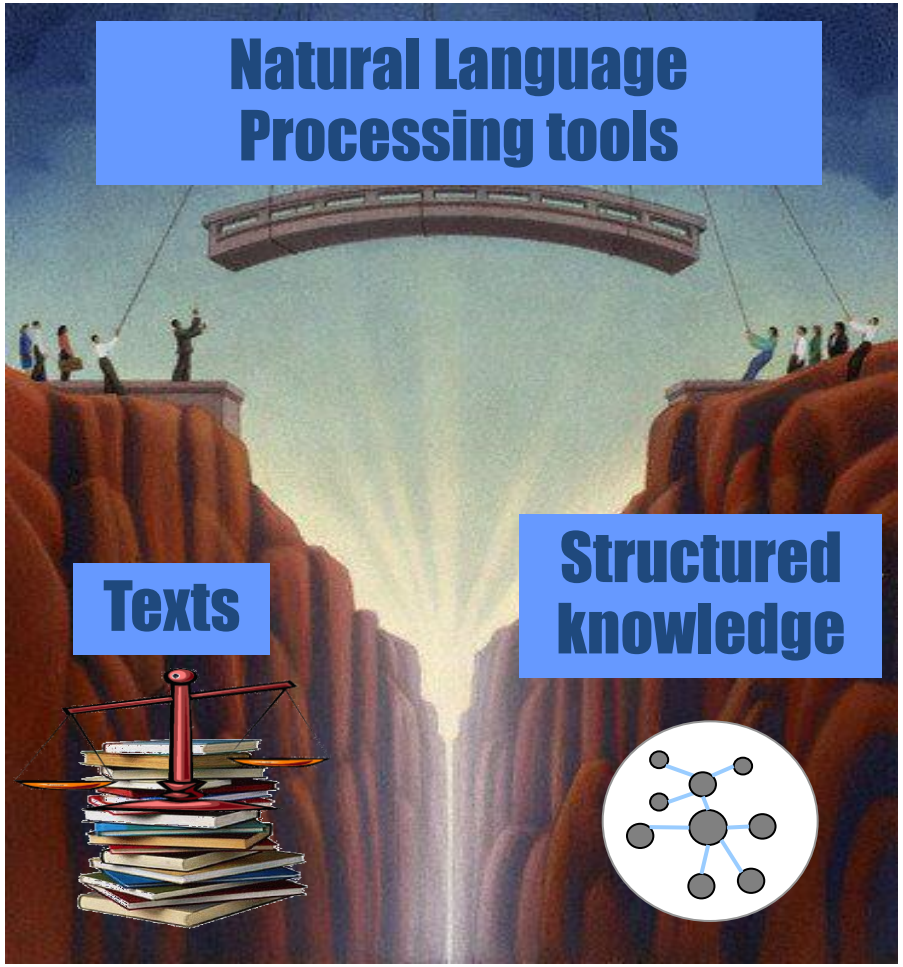
NATURAL LANGUAGE PROCESSING AND LEGAL KNOWLEDGE EXTRACTION

Simonetta Montemagni, Giulia Venturi

Istituto di Linguistica Computazionale «Antonio Zampolli» (ILC-CNR)

Summer School LEX2013 - 5th September 2013

Bridging the gap between text and knowledge: the crucial role of NLP tools



- Knowledge is mostly conveyed through text
  - Content access requires understanding the linguistic structure
- We need a bridge to overcome the gap between text and knowledge
- Technologies based on Natural Language Processing allows
  - accessing the domain-specific knowledge contained in texts
  - structuring the textual content

From text to knowledge: the main challenge in the legal domain

One of the main obstacles to progress in the field of artificial intelligence and law is the ***natural language barrier***

L. Thorne McCarty, International Conference on AI and Law (ICAIL-2007)

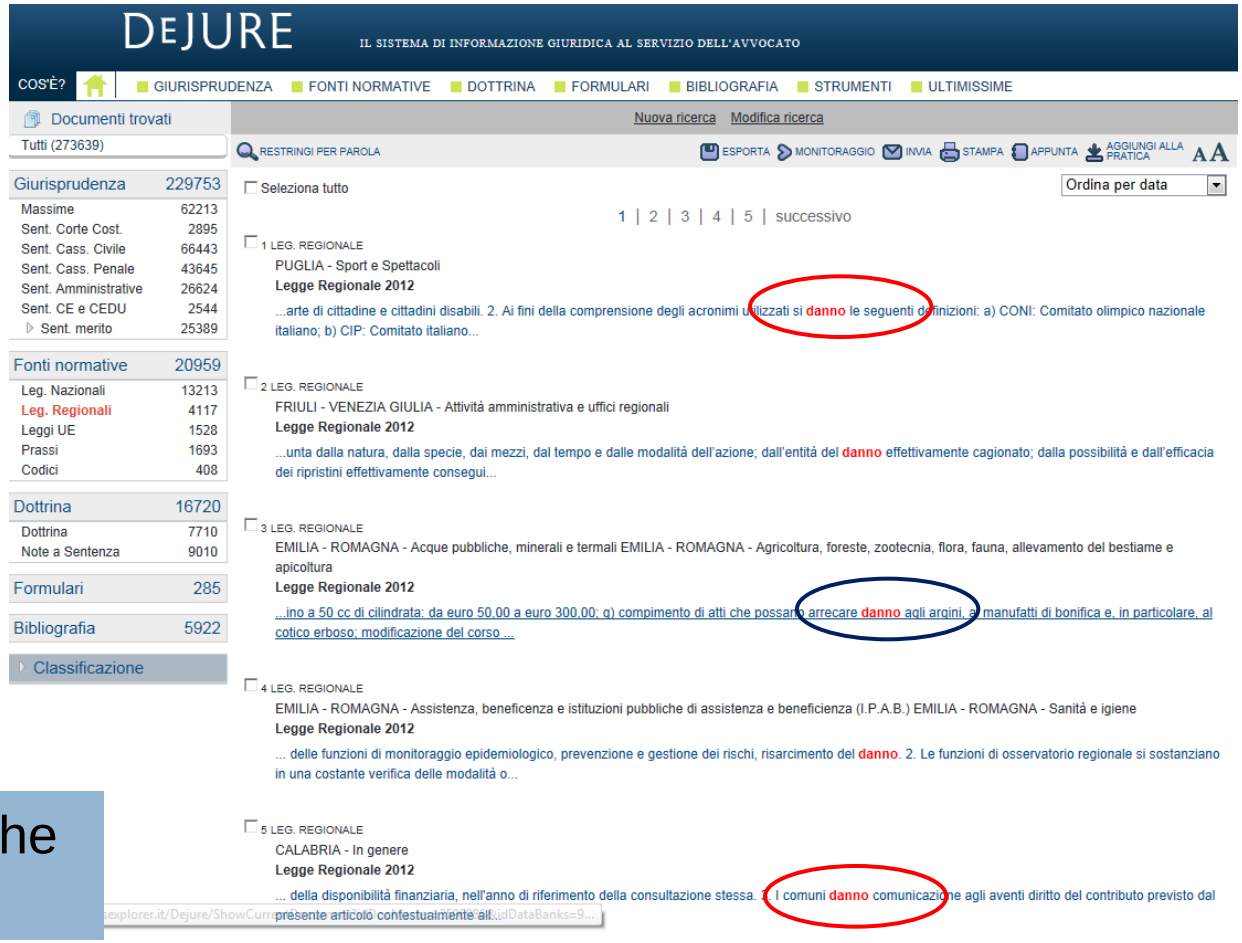
- Raw materials of the law are embodied in natural language (cases, statutes, regulations, etc.)
- Legal knowledge is heavily intertwined with natural language and common sense and therefore inherits all the hard problems that these imply
- Knowledge-based legal information systems need to access the content embedded in legal texts

IUSEXPLORER

- Legal search engine
  - gathering Italian different sources of law (case laws, legislation, jurisprundence, journals, etc.)

IUSEXPLORER: an example of word search query

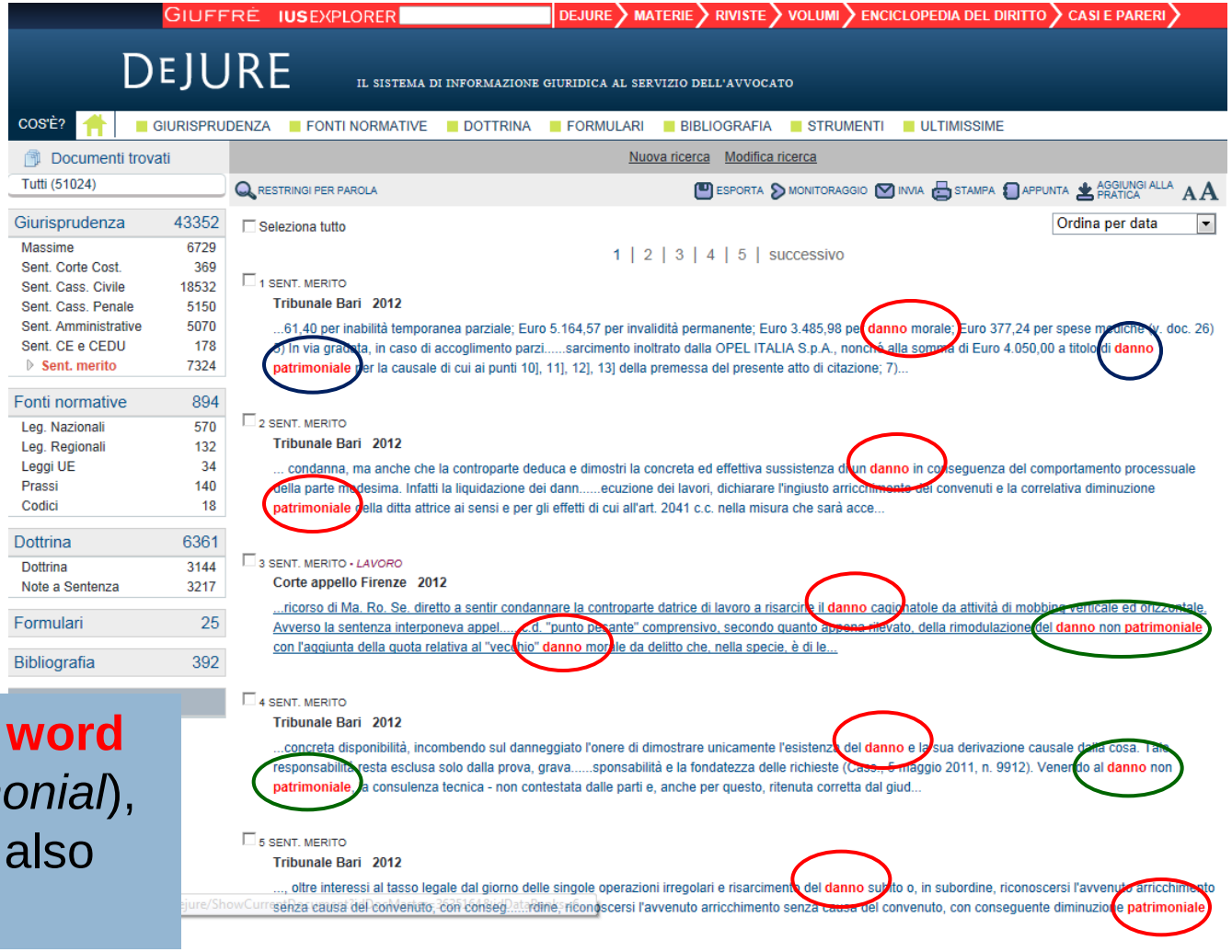
danno (damage)



Ambiguity between the **verb** and the **noun**

IUSEXPLORER: an example of word search query

danno patrimoniale (patrimonial damage)



It returns the **single word** (*damage* and *patrimonial*), the **multi-word** and also the **negation**

IUSEXPLORER

- Advanced search engine which provides customers with access to billions of searchable documents

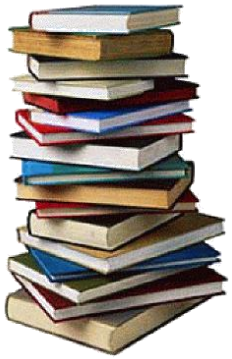
- It is still linguistically rudimentary
  - it does not exploit the potential offered by language technologies
  - it does not support semantic queries allowing an advanced access to documents

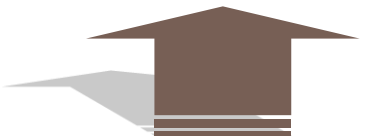
Need for increasingly sophisticated applications based on Natural Language Processing technologies for effectively accessing the content embedded in texts

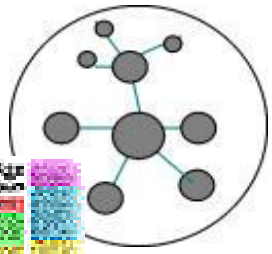
Summary

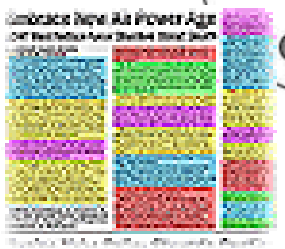
- From text to knowledge
  - The general approach
- Natural Language Processing tools
  - What and what for
- The main challenges of the legal domain
- Legal Knowledge Extraction
  - Identification and extraction of domain-relevant knowledge
  - Semantic annotation of legal texts

From text to knowledge: the general approach









Textual content
(implicit knowledge)

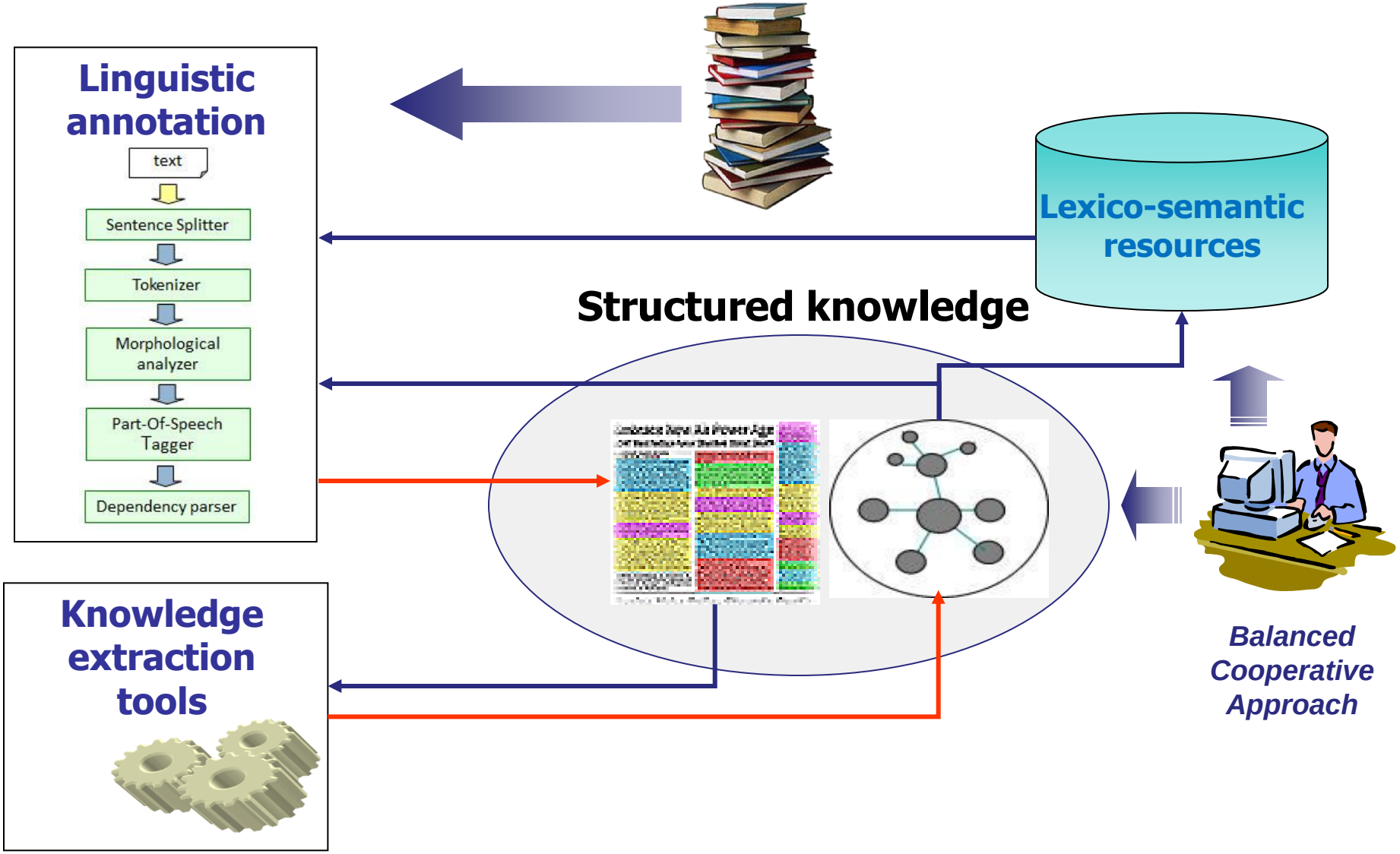
Linguistic annotation

Dynamic content structuring

Knowledge extraction

Structured knowledge
(explicit knowledge)

