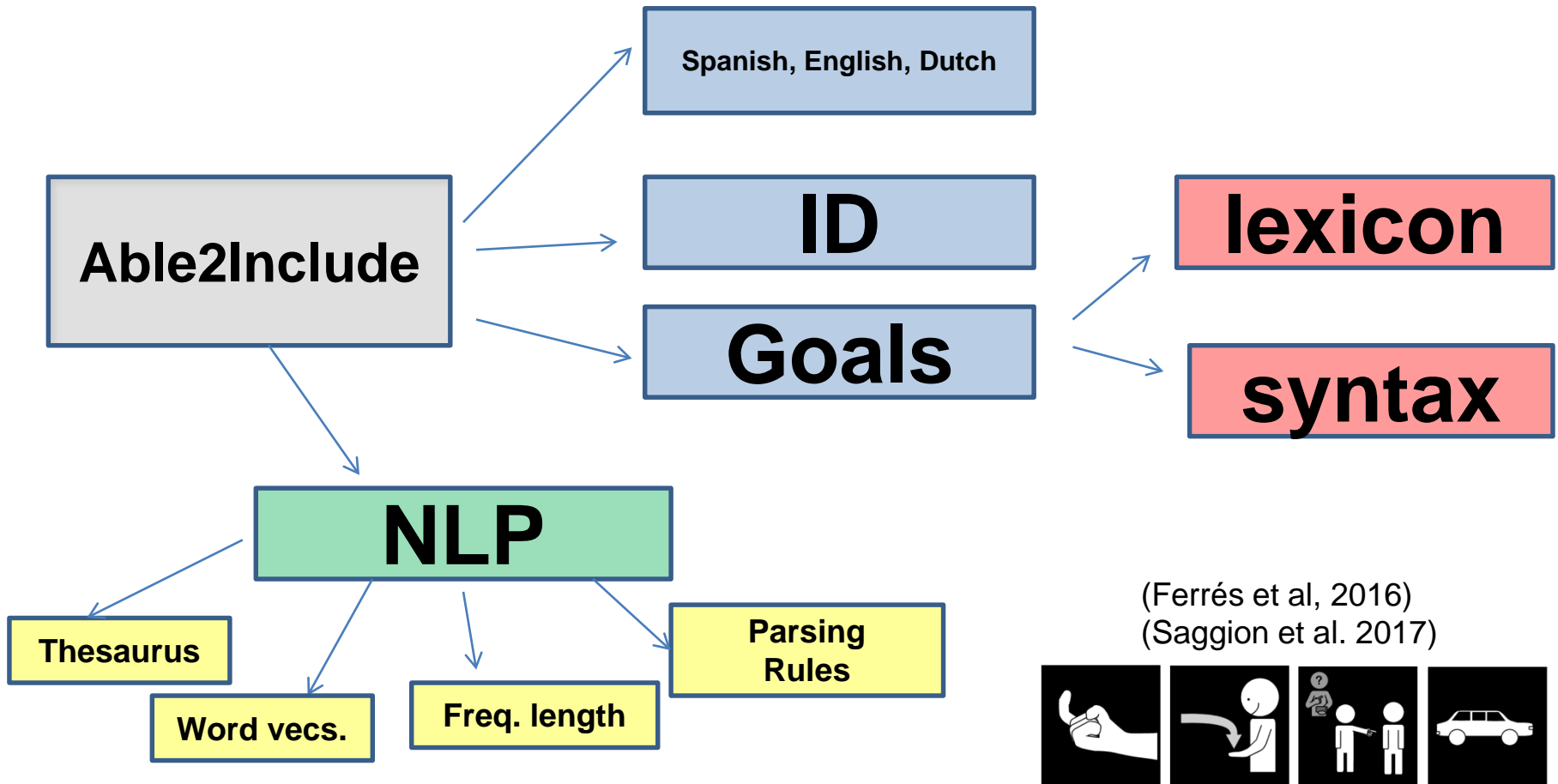
# Simplification Projects

---

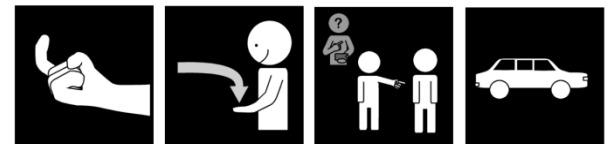


# Simplification Projects

---

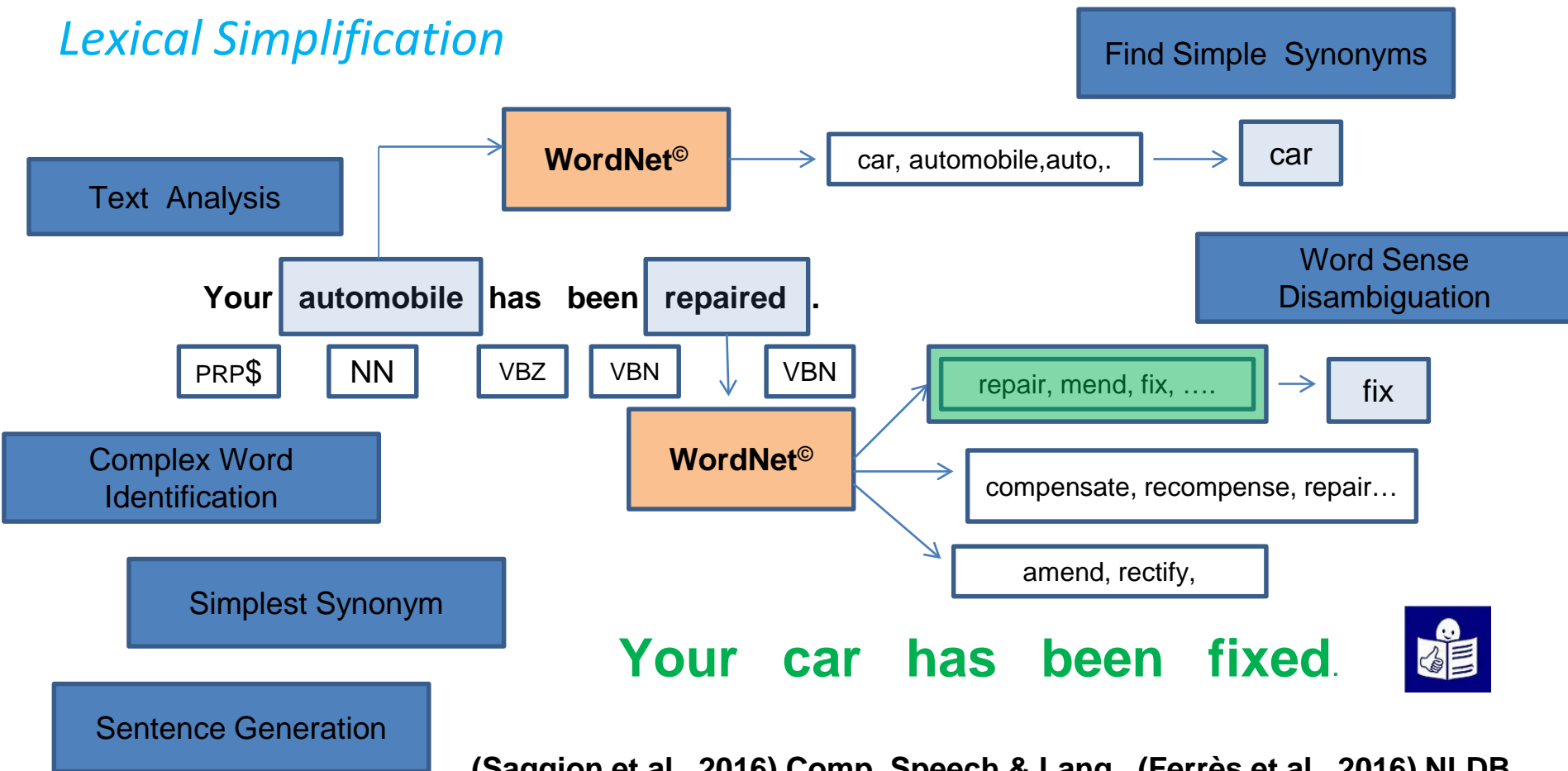


(Ferrés et al, 2016)  
(Saggion et al. 2017)



