NATURAL LANGUAGE PROCESSING AND LEGAL KNOWLEDGE EXTRACTION

Giulia Venturi

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

ItaliaNLP Lab – www.italianlp.it

Summer School LEX2015 - 10th September 2015

From text to knowledge: the starting point

- 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

From text to knowledge: the starting point

- 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

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

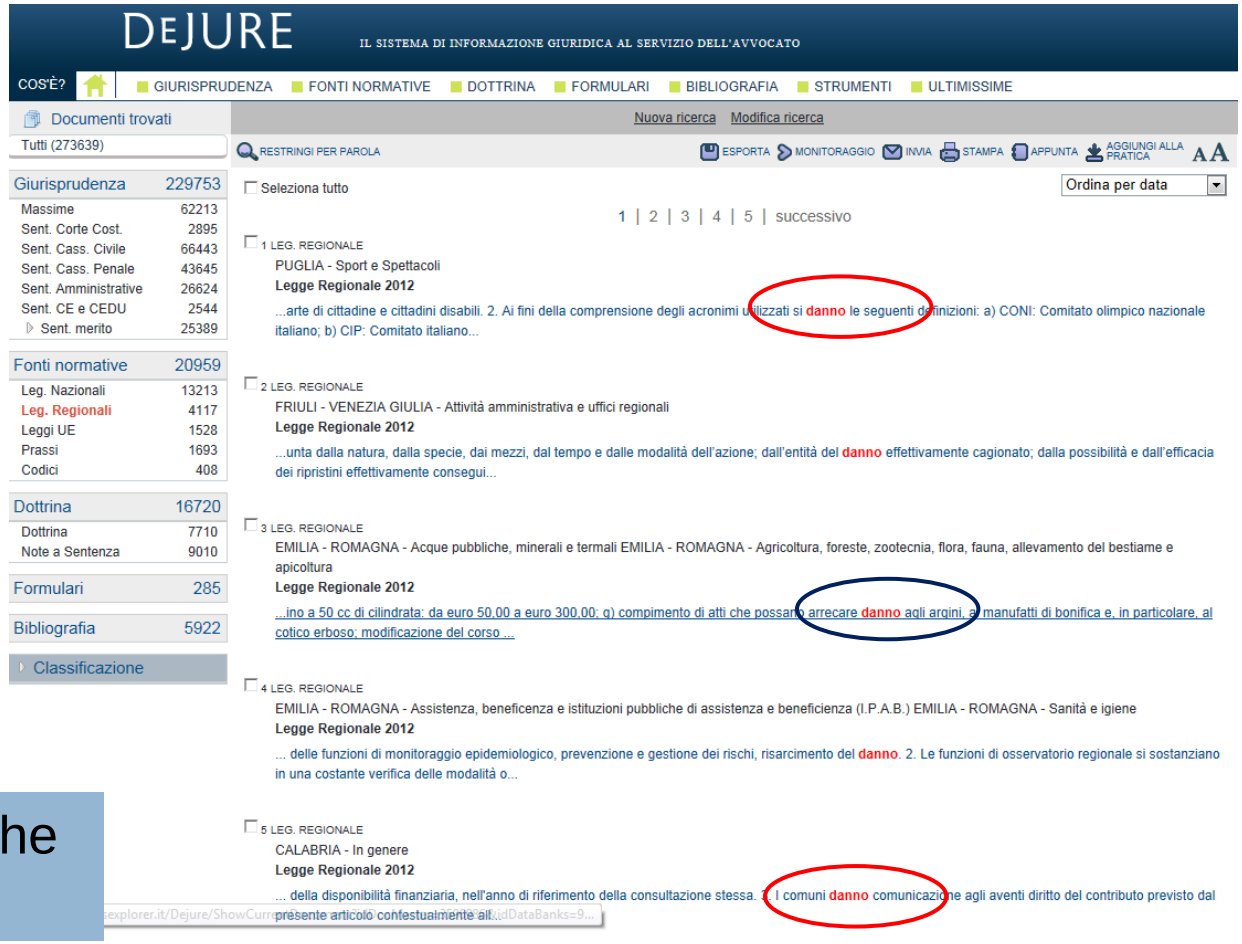
L. Thorne McCarty, International Conference on AI and Law (ICAIL-2007)

IUSEXPLORER

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

IUSEXPLORER: an example of word search query

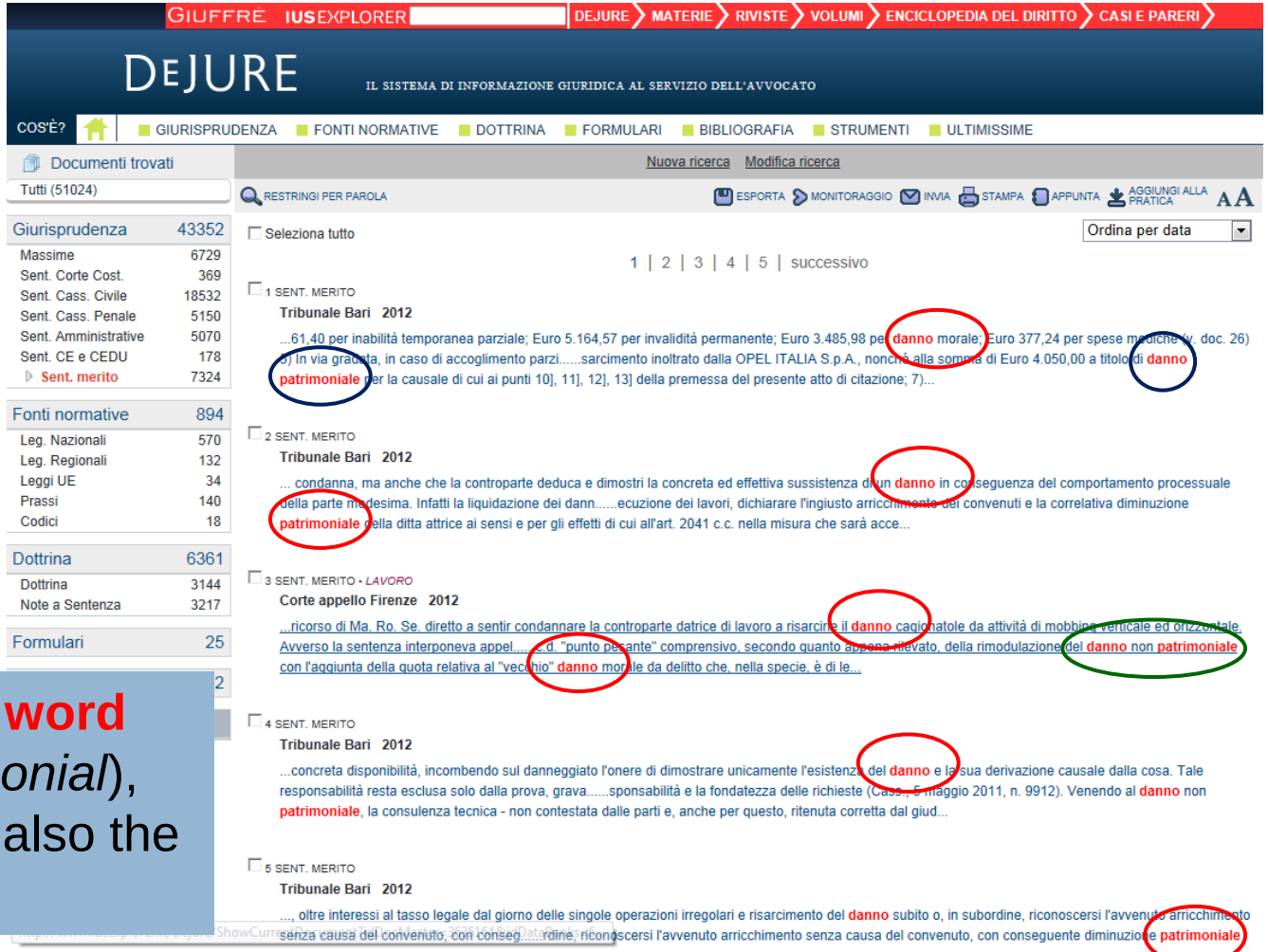
danno (damage)



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

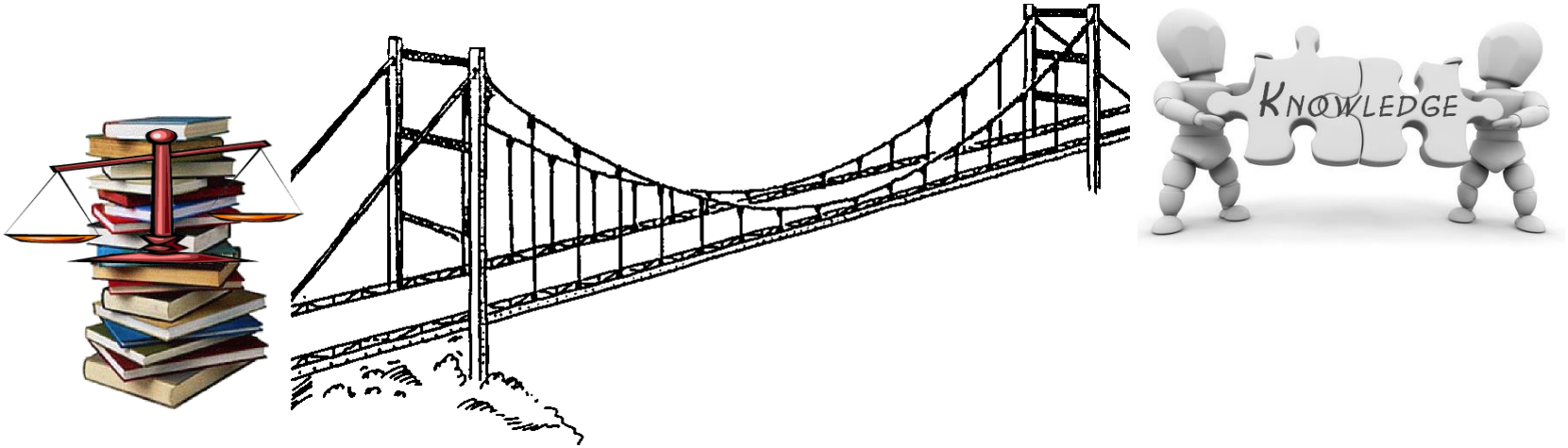
- It is still linguistically rudimentary
  - it does not exploit the potential offered by language technologies
  - it does not support semantic queries allowing an advanced access to documents

The need for increasingly sophisticated applications based on Natural Language Processing technologies and aimed at overcoming the knowledge acquisition bottleneck

Summary

- Natural Language Processing tools
  - What and what for
  - An example
- From text to knowledge
  - The general approach
  - The main challenges of the legal domain
- Legal Knowledge Extraction
  - What and what for
  - An example

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 allow
  - accessing the **linguistic** and **domain-specific** knowledge contained in texts
  - structuring the textual content