Natural Language Processing techniques and knowledge extraction



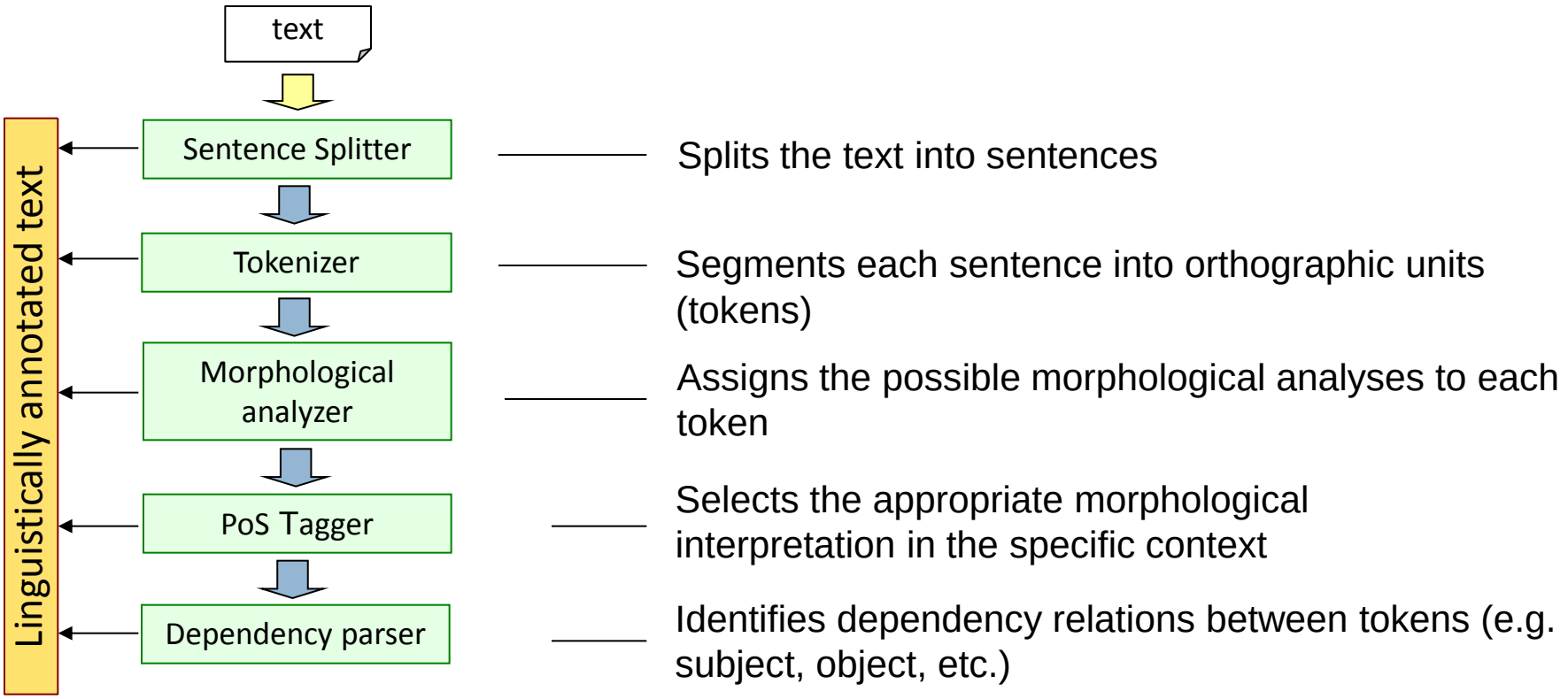
Linguistic annotation tools: what

- Linguistic annotation
  - the process in charge of reconstructing and making explicit the linguistic structure underlying texts

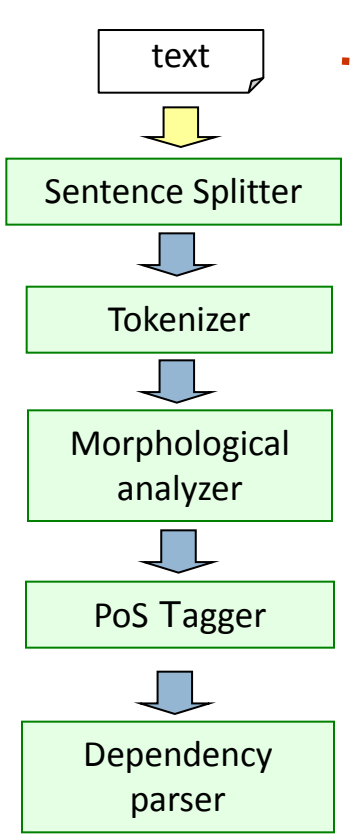
- State-of-the-art tools are based on machine-learning algorithms
  - Annotation process as probabilistic classification task

- Basic requirements
  - **robustness** to minimize failures due to lexical gaps, particularly complex linguistic constructions as well as ill-formed input
  - **accuracy** of achieved results
  - **efficiency** to deal with huge amounts of textual data
  - **portability** to different domains, textual genres, linguistic registers, other languages
  - **incrementality** of analysis

Linguistic annotation: an incremental process

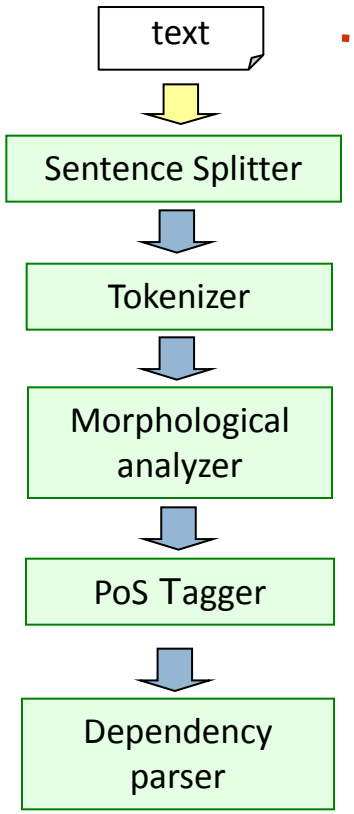


Linguistic annotation: an example



Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

Linguistic annotation: an example

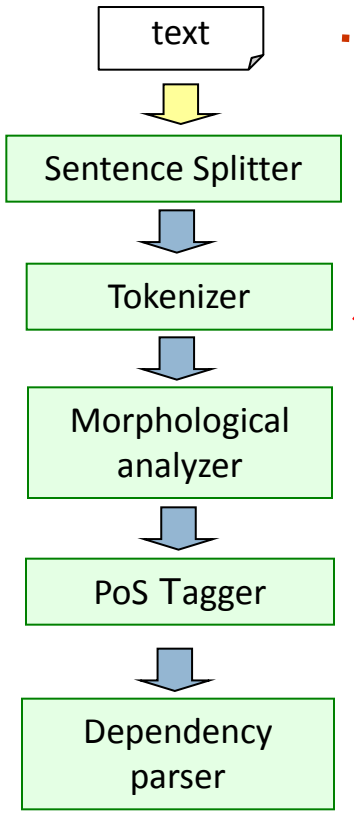


Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

- Il danno non poteva essere sottovalutato. (The damage could not be understimated.)

- Il sig. Rossi decise perciò di chiamare l'avvocato. (Mr. Rossi decided therefore to call the lawyer.)

Linguistic annotation: an example



Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

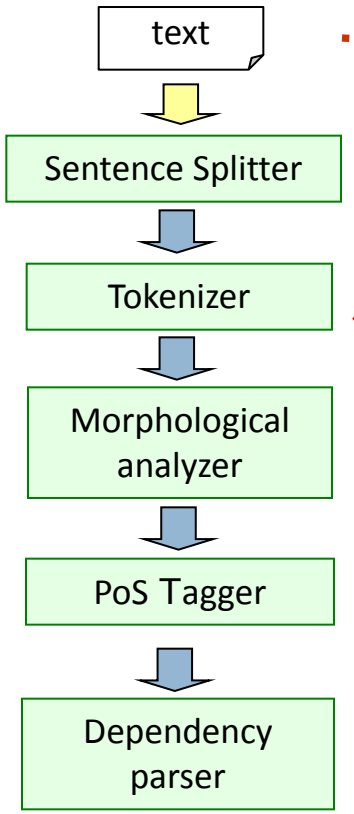
- Il danno non poteva essere sottovalutato. (The damage could not be understimated.)

- Il sig. Rossi decise perciò di chiamare l'avvocato. (Mr. Rossi decided therefore to call the lawyer.)

id	form
1	Il
2	danno
3	non
4	poteva
5	essere
6	sottovalutato

"CoNLL" tabular representation schema

Linguistic annotation: an example



Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

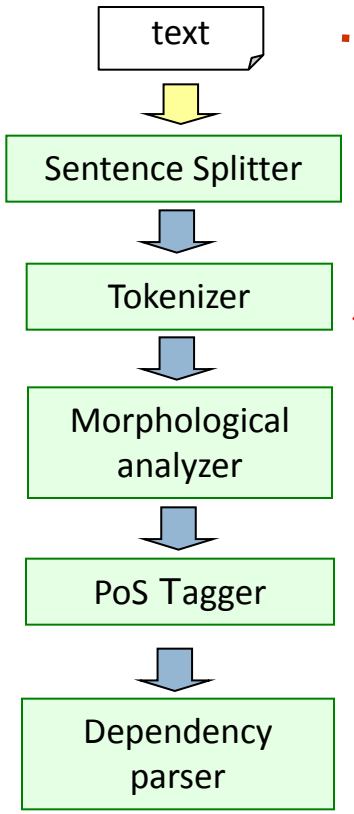
- Il danno non poteva essere sottovalutato. (The damage could not be understimated.)

- Il sig. Rossi decise perciò di chiamare l'avvocato. (Mr. Rossi decided therefore to call the lawyer.)

id	form	lemma	PoS	Feats
1	Il	il	RD	MS
2	danno	danno;dare	S;V	MS;P3IP
3	non	non	BN	NULL
4	poteva	potere	V	S3II
5	essere	essere	V	F
6	sottovalutato	sottovalutare	V	MSPR

"CoNLL" tabular representation schema

Linguistic annotation: an example



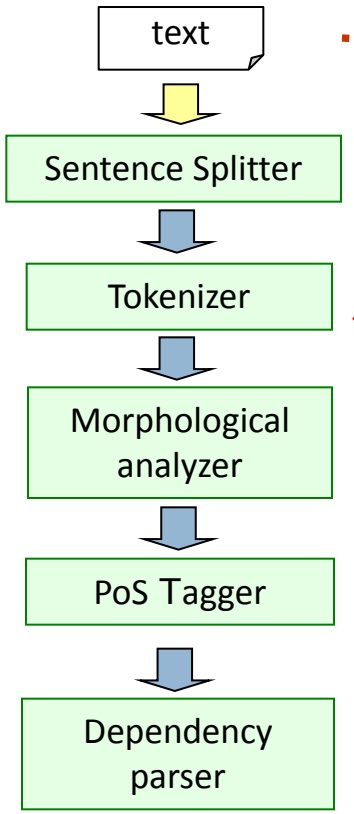
Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

- Il danno non poteva essere sottovalutato. (The damage could not be understimated.)
- Il sig. Rossi decise perciò di chiamare l'avvocato. (Mr. Rossi decided therefore to call the lawyer.)

id	form	lemma	PoS	Feats
1	Il	il	RD	MS
2	danno	danno	S	MS
3	non	non	BN	NULL
4	poteva	potere	V	S3II
5	essere	essere	V	F
6	sottovalutato	sottovalutare	V	MSPR

"CoNLL" tabular representation schema

Linguistic annotation: an example



Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

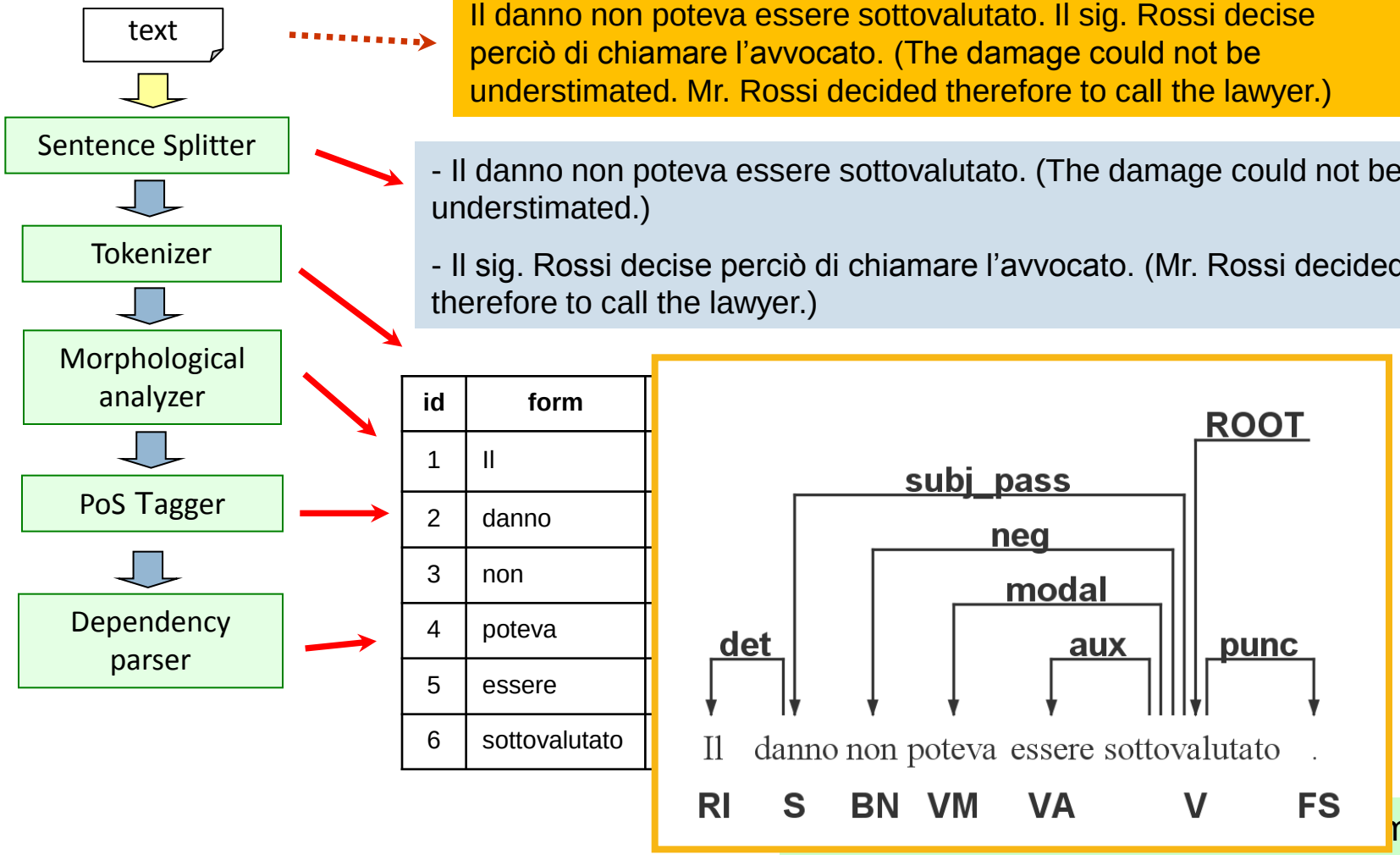
- Il danno non poteva essere sottovalutato. (The damage could not be understimated.)