# Unsupervised Lexical Simplification with Lexical Resources






## Lexical Simplification



(Saggion et al., 2016) *Comp. Speech & Lang.* (Ferrès et al., 2016) *NLDB*

# Context-based lexical replacement

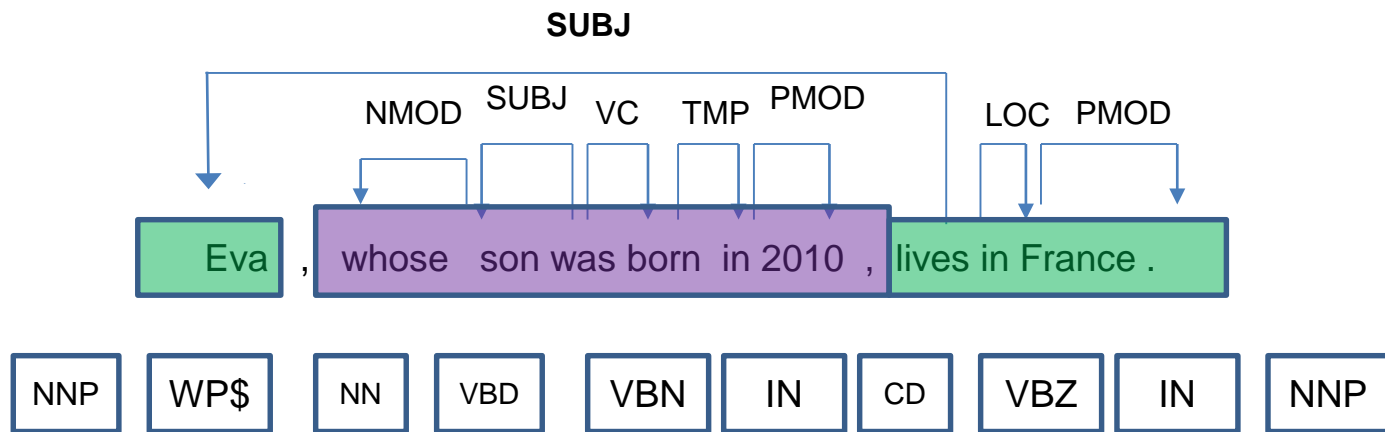
---

- Word inflexion and contexts are usually ignored
  - ... la **medicina** administrada... → ... la **remedio** administrada 
  - ... la **medicina** administrada... → ... el remedio administrado 
  - ...el **marit**... → ...el **home**... 
  - ...el **marit**... → ...l'**home**... 
  - ...le **sugerí**... → ...le **aconsejé**... 
- Robust (portable) morphological generator (Spanish, Catalan, Portuguese, Galician)
  - Rules+ Machine Learning

# Rule-based Syntactic Simplification

## Syntactic Simplification

Text Analysis



Simplification  
Grammars

Sentence  
Generation

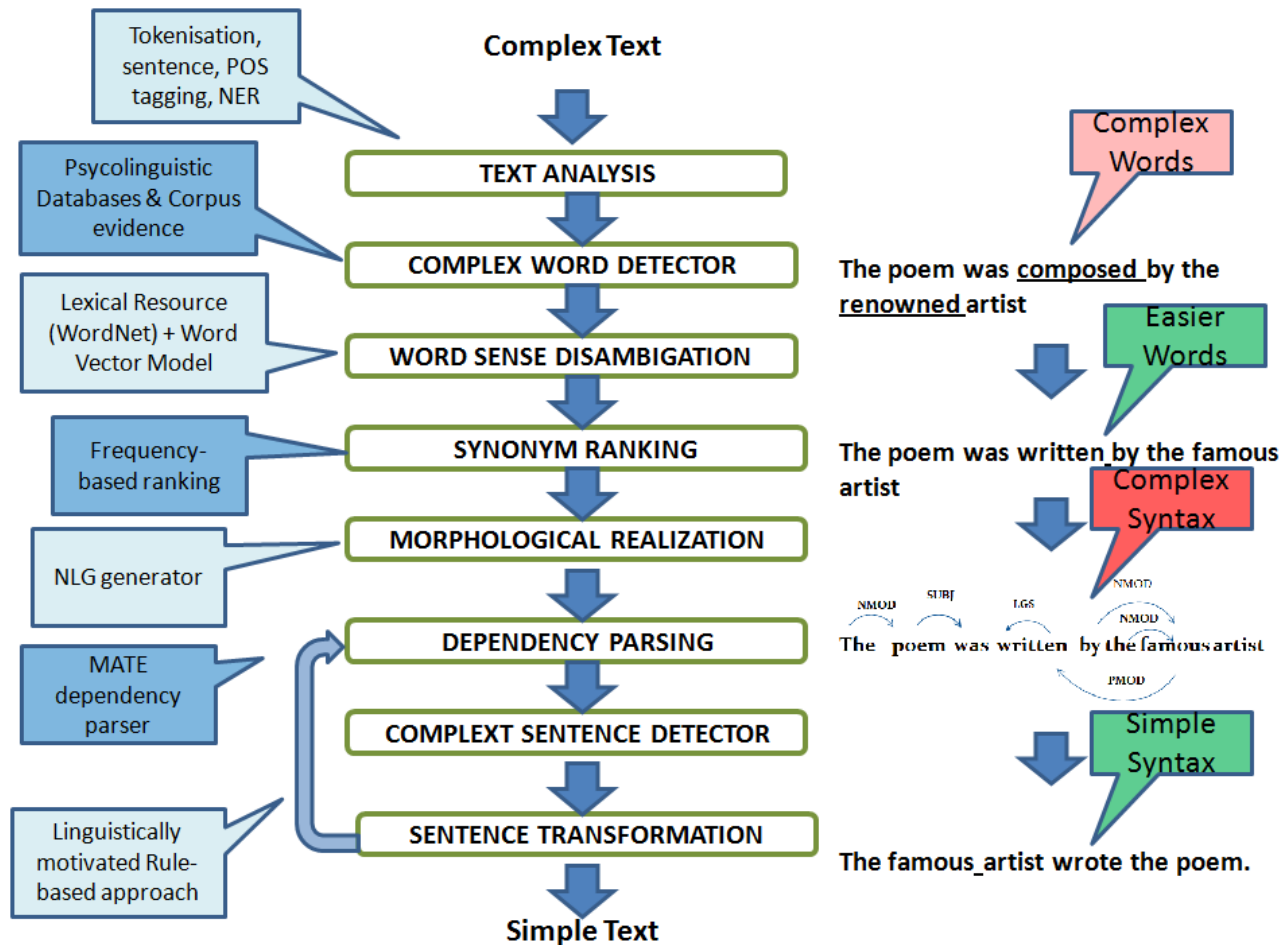
Eva lives in France .



The son of Eva was born in 2010 .

(Saggion et al., 2015) TACCESS ; (Ferrès et al., 2016) NLDB

# How a full system would look like....





# Recap

---

- Simplification projects usually had specific target users, although simplification was initially conceived as pre-processing for other NLP tasks
- In many occasions users/careers were available for testing the solution
- Projects were characterized by their multidisciplinaryity

# PART 4

## Resources

# 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 ↔ 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

# 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  $tf*idf + \text{cosine similarity}$
- 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.



# 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. (Bachelor of Fine Arts from the Polytechnic University of Valencia, Ana Juan is an illustrator, writer and painter.)</i>	<i>Ana Juan es ilustradora, escritora y pintora. Estudió Bellas Artes en la universidad de Valencia. (Ana Juan is an illustrator, writer and painter. She studied Fine Arts at the University of Valencia.)</i>
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”. ( 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>	<i>En 2011 se celebra el Año Internacional de la Química. (2011 marks the International Year of Chemistry.)</i>
3	<i>2011, AÑO INTERNACIONAL DE LA QUÍMICA. (2011, International Year of Chemistry. )</i>	<i>El 2011 es el Año Internacional de la Química. (2011 is the International Year of Chemistry.)</i>
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. ( 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>	<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. (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.)</i>

# Shared Tasks

---

- English Lexical Simplification (Specia et al., 2012)
  - <https://www.cs.york.ac.uk/semEval-2012/task1/index.html>
- Quality Assessment for Text Simplification (Stajner et al., 2016)
  - <http://qats2016.github.io/>
- Complex Word Identification – English (Paetzold and Specia, 2016)
  - <http://alt.qcri.org/semEval2016/task11/>
- Complex Word Identification – English, Spanish, German, French (Yimam et al., 2018)
  - <https://sites.google.com/view/cwisharedtask2018>

# 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>

# PorSimple 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> )
--------	---

# Recap

---

- Many linguistic resources for different languages
- Limited parallel data for text simplification (quality vs. quantity)
- Parallel data only for English, Italian, and Spanish