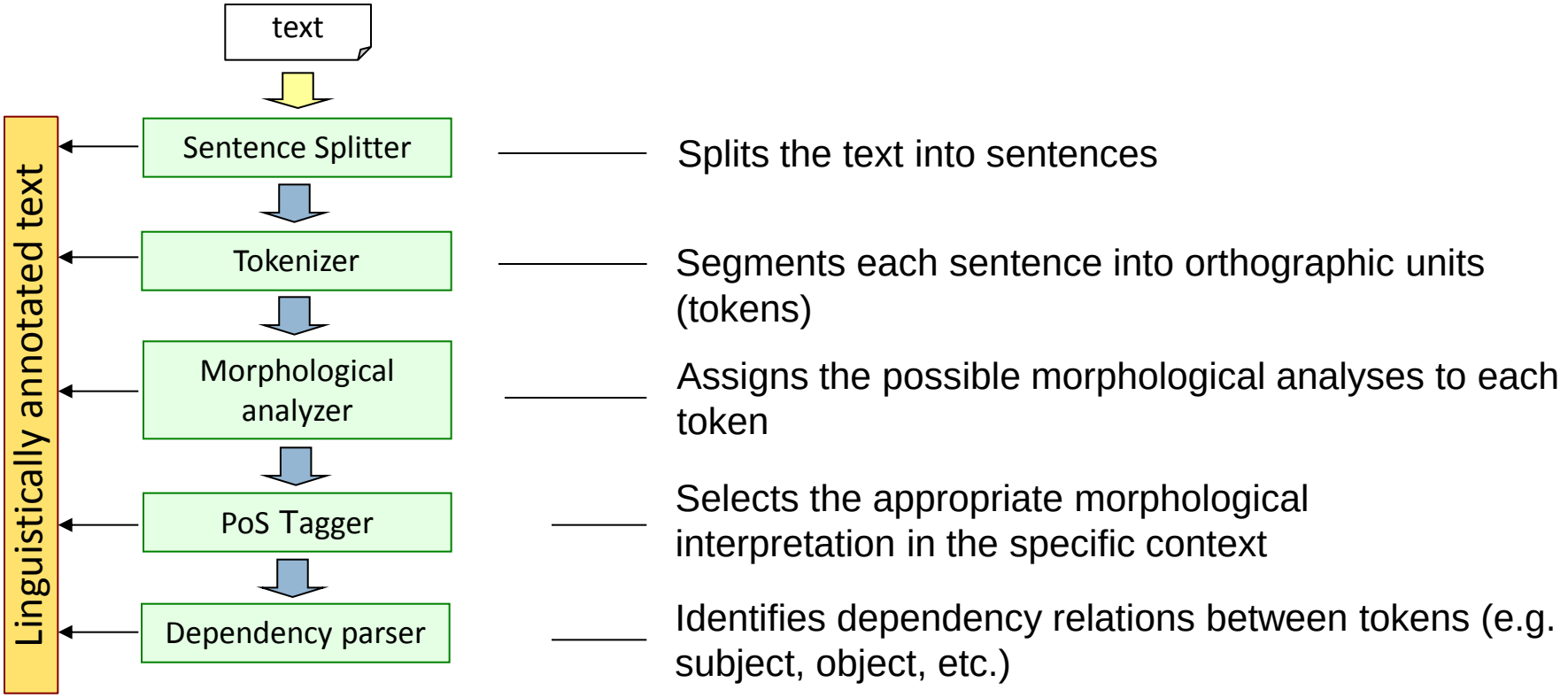
The Natural Language Processing tools: what

- Tools that enable computers to derive meaning from human or natural language input

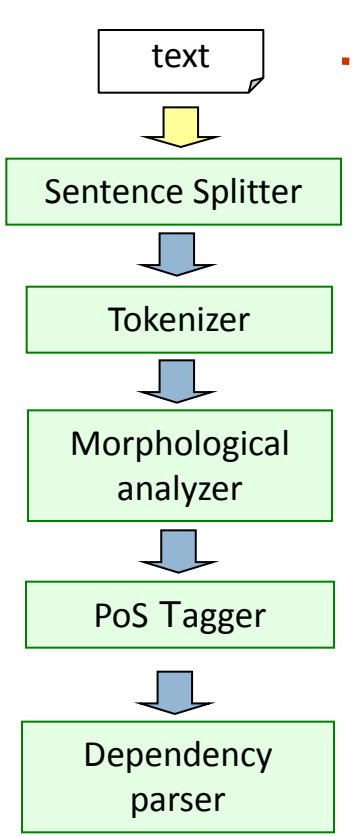
- They are a component of artificial intelligence, computer science and linguistics concerned with processing texts and making information accessible to computer applications

- They make machine-readable the linguistic structure implicitly embedded in texts
  - Automatic linguistic annotation process

Linguistic annotation: an incremental process

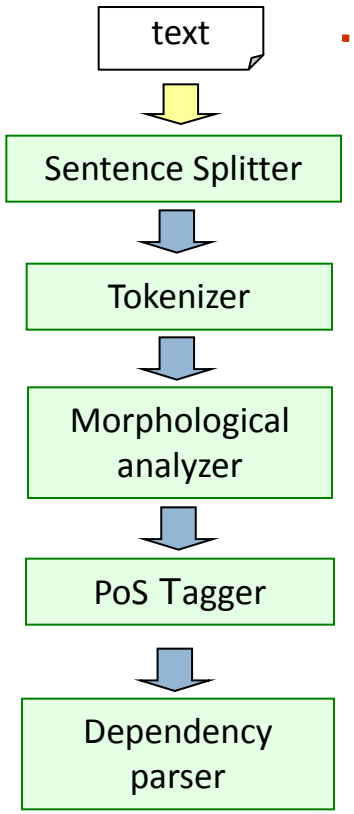


Linguistic annotation: an example



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

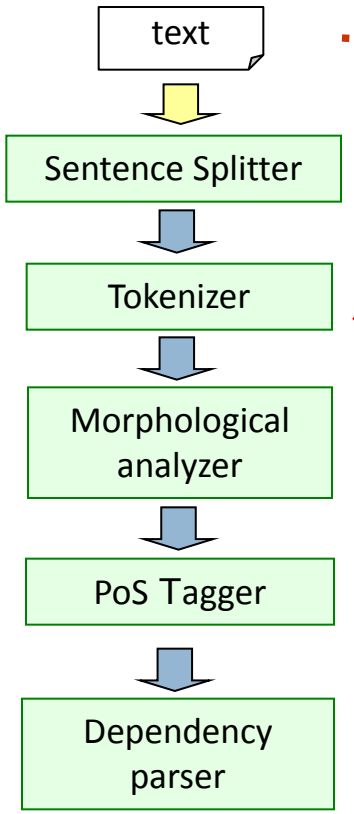


Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

- 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.)

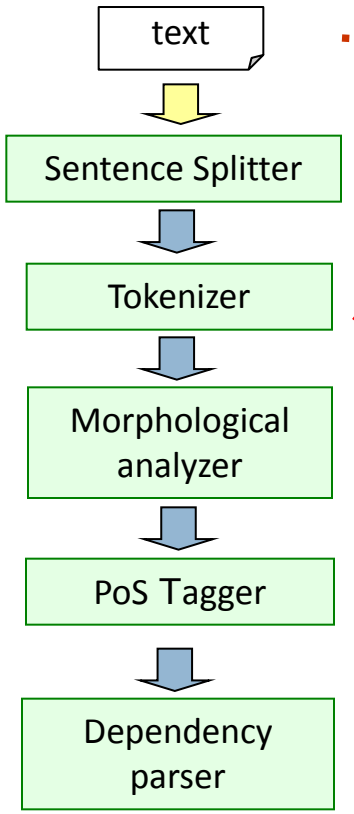
- Il danno non poteva essere sottovalutato. (The damage could not be understimated.)

- 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.)

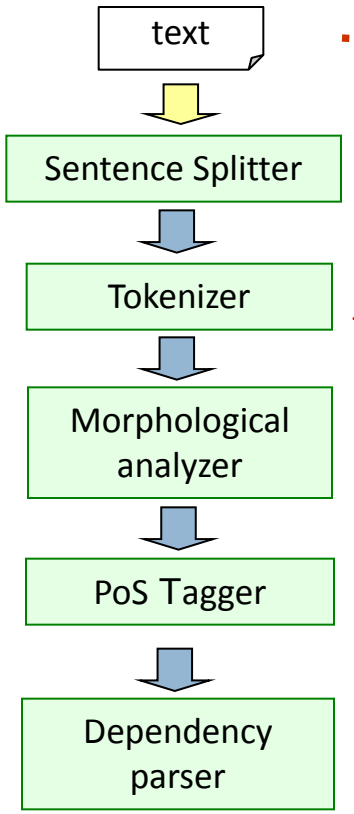
- Il danno non poteva essere sottovalutato. (The damage could not be understimated.)

- 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.)

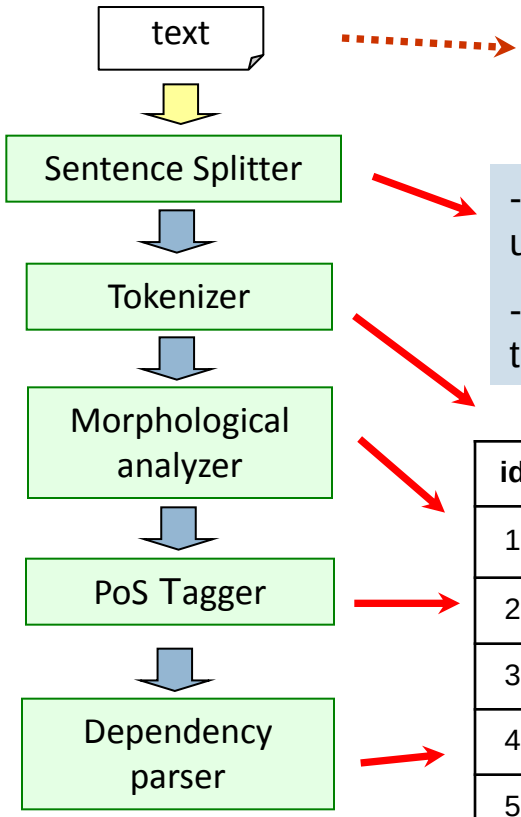
- Il danno non poteva essere sottovalutato. (The damage could not be understimated.)

- 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

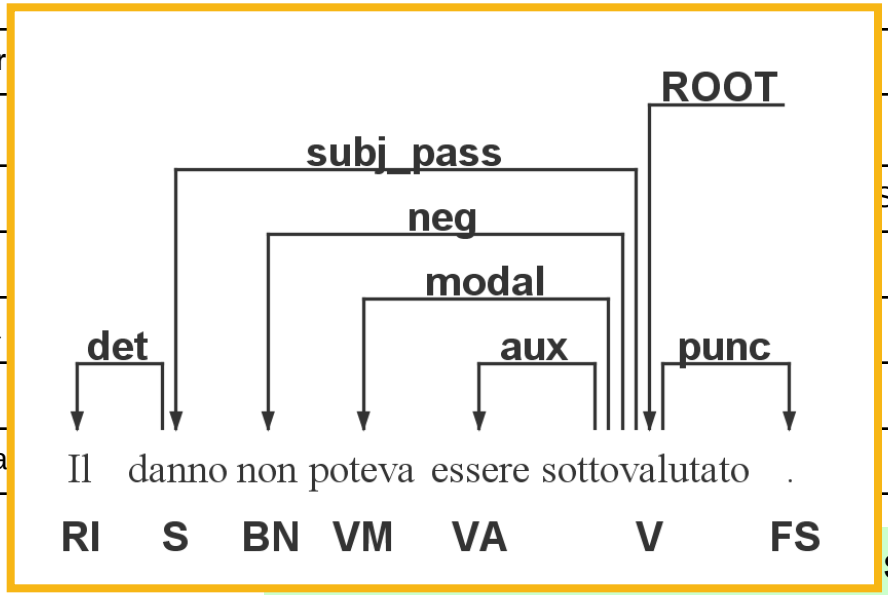


Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

- 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	for
1	Il
2	danno
3	non
4	poteva
5	essere
6	sottova



schema

The linguistic annotation tools @ ItaliaNLP Lab

- **LinguA** is a state-of-the-art linguistic annotation pipeline which combines rule-based and machine learning algorithms
  - developed by ILC and the University of Pisa

- Morpho-syntactic annotation (PoS tagger developed by Dell'Orletta, 2009)
  - Evalita 2009: accuracy = 96,34%
  - State-of-the-art for Italian
- Dependency syntactic annotation (DeSR parser, Attardi & Dell'Orletta, 2009)
  - Conll-2007: 81.3% LAS
  - Evalita 2009: 83.38% LAS
  - State-of-the-art for Italian

Demo at http://www.italianlp.it/demo/linguistic-annotation-tool/

Linguistic Annotation Pipeline

www.italianlp.it

Text | Sentence Splitting | Part of Speech Tagging | Syntactic Parsing | Syntactic Trees

LinguA is a linguistic annotation pipeline which combines rule-based and ML algorithms.