- Il sig. Rossi decise perciò di chiamare l'avvocato. (Mr. Rossi decided therefore to call the lawyer.)

id	form	lemma	PoS	Feats	head	DEP
1	Il	il	RD	MS	2	DET
2	danno	danno	S	MS	6	SUBJ_PASS
3	non	non	BN	NULL	6	NEG
4	poteva	potere	V	S3II	6	MODAL
5	essere	essere	V	F	6	AUX
6	sottovalutato	sottovalutare	V	MSPR	0	ROOT

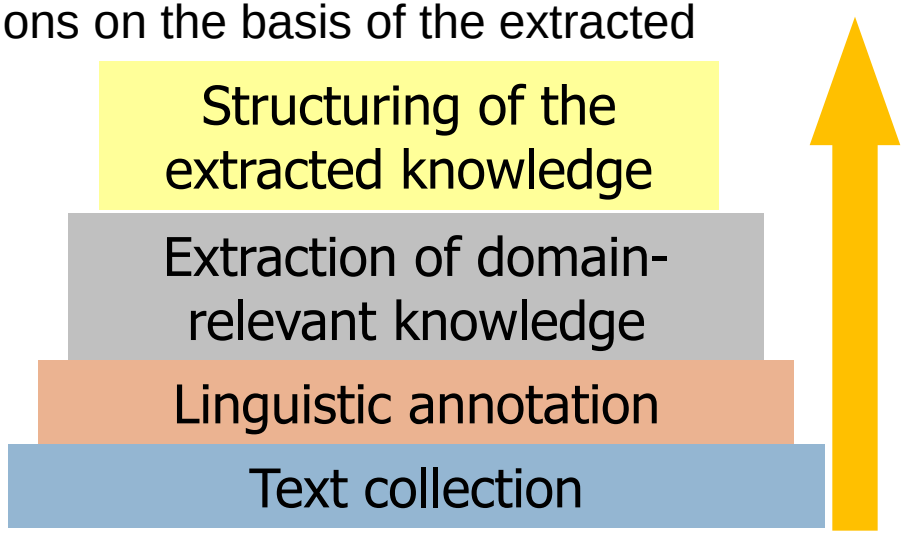
"CoNLL" tabular representation schema

Linguistic annotation: an example

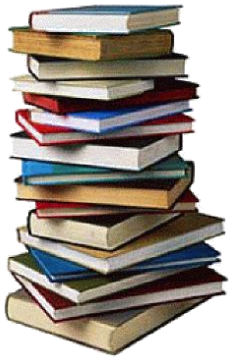


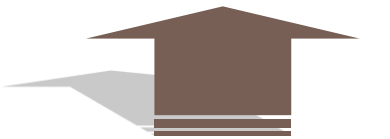
Linguistic annotation: what for

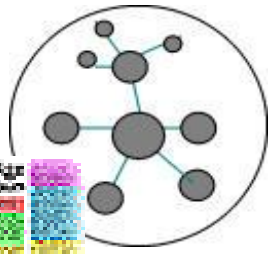
- Linguistic annotation plays a crucial role in accessing the content of texts by making it explicit the linguistic structure through which knowledge is encoded
- Starting point for several Knowledge Extraction tasks
  - extracting domain-relevant knowledge
  - structuring the extracted knowledge in semantic resources, e.g. lexicons, thesauri, domain-specific ontologies (*Ontology Learning*)
  - semantic indexing of text collections on the basis of the extracted knowledge
- Linguistic annotation and knowledge extraction
  - increasingly complex knowledge extraction tasks differentially exploit individual levels of linguistic annotation

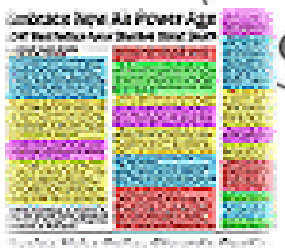


From text to knowledge: the general approach









Textual content
(implicit knowledge)

Linguistic annotation

Dynamic content structuring

Knowledge extraction

Structured knowledge
(explicit knowledge)

The legal domain: the main challenges

- The typical **knowledge acquisition bottleneck**
  - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure
- The **peculiarity of legal language and its impact on NLP tools**
  - Legal syntax is “convoluted and unnatural” (McCarty, NaLEA 2009) with respect to ordinary language
  - What is the performance of state-of-the-art NLP tools on legal texts?
- **Discriminate between legal and regulated domain knowledge**
  - By its very nature, law deals with behaviour in the world: domain independent concepts of law are tainted with concepts referring to the world the legal domain is about

The knowledge acquisition bottleneck

- Technologies in the area of knowledge management are typically confronted with the problem of processing linguistic structure
  - Particularly relevant in the legal domain, where law is strictly dependent on its linguistic expression
- Why legal language processing?
  - “Why parse statutes? To extract their logical structure, to refine the semantics of the domain, to develop a domain ontology” (McCarty, 2009)
- What are the domain-specific issues to be addressed when processing legal language?
  - Whether and to what extent legal language differs from ordinary language
  - Impact of recorded differences on the performance of NLP tools

The legal domain: the main challenges

- The typical **knowledge acquisition bottleneck**
  - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure

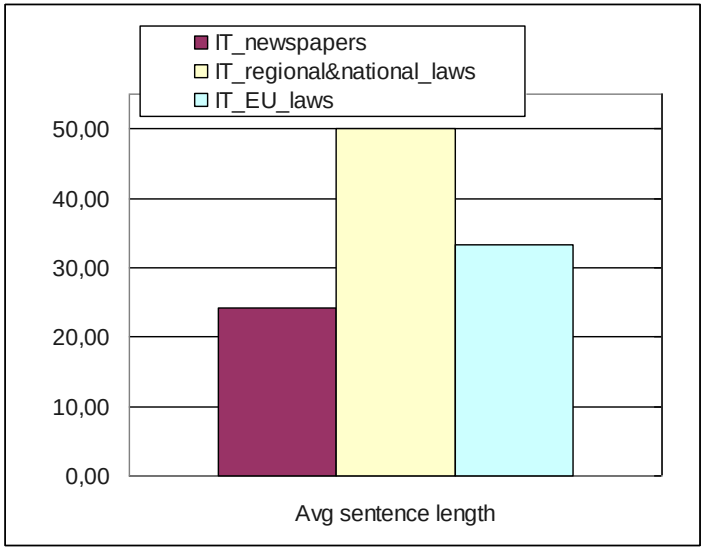
- The **peculiarity of legal language and its impact on NLP tools**
  - Legal syntax is “convoluted and unnatural” (McCarty, NaLEA 2009) with respect to ordinary language
  - What is the performance of state-of-the-art NLP tools on legal texts?

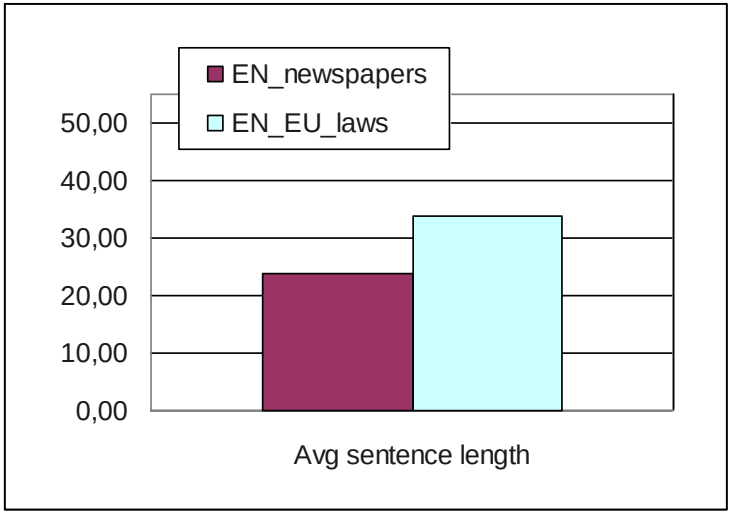
- **Discriminate between legal and regulated domain knowledge**
  - By its very nature, law deals with behaviour in the world: domain independent concepts of law are tainted with concepts referring to the world the legal domain is about

The peculiarity of legal language

- Legal texts differ significantly with respect to ordinary language texts
  - typically correlated with syntactic complexity

- Differences recorded at different annotation levels
  - **long sentences wrt newswire texts**



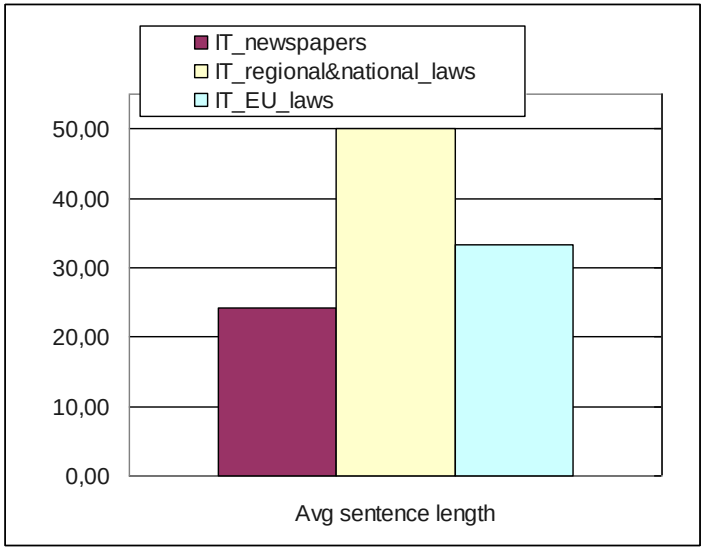


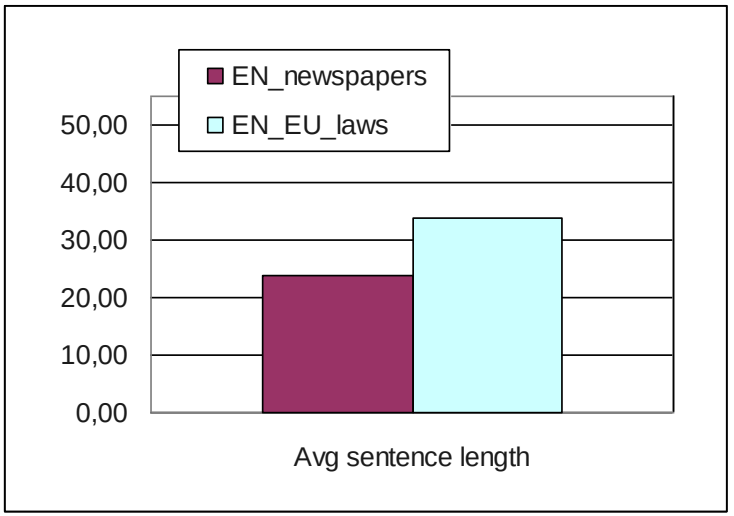
The peculiarity of legal language

- Legal texts differ significantly w
  language texts
  - typically correlated with syntactic c
- Differences recorded at different annotation levels
  - **long sentences wrt newswire texts**

Italian:
- a corpus of newspapers
- a collection of laws enacted by the European Commission, Italian State and Regions

English:
- a corpus of newspapers
- a collection of laws enacted by the European Commission





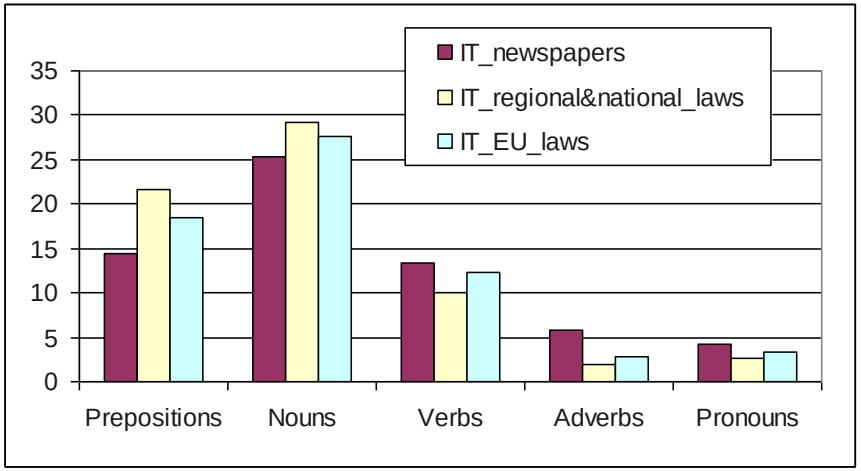
The peculiarity of legal language

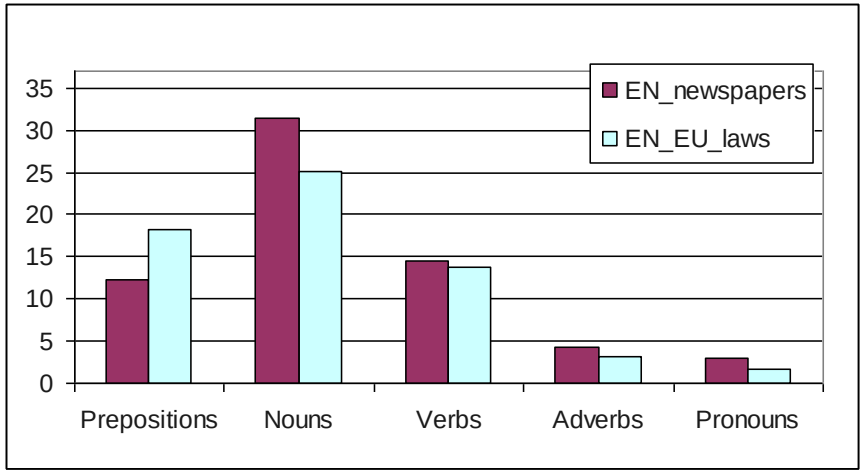
- Legal texts differ significantly with language texts
  - typically correlated with syntactic co

Italian:
- a corpus of newspapers
- a collection of laws enacted by the European Commission, Italian State and Regions

English:
- a corpus of newspapers
- a collection of laws enacted by the European Commission

- Differences recorded at different annotation levels
  - long sentences wrt newswire texts
  - **high % of prepositions and low % of verbs, adverbs, pronouns**





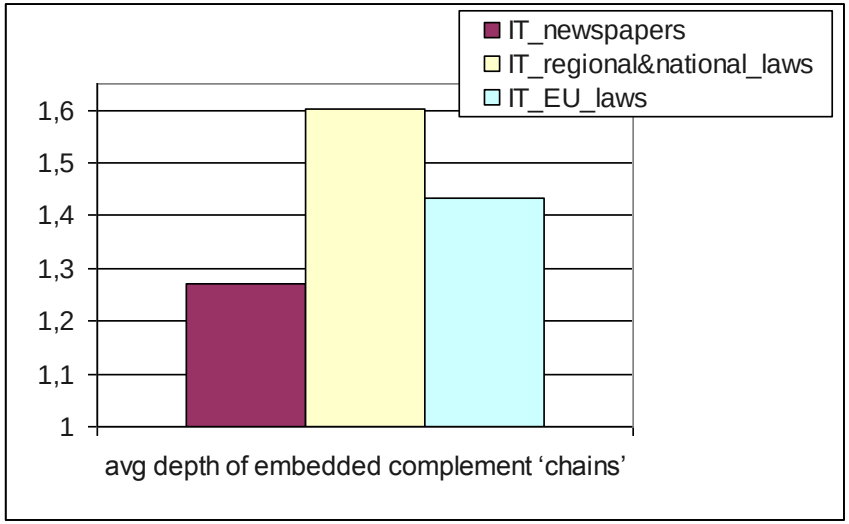
The peculiarity of legal language

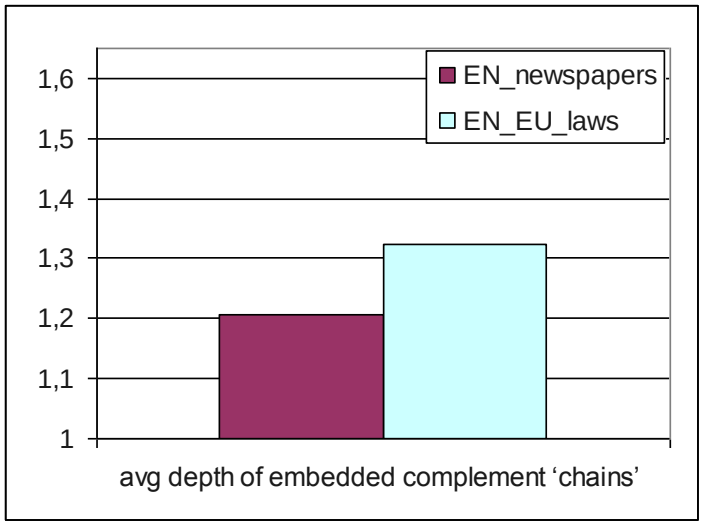
- Legal texts differ significantly with res
  texts
  - typically correlated with syntactic comple

