FONDAZIONE BRUNO KESSLER trentino italy **www.fbk.eu**

Tint 2.0
an All-inclusive Suite for NLP in Italian

Alessio Palmero Aprosio *aprosio@fbk.eu* **and Giovanni Moretti** *moretti@fbk.eu*

DH DIGITAL HUMANITIES

FONDAZIONE BRUNO KESSLER

Simpatico

Tint is a fast and extendable Natural Language Processing suite for Italian based on Stanford CoreNLP, released for the first time in 2016. The new release includes some improvements of the existing NLP modules, and a set of new text processing components for fine-grained linguistic analyses that were not available so far, including multi-word expression recognition, affix analysis, readability and classification of complex verb tenses.

Tint 2.0 Architecture Layers

Multi-format Output (XML, Json, CoNLL)

Original Tint Modules
Tokenization
Morphological analysis
Lemmatization
Keyword Extraction

Part-of-speech Tagging
Named-entity recognition
Dependency parsing

Wrappers
Entity linking
Geocoding
Temporal Expression

V2 Modules
Affixes annotation
Classification of verbal tenses
Text reuse
Readability

Stanford Core NLP

Tint development phylosophy

"Make it yourself using ingredients from other projects"

Home-made modules + V2 Modules

+

Stanford Built-in modules

+

Wrappers

=

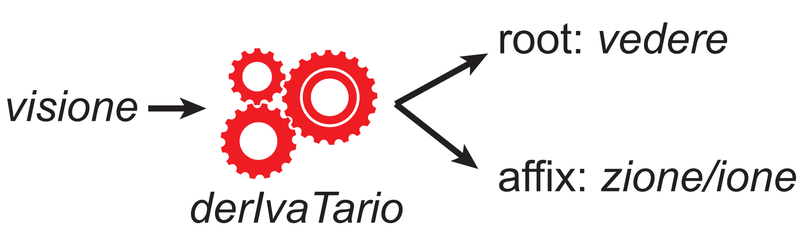
Tint 2.0

- Open source (GPL)
- Easy-to-use, versatile interface
- Written for interoperability
- Server mode (with web API)
- Someone (from Stanford!) maintains the core
- Included in Maven Central

V2 Modules

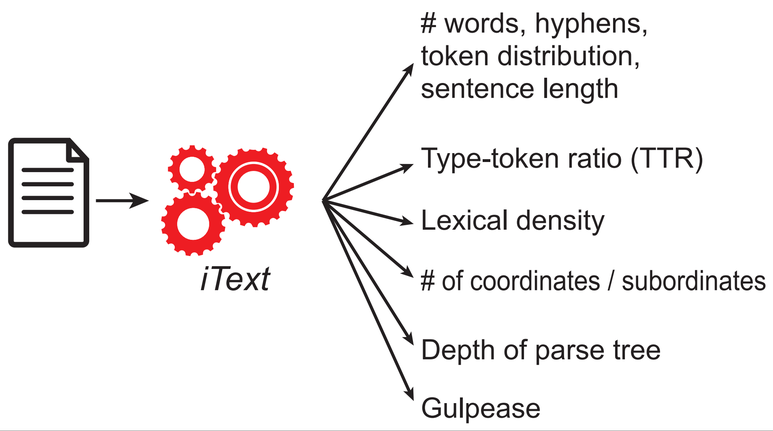
Affixes Annotator

Token level annotation



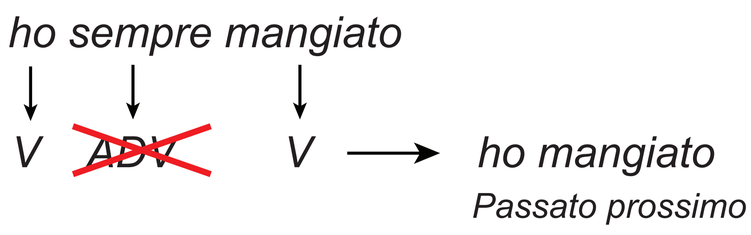
Readability Analysis

Sentence/Document level annotation



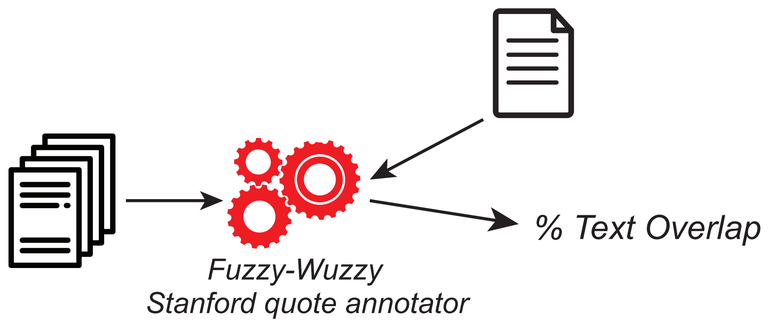
Verb Tense Classificator

Sentence level annotation (stand-off)



Text Reuse

Document level annotation



Evaluation

Some Evaluations

System	Speed (tok/sec)
Tint	**80,000**
Tanl API	30,000
TextPro 2.0	35,000

Tokenization and sentence splitting

System	Speed (tok/sec)	Accuracy
Tint	28,000	**98%**
Tanl API	20,000	n.a.
TextPro 2.0	20,000	96%
TreeTagger	**190,000**[16]	92%

Part-of-speech tagging

System	Speed (tok/sec)	Accuracy
Tint	97,000	96%
TextPro 2.0	9,000	96%
TreeTagger	**190,000**[16]	96%

Lemmatization

System	Speed	P	R	F_1
Tint	**30,000**	**84.37**	79.97	**82.11**
TextPro 2.0	4,000	81.78	**80.78**	81.28
Tanl API	16,000	72.89	52.50	61.04

Named entity recognition

System	Speed	LAS	UAS
Tint	**9,000**	84.67	87.05
TextPro 2.0	1,300	87.30	91.47
Tanl (DeSR)	900	**89.88**	**93.73**

Dependency parsing

http://tint.fbk.eu or https://github.com/dhfbk/tint