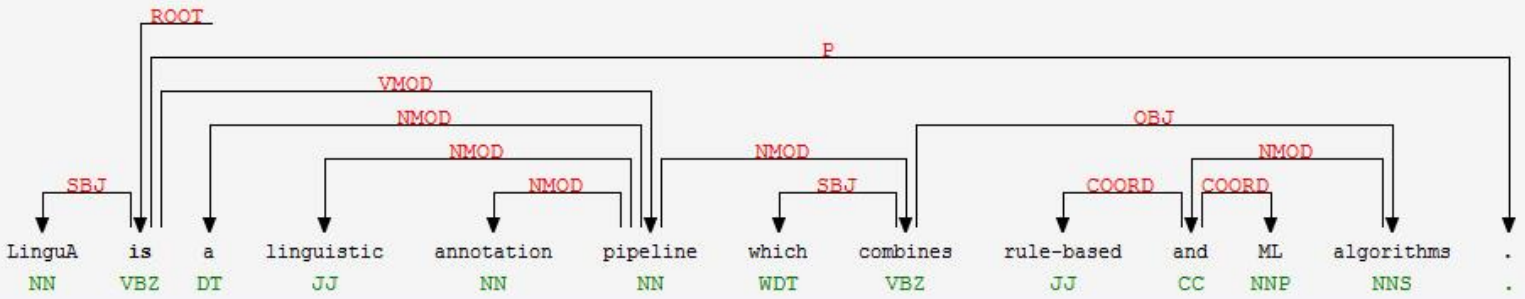
English text | Run | Clear

See your analysis or download here!

Linguistic Annotation Pipeline

www.italianlp.it

Text | Sentence Splitting | Part of Speech Tagging | Syntactic Parsing | Syntactic Trees

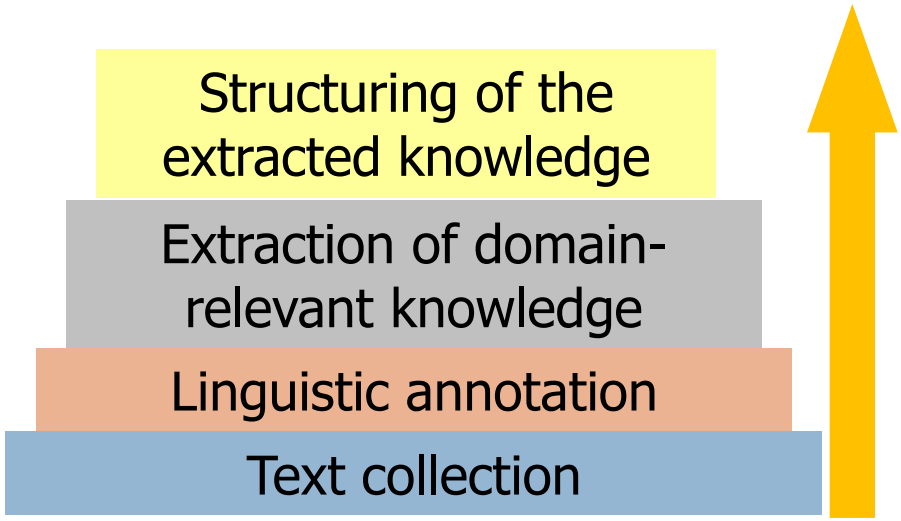


Thanks to Stefano Dei Rossi for the dependency graph visualization.

Linguistic annotation: what for

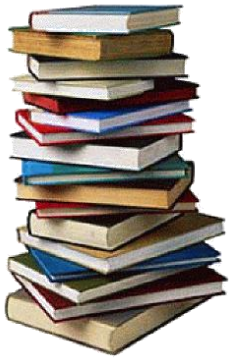
- Linguistic annotation plays a crucial role in accessing the content of texts by making it explicit the linguistic structure through which knowledge is encoded

- 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



From text to knowledge: the general approach

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



Textual content
(implicit knowledge)

Linguistic annotation

Dynamic content structuring

Knowledge extraction

Structured knowledge
(explicit knowledge)

From text to knowledge: the main challenges of the legal domain

- **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

From text to knowledge: the main challenges of the legal domain

- **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 and its impact on NLP tools

- Legal texts differ significantly with respect to ordinary language texts (e.g. newspapers)
  - typically correlated with syntactic complexity
- Dramatic drop of accuracy when NLP tools are tested on domains outside of the data from which they are trained or developed on (Gildea, 2001)
  - Key role of **natural language syntactic parsing** which represents a prerequisite for any advanced legal text processing task
- What is the performance of state-of-the-art NLP tools on legal texts?
  - A key issue for all NLP-based Legal Knowledge Extraction tasks

The peculiarity of legal language and its impact on NLP tools

- Recently, two initiatives 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 initiatives:
  - Domain Adaptation Track at Evalita 2011 – Italian
    - http://www.italianlp.it/software/evalita-2011-domain-adaptation-for-dependency-parsing/
  - SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts – Italian and English
    - http://www.italianlp.it/software/first-shared-task-on-dependency-parsing-of-legal-texts-at-splet-2012/

The peculiarity of legal language and its impact on NLP tools

- The Evalita 2011 results
  - for **dependency parsing**

Evalita 2011 – Domain Adaptation Task

Training	Test	Performance	Performance after Domain Adaptation
Newspaper 	Newspaper 	82.09% Labelled Attachment Score (LAS)	---
Newspaper 	Legal texts 	75.85% LAS	80.83% LAS

- 6.24 %

+ 5 %

From text to knowledge: the main challenges of the legal domain

- **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

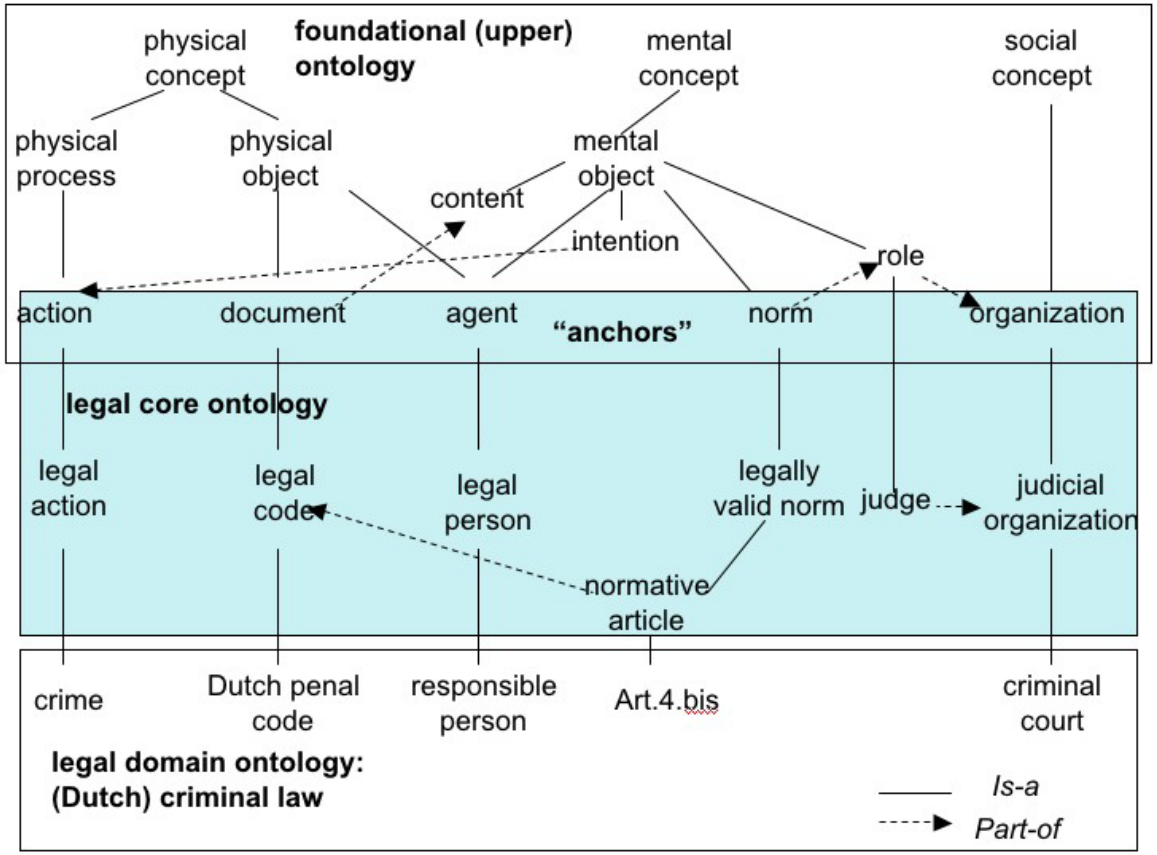
Discriminate between legal and regulated domain knowledge

- 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
  - e.g. *national provision, fundamental principle & hazardous substance, active ingredient*

- Discriminating between legal and regulated domain terms and/or concepts is key in constructing a legal semantic resource
  - It can be a helpful starting point for any further construction of domain-specific knowledge base where domain-relevant and the specific domain knowledge is kept separate
  - It is closely related to the reusability and interoperability issue