Italian:
- a corpus of newspapers
- a collection of laws enacted by the European Commission, Italian State and Regions

English:
- a corpus of newspapers
- a collection of laws enacted by the European Commission

- Differences recorded at different annc
  - long sentences wrt newswire texts
  - high % of prepositions and low % of verbs, adverbs, pronouns
  - **deep sequences of embedded prepositional complement chains**





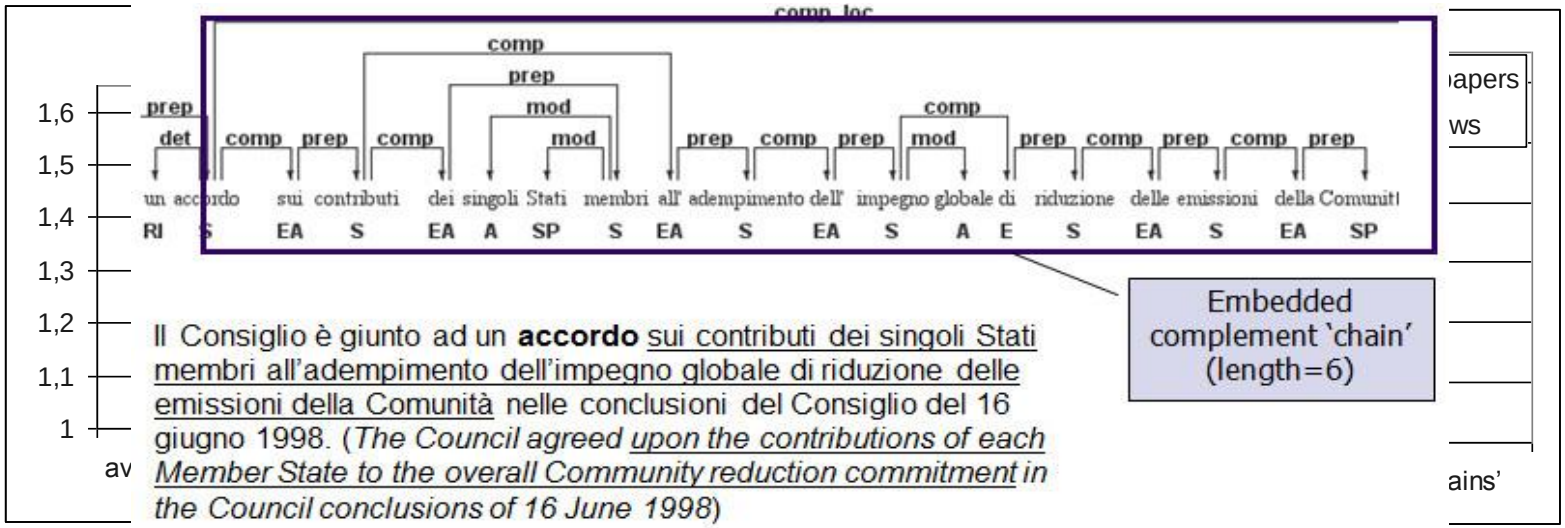
The peculiarity of legal language

- Legal texts differ significantly with res
  texts
  - typically correlated with syntactic comple

Italian:
- a corpus of newspapers
- a collection of laws enacted by the European Commission, Italian State and Regions

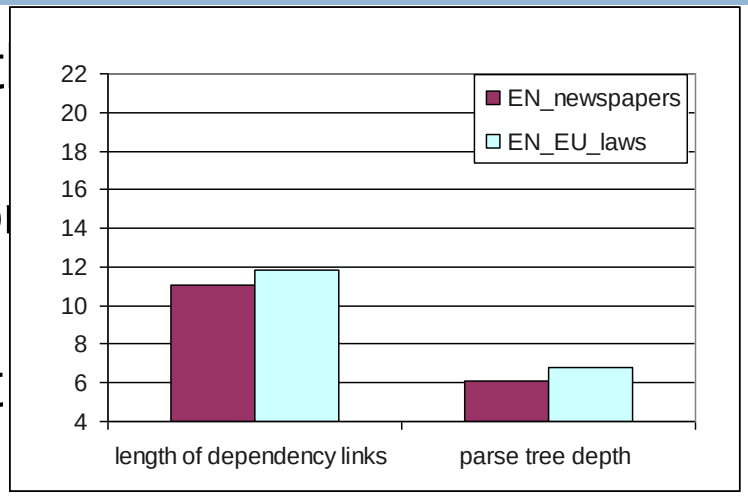
English:
- a corpus of newspapers
- a collection of laws enacted by the European Commission

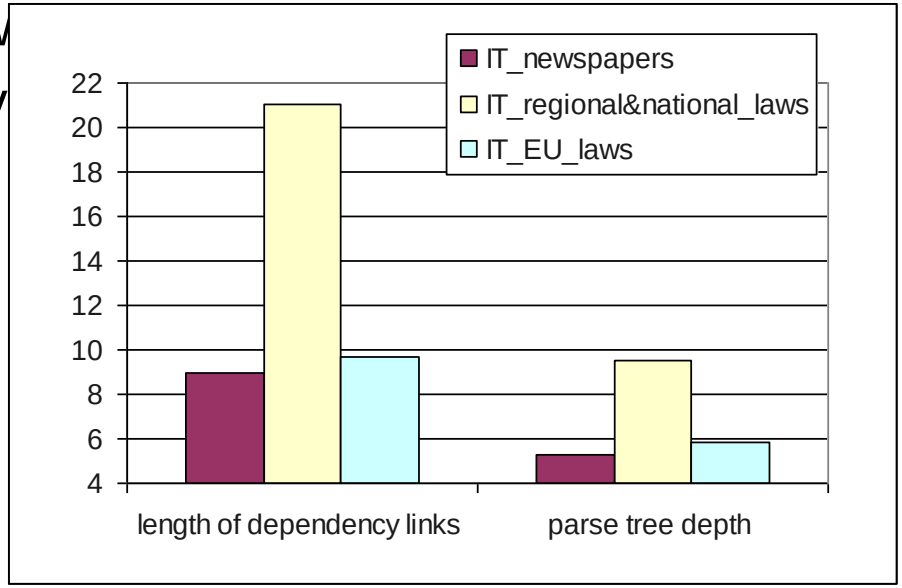
- Differences recorded at different anno
  - long sentences wrt newswire texts
  - high % of prepositions and low % of verbs, adverbs, pronouns
  - **deep sequences of embedded prepositional complement chains**



The peculiarity of legal language

- Legal texts differ significantly wit language texts
  - typically correlated with syntactic co
- Differences recorded at different
  - long sentences wrt newswire texts
  - high % of prepositions and lov
  - long sequences of consecutiv
  - **long dependency links**
  - **deep syntactic trees**





- Statistical parsers have a drop in accuracy when analyzing long distance dependencies (McDonald and Nivre, 2007)
- Parse tree depth is a well-known feature reflecting sentence complexity

The impact of legal language on NLP tools

- What is the performance of state-of-the-art NLP tools on legal texts?
  - A key issue for all NLP-based Knowledge Extraction tasks
  - Generally speaking, a dramatic drop of accuracy is reported when syntactic parsers are tested on domains outside of the data from which they are trained or developed on

- Recently, two initiatives focused on **dependency parsing** of legal texts which represents a prerequisite for any advanced legal text processing task
  - Domain Adaptation Track at Evalita 2011 – Italian
  - SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts – Italian and English
  - both aimed at
    - obtaining a clear idea of the current performance of state-of-the-art dependency parsing systems against legal texts
    - investigating techniques for adapting state-of-the-art dependency parsing systems to the legal domain

The impact of legal language on NLP tools

- Results of the Dependency Parsing subtask of the SPLeT-2012 *Shared Task on Dependency Parsing of Legal Texts*
  - Goal: testing the performance of general parsing systems on legal texts
- Accuracy results for Italian:

Participant System	Newspaper test	Reg/Nat legal test	EU legal test
1	82.36	75.88	83.08
2	82.90	74.03	81.93
3	81.43	75.55	81.58

- Accuracy results for English:

Participant System	Newspaper test	EU legal test
1	88.81	78.90

For both Italian and English, **lower** performance of parsing systems on legal texts wrt newspapers

Different performances across different **subvarietes of legal language**

- Significant drops on the IT regional and national texts
- 2 out of the 3 participant systems do not show a significant drop of accuracy when tested on EU legal texts

The legal domain: the main challenges

- The typical **knowledge acquisition bottleneck**
  - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure
- The **peculiarity of legal language and its impact on NLP tools**
  - Legal syntax is “convoluted and unnatural” (McCarty, NaLEA 2009) with respect to ordinary language
  - What is the performance of state-of-the-art NLP tools on legal texts?
- **Discriminate between legal and regulated domain knowledge**
  - By its very nature, law deals with behaviour in the world: domain independent concepts of law are tainted with concepts referring to the world the legal domain is about

Discriminate between legal and regulated domain knowledge

«As any legal source – legislation, contracts, precedence-law – reveals immediately: the majority of concepts in an individual source refers to specific domains of social activities. These domains are called 'world knowledge'.»

Breuker & Hoekstra 2004

- Domain-specific terms of law are tainted with terms referring to the world the legal domain is about
  - e.g. *national provision, fundamental principle & hazardous substance, active ingredient*

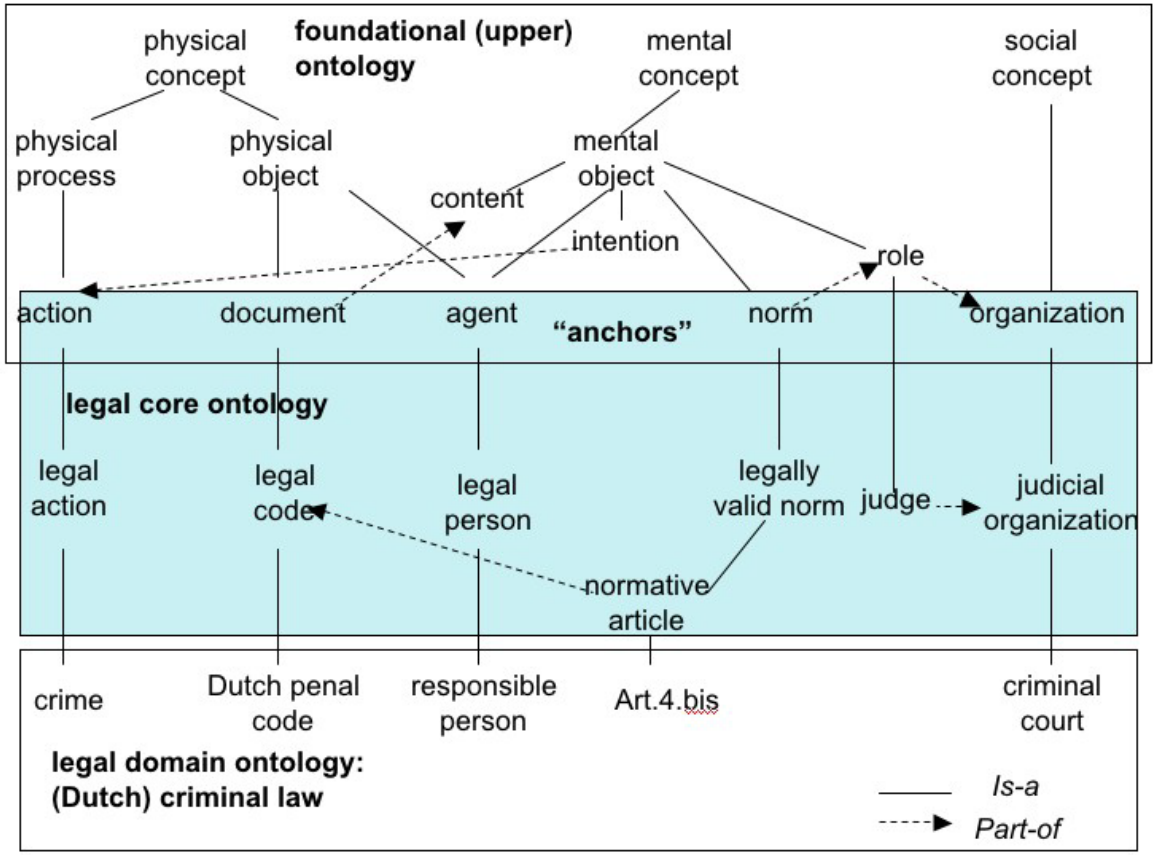
«Therefore it is not surprise that one may find that many legal ontologies are mixtures of epistemological and ontological perspectives.»

Breuker & Hoekstra 2004

- Discriminating between legal and regulated domain terms and/or concepts is key in constructing a legal semantic resource
  - It is closely related to the reusability and interoperability issue

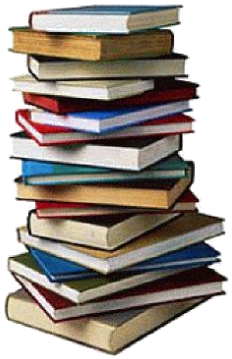
Discriminate between legal and regulated domain knowledge

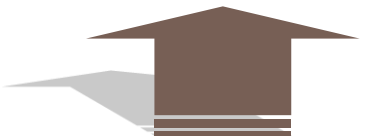
- According to the ontology design criteria, the level of generality in which concepts are organized is a distinctive characteristic
- Three different kinds of ontologies:
  - top or upper-level ontologies (general concepts)
  - core ontologies (top-level domain-specific concepts, e.g. legal)
  - domain-specific ontologies (which organize world knowledge)

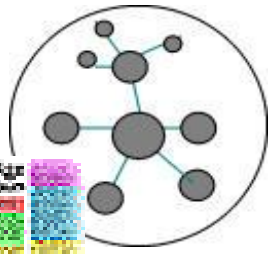


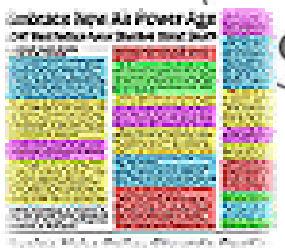
Breuker & Hoekstra 2004: LRI-Core layers: foundational and legal core share 'anchors' (high level concepts typical for law)

From text to knowledge: the general approach









Textual content
(implicit knowledge)

Linguistic annotation

Dynamic content structuring

Knowledge extraction

Structured knowledge
(explicit knowledge)

Legal Knowledge Extraction: focus on …

- **Identification, extraction and structuring** of domain-relevant knowledge
  - Goal: constructing semantic resources such as domain-specific ontologies or lexicons
- **Semantic annotation** of legal texts
  - Goal: content-based access and querying

