FONDAZIONE BRUNO KESSLER

DH DIGITAL HUMANITIES

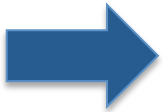
SIMPITIKI: a Simplification corpus for Italian

Sara Tonelli, *Alessio Palmero Aprosio, Francesca Saltori*
*{**satonelli**,aprosio,saltori}@fbk.eu*

Simpatico

Automatic text simplification

Qualora dal controllo dovesse emergere la non veridicità del contenuto della dichiarazione, il dichiarante decade dai benefici conseguiti sulla base della dichiarazione non veritiera



Chi rilascia una dichiarazione anche in parte falsa perde i benefici descritti

Issues:

- Approaches based on supervision are promising, but few resources other than **English**
- Need to understand which phenomena are involved to better understand how simplification can be automated and **evaluated**

Why not using Wikipedia edits?

- Inspired by Yatskar et al. (2010)
- Propose a **different approach** to create simplification corpora with respect to Brunato et al. (2015) but be consistent with their **annotation scheme**
- **Italian Wikipedia** is one of the most edited (35 active editors per million speakers)
- When sentences are modified, editors leave a **comment** to explain what they changed
- Data can be **re-distributed**

I Prussiani operarono l'isolamento delle linee telegrafiche

I Prussiani interruppero le comunicazioni via telegrafo

(*comment to edit: Frase semplificata*)

Corpus creation

- **Download** dump of Italian Wikipedia including history of every editing operation (> 60 million edits)
- Keep only pages edited with **comments** such as “semplice”, “semplificato”, …
- **Filter out** edits marked with “Template” and similar
- **Clean** markup
- **Automatically identify** pairs of text passages where changes were carried out (*DiffMatch* & *Patch library*)
- Final corpus dimension: **4,356 text pairs** with strings marked with deletion and insertion tags

Corpus annotation

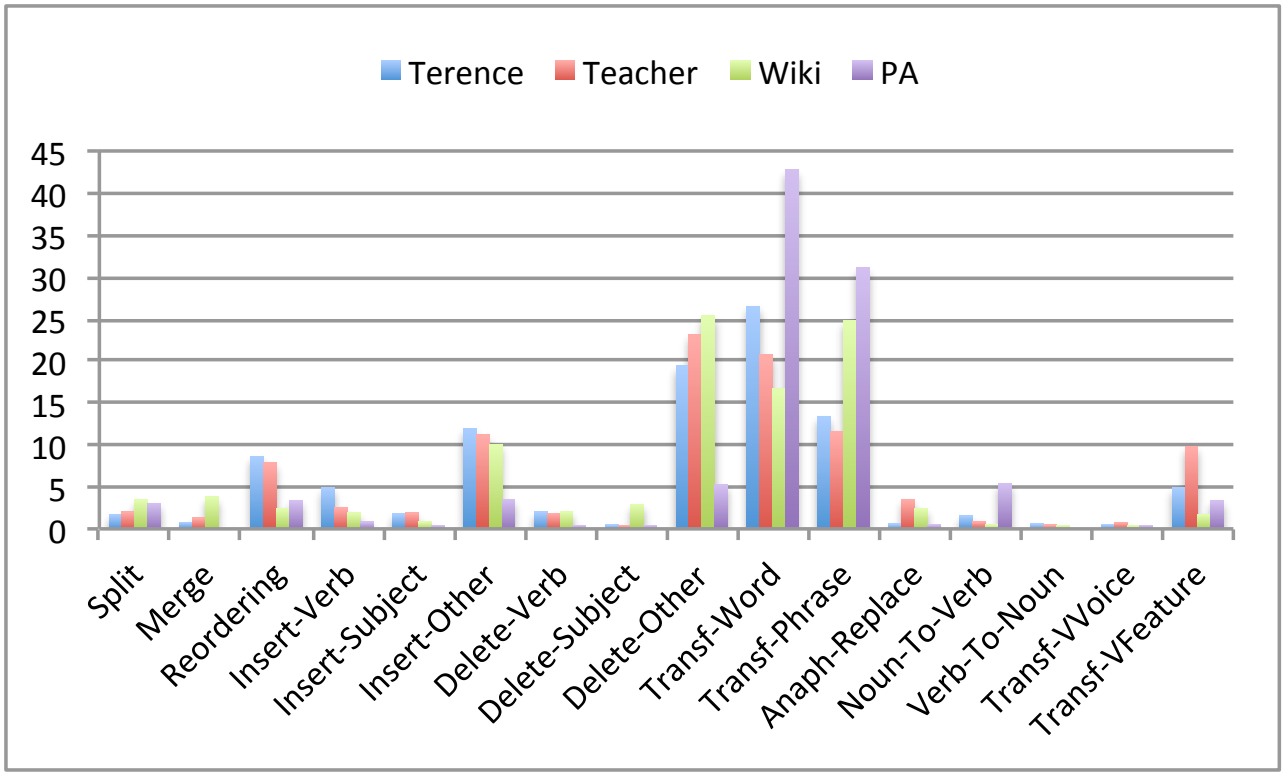
Skip changes that are not simplifications

Annotation scheme

"Current developments are also devoted to refining the annotation scheme, also by testing the suitability of this scheme for other corpora." (Brunato et al., 2015)

CLASS	SUBCLASS
Split	
Merge	
Reordering	
Insert	Verb / Subject / Other
Delete	Verb / Subject / Other
Transformation	Lexical substitution (word level)
	Lexical substitution (phrase level)
	Anaphoric replacement
	Noun to verb
	Verb to noun
	Verbal voice (passive, active)
	Verbal features (mood, tense)

Corpus statistics



Avg. simplifications per sentence pair:
Terence 2.1
Teacher 2.8
Wiki 1.6
PA 2.9

Comparison among **Terence** & **Teacher** corpus (Brunato et al., 2015), our Wikipedia-based corpus (**Wiki**) and a corpus we created by manually simplifying a set of documents in the public administration domain (**PA**)

What we make available

Everything!

- The corpus (1166 pairs)
- The code of the web-based annotation tool

`https://github.com/dhfbk/simpitiki`

To conclude

- We presented the freely available **SIMPITIKI** corpus
- Methodology and tools to extract a set of parallel pairs from Wikipedia edits: can be tested in **any language**
- Even if not much faster than manual simplification from scratch, the pairs extracted from **Wikipedia** have **different characteristics** compared to documents created in educational contexts
- Next step: **merge** all available simplification corpora for Italian to analyse the single simplification types and investigate type-specific **evaluations**

"We therefore advocate for a more informative evaluation that separates out each sub-task. We believe this will lead to more easily quantifiable metrics and possibly the development of automatic metrics." (Xu et al., 2016)

FONDAZIONE BRUNO KESSLER

DH DIGITAL HUMANITIES

Thanks!

European Commission | Horizon 2020 European Union funding for Research & Innovation