Discriminate between legal and regulated domain knowledge

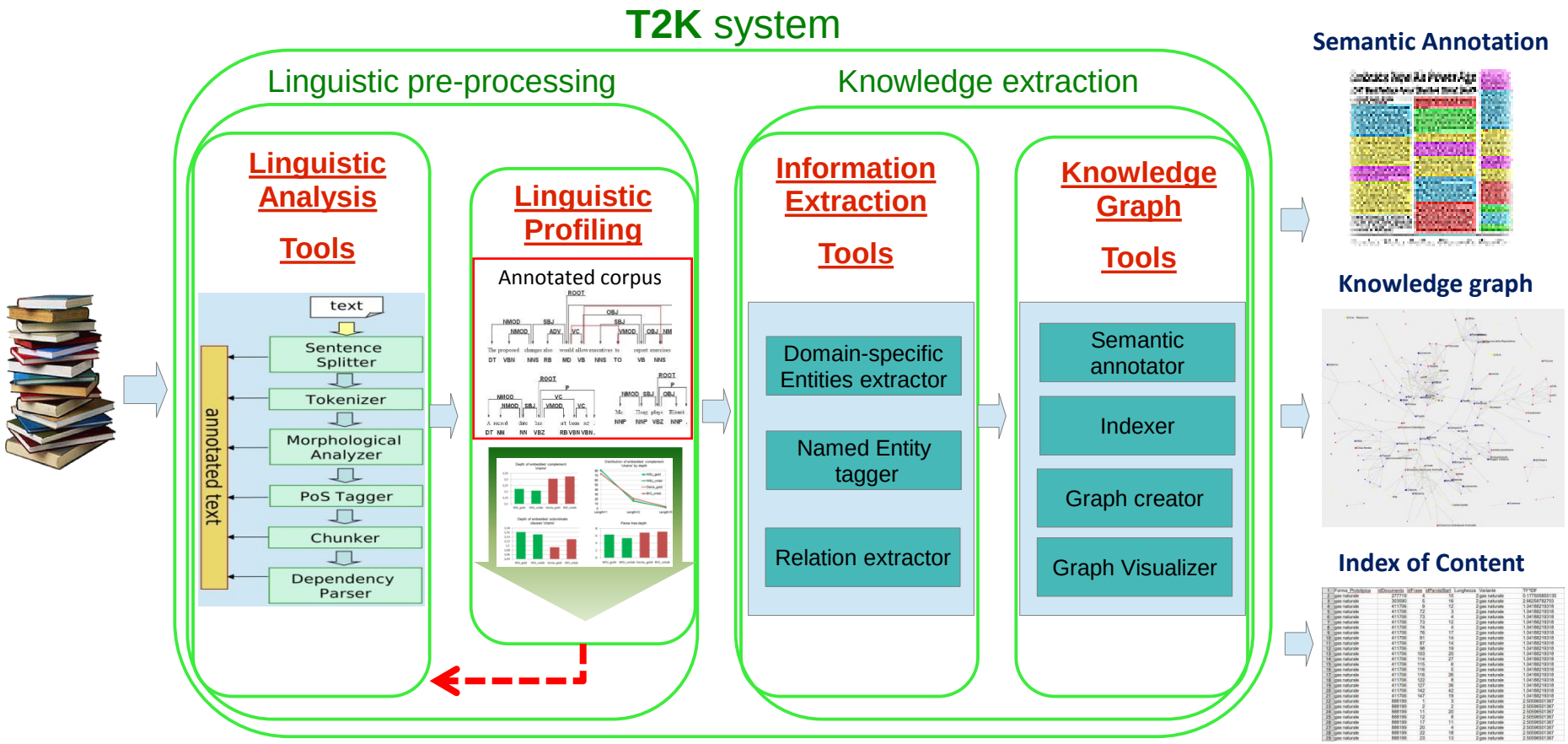
- According to the ontology design criteria, the level of generality in which concepts are organized is a distinctive characteristic
- Three different kinds of ontologies:
  - top or upper-level ontologies (general concepts)
  - core ontologies (top-level domain-specific concepts, e.g. legal)
  - domain-specific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



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

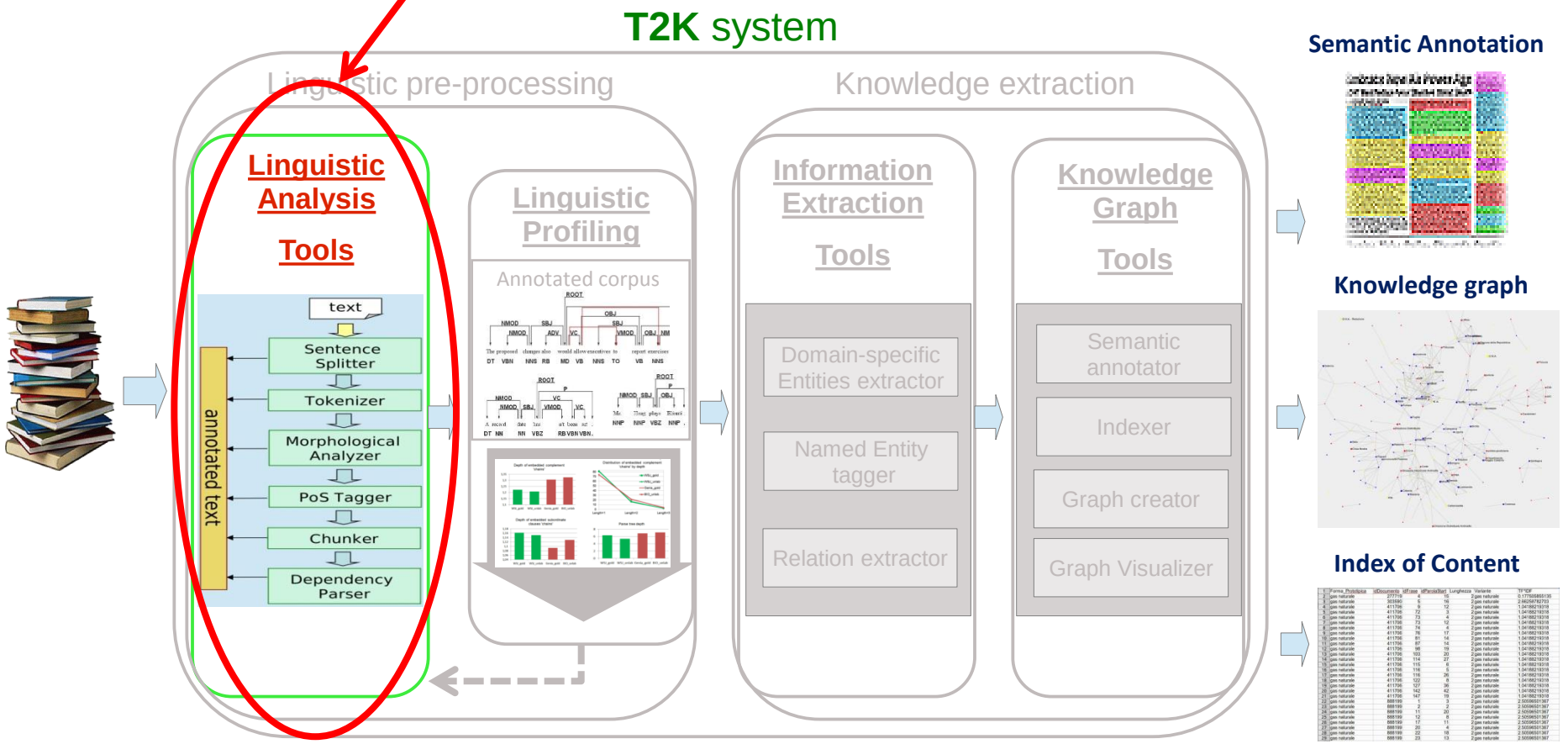
From text to knowledge @ ItaliaNLP Lab

T2K (Text-To-Knowledge) combines a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning which are dynamically integrated to provide an accurate representation of the domain-specific context of text corpora in different domains (Dell'Orletta et al., 2014)

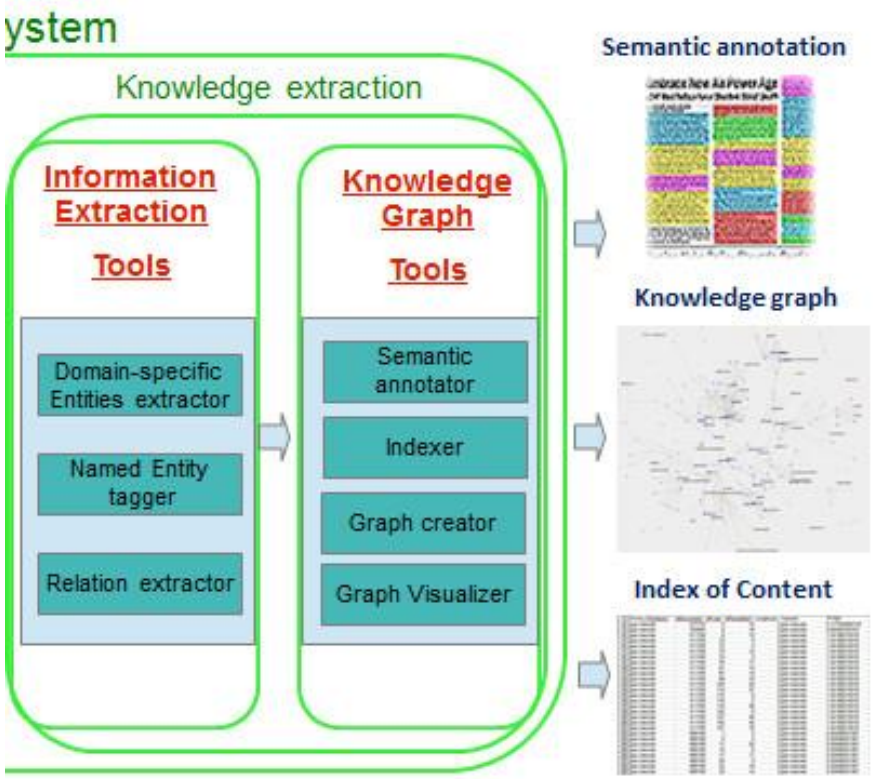


From text to knowledge @ ItaliaNLP Lab

In T2K the NLP tools were trained on two training sets: the ISST-TANL treebank consisting of newspaper articles and a syntactically annotated corpus of Italian legislative and administrative texts

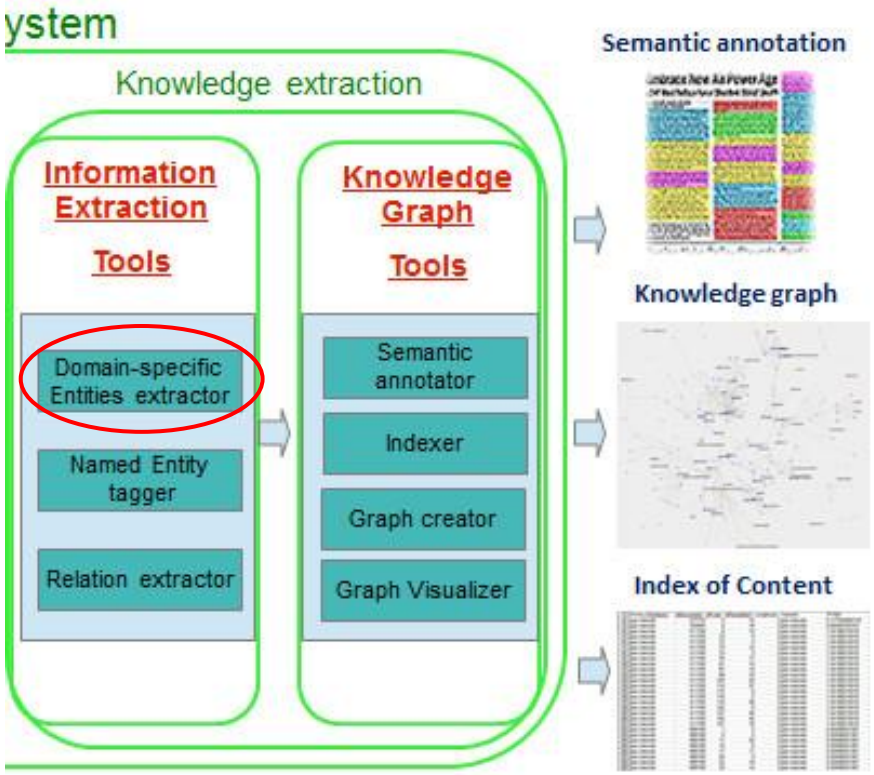


From text to knowledge @ ItaliaNLP Lab

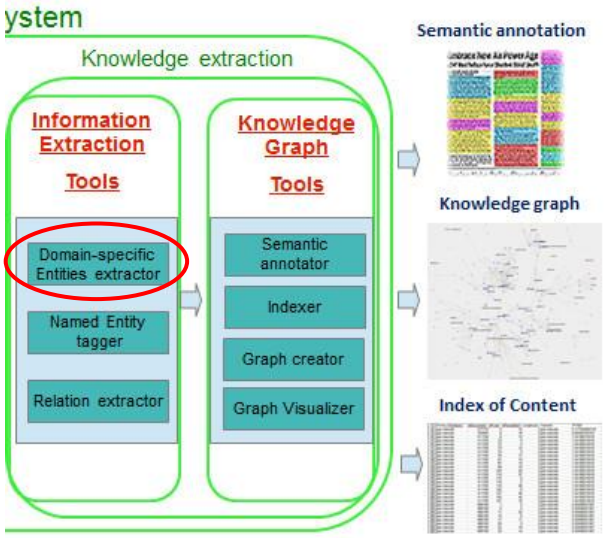


- The IE tools allow extracting
  - domain-specific entities (Bonin et al. 2010)
    - e.g. nominal terminology, verbs (both single- and multi-word expressions)
  - Named entities
    - i.e. Person, Location, Organization and Geopolitical
  - relations between the extracted entities
    - taxonomical
      - e.g. *health* ***research***, *international* ***research***, *cancer* ***research*** or ***research*** *projects,* ***research*** *infrastructure*
    - co-occurence within the same context and similarity on the basis of shared contexts
- They result in
  - multi-dimensional knowledge representation graph
  - document collection index and semantic annotation