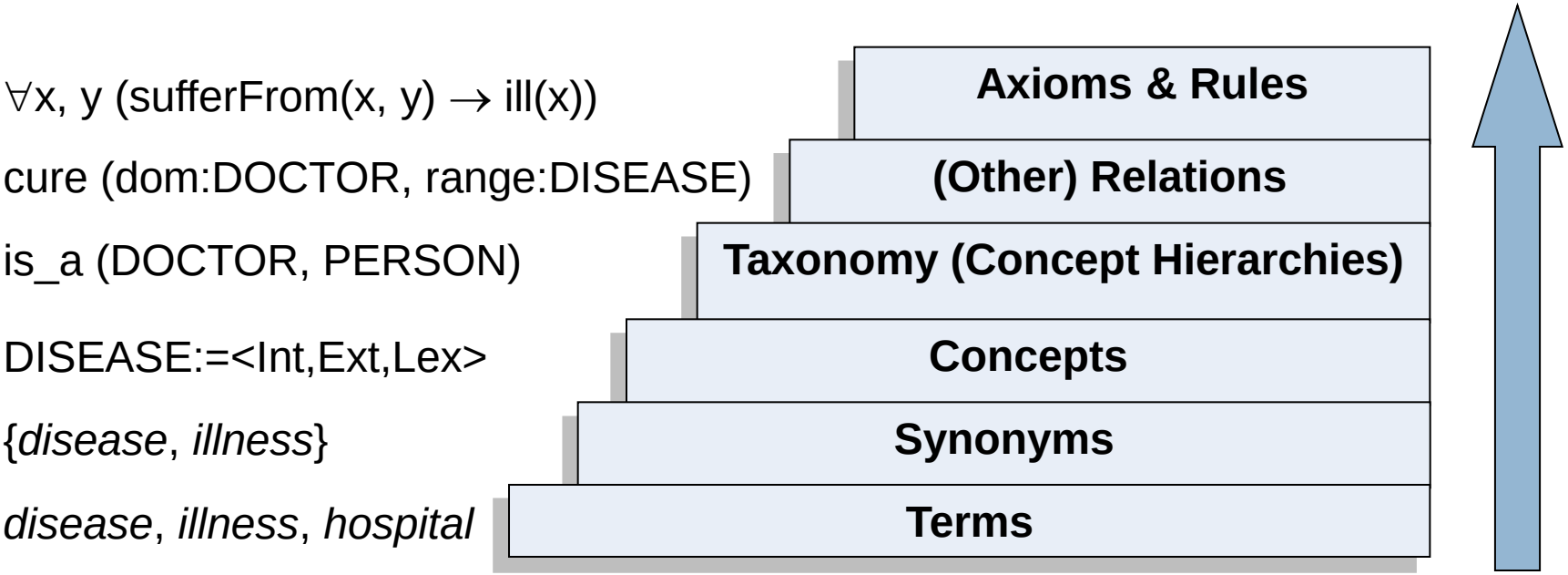
Legal Knowledge Extraction: focus on …

- **Identification, extraction and structuring** of domain-relevant knowledge
  - Goal: constructing semantic resources such as domain-specific ontologies or lexicons

Focus on the **Ontology Learning**
The construction of **Legal Ontologies** referred to as the «missing link» (Valente and Breuker, 2004) between Artificial Intelligence and Law and Legal Theory.
Key process since the emergence of the Semantic Web (Van Engers et al., 2008)

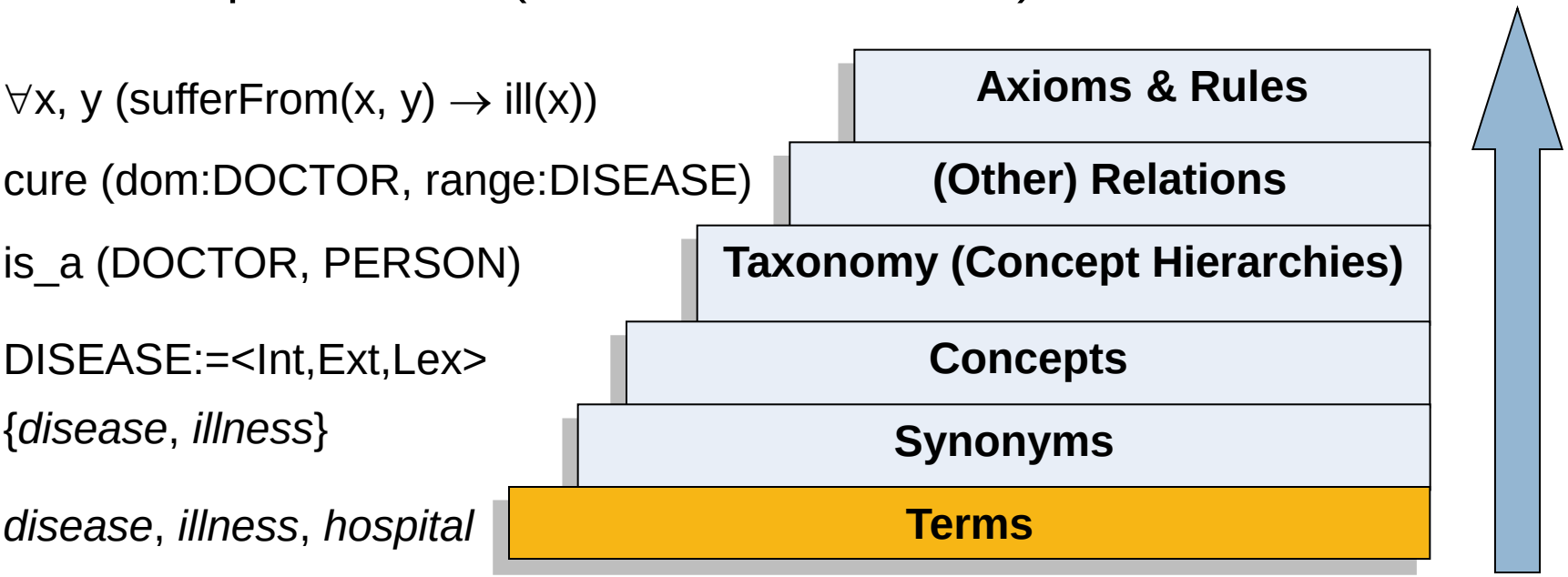
Ontology Learning

- The various steps of Ontology Learning from texts can be arranged in a “layer cake” of increasingly complex subtasks
  - (Buitelaar, Cimiano and Magnini, 2005)



Ontology Learning

- First step of each Ontology Learning process:
  - **Terminology Extraction**
  - «Terms are linguistic realizations of domain-specific concepts and are therefore central to further, more complex tasks» (Buitelaar et al., 2005)



Ontology Learning: Terminology Extraction

- Terms may consist of
  - a single wordform so-called **“simple”** (or one-word) terms
    - e.g. *artist*
  - two or more wordforms, called **“multi-word”** (or complex) terms
    - e.g. *art movement*

- Term extraction process articulated into two fundamental steps:
  - identifying term candidates from text
  - filtering through the candidates to separate *terms* from *non-terms*

- Different statistical measures are used
  - For the extraction of *simple* terms: frequency occurrence distribution, measures of statistical relevance such as TF/IDF (Term Frequency/Inverse Term Frequency), etc.
  - For the extraction of *multi-word* terms: association strength measures such as Mutual Information, C-NC Value, Log-likelihood, etc.

Ontology Learning

- The next step is the semantic structuring of the extracted terminology
  - definition of concepts and relations between them

$\forall x, y$ (sufferFrom(x, y) $\rightarrow$ ill(x))

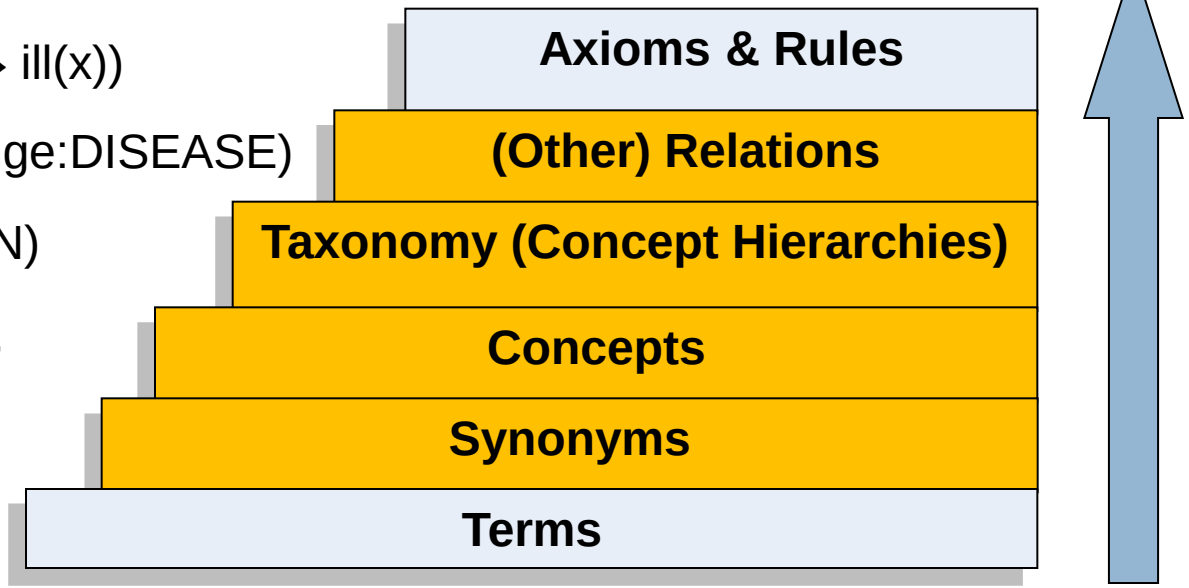
cure (dom:DOCTOR, range:DISEASE)

is_a (DOCTOR, PERSON)

DISEASE:=<Int,Ext,Lex>

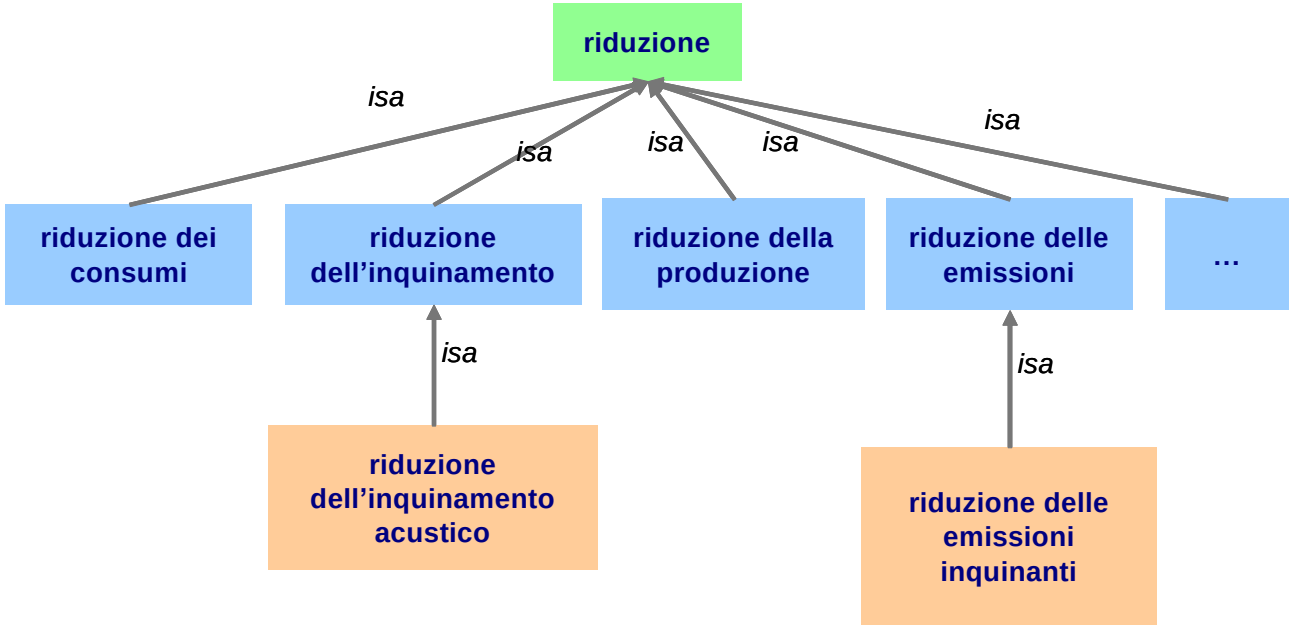
{*disease*, *illness*}

disease, *illness*, *hospital*



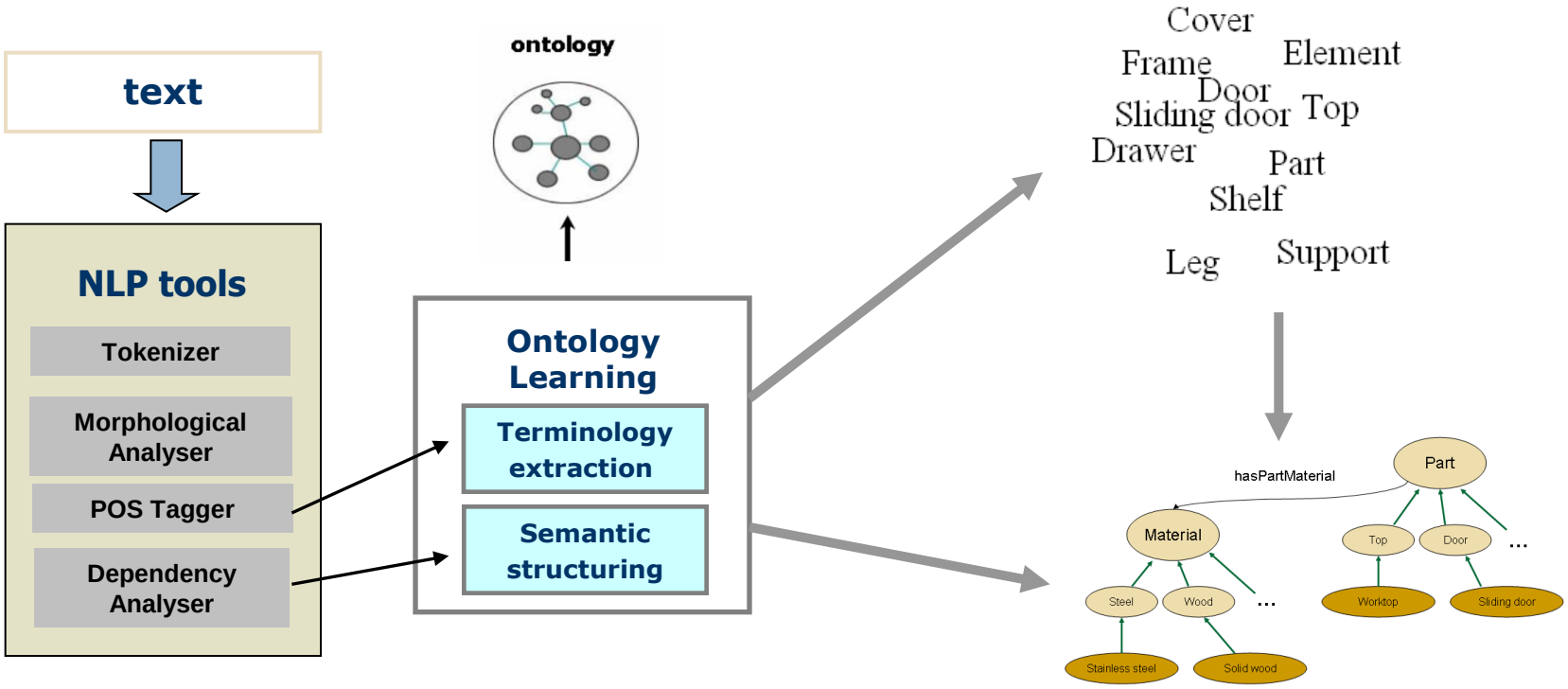
Ontology Learning: Semantic Structuring

- The extracted terms are organized into fragments of taxonomical chains
  - simple and multi-word terms are structured in a vertical hierarchy
  - on the basis of their internal linguistic structure (head sharing)



Ontology Learning: to sum up

- Knowledge extraction in two steps:
  - **Term Extraction**: detection of single and multi-word terms
  - **Semantic Structuring**: definition of concepts and relations between them



Ontology learning in the legal domain: so far …

- Overview of existing Legal Ontologies:
  - Núria Casellas, “Legal Ontology Engineering. Methodologies, Modeling Trends and the Ontology of Professional Judicial knowledge”, 2011