# PART 5

## Neural Approaches

# 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	<b>68.2</b>	<b>96.0</b>

# 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
<b>LIGHT-LS</b>	<b>0.540</b>

# 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%
<b>LIGHT-LS</b>	4.60	<b>3.76</b>	4.13	<b>68.6%</b>
Biran et al. Ch.	3.97	2.86	3.57	--
<b>LIGHT-LS Ch.</b>	<b>4.57</b>	<b>3.55</b>	<b>3.75</b>	--



# Example

---

Source	Sentence
Original	The <b>contrast</b> between a high level of education and a low level of political rights was <b>particularly</b> great in Aarau, and the city <b><u>refused</u></b> to send troops to <b>defend</b> the Bernese border.
Biran et al.	The <b>separate</b> 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 <b>especially</b> great in Aarau, and the city <b><u>asked</u></b> to send troops to <b>protect</b> 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	<b>1.000</b>
Kauchak	0.364	0.172	0.808
Glavas	0.456	0.197	0.741
<b>LS-NNS</b>	<b>0.464</b>	<b>0.226</b>	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

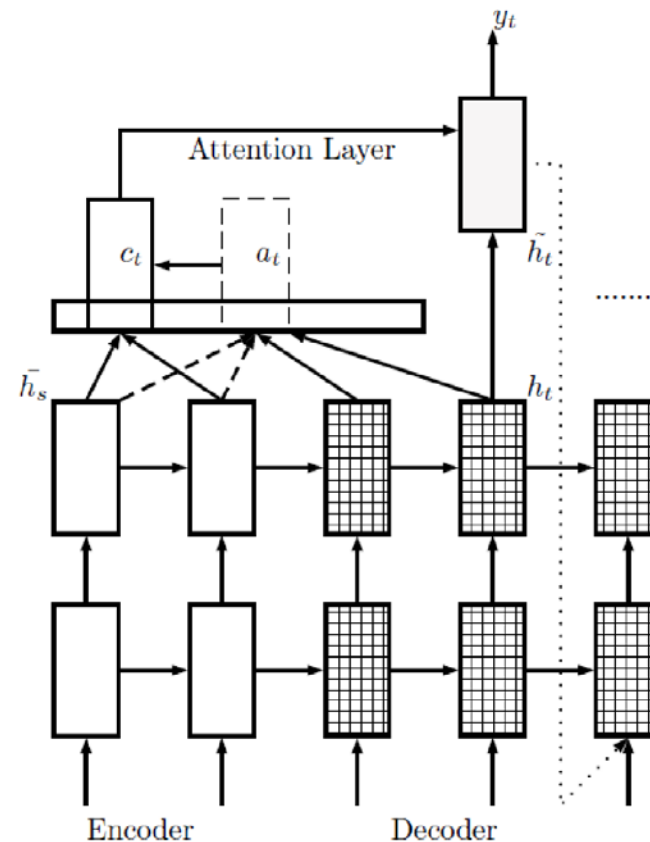
# Examples from EW-SEW (Hwang et al.)

---

Type	Original sentence	Simpler version
full	During the 13th century, gingerbread was brought to Sweden by German immigrants.	German immigrants brought it to Sweden during the 13th century.
partial	Gingerbread foods vary, ranging from a soft, moist loaf cake to something close to a ginger biscuit.	Gingerbread is a word which describes different sweet food products from soft cakes to a ginger biscuit.
partial	Humidity is the amount of water vapor in the air.	Humidity (adjective: humid) refers to water vapor in the air, but not to liquid droplets in fog, clouds, or rain.

# 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 (whole phrase is one change)
  - Correctness of changes (simpler and not damaging)
  - Grammaticality (1-5 scale, 3 annotators, native speakers)
  - Meaning preservation (1-5 scale, 3 annotators, native speakers)
  - Relative simplicity (semantic scale that corresponds to -2 to 2)

# 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

# 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

# Example

---

- Original:

Perry Saturn (with terry) defeated Eddie Guerrero (with chyna) **to win WWF European Championship (8:10); Saturn pinned Guerrero** after a diving elbow drop.

- NTS:

Perry Saturn **pinned Guerrero to win the WWF European Championship.**

# Recap

---

- NTS systems:
  - generate grammatical output
  - preserve the meaning well
  - lead to more changes on average
  - lead to higher percentage of correct changes
- Neural MT systems need to be adapted for TS:
  - adapted word embeddings
  - reranking the output (the default hypothesis is too conservative)
- Evaluation metric should be chosen depending on the target application or user.



# PART 6

## Fully Fledged Systems

# 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

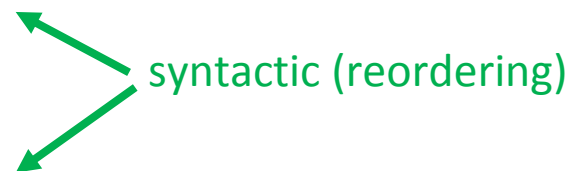
---

## 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 !!!!

#TextSimplification2019



## References

- [Agirre and Soroa(2009)] Eneko Agirre and Aitor Soroa. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics, 2009.
- [Aluisio et al.(2010)Aluisio, Specia, Gasperin, and Scarton] Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [Aluísio and Gasperin(2010)] Sandra Maria Aluísio and Caroline Gasperin. Fostering digital inclusion and accessibility: The PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [Amancio and Specia(2014)] Marcelo Adriano Amancio and Lucia Specia. An analysis of crowdsourced text simplifications. In *Proceedings of the Third Workshop on Predicting and Improving Text Readability for Target Reader Populations*, PITR, pages 123–130, Gothenburg, Sweden, 2014.
- [Anderson(1981)] Jonathan Anderson. Analysing the readability of English and non-English texts in the classroom with Lix. In *Proceedings of the Annual Meeting of the Australian Reading Association*, 1981.
- [Anula Rebollo(2008)] Ángel Alberto Anula Rebollo. Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad. In *La evaluación en el aprendizaje y la enseñanza del español como LE/L2, Pastor y Roca (eds.)*, pages 162–170, 2008.
- [Anula Rebollo(2009)] Ángel Alberto Anula Rebollo. Tipos de textos, complejidad lingüística y facilitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61, 2009.
- [Aranzabe et al.(2012)Aranzabe, de Ilarraza, and Gonzalez-Dios] María Jesús Aranzabe, Arantza Diaz de Ilarraza, and Itziar Gonzalez-Dios. First approach to automatic text simplification in Basque. In *Proceedings of the First Workshop on Natural Language Processing for Improving Textual Accessibility, NLP4ITA*, pages 1–8, 2012.
- [Aswani et al.(2007)Aswani, Tablan, Bontcheva, and Cunningham] Niraj Aswani, Valentin Tablan, Kalina Bontcheva, and Hamish Cunningham. Indexing and Querying Linguistic Metadata and Document Content. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, volume 292 of *Current Issues in Linguistic Theory*, pages 35–44. John Benjamins, Amsterdam & Philadelphia, 2007.



- [Baeza-Yates et al.(2015)Baeza-Yates, Rello, and Dembowski] Ricardo A. Baeza-Yates, Luz Rello, and Julia Dembowski. CASSA: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1380–1385, Denver, Colorado, USA, May 31 - June 5 2015.
- [Barbu et al.(2013)Barbu, Martín-Valdivia, and Ureña López] Eduard Barbu, Maria Teresa Martín-Valdivia, and Luis Alfonso Ureña López. Open Book: a tool for helping ASD users’ semantic comprehension. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility, NLP4ITA*, pages 11–19, Atlanta, Georgia, June 2013. Association for Computational Linguistics.
- [Barlacchi and Tonelli(2013)] Gianni Barlacchi and Sara Tonelli. ERNESTA: A sentence simplification tool for children’s stories in italian. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, volume 2 of *CICLING*, pages 476–487, 2013.
- [Barthe et al.(1999)Barthe, Juaneda, Leseigneur, Loquet, Morin, Escande, and Vayrette] Kathy Barthe, Claire Juaneda, Dominique Leseigneur, Jean-Claude Loquet, Claude Morin, Jean Escande, and Annick Vayrette. GIFAS rationalized French: A controlled language for aerospace documentation in French. *Technical Communication*, 46(2):220–229, 1999.
- [Barzilay and Elhadad(2003)] Regina Barzilay and Noemie Elhadad. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 25–32, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
- [Barzilay and Lapata(2008)] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, March 2008. ISSN 0891-2017.
- [Bautista and Saggion(2014)] Susana Bautista and Horacio Saggion. Making numerical information more accessible: The implementation of a Numerical Expression Simplification System for Spanish. *Special issue of the International Journal of Applied Linguistics*, 165(2):299–323, 2014.
- [Bautista et al.(2011)Bautista, León, Hervás, and Gervás] Susana Bautista, Carlos León, Raquel Hervás, and Pablo Gervás. Empirical identification of text simplification strategies for reading-impaired people. In *European Conference for the Advancement of Assistive Technology*, Maastricht, the Netherlands, September 2011.
- [Bengio et al.(2003)Bengio, Ducharme, Vincent, and Janvin] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3: 1137–1155, March 2003. ISSN 1532-4435.
- [Benjamin(2012)] Rebekah George Benjamin. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty.