Terminology Extraction



TERMINOLOGY EXTRACTION



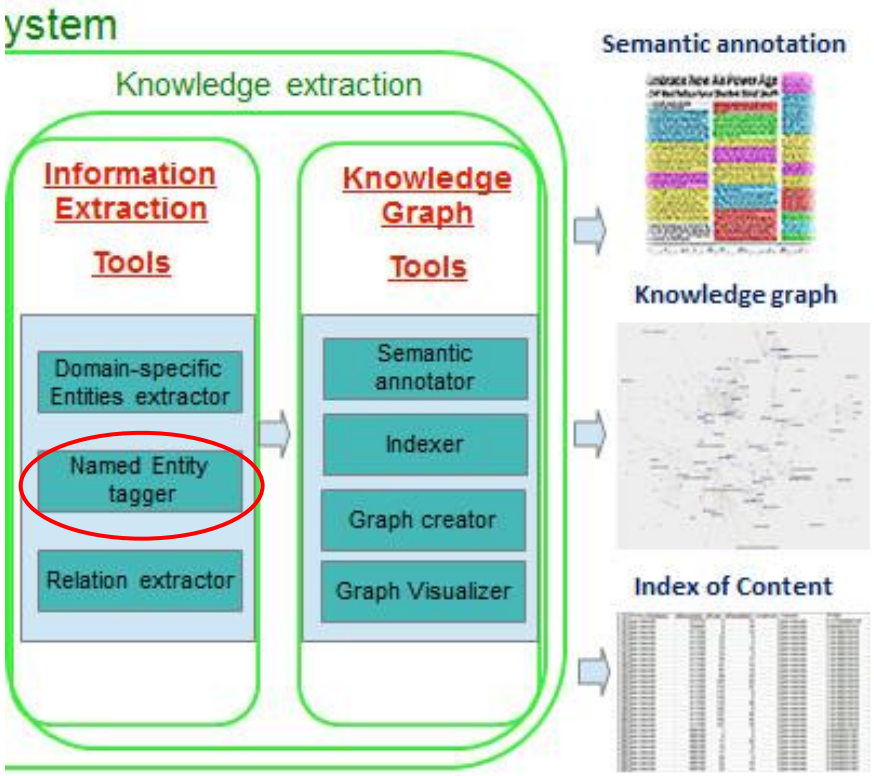
Input corpus: a collection of European Italian Directives on consumer protection

Lemma of Term	Domain Relevance	Frequency
trade service	100.00%	421
Sub-contracted operation	100.00%	398
custom authority	100.00%	326
wholesale trade service	100.00%	154
personal data	100.00%	151
contract basis	100.00%	120
third country	100.00%	117
convention	100.00%	584
national law	100.00%	112
tariff information	100.00%	102
custom debt	100.00%	100
travel document	100.00%	86
import duty	100.00%	82
free circulation	100.00%	82

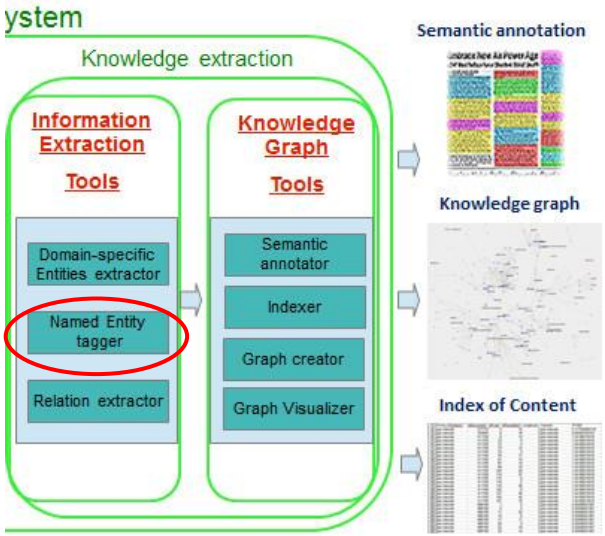
Discriminate between legal and regulated domain knowledge

- T2K handles this challenge thanks to
  - a **multilayred contrastive approach** to entity extraction
    - The domain relevance of entities is assessed on the basis of the contrastive distribution of relevant candidate entities across an input corpus and a different corpus
    - The contrastive analysis is iterated twice:
      - against a top list of open-domain entities (e.g. from newspapers) to prune common entities (e.g. *following day*)
      - against a top list of entities from e.g. a different regulated domain to discriminate legal and regulated-domain entities
  - a **new term ranking function** suitable for handling variation in low frequency events
    - E.g. in the legal texts, regulated-domain entities have low frequency and they are sparse

Named Entity Extraction



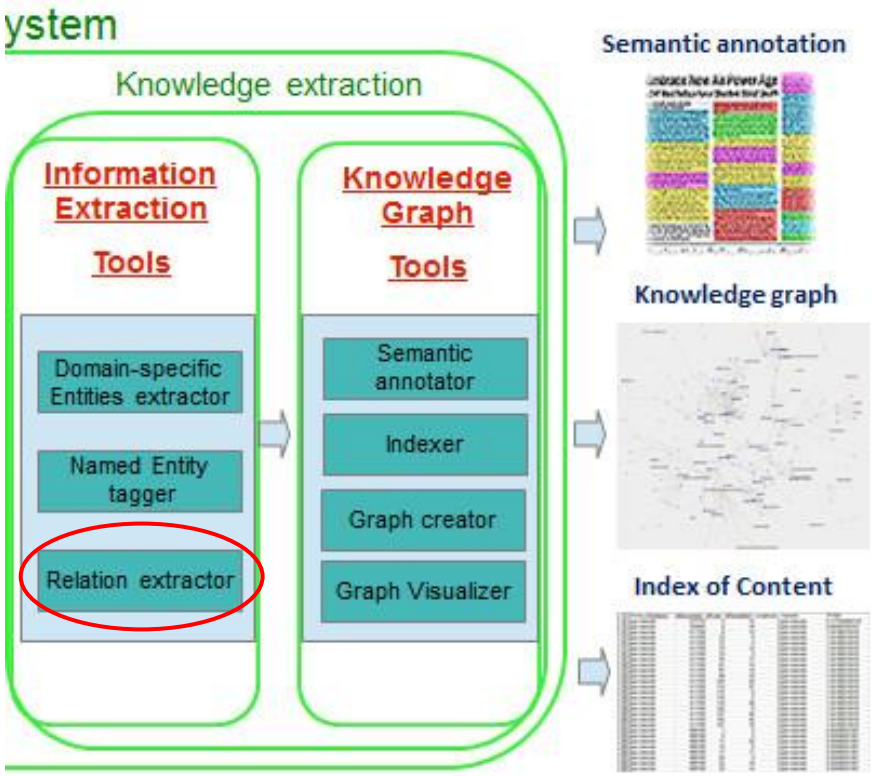
NAMED ENTITY EXTRACTION



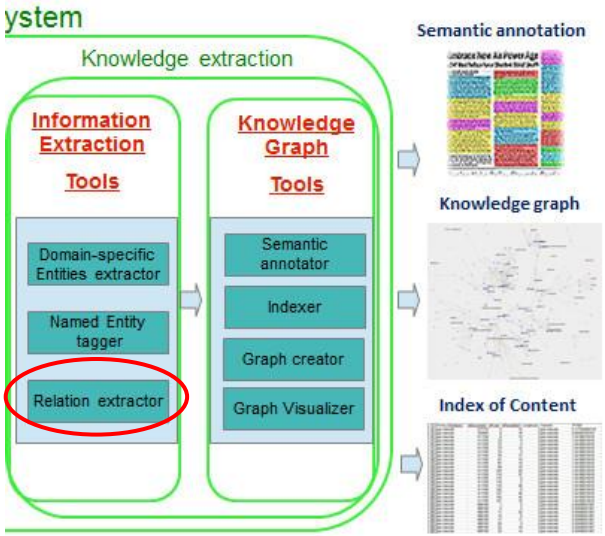
Input corpus: a collection of European Italian Directives on consumer protection

Entity	Class	Frequency	Frequency (%)
Commission	Organization	980.0	5.55
EC	Organization	362.0	2.05
EEC	Organization	341.0	1.93
European Union	Organization	311.0	1.76
European Parliament	Organization	184.0	1.04
EU	Organization	142.0	0.80
Brussels	Location	118.0	0.67
European Community	Organization	110.0	0.62
Council of Europe	Organization	77.0	0.44
Europe	Location	72.0	0.41
Rome	Location	66.0	0.37
Consultative Committee	Organization	65.0	0.37
Council	Organization	62.0	0.35
Euratom	Organization	62.0	0.35
Schengen	Location	59.0	0.33
Romania	Location	58.0	0.33
Luxembourg	Location	50.0	0.28
Association Council	Organization	45.0	0.25
Management Board	Organization	45.0	0.25
Bulgaria	Location	45.0	0.25
Schengen Information System	Organization	43.0	0.24
Stabilisation and Association Council	Organization	42.0	0.24
EUROPEAN UNION	Organization	40.0	0.23
European Council	Organization	37.0	0.21
Ireland	Location	34.0	0.19
France	Location	33.0	0.19
Northern Ireland	Location	30.0	0.17
Republic of Cyprus	Location	30.0	0.17
Denmark	Location	28.0	0.16
Federal Republic of Germany	Organization	27.0	0.15
Austria	Location	27.0	0.15

Relation Extraction



Relation Extraction



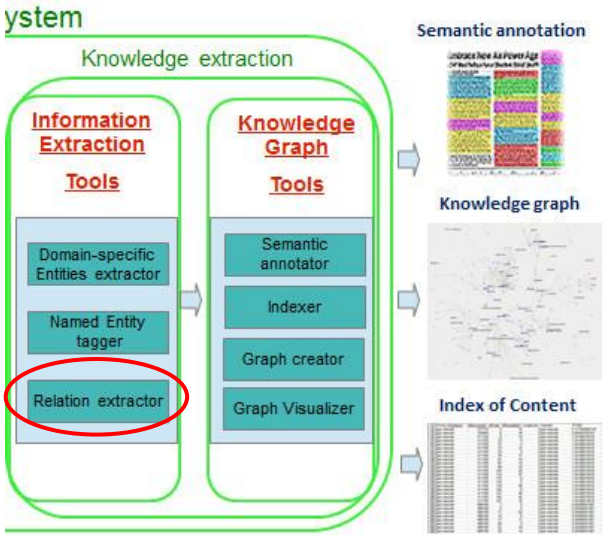
E.g.: terms in relation with *imaging cerebrale* in criminal case law texts

imaging cerebrale (*brain imaging*)	
genetica molecolare (*molecular genetics*)	quadro clinico (*medical case*)
difesa (*defense*)	comportamenti illeciti (*illegal behaviours*)
valutazione (*evaluation*)	nesso causale (*causal relationship*)
colloqui clinici (*clinical interviews*)	apporto tecnico (*technical contribution*)
emergenze psichiatriche (*psychiatric emergencies*)	sfera psichica (*psychic sphere*)
accertamenti psichiatrici (*psychiatric inspections*)	imputata (*defendant*)