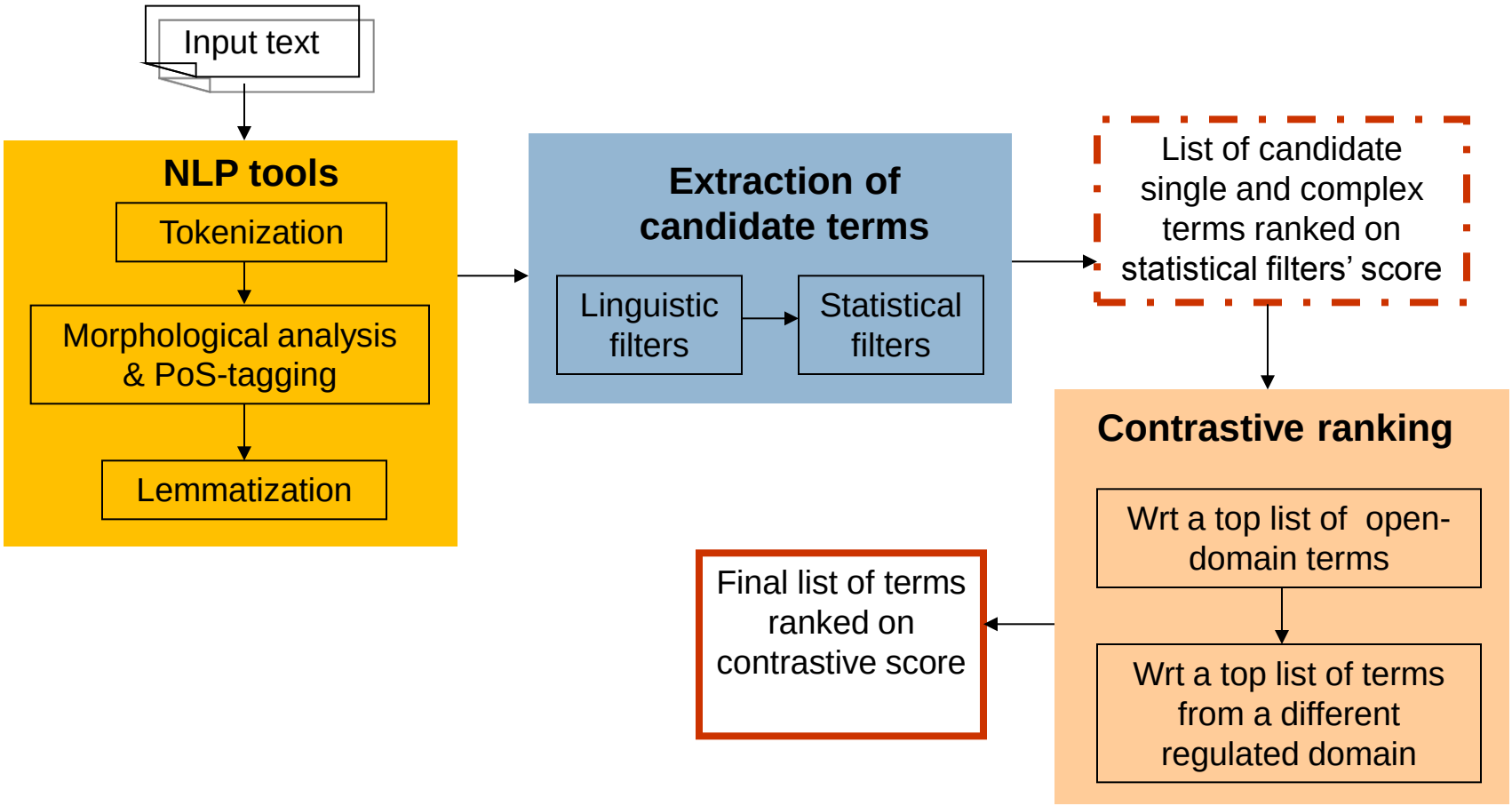
- Approaches to semi-automatically induce legal domain ontologies from texts
  - focus on definitions in German court decisions from which legal concepts are identified together with relevant terminology and relations
    - Walter and Pinkal (2006)
  - extraction of domain relevant terminology from which domain relevant concepts are derived together with relations linking them
    - Lame (2000, 2005): French
    - Saias and Quaresma (2005): Portuguese
    - Völker et al. (2008): Spanish
    - Lenci at al. (2009): Italian
  - ontology modelling
    - LKIF Core ontology (Hoekstra et al., 2007)
    - LOIS (Peters et al., 2005)
    - OPJK (Casellas, 2008)
    - DALOS (Agnoloni et al., 2009)

Ontology Learning: exemplifying Terminology Extraction

- Focus on the term extractor developed by ItaliaNLP Lab at ILC-CNR (Bonin et al., 2010)
  - It follows a multilayered and contrastive approach to overcome the need to discriminate between legal and world knowledge
    - It singles out legal terms, e.g. *law, legislative decree* (legal knowledge), from regulated-domain terms, e.g. *consumer, hazardous substance* (*world knowledge*)
  - Tested in different case studies
    - Corpus of environmental laws (Bonin et al., 2010)
      - EU Directives (394,088 tokens)
    - Case Law corpus (LIDER-Lab, Scuola Superiore Sant'Anna, Pisa)
      - Case law on personal offence (1,206,831 tokens)
    - Case Law corpus (Lazari & Venturi, 2012)
      - Case law on state liability (933,077 tokens)

Ontology Learning: exemplifying Terminology Extraction

- The multi-layered architecture developed by the ItaliaNLP Lab



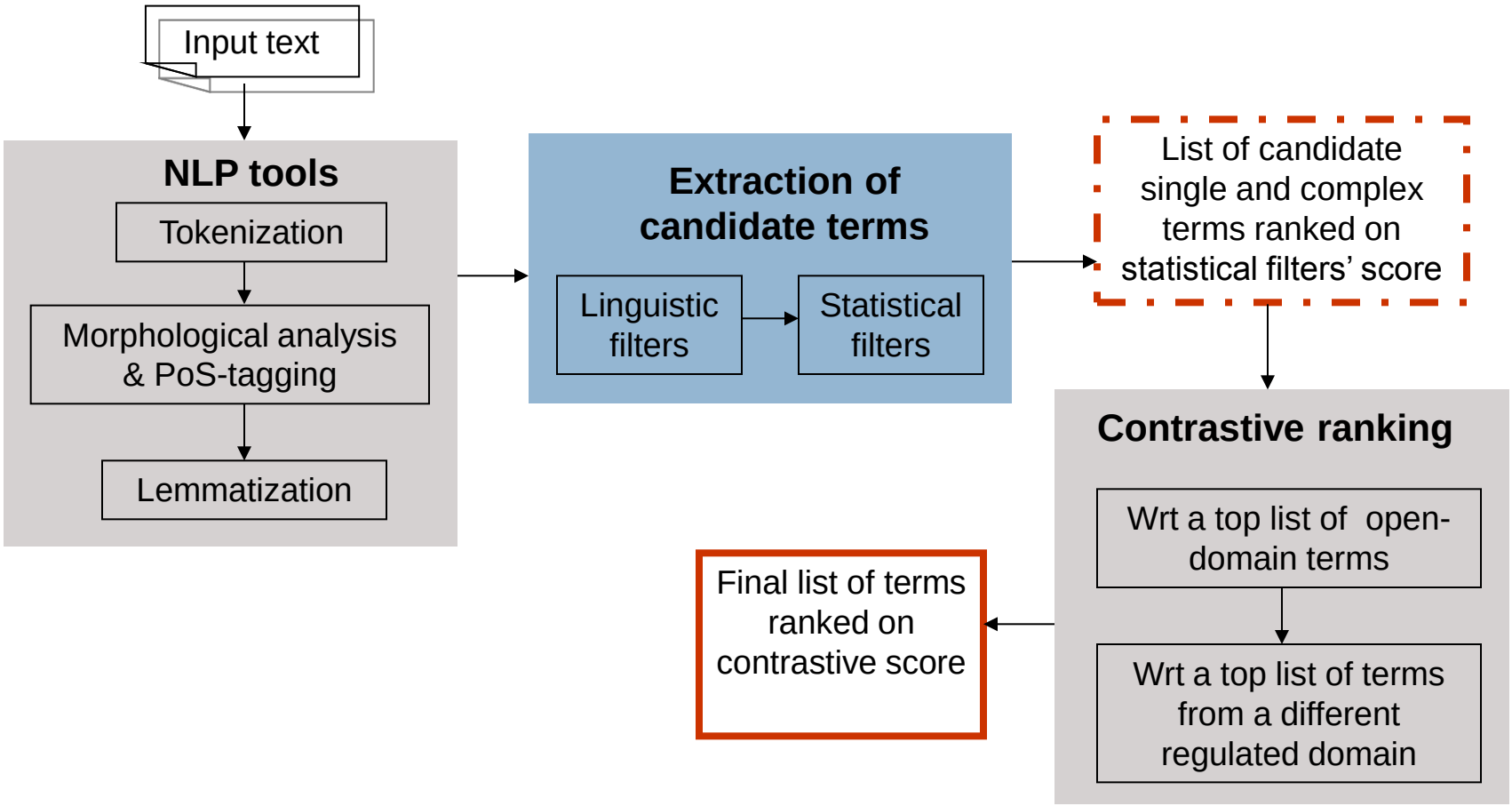
Ontology Learning: exemplifying Terminology Extraction

- Linguistic annotation until the Part-Of-Speech and Lemmatization levels
  - E.g. Il piano nazionale di riduzione delle emissioni in nessun caso può esonerare un impianto dal rispetto della pertinente normativa comunitaria, compresa la direttiva 96/61/CE (*The national emission reduction plan may under no circumstances exempt a plant from the provisions laid down in relevant Community legislation, including inter alia Directive 96/61/EC*)

Forma	Lemma	CPoSTag	PosTag	Tratti morfologici	Forma	Lemma	CPoSTag	PosTag	Tratti morfologici
Il	il	R	RD	num=s|gen=m	un	un	R	RI	num=s|gen=m
piano	piano	S	S	num=s|gen=m	impianto	impianto	S	S	num=s|gen=m
nazionale	nazionale	A	A	num=s|gen=n	dal	da	E	EA	num=s|gen=m
di	di	E	E	_	rispetto	rispetto	S	S	num=s|gen=m
riduzione	riduzione	S	S	num=s|gen=f	della	di	E	EA	num=s|gen=f
delle	di	E	EA	num=p|gen=f	pertinente	pertinente	A	A	num=s|gen=n
emissioni	emissione	S	S	num=p|gen=f	normativa	normativa	S	S	num=s|gen=f
in	in	E	E	_	comunitaria	comunitario	A	A	num=s|gen=f
nessun	nessun	D	DI	num=s|gen=m	,	,	F	FF	_
caso	caso	S	S	num=s|gen=m	compresa	comprendere	V	V	num=s|mod=p|gen=f
può	potere	V	VM	num=s|per=3|mod=i|ten=p	la	il	R	RD	num=s|gen=f
esonerare	esonerare	V	V	mod=f	direttiva	direttiva	S	S	num=s|gen=f
					96/61/CE.	96/61/CE.	S	SP	_

Ontology Learning: exemplifying Terminology Extraction

- The multi-layered architecture developed by the ItaliaNLP Lab



Ontology Learning: exemplifying Terminology Extraction

- **Single terms**
  - Linguistic filters:
    - nouns, e.g. *impianto* (plant), *direttiva* (directive)
  - Statistical filters:
    - frequency of occurrence in the input text

impianto	1,570796318
amministratore	1,570796316
emissione	1,570796316
gas	1,570796316
sostanza	1,570796316
energia	1,570796316
serra	1,570796313
produzione	1,570796312
deposito	1,570796308
tabella	1,570796306
riduzione	1,570796305
stoccaggio	1,570796304
veicolo	1,570796304
quota	1,5707963
protocollo	1,5707963
fonte	1,570796297
costruttore	1,570796297
elettricità	1,570796297
inquinamento	1,570796297
autovettura	1,570796295
aria	1,570796294
strategia	1,57079629
unità	1,570796289
carbonio	1,570796289
quantità	1,570796288
acqua	1,570796287
gestore	1,570796285
misurazione	1,570796285
conte	1,570796284
trasporto	1,570796283

Corpus of European directives in the environmental domain (Bonin et al., 2010)

Ontology Learning: exemplifying Terminology Extraction

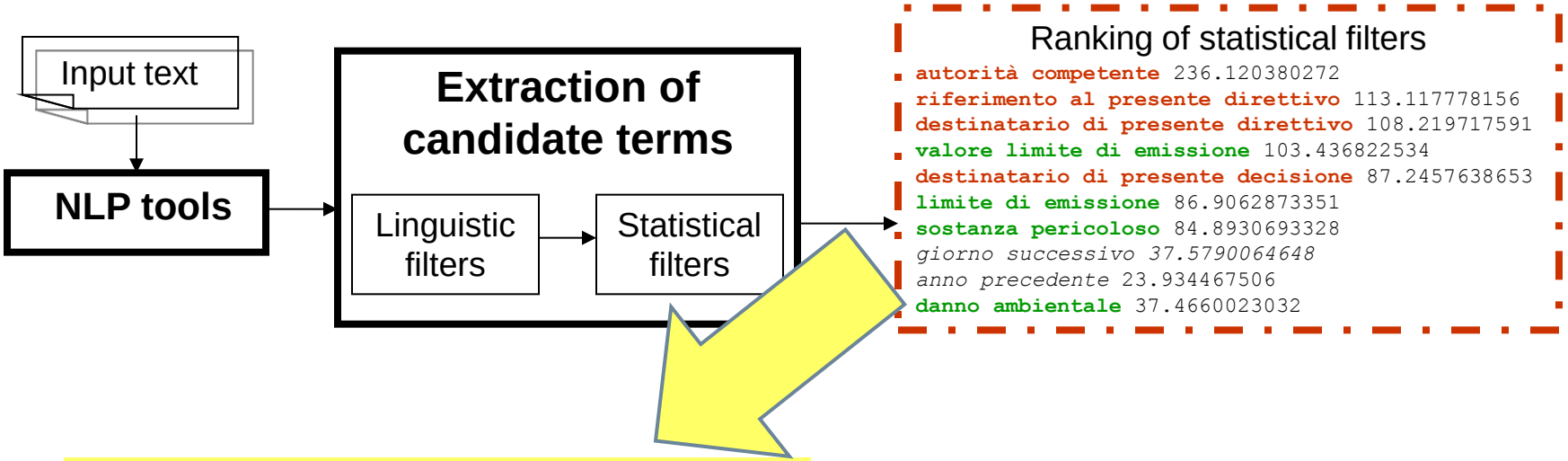
- **Single terms**
  - Linguistic filters:
    - nouns, e.g. *impianto* (plant), *direttiva* (directive)
  - Statistical filters:
    - frequency of occurrence in the input text
- **Multi-word terms**
  - Linguistic filters:
    - noun+preposition+noun, e.g. *riduzione di emissione* (emission reduction); noun+adjective (S+A), e.g. *piano nazionale* (national plan), *normativa comunitaria* (Community legislation)
  - Statistical filters:
    - C-NC Value (Frantzi & Ananiadou 1999), assessing the likelihood for a term of being a well-formed and relevant multi-word term

Corpus of European directives in the environmental domain (Bonin et al., 2010)

impianto	1,570796318
amministratore	1,570796316
emission	
gas	
sostanza	
energia	
serra	
produzion	
deposito	
tabella	
riduzione	
stoccagg	
veicolo	
quota	
protocoll	
fonte	
costruttor	
elettricità	
inquinam	
autovettu	
aria	
strategia	
unità	
carbonio	
quantità	
acqua	
gestore	
misurazi	
conte	
trasporto	

gas a effetto serra	505,722933
norma di articolo	481,0415423
emissione di gas a effetto serra	428,9508281
amministratore di registro	421,4184853
gas a effetto	395,1409139
effetto serra	326,6256871
riduzione di emissione	322,2677274
emissione di gas	305,4627825
parlamento europeo	282,4679776
energia da fonte rinnovabile	265,7397475
piano nazionale di assegnazione	220,2137528
autorità competente	216,3398553
energia da fonte	211,2850303
conto di deposito	200,1239556
cambiamento climatico	195,1698283
paese in via di sviluppo	190,1649889
quota di emissione	184,0395947
fonte energetico rinnovabile	169,4860705
fonte rinnovabile	168,9366581
qualità di aria	163,1458593
tabella relativo al piano nazionale	135,7103792
procedura di regolamentazione con controllo	132,5308836
emissione specifico	129,2489984
amministratore centrale	121,1702383
fonte energetico	117,2390528
sistema comunitario	116,0920768
piano nazionale	112,9551689
parte di presente protocollo	112,3390153
sito di stoccaggio	112,0166593
presente protocollo	108,5429485