*Educational Psychology Review*, 24(1):63–88, March 2012. ISSN 1040-726X.

- [Bernhard et al.(2012)Bernhard, De Viron, Moriceau, and Tannier] Delphine Bernhard, Louis De Viron, Véronique Moriceau, and Xavier Tannier. Question generation for French: Collating parsers and paraphrasing questions. *Dialogue and Discourse*, 3:43–74, 2012.
- [Biran et al.(2011)Biran, Brody, and Elhadad] Or Biran, Samuel Brody, and Noémie Elhadad. Putting it simply: A context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 496–501, Portland, Oregon, USA, 2011.
- [Bohnet(2009)] Bernd Bohnet. Efficient parsing of syntactic and semantic dependency structures. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL, pages 67–72, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [Bohnet et al.()]Bohnet, Langjahr, and Wanner] Bernd Bohnet, Andreas Langjahr, and Leo Wanner. A development environment for an MTT-based sentence generator.
- [Bosque Muñoz and Demonte Barreto(1999)] Ignacio Bosque Muñoz and Violeta Demonte Barreto. *Gramática descriptiva de la lengua española*. Real Academia Española, 1999.
- [Bott and Saggion(2011a)] Stefan Bott and Horacio Saggion. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 20–26, Stroudsburg, PA, USA, 2011a. Association for Computational Linguistics. ISBN 9781937284053.
- [Bott and Saggion(2011b)] Stefan Bott and Horacio Saggion. Spanish text simplification: An exploratory study. *Procesamiento del Lenguaje Natural*, 47:87–95, 2011b.
- [Bott et al.()]Bott, Rello, Drndarević, and Saggion] Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. Can Spanish be simpler? LexSiS: Lexical simplification for spanish.
- [Bott et al.(2012)Bott, Saggion, and Mille] Stefan Bott, Horacio Saggion, and Simon Mille. Text simplification tools for Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, LREC, pages 1665–1671, 2012.
- [Bouayad-Agha et al.(2009)Bouayad-Agha, Casamayor, Ferraro, Mille, Vidal, and Wanner] Nadjat Bouayad-Agha, Gerard Casamayor, Gabriela Ferraro, Simon Mille, Vanesa Vidal, and Leo Wanner. Improving the comprehension of legal documentation: The case of patent claims. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 78–87. ACM, 2009. ISBN 978-1-60558-597-0.

- [Brants and Franz(2006)] Thorsten Brants and Alex Franz. Web 1T 5-gram version 1 ldc2006t13. Web download. <https://catalog.ldc.upenn.edu/LDC2006T13>, 2006.
- [Breiman(2001)] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [Briscoe and Carroll(1995)] Ted Briscoe and John A. Carroll. Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In *Proceedings of the 4th International Workshop on Parsing Technologies*, pages 48–58, Prague, Czech Republic, 1995.
- [Brooke et al.(2012)Brooke, Tsang, Jacob, Shein, and Hirst] Julian Brooke, Vivian Tsang, David Jacob, Fraser Shein, and Graeme Hirst. Building readability lexicons with unannotated corpora. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, PITR, pages 33–39, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [Brouwers et al.(2014)Brouwers, Bernhard, Ligozat, and François] Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. Syntactic sentence simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, PITR, pages 47–56, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.
- [Bruce et al.(1981)Bruce, Rubin, and Starr] Bertram C. Bruce, Ann D. Rubin, and Kathleen S. Starr. Why readability formulas fail. *IEEE Transactions on Professional Communication*, PC-42:50–52, 1981.
- [Brunato et al.(2016)Brunato, Cimino, Dell’Orletta, and Venturi] Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, and Giulia Venturi. PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361, Austin, Texas, November 2016. Association for Computational Linguistics.
- [Burga et al.(2013)Burga, Codina, Ferraro, Saggion, and Wanner] Alicia Burga, Joan Codina, Gabriela Ferraro, Horacio Saggion, and Leo Wanner. The challenge of syntactic dependency parsing adaptation for the patent domain. In *Proceedings of the ESSLI Workshop on Extrinsic Parse Improvement*, 2013.
- [Burges(1998)] Christopher J. C. Burges. A tutorial on Support Vector Machines for pattern recognition. *Data Mining Knowledge Discovery*, 2(2): 121–167, June 1998. ISSN 1384-5810.
- [Burstein et al.(2013)Burstein, Sabatini, Shore, Moulder, and Lentini] Jill Burstein, John Sabatini, Jane Shore, Brad Moulder, and Jennifer Lentini. A user study: Technology to increase teachers’ linguistic awareness to improve instructional language support for English language learners. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, NLP4ITA, pages 1–10, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

- [Callison-Burch et al.(2011)Callison-Burch, Philipp Koehn, Monz, and Zaidan] Chris Callison-Burch, Philipp Philipp Koehn, Christof Monz, and Omar Zaidan. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, 2011.
- [Canning et al.(2000)Canning, Tait, Archibald, and Crawley] Yvonne Canning, John Tait, Jackie Archibald, and Ros Crawley. Cohesive generation of syntactically simplified newspaper text. In Petr Sojka, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 1902 of *Lecture Notes in Computer Science*, pages 145–150. Springer Berlin Heidelberg, 2000. ISBN 978-3-540-41042-3.
- [Carroll et al.(1998)Carroll, Minnen, Canning, Devlin, and Tait] John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI’98 Workshop on Integrating AI and Assistive Technology*, pages 7–10, 1998.
- [Celex(1993)] Celex. The CELEX lexical database. Centre for Lexical Information, Max Planck Institute for Psycholinguistics, 1993.
- [Chandrasekar and Srinivas(1997)] R. Chandrasekar and B. Srinivas. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10:183–190, 1997.
- [Chandrasekar et al.(1996)Chandrasekar, Doran, and Srinivas] R. Chandrasekar, Christine Doran, and B. Srinivas. Motivations and methods for text simplification. In *16th International Conference on Computational Linguistics*, pages 1041–1044, 1996.
- [Charniak(2000)] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL, pages 132–139, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
- [Church and Hanks(1990)] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, March 1990. ISSN 0891-2017.
- [Clarke and Lapata(2006)] James Clarke and Mirella Lapata. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 377–384, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- [Collins-Thompson(2014)] Kevyn Collins-Thompson. Computational assessment of text readability. A survey of current and future research. *Special issue of the International Journal of Applied Linguistics*, 165(2):97–135, 2014.

- [Collins-Thompson and Callan(2004)] Kevyn Collins-Thompson and James P. Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 193–200, 2004.
- [Collins-Thompson et al.(2011)Collins-Thompson, Bennett, White, de la Chica, and Sontag] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. Personalizing Web search results by reading level. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 403–412, 2011.
- [Colman(2016)] Andrew M. Colman. *Oxford Dictionary of Psychology (On-line Version)*. Oxford University Press, New York, 4rd edition edition, 2016.
- [Coster and Kauchak(2011)] William Coster and David Kauchak. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284053.
- [Crossley et al.(2007)Crossley, Dufty, McCarthy, and McNamara] Scott A. Crossley, David F. Dufty, Philip M. McCarthy, and Danielle S. McNamara. Toward a new readability: A mixed model approach. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pages 197–202, 2007.
- [Crystal(1987)] David Crystal. *The Cambridge Encyclopedia of Language*. Cambridge University Press, Cambridge, England, first edition, 1987.
- [Cunningham et al.(2000)Cunningham, Maynard, and Tablan] Hamish Cunningham, Diana Maynard, and Valentin Tablan. JAPE: a Java annotation patterns engine (second edition). Research Memorandum CS-00-10, Department of Computer Science, University of Sheffield, November 2000.
- [Curran et al.(2007)Curran, Clark, and Bos] James Curran, Stephen Clark, and Johan Bos. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, ACL, pages 33–36, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [Dale and Chall(1948a)] Edgar Dale and Jeanne S. Chall. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28, 1948a. ISSN 15554023.
- [Dale and Chall(1948b)] Edgar Dale and Jeanne S. Chall. The concept of readability. *Elementary English*, 23(24), 1948b.
- [Davison et al.(1980)Davison, Kantor, Kantor, Hermon, Lutz, and Salzillo] Alice Davison, Robert N. Kantor, Jean Hannah Kantor, Gabriela Hermon, Richard Lutz, and Robert Salzillo. Limitations of readability formulas in

guiding adaptations of texts. Technical Report 162, University of Illinois Center for the Study of Reading, Urbana, 1980.