Input corpus: a collection of Italian case laws concerning the use of neuroscience in the Italian courtrooms

Relation Extraction

E.g.: terms in relation with *imaging cerebrale* in criminal case law texts



imaging cerebrale (*brain imaging*)	
genetica molecolare (*molecular genetics*)	quadro clinico (*medical case*)
difesa (*defense*)	comportamenti illeciti (*illegal behaviours*)
valutazione (*evaluation*)	nesso causale (*causal relationship*)
colloqui clinici (*clinical interviews*)	apporto tecnico (*technical contribution*)
emergenze psichiatriche (*psychiatric emergencies*)	sfera psichica (*psychic sphere*)
accertamenti psichiatrici (*psychiatric inspections*)	imputata (*defendant*)

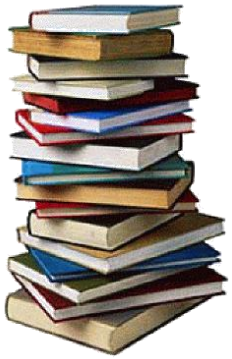
e.g. context

Input corpus: a collection of Italian case laws concerning the use of neuroscience in the Italian courtrooms

Sia le emergenze psichiatriche , completate dalle risultanze dell' imaging cerebrale e di genetica molecolare , che quelle processuali consentono di rilevare gravi segni di disfunzionalità psichica , eterogenei ma convergenti nell' indicare un nesso causale tra i disturbi dell' imputata ed i suoi comportamenti illeciti

From text to knowledge: the general approach

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



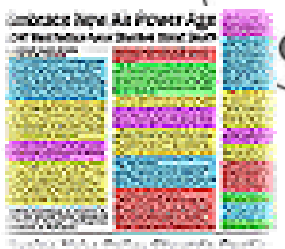
Textual content
(implicit knowledge)

Dynamic content structuring

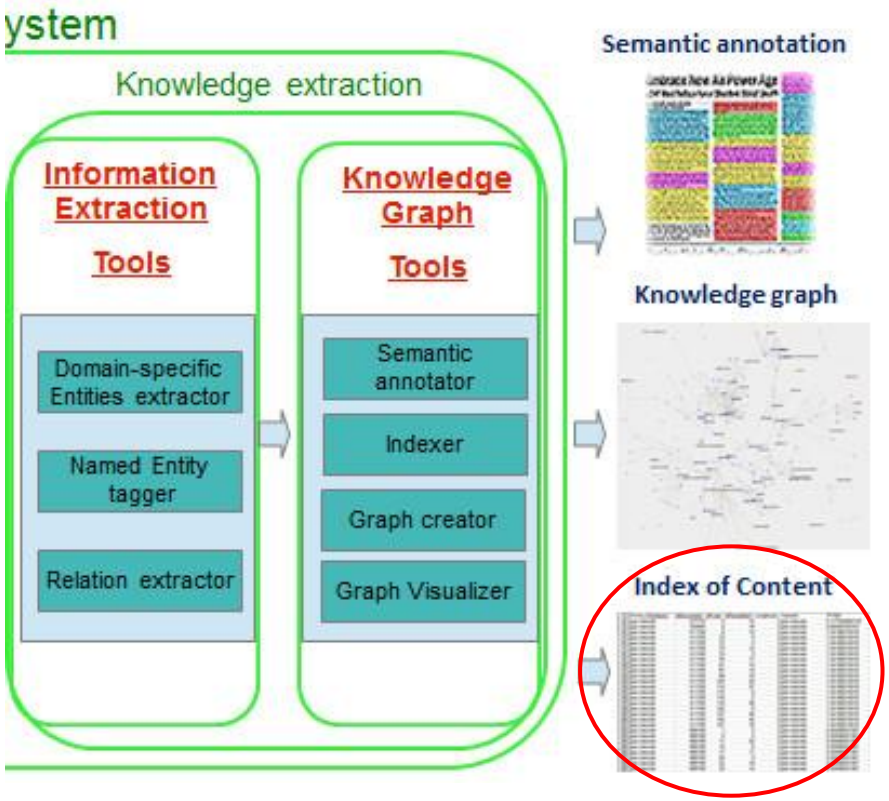
Linguistic annotation

Knowledge extraction

Structured knowledge
(explicit knowledge)



Document indexing



Document indexing

The acquired knowledge (e.g. terms, named entities) is used for document indexing on the basis of the extracted domain-specific knowledge

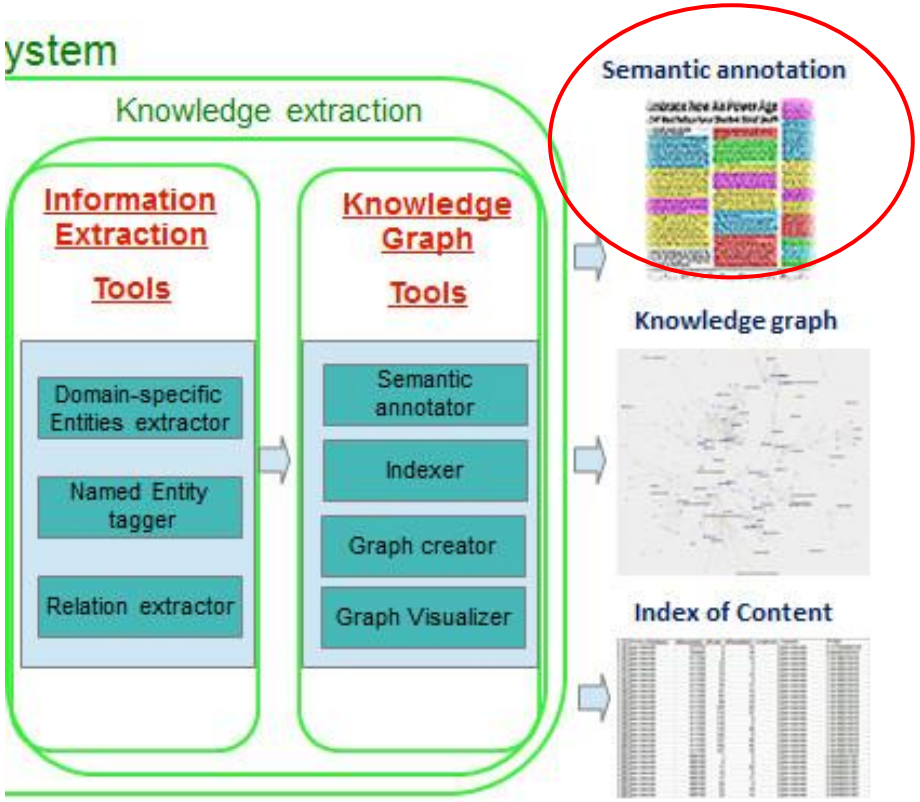
e.g.

Term	Document	TF*IDF
risonanza magnetica	Penale/merito/merito/massime/Cort_Assis_App.Trieste, pen. , mass. , 01-10-2009.txt	0.141522742511
risonanza magnetica	Penale/Cassazione/Cassazione penale/2012/Cass_pen_Sez_I, Sent. , (ud_25-10-2012) 21-11-2012, n_45559.txt	0.0637848980331
risonanza magnetica	Penale/merito/merito/2007/Cort.Assis.Treviso, pen, Sent. , 12-11-2007.txt	0.0611990237885
risonanza magnetica	Penale/Cassazione/Cassazione penale/2012/Cass_pen_Sez_I, Sent. , (ud_16-12-2011) 03-05-2012, n_16281.txt	0.0576907994949
risonanza magnetica	Penale/Cassazione/Cassazione penale/2010/Cass_pen_Sez_IV, (ud_20-11-2009) 14-01-2010, n_1489.txt	0.0485220831466
risonanza magnetica	Penale/Cassazione/Cassazione penale/2006/Cass_pen_Sez_III, (ud_21-06-2006) 10-10-2006, n_33974.txt	0.0358002194494
risonanza magnetica	Penale/Cassazione/Cassazione penale/2007/Cass_pen_Sez_I, Sent. , (ud_13-12-2006) 02-03-2007, n_9173.txt	0.0334223450948
risonanza magnetica	Penale/Cassazione/Cassazione penale/2010/Cass_pen_Sez_III, Sent. , (ud_18-03-2000) 11-05-2010, n_17955.tx	0.0314494983357
risonanza magnetica	Penale/merito/merito/2006/Trib_ord. ,Genova, pen. , Sez_III, 21-07-2006.txt	0.0301915184023
risonanza magnetica	Penale/Cassazione/Cassazione penale/2006/Cass_pen_Sez_V, Sent. , (ud_18-05-2006) 14-11-2006, n_37452.tx	0.0190282679006
risonanza magnetica	Penale/Cassazione/Cassazione penale/2013/Cass_pen_Sez_III, Sent. , (ud_04-12-2012) 22-01-2013, n_3258.txt	0.0174854353681
risonanza magnetica	Penale/Cassazione/Cassazione penale/2012/Cass_pen_Sez_I, Sent. , (ud_03-07-2012) 20-07-2012, n_29707.txt	0.0159462245083
risonanza magnetica	Penale/merito/merito/2013/Trib.ord. , Venezia, pen. ,Sent. ,08-04-2013_utf8.txt	0.015652284886
risonanza magnetica	Penale/Cassazione/Cassazione penale/2012/Cass_pen_Sez_IV, Sent. , (ud_01-12-2011) 31-01-2012, n_3986.txt	0.0147036615596
risonanza magnetica	Penale/merito/merito/2009/Cort_Assise di App. , Trieste, pen. , Sen. , 18-09-2009, n_5.txt	0.0117021389156
risonanza magnetica	Penale/Cassazione/Cassazione penale/2007/Cass_pen_Sez_II, Sent. , (ud_17-04-2007) 02-05-2007, n_16632.txt	0.00817459884539

Input corpus: a collection of Italian case laws concerning the use of neuroscience in the Italian courtrooms

The term *risonanza magnetica* (magnetic resonance) occurs both in lower courts and in the Court of Cassation but it is 'more relevant' in the criminal case resolved by the Trieste ordinary tribunal in 2009

Semantic annotation



Semantic annotation

The acquired knowledge (e.g. terms, named entities) is projected back onto the corpus

e.g.