Ontology Learning: exemplifying Terminology Extraction

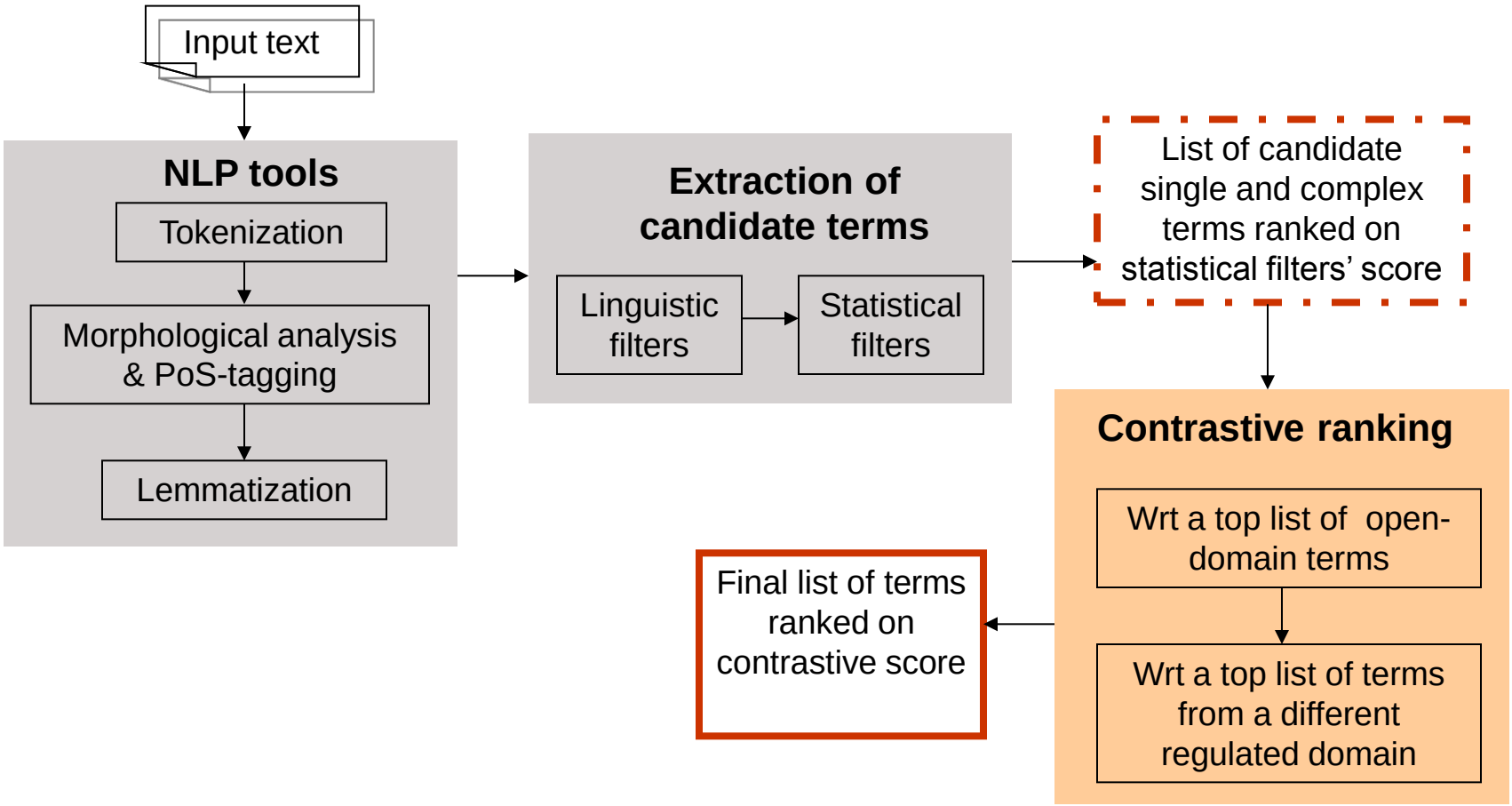


Output of the statistical filters:

Open domain terms, **legal domain terms**, **domain-specific terms** (belonging to the environmental domain) are mixed

Ontology Learning: exemplifying Terminology Extraction

- The multi-layered architecture developed by the ItaliaNLP Lab



Ontology Learning: exemplifying Terminology Extraction

Output of the 1st contrastive phase:

Open domain terms are pruned, but **legal domain terms**, **domain-specific terms** (belonging to the environmental domain) are still mixed

Ranking of statistical filters

autorità competente 236.120380272
riferimento al presente direttivo 113.117778156
destinatario di presente direttivo 108.219717591
valore limite di emissione 103.436822534
destinatario di presente decisione 87.2457638653
limite di emissione 86.9062873351
sostanza pericoloso 84.8930693328
giorno successivo 37.5790064648
anno precedente 23.934467506
danno ambientale 37.4660023032

Contrastive ranking

1st contrastive phase

valore limite 1.57079632502
destinatario di presente 1.57079632361
limite di emissione 1.57079632309
valore limite di emissione 1.57079632286
sostanza pericoloso 1.57079632218
aria ambiente 1.57079632135
riferimento al presente direttivo 1.57079632044
autorità competente 1.57079632041
destinatario di presente direttivo 1.57079631994

Contrast against a top list of terms from a general language corpus (newspaper)

Ontology Learning: exemplifying Terminology Extraction

Output of the 2nd contrastive phase:

legal domain terms are singled out by **domain-specific terms** (belonging to the environmental domain)

Ranking of statistical filters

autorità competente 236.120380272
riferimento al presente direttivo 113.117778156
destinatario di presente direttivo 108.219717591
valore limite di emissione 103.436822534
destinatario di presente decisione 87.2457638653
limite di emissione 86.9062873351
sostanza pericoloso 84.8930693328
giorno successivo 37.5790064648
anno precedente 23.934467506
danno ambientale 37.4660023032

Contrastive ranking

1st contrastive phase

valore limite 1.57079632502
destinatario di presente 1.57079632361
limite di emissione 1.57079632309
valore limite di emissione 1.57079632286
sostanza pericoloso 1.57079632218
aria ambiente 1.57079632135
riferimento al presente direttivo 1.57079632044
autorità competente 1.57079632041
destinatario di presente direttivo 1.57079631994

Contrast against a top list of terms from a general language corpus (newspaper)

Contrast against a top list of terms from a corpus of European directives regulating a different domain (consumer protection)

Final term list (2nd contrastive phase)

sostanza pericoloso 1.57079625565
salute umano 1.57079624903
sviluppo sostenibile 1.57079623794
principio attivo 1.57079622006
inquinamento atmosferico 1.57079621766
……
norma nazionale 1.57079084047
testo di disposizione 1.57078547573
testo di disposizione essenziale 1.57078274091
disposizione nazionale 1.57078159756
funzionamento di mercato interno 1.57079632044

Ontology Learning: using extracted terminology to build a legal ontology

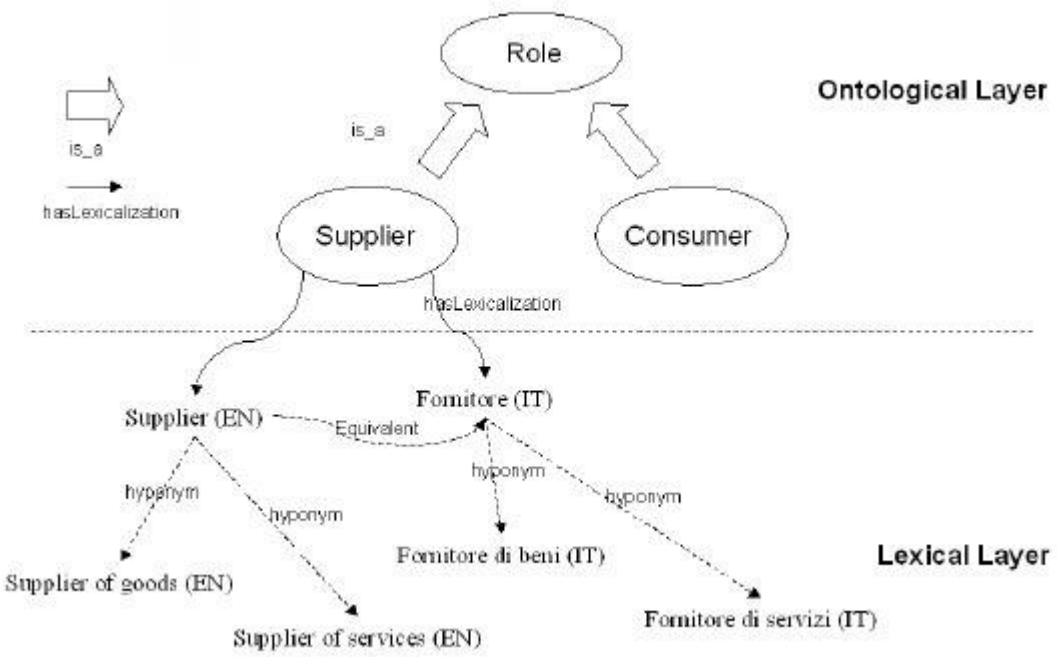
- The DALOS (*Drafting Legislation with Ontology–based Support*) European project (Agnoloni et al., 2009)
  - Aimed at
    - providing law-makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting
    - enhancing accessibility and alignment of legislation at European level

- Architecture of the DALOS Knowledge Organization System (*DALOS ontology*)
  - the **Ontological layer**, containing the conceptual modelling at a language independent level
  - the **Lexical layer**, containing multi-lingual terminology conveying the concepts represented at the Ontological layer

Ontology Learning: using extracted terminology to build a legal ontology

- The DALOS (*Drafting Legislation with Ontology–based Support*) project

- **Lexical layer**
  - Terms are
    - automatically extracted from a corpus of Consumer Protection laws
    - automatically organized into taxonomical structures
    - linked to their translation equivalent
- **Ontological layer**
  - Domain-specific concepts and their relationships manually defined by domain experts



Legal Knowledge Extraction: focus on …

- **Identification, extraction and structuring** of domain-relevant knowledge
  - Goal: constructing semantic resources such as domain-specific ontologies or lexicons
- **Semantic annotation** of legal texts
  - Goal: content-based access and querying

Semantic annotation of legal texts: towards a virtuous circle

Textual content
(implicit knowledge)

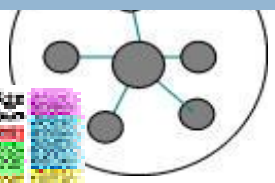
Incremental process of annotation–acquisition–annotation:
knowledge acquired from linguistically–annotated texts is projected back onto texts for extra linguistic information to be annotated and further knowledge layers to be extracted

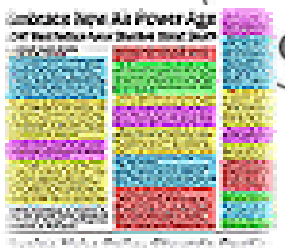
Dynamic content structuring

Linguistic annotation

Knowledge extraction

Structured knowledge
(explicit knowledge)





Semantic annotation of legal texts: what for

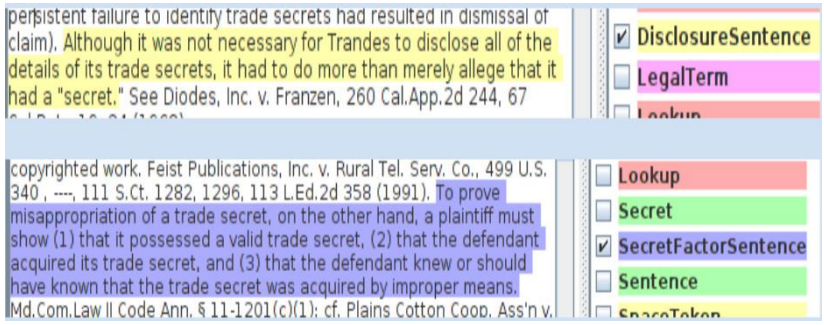
- Tasks requiring NLP-enabled knowledge extraction
  - Legal Argumentation Mining
  - Legal case elements and factors Extraction
  - Legal Text Summarization
  - Court decision Structuring
  - Legal Metadata Extraction
  - Legal definition Extraction
  - Legal citation Extraction
  - Legal Information Retrieval
  - …

Semantic annotation of legal texts: example (1)

- Legal case elements and factors Extraction for Legal Argumentation Mining
  - Adam Wyner (tomorrow morning)

- NLP tools used to make explicit relevant legal facts and legal roles starting from their linguistic realization in a collection of legal cases
  - E.g. the Appellee, Defendant, Plaintiff, etc.
  - E.g. the Disclosure-in-Negotiation fact (i.e. the fact that the plaintiff disclosed information during negotiation with defendant)

```
<ExceptionClause2>Except as <Verb>specified</Verb> in paragraph c</
ExceptionClause2>, <AgentNP>you, an establishment that <Verb>collects</Verb>
blood,</AgentNP> <Obligation>must</Obligation> <Verb>test</Verb>
<ThemeNP>each donation of human blood that <Passive><Verb>is</Verb>
<Verb>intended</Verb></Passive> for use in preparing a product for evidence
of infection due to the following communicable disease agents</ThemeNP>:
```

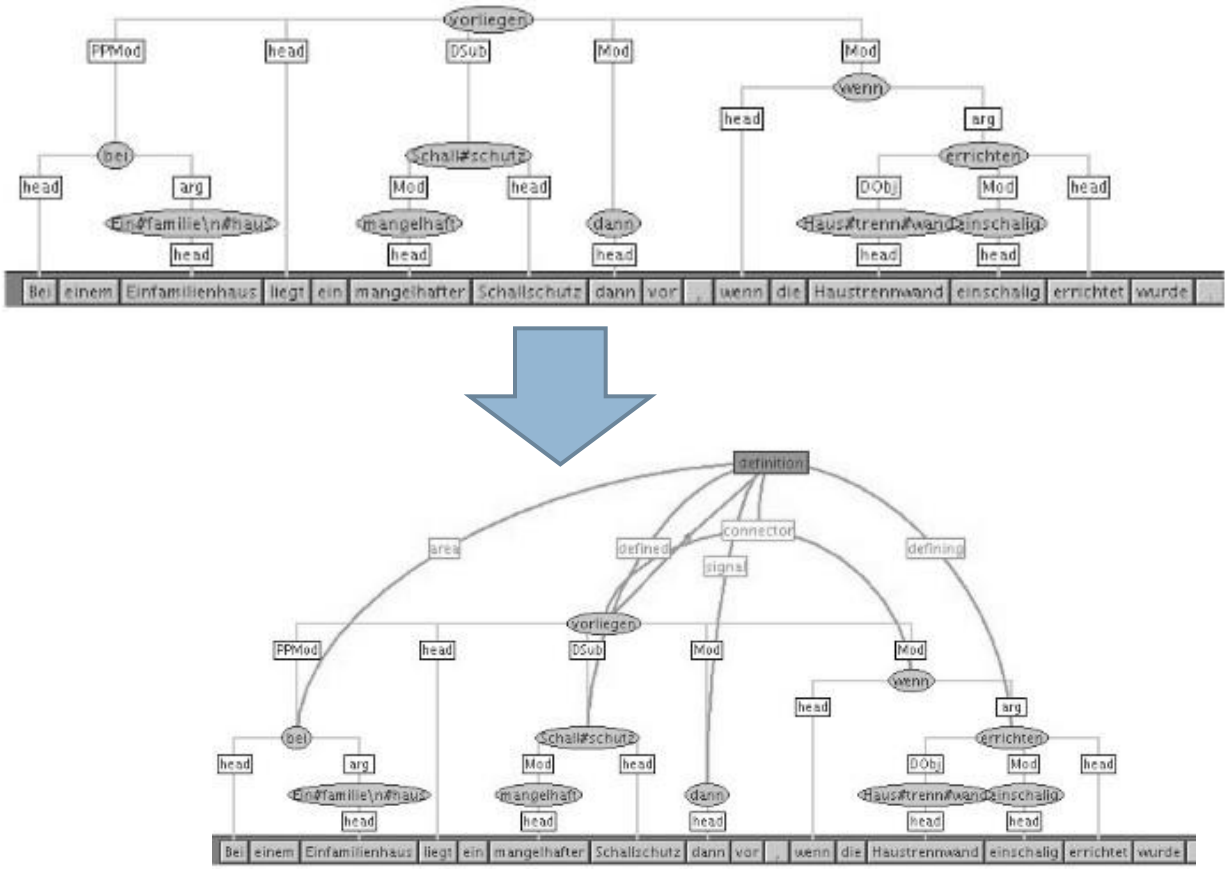


The annotation are the building blocks of a language of formal rules

Semantic annotation of legal texts: example (2)