- [De Belder(2014)] Jan De Belder. *Integer Linear Programming for Natural Language Processing*. PhD thesis, Informatics Section, Department of Computer Science, Faculty of Engineering Science, March 2014.
- [De Belder and Moens(2012)] Jan De Belder and Marie-Francine Moens. A dataset for the evaluation of lexical simplification. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7182 of *Lecture Notes in Computer Science*, pages 426–437. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-28600-1.
- [De Belder et al.(2010)De Belder, Deschacht, and Moens] Jan De Belder, Koen Deschacht, and Marie-Francine Moens. Lexical simplification. In *Proceedings of the 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*, Kortrijk, Belgium, 25-27 May 2010.
- [Dell’Orletta et al.(2011)Dell’Orletta, Montemagni, and Venturi] Felice Dell’Orletta, Simonetta Montemagni, and Giulia Venturi. READ-IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT, pages 73–83, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [Dell’Orletta et al.(2014a)Dell’Orletta, Montemagni, and Venturi] Felice Dell’Orletta, Simonetta Montemagni, and Giulia Venturi. Assessing document and sentence readability in less resourced languages and across textual genres. *Special Issue of the International Journal of Applied Linguistics*, 165(2):163–193, 2014a.
- [Dell’Orletta et al.(2014b)Dell’Orletta, Wieling, Venturi, Cimino, and Montemagni] Felice Dell’Orletta, Martijn Wieling, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. Assessing the readability of sentences: Which corpora and features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, Baltimore, Maryland, June 2014b. Association for Computational Linguistics.
- [Dempster et al.(1977)Dempster, Laird, and Rubin] A. P. Dempster, M. N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–22, 1977.
- [Devlin and Tait(1998)] Siobhan Devlin and John Tait. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173, 1998.
- [Devlin and Unthank(2006)] Siobhan Devlin and Gary Unthank. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 225–226, New York, NY, USA, 2006. ACM. ISBN 1-59593-290-9.

- [Dietterich(2000)] Thomas G. Dietterich. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 1–15, London, UK, 2000. Springer-Verlag. ISBN 3-540-67704-6.
- [Ding and Palmer(2005)] Yuan Ding and Martha Palmer. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL, pages 541–548, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [Dodgington(2002)] George Dodgington. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the 2nd Conference on Human Language Technology Research*, pages 138–145, San Diego, 2002.
- [Dras(1999)] Mark Dras. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. PhD thesis, Macquarie University, 1999.
- [Drndarević and Saggion(2012a)] Biljana Drndarević and Horacio Saggion. Reducing text complexity through automatic lexical simplification: An empirical study for Spanish. *Procesamiento del Lenguaje Natural*, 49:13–20, 2012a.
- [Drndarević and Saggion(2012b)] Biljana Drndarević and Horacio Saggion. Towards automatic lexical simplification in Spanish: An empirical study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, PITR, pages 8–16, Montréal, Canada, June 2012b. Association for Computational Linguistics.
- [Drndarević et al.(2012)Drndarević, Štajner, and Saggion] Biljana Drndarević, Sanja Štajner, and Horacio Saggion. Reporting simply: A lexical simplification strategy for enhancing text accessibility. In *Proceedings of the Easy-to-read on the Web Symposium*, 2012.
- [Drndarević et al.(2013)Drndarević, Štajner, Bott, Bautista, and Saggion] Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. Automatic text simplification in Spanish: A comparative evaluation of complementing modules. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 488–500, 2013.
- [DuBay(2004)] William H. DuBay. The principles of readability. *Impact Information*, pages 1–76, 2004.
- [Eisner(2003)] Jason Eisner. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, ACL, pages 205–208. The Association for Computer Linguistics, 2003.
- [Elhadad(2006)] Noemie Elhadad. Comprehending technical texts: Predicting and defining unfamiliar terms. In *AMIA Annual Symposium Proceedings*, pages 239–243, 2006.

- [Eskenazi et al.(2013)Eskenazi, Lin, and Saz] Maxine Eskenazi, Yibin Lin, and Oscar Saz. Tools for non-native readers: The case for translation and simplification. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, NLP4ITA, pages 20–28, Atlanta, Georgia, June 2013. Association for Computational Linguistics.
- [Evans et al.(2014)Evans, Orasan, and Dornescu] Richard Evans, Constantin Orasan, and Iustin Dornescu. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, PITR, pages 131–140, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.
- [Evans(2011)] Richard J. Evans. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing*, 26(4):371–388, 2011.
- [Fajardo et al.(2014)Fajardo, Ávila, Ferrer, Tavares, Gómez, and Hernández] Inmaculada Fajardo, Vicenta Ávila, Antonio Ferrer, Gema Tavares, Marcos Gómez, and Ana Hernández. Easy-to-read texts for students with intellectual disability: Linguistic factors affecting comprehension. *Journal of Applied Research in Intellectual Disabilities*, 27(3):212–225, 2014. ISSN 1468-3148.
- [Faruqui et al.(2015)Faruqui, Dodge, Jauhar, Dyer, Hovy, and Smith] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, pages 1606–1615, Denver, Colorado, May-June 2015. Association for Computational Linguistics.
- [Febowitz and Kauchak(2013)] Dan Febowitz and David Kauchak. Sentence simplification as tree transduction. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, PITR, pages 1–10, Sofia, Bulgaria, 2013.
- [Feng et al.(2009)Feng, Elhadad, and Huenerfauth] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 229–237, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [Feng et al.(2010)Feng, Jansche, Huenerfauth, and Elhadad] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING, pages 276–284, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [Ferraro(2012)] Gabriela Ferraro. *Towards deep content extraction from specialized discourse : The case of verbal relations in patent claims*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2012.



- [Ferrés et al.(2015)Ferrés, Marimon, and Saggion] Daniel Ferrés, Monserrat Marimon, and Horacio Saggion. A Web-based text simplification system for English. *Procesamiento del Lenguaje Natural*, 55:191–194, 2015.
- [Ferrés et al.(2016)Ferrés, Marimon, Saggion, and AbuRa’ed] Daniel Ferrés, Montserrat Marimon, Horacio Saggion, and Ahmed AbuRa’ed. YATS: Yet another text simplifier. In *Proceedings of the 21st International Conference on Applications of Natural Language to Information Systems*, pages 335–342, 2016.
- [Filippova and Strube(2008)] Katja Filippova and Michael Strube. Sentence fusion via dependency graph compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 177–185, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.
- [Flesch(1949)] Rudolf Flesch. *The Art of Readable Writing*. Harper, New York, 1949.
- [Flor and Klebanov(2014)] Michael Flor and Beata Beigman Klebanov. Associative lexical cohesion as a factor in text complexity. *ITL - International Journal of Applied Linguistics* 165:2, 165(2):223–258, 2014.
- [Freyhoff et al.(1998)Freyhoff, Hess, Kerr, Tronbacke, and Van Der Veken] Geert Freyhoff, Gerhard Hess, Linda Kerr, Bror Tronbacke, and Kathy Van Der Veken. *Make it Simple, European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability*. ILSMH European Association, Brussels, 1998.
- [Gala et al.(2013)Gala, François, and Fairon] Núria Gala, Thomas François, and Cédric Fairon. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *Proceedings of the eLex 2013 Conference: Electronic lexicography in the 21st century: thinking outside the paper*, Tallinn, Estonia, 2013.
- [Gale et al.(1992)Gale, Church, and Yarowsky] William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237, 1992. ISBN 1-55860-272-0.
- [Gasperin et al.(2009a)Gasperin, Maziero, Specia, Pardo, and Aluisio] Caroline Gasperin, Erick Maziero, Lucia Specia, Thiago Pardo, and Sandra M. Aluisio. Natural language processing for social inclusion: A text simplification architecture for different literacy levels. In *Proceedings of SEMISH-XXXVI Seminário Integrado de Software e Hardware*, pages 387–401, 2009a.
- [Gasperin et al.(2009b)Gasperin, Specia, Pereira, and Aluisio] Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Sandra M. Aluisio. Learning when to simplify sentences for natural text simplification. In *Encontro Nacional de Inteligência Artificial*, pages 809–818, 2009b.