Input corpus: a collection of Italian case laws on state liability

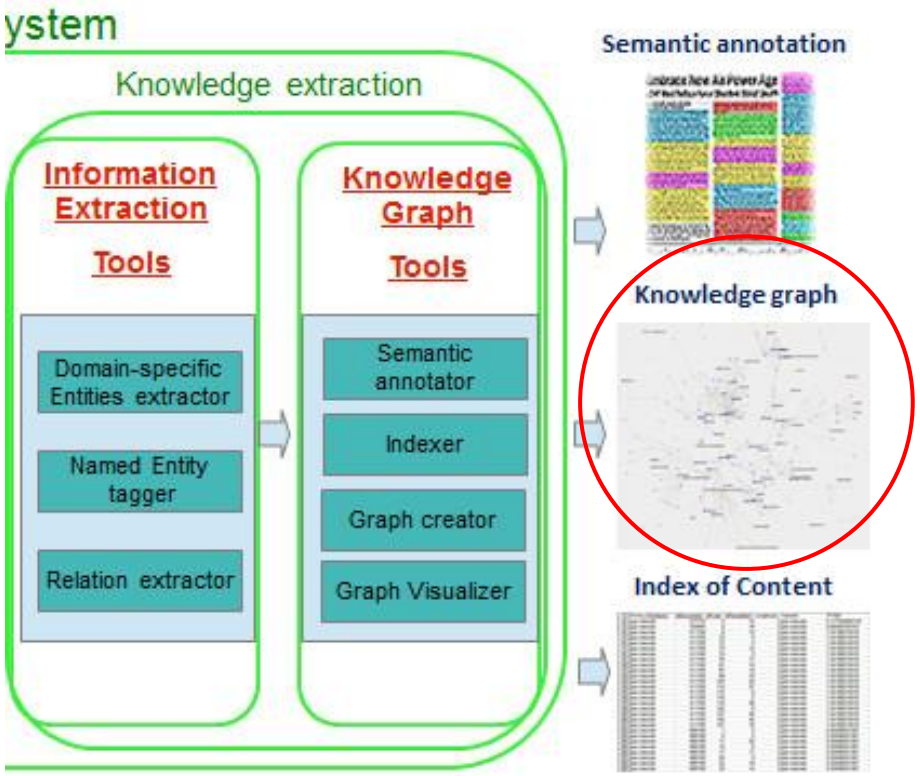
La sentenza ritiene azionato, pur in assenza di espressa qualificazione in tal senso nell'atto introduttivo del giudizio, il **diritto al risarcimento del danno**, ex art. 2043 c.c., per **violazione dell'obbligo** dello Stato di dare attuazione alle **direttive comunitarie** che imponevano di remunerare adeguatamente il medico per la frequenza di un **corso di specializzazione**; considera comprovato, in assenza di contestazioni specifiche, che il C. avesse superato il corso di formazione quadriennale, come da attestazione del 5.11.1992, con frequenza a tempo pieno e senza svolgimento di attività libero-professionale; dichiara inammissibile l'eccezione di **prescrizione quinquennale** sollevata dall'amministrazione ed accolta dal primo giudice, sul rilievo che era stata formulata, senza le necessarie allegazioni in fatto e diritto, con riferimento all'art. 2948 c.c., n. 4, in termini, quindi, non pertinenti al rapporto giuridico dedotto in giudizio, atteso che non si trattava di rapporto di impiego pubblico (prospettazione su cui si fondava il **difetto di giurisdizione** ordinaria, eccepito dall'amministrazione in primo grado) e di **responsabilità contrattuale**; liquida il risarcimento nell'importo di L. 13.000.000 annue (Euro 6.713,93) secondo il parametro fornito dalla L. n. 370 del 1999, art. 1, comma 1 (**borsa di studio** annuale per i medici ammessi presso le università alle **scuole di specializzazione** in medicina dall'anno accademico 1983-1984 all'anno accademico 1990-1991, in attuazione di giudicati amministrativi), con l'aggiunta della rivalutazione monetaria e degli interessi legali dalla maturazione del credito, fissata alla data del 5 novembre 1992.

Semantic annotation

The semantically annotated corpus can be used by a search engine to retrieve the text spans containing the information searched for

e.g.

Knowledge graph

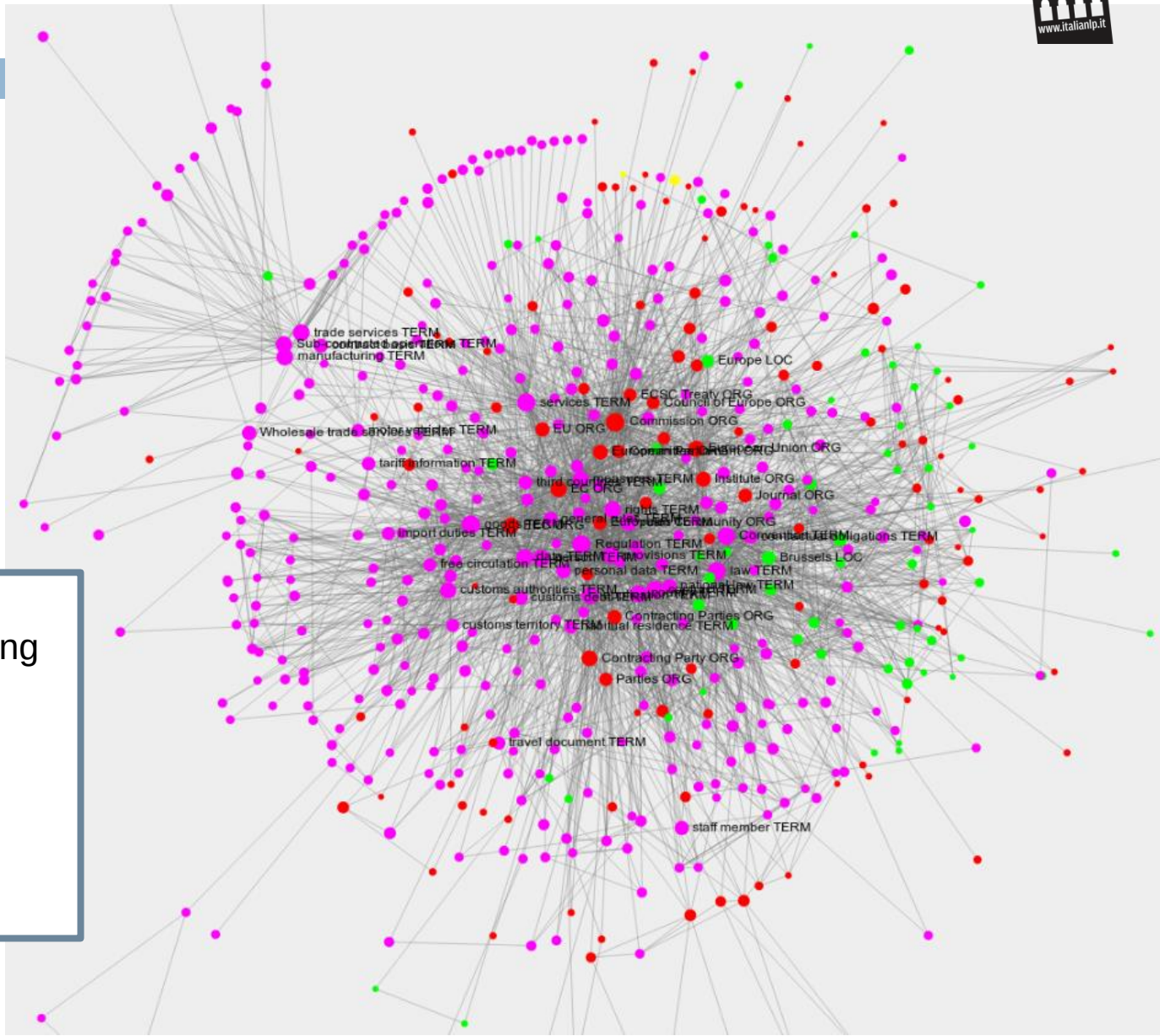


Knowledge graph

Input corpus: a collection of European Italian Directives on consumer protection

e.g.

In T2K the extracted information interact resulting in a multidimensional knowledge representation graph creating the prerequisites for sophisticated text mining processes

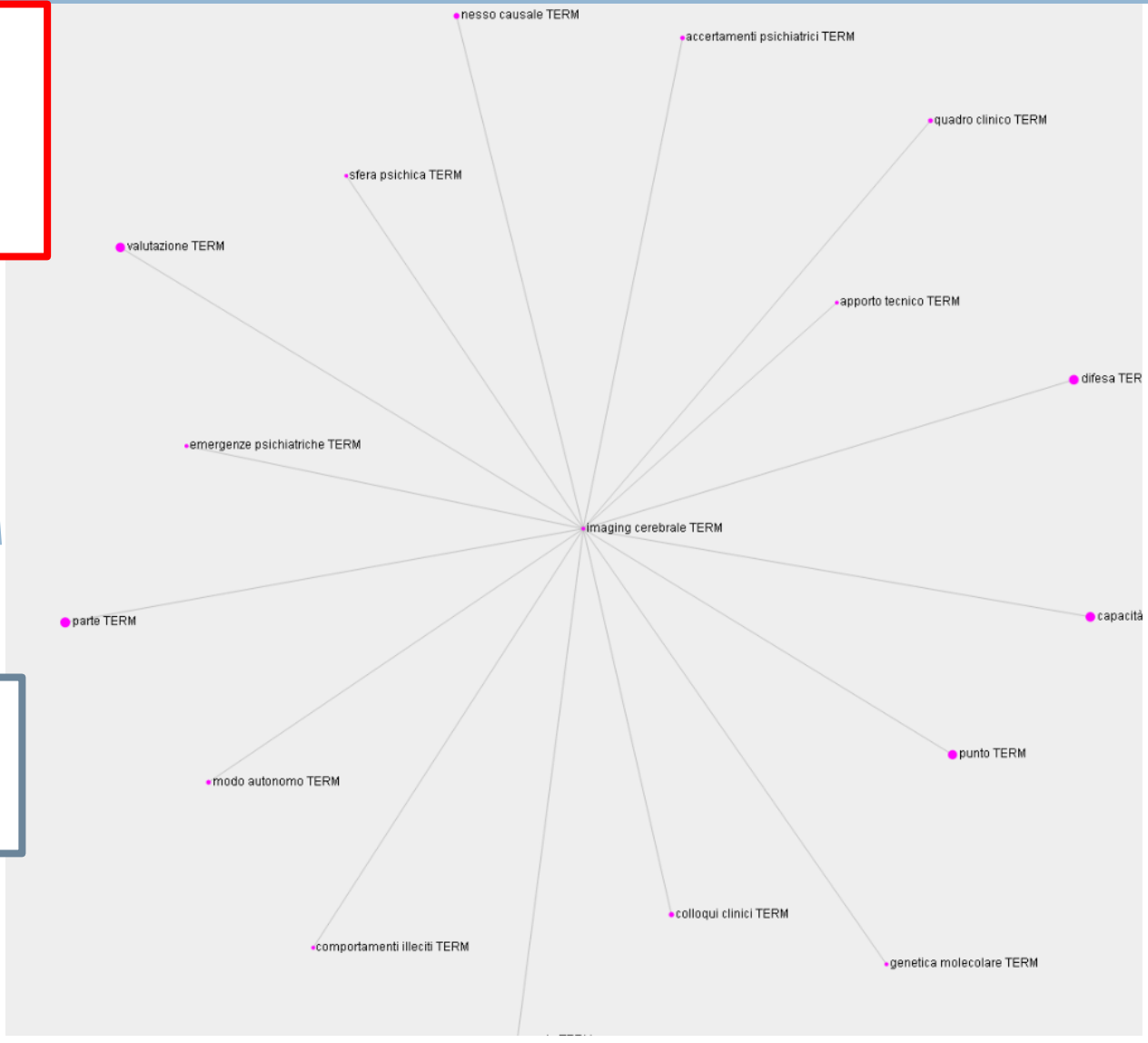


Knowledge graph

Input corpus: a collection of Italian case laws concerning the use of neuroscience in the Italian courtrooms

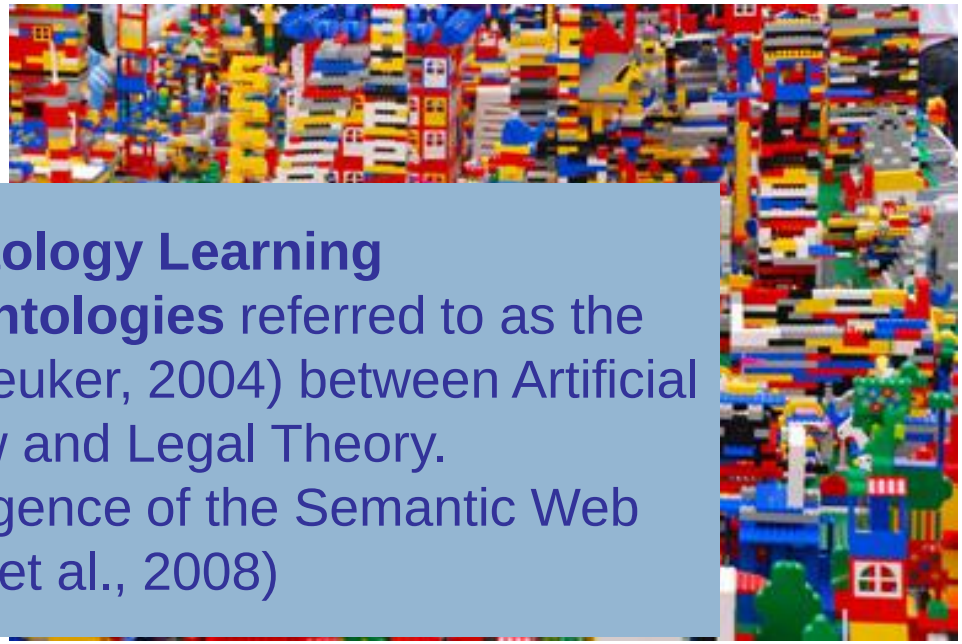
e.g.