- Legal definition Extraction
  - Walter and Pinkal, 2006: from German court decisions

- NLP tools are used to identify legal definitions on the basis of the linguistic realization of *definiendum* and *definiens*
  - "One-family row-houses have *insufficient noise insulation* if the separating wall is one-layered"
- The linguistic structure is transformed to a semantic representation by a series of heuristic rules
- Promising step for Ontology Learning purposes



Semantic annotation of legal texts: example (3)

- Legal Metadata Extraction
  - Focus on MELT (*Metadata Extraction from Legal Texts*) system jointly developed by ILC and ITTIG
    - It combines
      - a set of tools which transform a plain text in XML, detect references and classify provisions (i.e. xmLeges tools)
      - a suite of NLP tools for the analysis of Italian texts
    - It aims at supporting the consolidation of legislative texts process (in force law)
    - It provides a formalized representation of textual amendments by a metadata set
      - Repeal, substitution and integration
    - The text modification is performed on the metadata interpretation

Semantic annotation of legal texts: example (3)

Metadata	Description
pos	information on the amending provision
pos:xlink	ID reference to the amending provision
norm	information on the norm to be amended
norm:xlink	URN reference to the norm
– pos	further information on the norm
– pos:xlink	URN reference to the norm with the partition ID
– – border	information on further narrower container
– – border:type	container type (e.g. point, "alinea", period, etc.)
– – border:num	container label expressed by a number or a letter
– – border:ord	container position expressed by an ordinal (e.g. 2nd) or a relative (e.g. last) number
position	information on the specific modifying point within the narrowest container
– pos	information on a string (quoted) and a bound of the deleting or inserting point
– pos:xlink	ID reference to the string, a bound of which is the beginning of the modifying text
– pos:where	specific bound of the string or container (before, after, start, end)

novellando	information on the outgoing text
– type	information on the "novellando" type
– type:value	"novellando" type (e.g. article, paragraph, "alinea", period, words, etc.)
– pos	information on the outgoing string (in quotes)
– pos:xlink	ID reference to the string that is either the outgoing text, or the beginning or ending of the outgoing text
– – role	information on the meaning of the string
– – role:value	string role: beginning (from) or ending (up to) of the outgoing text
novella	information on the incoming text
– type	information on the "novella" type
– type:value	novella type (e.g. article, paragraph, "alinea", period, words, etc.)
– pos	information on the incoming string (quoted)
– pos:xlink	ID reference to the incoming string

Semantic annotation of legal texts: example (3)

- Legal Metadata Extraction
  - Focus on MELT (*Metadata Extraction from Legal Texts*) system jointly developed by ILC and ITTIG
  - An example
    - "All'**articolo 1, comma 1, della legge 8 febbraio 2001, n. 12**, la lettera d) è abrogata" (In **article 1, paragraph 1, of the act 8 February 2001, n. 12**, letter d) is repealed)

Semantic annotation of legal texts: example (3)

- Legal Metadata Extraction
  - Focus on MELT (*Metadata Extraction from Legal Texts*) system jointly developed by ILC and ITTIG
  - An example
    - "All'**REF mod31-rif2#art1-com1**, la lettera d) è abrogata" (In REF mod31-rif2#art1-com1, letter d) is repealed)

Semantic annotation of legal texts: example (3)

- Legal Metadata Extraction
  - Focus on MELT (*Metadata Extraction from Legal Texts*) system jointly developed by ILC and ITTIG
  - An example
    - The sentence was linguistically analyzed ay a shallow syntactic level of analysis

```
[[CC:P_C][PREP:A#E][DET:LO#RD@MS][POTGOV:REF#SP@NN]]
[[CC:U_C][FORM:MOD31-RIF1#ART1-COM1]] [[CC:PUNC_C][PUNCTYPE: ,#@]]
[[CC:N_C][DET:LO#RD@FS][AGR:@FS][POTGOV:LETTERA#S@FS]]
[[CC:N_C][AGR:@FS][POTGOV:D#S@FS]]
[[CC:PUNC_C][PUNCTYPE: )#@]]
[[CC:FV_C][AUX:ESSERE#V@S3IP][POTGOV:ABROGARE#V@FSPR]]
[[CC:PUNC_C][PUNCTYPE: .#@]]
```

Semantic annotation of legal texts: example (3)

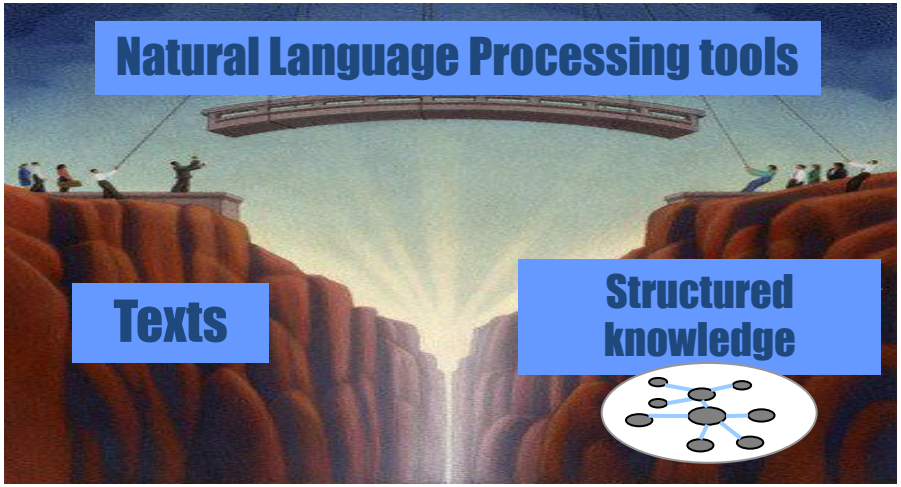
- Legal Metadata Extraction
  - Focus on MELT (*Metadata Extraction from Legal Texts*) system jointly developed by ILC and ITTIG
  - An example
    - The annotation of informative metadata was carried out by a finite-state compiler which uses a specialized grammar covering the amendment types considered on the basis of patterns formalized in terms of regular expressions operating over sequences of chunks

```
<repeal>
In <norm>REF mod31-rif1#art1-com1#</norm>, <border>
<border:type>letter</border:type>
<border:num>d)</border:num></border> is repealed.
</repeal>
```

Conclusion

- Natural Language Processing techniques represent a key ingredient for Legal Knowledge Management

Knowledge Creation:
Legal Ontologies and Lexicons



Knowledge Use:
"Intelligent" content access

Hopefully, thanks to NLP Legal Search Engines will be able to access the content embedded in texts more effectively

IUS EXPLORER

Conclusion

*One of the main obstacles to progress in the field of artificial intelligence and law is the **natural language barrier***

L. Thorne McCarty, International Conference on AI and Law (ICAIL-2007)

Natural Language Processing
combined with
Knowledge Extraction techniques
can help removing or at least penetrating
the natural language barrier in the AI&Law field

Credits

- The NLP tools and techniques have been developed in the framework of the activities of the people of *ItaliaNLP Lab* at the Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC-CNR)
  - http://www.italianlp.it/

- Special thanks to Felice Dell’Orletta

On-line demos

- Linguistic analysis of Italian and English texts
  - http://www.ilc.cnr.it/dylanlab/index.php?page=texttools&hl=it_IT&tmid=tm_source
- Term extraction from Italian and English texts
  - http://www.ilc.cnr.it/dylanlab/index.php?page=texttools&hl=it_IT&tmid=tm_term_extractor

References

- **Ontology Learning**
  - Buitelaar, P. Cimiano, P., Magnini, B. (Eds.) Ontology Learning from Text: Methods, Evaluation and Applications. Frontiers in Artificial Intelligence and Applications Series, Vol. 123, IOS Press, July 2005.
  - Cimiano, Philipp Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. 2006, XXVIII, Springer, 2006.
  - Maedche, Alexander Ontology Learning for the Semantic Web. Kluwer Academic Publishers, 2002.

- **Ontology Learning in the legal domain (1)**
  - Núria Casellas, “Legal Ontology Engineering. Methodologies, Modeling Trends and the Ontology of Professional Judicial knowledge”. Law, Governance and Technology Series (Vol. 3), Springer-Verlag, 2011
  - Valente A. and Breuker J., Ontologies: the missing link between Legal Theory and AI & Law. In A. Soeteman (ed.) Legal Knowledge Based Systems JURIX 94, pp. 139-149, 1994.
  - Van Engers T., Boer A., Breuker J., Valente A., Winkels R., Ontologies in the legal domain. In Chen et al. (eds.) Digital Government. E-Government Research, Case studies and Implementation, Springer, pp. 233-261, 2008.
  - Cimiano, P., Völker, J. Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In Montoyo et al. (eds.), Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB). pp. 227–238. Springer, Alicante, Spain, June 2005.
  - Lame, G. Knowledge acquisition from texts towards an ontology of French law. in Proceedings of the International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks (EKAW-2000), Juan-les-Pins, 2000.
  - Lame, G. Using NLP techniques to identify legal ontology components: concepts and relations. In Benjamins et al. (eds.), Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications. Lecture Notes in Computer Science, Volume 3369: 169–184, 2005.
  - Saias, J. and P. Quaresma A Methodology to Create Legal Ontologies in a Logic Programming Based Web Information Retrieval System. In Benjamins et al. (eds.), Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications. Lecture Notes in Computer Science, Volume 3369: 185–200, 2005.

References

- **Ontology Learning in the legal domain (2)**
  - Breuker J and R. Hoekstra, Epistemology and Ontology in Core Ontologies: FOLaw and LRI-Core, two core ontologies for law, in Proceedings of the Workshop on Core Ontologies in Ontology Engineering (EKAW04), Northamptonshire, UK, pp. 15-27, 2004.
  - T. Agnoloni, L. Bacci, E. Francesconi, W. Peters, S. Montemagni, G. Venturi, 2008, A two-level knowledge approach to support multilingual legislative drafting, in Joost Breuker, Pompeu Casanovas, Michel C.A. Klein, Enrico Francesconi (eds.), Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Frontiers in Artificial Intelligence and Applications, Springer, Volume 188, 2008, pp. 177-198.
  - Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, Giulia Venturi, 2008, Ontology learning from Italian legal texts, in Joost Breuker, Pompeu Casanovas, Michel C.A. Klein, Enrico Francesconi (eds.), Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Frontiers in Artificial Intelligence and Applications, Springer, Volume 188, 2008, pp. 75-94.
  - Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds.), 2010, SEMANTIC PROCESSING OF LEGAL TEXTS, LNCS 6036, Springer
    - PART II - Legal Text Processing and Construction of Knowledge Resources
      - Automatic Identification of Legal Terms in Czech Law Texts by Karel Pala, Pavel Rychlý, Pavel Šmerk
      - Integrating a Bottom-Up and Top-Down Methodology for Building Semantic Resources for the Multilingual Legal Domain by Enrico Francesconi, Simonetta Montemagni, Wim Peters, Daniela Tiscornia
      - Ontology Based Law Discovery by Alessio Bosca, Luca Dini
      - Multilevel Legal Ontologies by Gianmaria Ajani, Guido Boella, Leonardo Lesmo, Marco Martin, Alessandro Mazzei, Daniele P. Radicioni, Piercarlo Rossi
  - Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni, 2010, Singling out Legal Knowledge from World Knowledge. An NLP-based approach, In Proceedings of LOAIT 2010, Fiesole, 7 July 2010

References

- **Legal Language Processing (computational linguistics literature)**

  - Dell'Orletta F., Marchi S., Montemagni S., Plank B. e Venturi G. (2012), The SPLeT–2012 Shared Task on Dependency Parsing of Legal Texts. In Proceedings of the 4th Workshop on "Semantic Processing of Legal Texts", held in conjunction with LREC 2012, Istanbul, Turkey, 27th May
  - Giuseppe Attardi, Daniele Sartiano and Maria Simi (2012), Active Learning for Domain Adaptation of Dependency Parsing on Legal Texts. In Proceedings of the 4th Workshop on "Semantic Processing of Legal Texts"
  - Alessandro Mazzei, Cristina Bosco (2012), Simple Parser Combination. In Proceedings of the 4th Workshop on "Semantic Processing of Legal Texts"
  - Dell'Orletta F., Marchi S., Montemagni S., Venturi G., Agnoloni T. e Francesconi E. (2012), Domain Adaptation for Dependency Parsing at Evalita 2011. In Working Notes of EVALITA 2011, 24th–25th January 2012, Rome, Italy, ISSN 2240–5186
  - Barbara Plank and Anders Søgaard (2012), Experiments in newswire-to-law adaptation of graph-based dependency parsers. In Working Notes of EVALITA 2011, 24th–25th January 2012, Rome, Italy, ISSN 2240–5186
  - Giuseppe Attardi, Maria Simi, and Andrea Zanelli (2012), Domain Adaptation by Active Learning . In Working Notes of EVALITA 2011, 24th–25th January 2012, Rome, Italy, ISSN 2240–5186
  - Venturi G. (2012), Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts. In Proceedings of the 4th Workshop on "Semantic Processing of Legal Texts", held in conjunction with LREC 2012, Istanbul, Turkey, 27th May
  - Venturi G. (2008), Parsing Legal texts. A contrastive study with a View to Knowledge Management Applications. In Proceedings of the "Ist Workshop Semantic Processing of Legal Texts", held in conjunction with LREC, Marrakech, Morocco, May 26–1,
  - Venturi G. (2012) Lingua e diritto: una prospettiva linguistico-computazionale, tesi di dottorato dell'Università di Torino, ottobre 2012.

References

- **Semantic Processing of legal texts**
  - Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds.), 2010, SEMANTIC PROCESSING OF LEGAL TEXTS, LNCS 6036, Springer
  - Bartolini R., Lenci A., Montemagni S., Pirrelli V., Soria C., Automatic classification and analysis of provisions in legal texts: a case study. In A.R. Meersman, Z. Tari, A. Corsaro (eds.) On the Move to Meaningful Internet Systems (OTM 2004 Workshops), LNCS, vol. 3292, pp. 593-604. Springer-Verlag, 2004.
  - Mazzei A., Radicioni D.P. and Brighi R., NLP-based extraction of modificatory provisions semantics. In Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009), pp. 50-57, Barcelona, Spain, 2009.
  - Spinosa P.L., Giardiello G., Cherubini M., Marchi S., Venturi G. and Montemagni S., NLP-based metadata extraction for legal text consolidation. In Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009), pp. 40-49, Barcelona, Spain, 2009.
  - Walter S. and Pinkal M., Definitions in Court Decisions. Automatic Extraction and Ontology Acquisition. In Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web, IOS Press, pp. 95-113, 2009.
  - Wyner A., Towards annotating and extracting textual legal case elements. In Proceedings of the IV Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2010), pp. 9-18, Fiesole, Italia, 2010.
  - Wyner A. and Peters W., Lexical semantics and expert legal knowledge towards the identification of legal case factors. In Proceedings of Legal Knowledge and Information Systems (JURIX) Conference, pp. 127-136, Liverpool, United Kingdom, 2010a. IOS Press.
  - Wyner A. and van Engers T., From argument in natural language to formalised argumentation: Components, prospects and problems. In Proceedings of the Worskhop on Natural Language Engineering of Legal Argumentation (NaLEA2009), Barcelona, Spain, 2009.
  - Zhang P. and Koppaka L., Semantics-base legal citation network. In Proceedings of the 11° International Conference on Artificial Intelligence and Law, pp. 123-130, 2007.