- [Glavaš and Štajner(2013)] Goran Glavaš and Sanja Štajner. Event-centered simplification of news stories. In *Recent Advances in Natural Language Processing, RANLP 2013, 9-11*, pages 71–78, 2013.
- [Glavaš and Štajner(2015)] Goran Glavaš and Sanja Štajner. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, pages 63–68, July 26-31 2015.
- [Gonzalez-Dios et al.(2014)Gonzalez-Dios, Aranzabe, de Ilarraza, and Salaberri] Itziar Gonzalez-Dios, Maria Jesús Aranzabe, Arantza Diaz de Ilarraza, and Haritz Salaberri. Simple or complex? Assessing the readability of Basque texts. *Proceedings of the 5th International Conference on Computational Linguistics*, pages 334–344, 2014.
- [Graesser et al.(2004)Graesser, McNamara, Louwerse, and Cai] Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202, May 2004.
- [Gunning(1952)] Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.
- [Hall et al.(2009)Hall, Frank, Holmes, Pfahringer, Reutemann, and Witten] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145.
- [Halliday and Hasan(1976)] Michael A.K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.
- [Heilman and Smith(2010)] Michael Heilman and Noah A. Smith. Good question! Statistical ranking for question generation. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, pages 609–617, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5.
- [Heilman et al.(2007)Heilman, Collins, and Callan] Michael J. Heilman, Kevyn Collins, and Jamie Callan. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the Human Language Technology Conference*, 2007.
- [Horn et al.(2014)Horn, Manduca, and Kauchak] C. Horn, C. Manduca, and D. Kauchak. Learning a lexical simplifier using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL*, pages 458–463, 2014.
- [Inui et al.(2003)Inui, Fujita, Takahashi, Iida, and Iwakura] Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. Text

- simplification for reading assistance: A project note. In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*, pages 9–16, 2003.
- [Jauhar and Specia(2012)] Kumar Sujay Jauhar and Lucia Specia. UOW-SHEF: SimpLex – Lexical simplicity ranking based on contextual and psycholinguistic features. In *Proceedings of the Sixth International Workshop on Semantic Evaluation, SEMEVAL*, pages 477–481. Association for Computational Linguistics, 2012.
- [Joachims(1998)] Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning, ECML*, pages 137–142, London, UK, 1998. Springer-Verlag. ISBN 3-540-64417-2.
- [Joachims(2006)] Thorsten Joachims. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, pages 217–226, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5.
- [Jonnalagadda et al.(2009)] Jonnalagadda, Tari, Hakenberg, Baral, and Gonzalez] Siddhartha Jonnalagadda, Luis Tari, Jörg Hakenberg, Chitta Baral, and Graciela Gonzalez. Towards effective sentence simplification for automatic processing of biomedical text. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, pages 177–180, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [Kajiwara and Komachi(2016)] Tomoyuki Kajiwara and Mamoru Komachi. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 1147–1158, Osaka, Japan, December 11-16 2016.
- [Kamp(1981)] Hans Kamp. A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof, editors, *Formal Methods in the Study of Language*. 1981.
- [Kauchak(2013)] David Kauchak. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 1537–1546. The Association for Computer Linguistics, 2013. ISBN 978-1-937284-50-3.
- [Kennedy and Eberhart(1995)] James Kennedy and Russell C. Eberhart. Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 1942–1948, 1995.
- [Keselman et al.(2007)] Keselman, Slaughter, Arnott-Smith, Kim, Browne, and Zeng-Treitler] A. Keselman, L. Slaughter, C. Arnott-Smith, G. Kim, H. Divita, C. Browne, A. Tsai, and Q. Zeng-Treitler. Towards consumer-friendly PHRs: Patients experience with reviewing their health records. In *AMIA Annual Symposium Proceedings*, pages 399–403, 2007.

- [Keskisärkkä(2012)] Robin Keskisärkkä. Automatic text simplification via synonym replacement. Master’s thesis, Linköping University, 2012.
- [Kincaid et al.(2012)] Kincaid, Fishburne, Rogers, and Chissom] J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (Automated Readability Index, Fog count and Flesch Reading Ease Formula) for Navy enlisted personnel. Technical report, Naval Technical Training Command.
- [Klebanov et al.(2004)] Klebanov, Knight, and Marcu] Beata Beigman Klebanov, Kevin Knight, and Daniel Marcu. Text simplification for information-seeking applications. In *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, OTM Confederated International Conferences*, pages 735–747, Agia Napa, Cyprus, October 25-29 2004.
- [Klein and Manning(2003)] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL, pages 423–430, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
- [Klerke and Søgaard(2013)] Sigrid Klerke and Anders Søgaard. Simple, readable sub-sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop*, ACL, pages 142–149. The Association for Computational Linguistics, 2013.
- [Knight and Marcu(2002)] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, July 2002. ISSN 0004-3702.
- [Koehn et al.(2007)] Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics - Interactive Poster and Demonstration Sessions*, ACL, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
- [Krovetz(1998)] Robert Krovetz. More than One Sense Per Discourse. In *NEC Princeton NJ Labs., Research Memorandum*, 1998.
- [Kuperman et al.(2012)] Kuperman, Stadthagen-Gonzalez, and Brysbaert] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990, 2012.
- [Lal and Rürger(2002)] Partha Lal and Stefan Rürger. Extract-based summarization with simplification. In *Proceedings of the 2002 Document Understanding Conferences*, 2002.

- [Landauer and Dutnais(1997)] Thomas K. Landauer and Susan T. Dutnais. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.
- [Laufer and Nation(1999)] Batia Laufer and Paul Nation. A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1):33–51, 1999.
- [Lavelli et al.(2009)Lavelli, Hall, Nilsson, and Nivre] Alberto Lavelli, Johan Hall, Jens Nilsson, and Joakim Nivre. Maltparser at the EVALITA 2009 dependency parsing task. In *Proceedings of EVALITA*, 2009.
- [Lavoie and Rambow(1997)] Benoit Lavoie and Owen Rambow. A fast and portable realizer for text generation systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington DC*, pages 265–268, 1997.
- [Lin(2004)] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization*, Barcelona, 2004.
- [Lin and Hovy(2003)] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 71–78, Edmonton, Canada, 2003.
- [Lin(2003)] Dekang Lin. Dependency-based evaluation of Minipar. In Anne Abeill, editor, *Treebanks: Building and Using Parsed Corpora*, volume 20, pages 317–329. Springer Netherlands, Dordrecht, 2003.
- [Lin et al.(2012)Lin, Michel, Aiden, Orwant, Brockman, and Petrov] Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. Syntactic annotations for the Google Books Ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, ACL, pages 169–174, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [Malmasi et al.(2016)Malmasi, Dras, and Zampieri] Shervin Malmasi, Mark Dras, and Marcos Zampieri. LTG at SemEval-2016 task 11: Complex word identification with classifier ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 996–1000, 2016.
- [Mann and Thompson(1988)] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.
- [Manning et al.(2008)Manning, Raghavan, and Schtze] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

- [Marimon et al.(2015)Marimon, Saggion, and Ferrés] Montserrat Marimon, Horacio Saggion, and Daniel Ferrés. Porting a methodology for syntactic simplification from English to Spanish. In *Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software (IJCAI 2015)*, 2015.
- [Martín-Valdivia et al.(2014)Martín-Valdivia, Cámara, Barbu, López, Moreda, and Lloret] Maria Teresa Martín-Valdivia, Eugenio Martínez Cámara, Eduard Barbu, Luis Alfonso Ureña López, Paloma Moreda, and Elena Lloret. Proyecto FIRST (Flexible Interactive Reading Support Tool): Desarrollo de una herramienta para ayudar a personas con autismo mediante la simplificación de textos. *Procesamiento del Lenguaje Natural*, 53:143–146, 2014.
- [Maynard et al.(2002)Maynard, Tablan, Cunningham, Ursu, Saggion, Bontcheva, and Wilks] Diana Maynard, Valentin Tablan, Hamish Cunningham, Cristian Ursu, Horacio Saggion, Kalina Bontcheva, and Yorick Wilks. Architectural elements of language engineering robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274, 2002.
- [McCarthy and Navigli(2009)] Diana McCarthy and Roberto Navigli. The English lexical substitution task. *Language Resources and Evaluation*, 43(2): 139–159, 2009.
- [McLaughlin(1969)] Harry G. McLaughlin. SMOG grading - a new readability formula. *Journal of Reading*, pages 639–646, May 1969.
- [Medero and Ostendorf(2011)] Julie Medero and Mari Ostendorf. Identifying targets for syntactic simplification. In *Proceedings of the International Workshop on Speech and Language Technology in Education*, pages 69–72, Venice, Italy, August 24-26 2011.
- [Mel’čuk(1988)] Igor Mel’čuk. *Dependency Syntax: Theory and Practice*. State University of New York Press, 1988.
- [Mikolov et al.(2013)Mikolov, Yih, and Zweig] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association of Computational Linguistics*, NAACL, pages 746–751, Atlanta, Georgia, USA, June 9-14 2013.
- [Miller et al.(1990)Miller, Beckwith, Fellbaum, Gross, and Miller] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katheine J. Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.
- [Mitchell(1997)] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, first edition, 1997. ISBN 0070428077, 9780070428072.
- [Narayan and Gardent(2014)] Shashi Narayan and Claire Gardent. Hybrid simplification using deep semantics and machine translation. In *Proceedings*

of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL, pages 435–445, Baltimore, MD, USA, June 22-27 2014.