The sub-graph of *imaging cerebrale* in criminal case law texts



To sum up: from bricks of knowledge to a domain ontology

To sum up: from bricks of knowledge to a domain ontology



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)

Approaches to Ontology Design and Development: top-down vs bottom-up

TOP-DOWN

ontology construction starts by modeling top level concepts, which are then subsequently refined

this approach is typically carried out manually by domain experts and leads to a high-quality engineered ontology

Approaches to Ontology Design and Development: top-down vs bottom-up

TOP-DOWN

ontology construction starts by modeling top level concepts, which are then subsequently refined

this approach is typically carried out manually by domain experts and leads to a high-quality engineered ontology

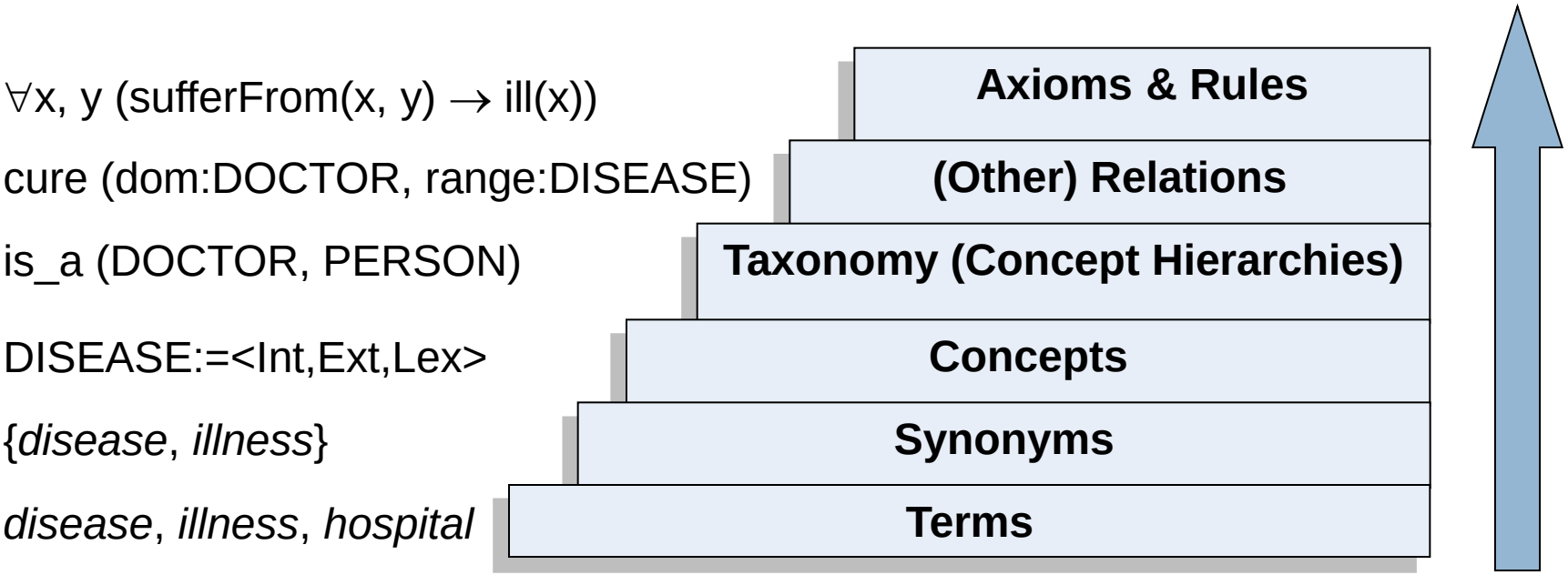
it starts from the assumption that most concepts and conceptual structures of the domain are contained in documents

the terminological and conceptual knowledge contained in document collections is semi-automatically extracted from texts, thus creating the basis for ontology construction

BOTTOM-UP

Ontology Learning: an incremental process

- 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: an example

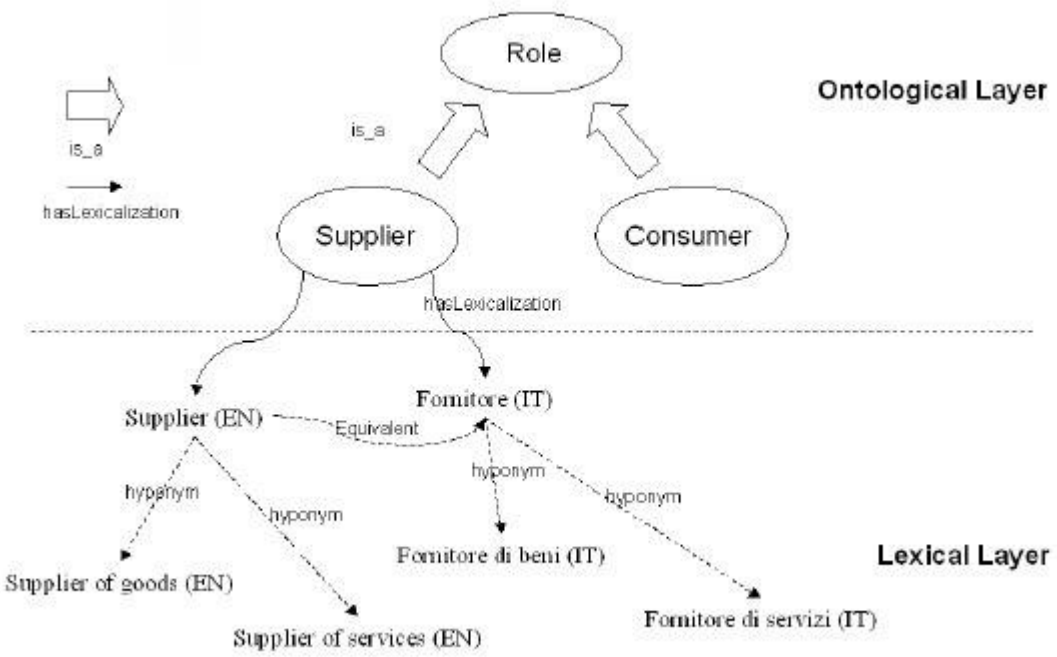
- The DALOS (*Drafting Legislation with Ontology–based Support*) European project (Agnoloni et al., 2009)
  - Aimed at
    - providing law-makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting
    - enhancing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

- 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



Conclusion

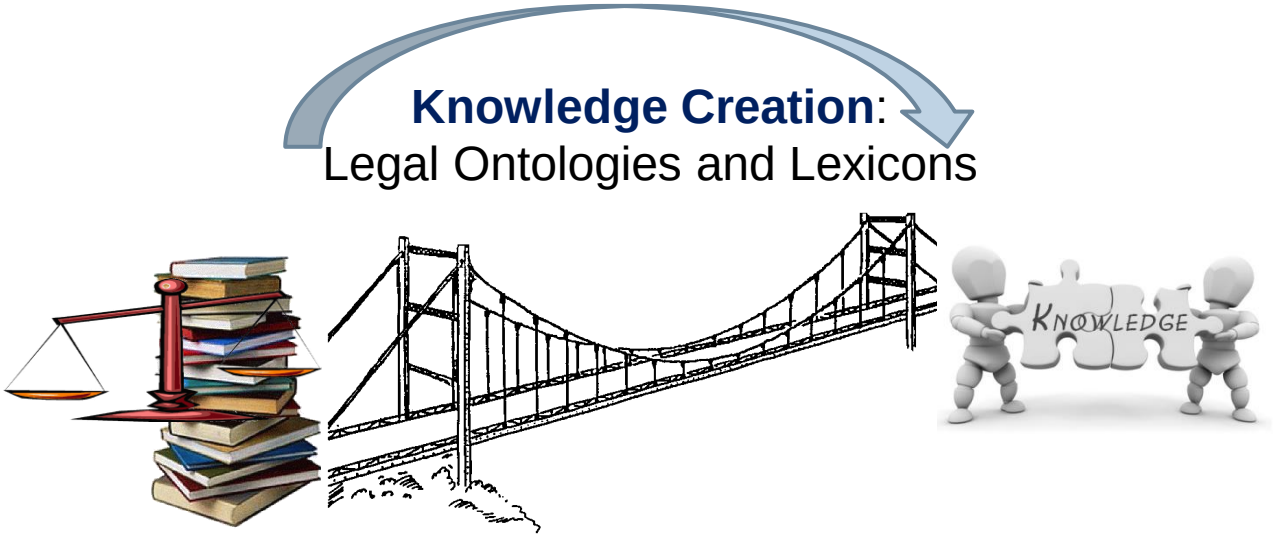
One of the main obstacles to progress in the field of artificial intelligence and law is the ***natural language barrier***

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

Conclusion

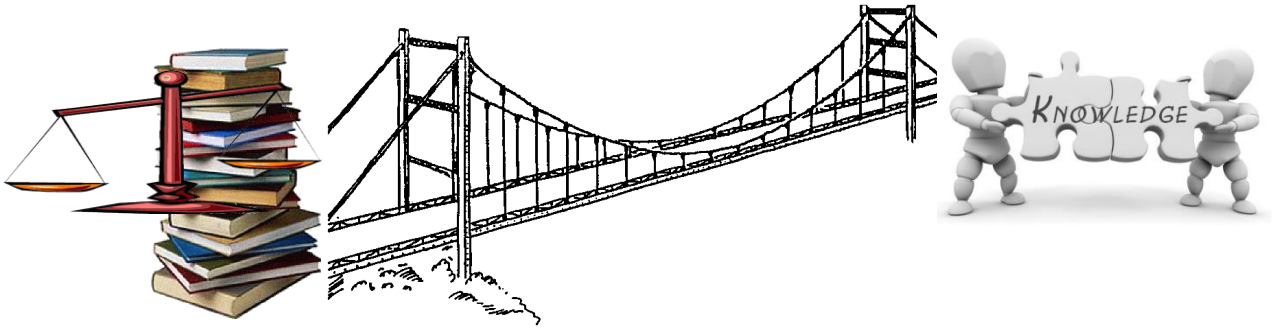
- Natural Language Processing techniques represent a key ingredient for Legal Knowledge Extraction and Management



Conclusion

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

Credits

- The NLP tools and techniques have been developed in the framework of the activities of the *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.italianlp.it/demo/linguistic-annotation-tool/
- Text-To-Knowledge (T2K)
  - http://www.italianlp.it/demo/t2k-text-to-knowledge/

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
  - Dell'Orletta F., Venturi G., Cimino A., Montemagni S. (2014) "T2K²: a System for Automatically Extracting and Organizing Knowledge from Texts". In Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014), 26-31 May, Reykjavik, Iceland.

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.