- [Navarro(2001)] Gonzalo Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, March 2001. ISSN 0360-0300.
- [Nisioi and Nauze(2016)] Sergiu Nisioi and Fabrice Nauze. An ensemble method for quality assessment of text simplification. In *Proceedings of the Workshop & Shared Task on Quality Assessment for Text Simplification (QATS)*, Portoroz, Slovenia, 2016.
- [Norbury(2005)] C.F. Norbury. Barking up the wrong tree? lexical ambiguity resolution in children with language impairments and autistic spectrum disorders. *Journal of Experimental Child Psychology*, 90:142–171, 2005.
- [Och and Ney(2003)] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March 2003. ISSN 0891-2017.
- [Ogden(1937)] Charles Kay Ogden. *Basic English: A General Introduction with Rules and Grammar*. Paul Treber, London, 1937.
- [Ong et al.(2007)Ong, Damay, Lojico, Lu, and Tarantan] Ethel Ong, Jerwin Damay, Gerard Lojico, Kimberly Lu, and Dex Tarantan. Simplifying text in medical literature. *Journal of Research in Science, Computing and Engineering*, 4(1):37–47, 2007.
- [Padró et al.(2010)Padró, Collado, Reese, Lloberes, and Castellón] Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. FreeLing 2.1: Five years of open-source language processing tools. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC, Valletta, Malta, May 19-21 2010.
- [Paetzold and Specia(2015)] Gustavo Paetzold and Lucia Specia. LEXenstein: A framework for lexical simplification. In *Proceedings of the ACL-IJCNLP 2015 System Demonstrations*, pages 85–90, Beijing, China, 2015.
- [Paetzold and Specia(2016a)] Gustavo Paetzold and Lucia Specia. SemEval 2016 Task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SEMEVAL*, pages 560–569, San Diego, California, June 2016a. Association for Computational Linguistics.
- [Paetzold and Specia(2016b)] Gustavo Paetzold and Lucia Specia. SV000gg at SemEval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SEMEVAL*, pages 969–974, San Diego, CA, USA, June 16-17 2016b.
- [Paetzold(2016)] Gustavo Henrique Paetzold. *Lexical Simplification for Non-Native English Speakers*. PhD thesis, The University of Sheffield, 2016.

- [Palotti et al.(2015)Palotti, Zuccon, and Hanbury] João Rafael de Moura Palotti, Guido Zuccon, and Allan Hanbury. The influence of pre-processing on the estimation of readability of Web documents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM*, pages 1763–1766, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3794-6.
- [Papineni et al.(2002)Papineni, Roukos, Ward, and Zhu] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
- [Pastra and Saggion(2003)] Katerina Pastra and Horacio Saggion. Colouring summaries BLEU. In *Proceeding of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, pages 35–42, April 14th 2003.
- [Pennington et al.(2014)Pennington, Socher, and Manning] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1532–1543, Doha, Qatar, October 25-29 2014.
- [Petersen and Ostendorf(2007)] Sarah E. Petersen and Mari Ostendorf. Text simplification for language learners: A corpus analysis. In *Proceedings of the Workshop on Speech and Language Technology in Education, SLaTE*, pages 69–72, Farmington, PA, USA, October 1-3 2007.
- [Pianta et al.(2008)Pianta, Girardi, and Zanolì] Emanuele Pianta, Christian Girardi, and Roberto Zanolì. The TextPro tool suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC, Marrakech, Morocco*, May 2008.
- [Pitler and Nenkova(2008)] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 186–195, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.
- [Quinlan(1993)] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. ISBN 1-55860-238-0.
- [Quinlan(1992)] P. Quinlan. *The Oxford Psycholinguistic Database*. Oxford University Press, 1992.
- [Quirk et al.(1985)Quirk, Greenbaum, Leech, and Svartvik] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language*. Longman Inc. New York, 1985.
- [Rello(2014)] Luz Rello. *DysWebxia. A Text Accessibility Model for People with Dyslexia*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2014.



- [Rello et al.(2013a)Rello, Baeza-Yates, Bott, and Saggion] Luz Rello, Ricardo A. Baeza-Yates, Stefan Bott, and Horacio Saggion. Simplify or help?: text simplification strategies for people with dyslexia. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility, W4A*, Rio de Janeiro, Brazil, May 13-15 2013a.
- [Rello et al.(2013b)Rello, Baeza-Yates, Dempere-Marco, and Saggion] Luz Rello, Ricardo A. Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Proceedings of the International Conference on Human-Computer Interaction (Part IV)*, INTERACT, pages 203–219, Cape Town, South Africa, September 2-6 2013b.
- [Rello et al.(2013c)Rello, Bautista, Baeza-Yates, Gervás, Hervás, and Saggion] Luz Rello, Susana Bautista, Ricardo A. Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. One half or 50%? An eye-tracking study of number representation readability. In *Proceedings of the International Conference on Human-Computer Interaction (Part IV)*, INTERACT, pages 229–245, Cape Town, South Africa, September 2-6 2013c.
- [Rodríguez Diéguez et al.(1993)Rodríguez Diéguez, Moro Berihuete, and Cabero Pérez] José Rodríguez Diéguez, Pilar Moro Berihuete, and Maria Cabero Pérez. Ecuaciones de predicción de lecturabilidad. *Enseñanza: anuario interuniversitario de didáctica*, 10-11:47–64, 1993. ISSN 2386-3927.
- [Ronzano et al.(2016)Ronzano, AbuRa’ed, Anke, and Saggion] Francesco Ronzano, Ahmed AbuRa’ed, Luis Espinosa Anke, and Horacio Saggion. TALN at SemEval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SEMEVAL*, pages 1011–1016, San Diego, CA, USA, June 16-17 2016.
- [Saggion et al.(.)Saggion, Gómez-Martínez, Etayo, Anula, and Bourg] Horacio Saggion, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. Text simplification in Simplext: Making text more accessible.
- [Saggion et al.(2015a)Saggion, Marimon, and Ferrés] Horacio Saggion, Montserrat Marimon, and Daniel Ferrés. Simplificación automática de textos para la accesibilidad de colectivos con discapacidad: experiencias para el español y el inglés. In *IX Jornadas Científicas Internacionales de Investigación sobre Personas con Discapacidad*, Salamanca, Spain, 2015a.
- [Saggion et al.(2015b)Saggion, Štajner, Bott, Mille, Rello, and Drndarević] Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarević. Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14, 2015b.

- [Saggion et al.(2016)Saggion, Bott, and Rello] Horacio Saggion, Stefan Bott, and Luz Rello. Simplifying words in context. Experiments with two lexical resources in Spanish. *Computer Speech & Language*, 35:200–218, 2016.
- [Sahlgren(2006)] Magnus Sahlgren. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. PhD thesis, Stockholm University, 2006.
- [Schmid(1994)] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.
- [Schrijver(1986)] Alexander Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1986. ISBN 0-471-90854-1.
- [Schwam and Ostendorf(2005)] Sarah E. Schwam and Mari Ostendorf. Reading level assessment using Support Vector Machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL, pages 523–530, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [Seretan(2012)] Violeta Seretan. Acquisition of syntactic simplification rules for French. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.
- [Shalev-Shwartz et al.(2007)Shalev-Shwartz, Singer, and Srebro] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 807–814, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3.
- [Shardlow(2013)] Matthew Shardlow. A comparison of techniques to automatically identify complex words. In *Proceedings of the ACL Student Research Workshop*, pages 103–109. The Association for Computer Linguistics, 2013.
- [Shardlow(2014)] Matthew Shardlow. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.
- [Si and Callan(2001)] Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM, pages 574–576, New York, NY, USA, 2001. ACM.

- [Siddharthan(2002)] Advaith Siddharthan. An architecture for a text simplification system. In *Proceedings of the Language Engineering Conference*, pages 64–71, 2002.
- [Siddharthan(2003)] Advaith Siddharthan. Preserving discourse structure when simplifying text. In *Proceedings of the 2003 European Natural Language Generation Workshop*, pages 103–110, 2003.
- [Siddharthan(2006)] Advaith Siddharthan. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109, 2006. ISSN 1570-7075.
- [Siddharthan(2011)] Advaith Siddharthan. Text simplification using typed dependencies: A comparison of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, ENLG, pages 2–11, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [Siddharthan and Mandya(2014)] Advaith Siddharthan and Angrosh Mandya. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 722–731, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.
- [Siddharthan et al.(2004)] Siddharthan, Nenkova, and McKeown] Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [Snover et al.(2006)] Snover, Dorr, Schwartz, Micciulla, and Makhoul] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, August 2006.
- [Snover et al.(2009)] Snover, Madnani, Dorr, and Schwartz] Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece, 2009.
- [Spaulding(1956)] Seth Spaulding. A Spanish readability formula. *The Modern Language Journal*, 40:433–441, 1956.
- [Specia(2010)] Lucia Specia. Translating from complex to simplified sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*, PROPOR, pages 30–39, Porto Alegre, RS, Brazil, April 27–30 2010.

- [Specia et al.(2012)Specia, Jauhar, and Mihalcea] Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. SemEval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics, SemEval*, pages 347–355, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [Srinivas(1997)] B. Srinivas. Performance evaluation of supertagging for partial parsing. In *Proceedings of the Fifth International Workshop on Parsing Technology*, Boston, USA, 1997.
- [Štajner et al.(2016)Štajner, Popovič, and Béchara] Sanja Štajner, Maja Popovič, and Hanna Béchara. Quality estimation for text simplification. In *Proceedings of the Workshop & Shared Task on Quality Assessment for Text Simplification (QATS)*, Portoroz, Slovenia, 2016.
- [Tai et al.(2015)Tai, Socher, and Manning] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured Long Short-term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, pages 1556–1566, Beijing, China, July 26-31 2015.
- [Štajner et al.(2012)Štajner, Evans, Orasan, , and Mitkov] S. Štajner, R. Evans, C. Orasan, , and R. Mitkov. What can readability measures really tell us about text complexity? In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility, NLP4ITA*, Istanbul, Turkey, May 27 2012.
- [Vajjala and Meurers(2014)] Sownya Vajjala and Detmar Meurers. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL - International Journal of Applied Linguistics* 165:2, 165(2):194–222, 2014.
- [Vandeghinste and Schuurman()] Vincent Vandeghinste and Ineke Schuurman. Linking pictographs to synsets: Sclera2Cornetto.
- [Vandeghinste et al.(2015)Vandeghinste, Schuurman, Sevens, and Van Eynde] Vincent Vandeghinste, Ineke Schuurman, Leen Sevens, and Frank Van Eynde. Translating text into pictographs. *Natural Language Engineering*, pages 1–28, 2015.
- [Štajner(2014)] Sanja Štajner. Translating sentences from 'original' to 'simplified' Spanish. *Procesamiento del Lenguaje Natural*, 53:61–68, 2014.
- [Štajner and Saggion(2013a)] Sanja Štajner and Horacio Saggion. Readability indices for automatic evaluation of text simplification systems: A feasibility study for Spanish. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 374–382, 2013a.
- [Štajner and Saggion(2013b)] Sanja Štajner and Horacio Saggion. Adapting text simplification decisions to different text genres and target users. *Procesamiento del Lenguaje Natural*, 51:135–142, 2013b.

- [Štajner et al.(2013)Štajner, Drndarević, and Saggion] Sanja Štajner, Biljana Drndarević, and Horacio Saggion. Eliminación de frases y decisiones de división basadas en corpus para simplificación de textos en español. *Computación y Sistemas*, 17(2), 2013.
- [Štajner et al.(2014)Štajner, Evans, and Dornescu] Sanja Štajner, Richard Evans, and Iustin Dornescu. Assessing conformance of manually simplified corpora with user requirements: The case of autistic readers. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society*, pages 53–63, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.
- [Štajner et al.(2015)Štajner, Béchara, and Saggion] Sanja Štajner, Hannah Béchara, and Horacio Saggion. A deeper exploration of the standard PB-SMT approach to text simplification and its evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, ACL, pages 823–828, Beijing, China, July 26-31 2015.
- [Štajner et al.(2016)Štajner, Popovič, Saggion, Specia, and Fishel] Sanja Štajner, Maja Popovič, Horacio Saggion, Lucia Specia, and Mark Fishel. Shared task on quality assessment for text simplification. In *Proceedings of the Workshop & Shared Task on Quality Assessment for Text Simplification (QATS)*, Portoroz, Slovenia, 2016.
- [Vu et al.(2014)Vu, Tran, and Pham] Tu Thanh Vu, Giang Binh Tran, and Son Bao Pham. Learning to simplify children stories with limited data. In *Intelligent Information and Database Systems*, LNAI, pages 31–41. Springer International Publishing, Switzerland, 2014.
- [W3C(2008)] W3C. *Web Content Accessibility Guidelines (WCAG) 2.0*, 2008. URL <http://www.w3.org/TR/WCAG20/>.
- [Walker et al.(2011)Walker, Siddharthan, and Starkey] Andrew Walker, Advait Siddharthan, and Andrew Starkey. Investigation into human preference between common and unambiguous lexical substitutions. In *Proceedings of the 13th European Workshop on Natural Language Generation*, ENLG, pages 176–180, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [Williams and Reiter(2005)] Sandra Williams and Ehud Reiter. Generating readable texts for readers with low basic skills. In *Proceedings of the 10th European Workshop on Natural Language Generation*, 2005.
- [Woodsend and Lapata(2011)] Kristian Woodsend and Mirella Lapata. Learning to simplify sentences with Quasi-synchronous grammar and integer programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 409–420, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4.

- [Wrobel(2016)] Krzysztof Wrobel. PLUJAGH at SemEval-2016 task 11: Simple system for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SEMEVAL*, pages 953–957, San Diego, CA, USA, June 16-17 2016.
- [Wu et al.(2012)Wu, Torii, Vijay-Shanker, Tudor, and Peng] Cathy H. Wu, Manabu Torii, K. Vijay-Shanker, Catalina O. Tudor, and Yifan Peng. iSimp: A sentence simplification system for biomedical text. In *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine, BIBM*, pages 1–6, Washington, DC, USA, 2012. IEEE Computer Society. ISBN 978-1-4673-2559-2.
- [Wu et al.(2010)Wu, Burges, Svore, and Gao] Qiang Wu, Christopher J. Burges, Krysta M. Svore, and Jianfeng Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, June 2010. ISSN 1386-4564.
- [Wubben et al.(2012)Wubben, van den Bosch, and Krahmer] Sander Wubben, Antal van den Bosch, and Emiel Krahmer. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL*, pages 1015–1024, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [Xu et al.(2015)Xu, Callison-Burch, and Napoles] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.
- [Yamada and Knight(2001)] Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, ACL*, pages 523–530, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.
- [Yaneva et al.(2016a)Yaneva, Temnikova, and Mitkov] Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. A corpus of text data and gaze fixations from autistic and non-autistic adults. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference, LREC*, 2016a.
- [Yaneva et al.(2016b)Yaneva, Temnikova, and Mitkov] Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference, LREC*, 2016b.
- [Yatskar et al.(2010)Yatskar, Pang, Danescu-Niculescu-Mizil, and Lee] Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 365–368, 2010.

- [Zajic et al.(2004)Zajic, Dorr, and Schwartz] David Zajic, Bonnie Dorr, and Richard Schwartz. BBN/UMD at DUC-2004: Ropiary. In *Proceedings of Document Understanding Conference*, 2004.
- [Zeng-Treitler et al.(2007)Zeng-Treitler, Goryachev, Kim, Keselman, and Rosendale] Qing Zeng-Treitler, Sergey Goryachev, Hyeoneui Kim, Alla Keselman, and Douglas Rosendale. Making texts in electronic health records comprehensible to consumers: A prototype translator. In *AMIA Annual Symposium Proceedings*, pages 846–850, 2007.
- [Zhang et al.(2004)Zhang, Vogel, and Waibel] Ying Zhang, Stephan Vogel, and Alex Waibel. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC. European Language Resources Association, 2004.
- [Zhu et al.(2010)Zhu, Bernhard, and Gurevych] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING, pages 1353–1361, Beijing, China, August 2010.