Ontology Learning in the legal domain. An introduction

Giulia Venturi
giulia.venturi@ilc.cnr.it
Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC- CNR)
Scuola Superiore Sant’Anna, Pisa

Simonetta Montemagni
simonetta.montemagni@ilc.cnr.it
Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC- CNR)

Summer School LEX2012 – 13th September 2012

Summary

- Part 1
  - Ontology Learnig: basics
    - Pros and cons of the current approaches to Ontology construction
    - Ontology Learning steps
    - Evaluation
- Part 2
  - Ontology Learning in the legal domain
    - Why
    - Main challenges
    - Case studies carried out in the legal domain
    - Open issues

Part 1

Ontology Learning: basics

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

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

Top-down vs bottom-up approaches to Ontology construction: pros and cons

- Top-down approach
  - **Pros**
    - Top-down ontologies may be reused across different application scenarios, and can serve as a starting point for developing new ontologies
  - **Cons**
    - they necessarily require an expert-based approach
    - their development is costly in terms of both time and effort
    - their coverage is typically rather restricted, and this is a disadvantage when they are used in the framework of real knowledge management applications
    - linking of textual information to the ontology, which requires linguistic knowledge about the terminology used to convey domain-specific concepts
    - the highly dynamic and constantly evolving nature of ontologies in different domains, including the legal one, requires continuous updating and refinement

Top-down vs bottom-up approaches to Ontology construction: pros and cons

- Bottom-up approach
  - **Pros**
    - ontological knowledge discovered at a larger scale and a faster pace
    - it can support
      - the semi-automatic construction of ontologies
      - the refining and expanding of existing ontologies by incorporating new knowledge emerging from texts
      - the detection and revision of human-introduced biases and inconsistencies
    - it creates the prerequisites for the alignment between the ontology and texts
      - with ontologies boostrapped from texts the linking with textual information is made easier
  - **Cons**
    - a bottom-up approach results in a very high level of detail which makes it difficult to spot commonality between related concepts and increases the risk of inconsistencies

The "middle-out" approach to Ontology construction

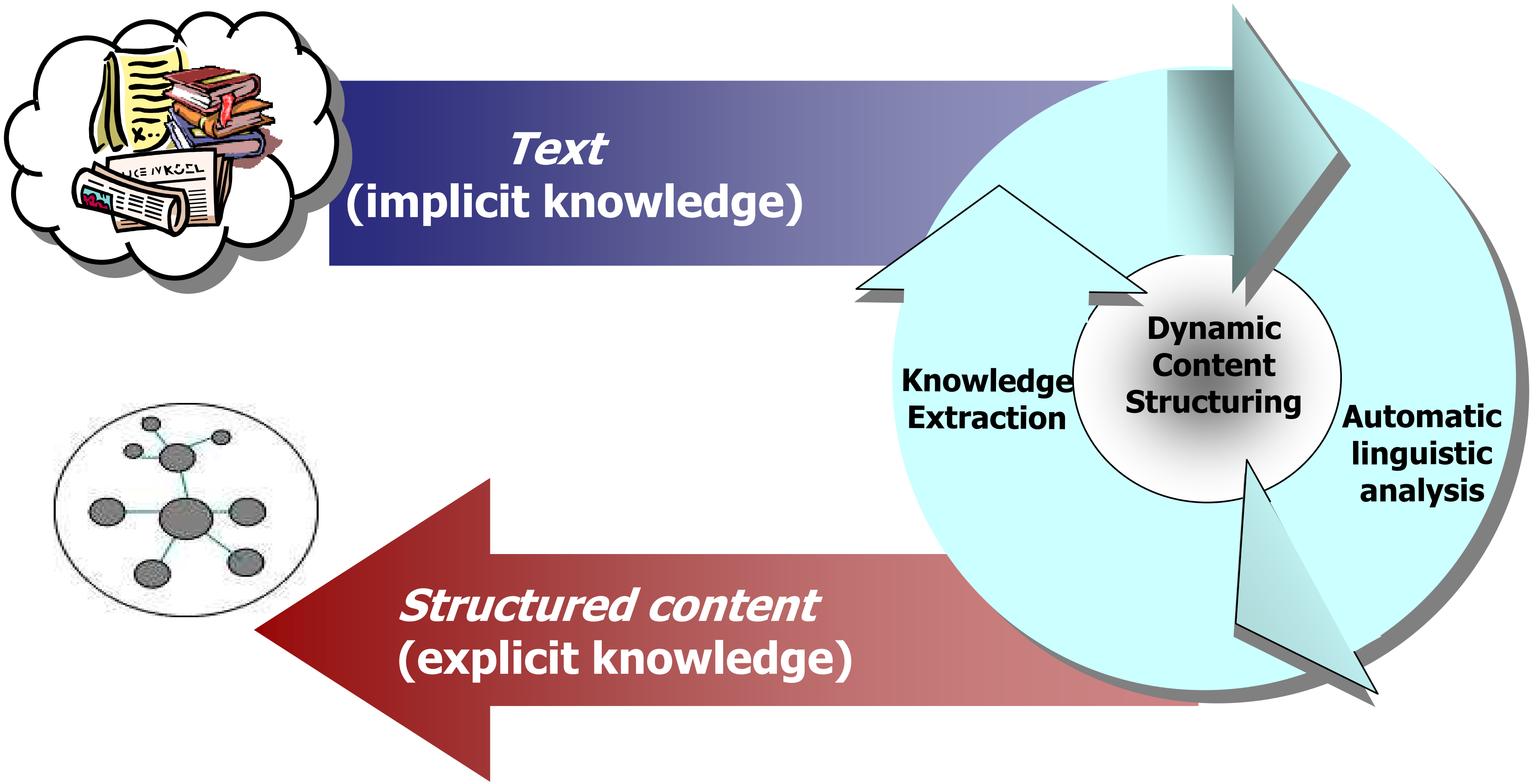
- So far:
  - complementarity of the top-down and bottom-up approaches
- A widely acknowledged need:
  - any comprehensive domain ontology needs to be built both top-down and bottom-up
    - only by proceeding in this way, the resulting ontology reflects domain knowledge and is at the same time anchored to texts
- A possible solution:
  - the **"middle-out" approach** based on the combination of top-down and bottom-up ontology modelling
- How?
  - through the support of **automatic document analysis**
    - relevant lexical entries and relations holding between them are extracted semi-automatically from available documents
  - the (semi-)automatic support in ontology development is nowadays referred to as **Ontology Learning**

Ontology Learning as a support to ontology construction

- Ontology Learning represents a promising line of research
  - Knowledge acquisition from texts is the basis for the construction and/or extension of ontologies
  - The learning process is typically carried out by combining **Natural Language Processing** (NLP) technologies with **Machine Learning** (ML) techniques

- Pros
  - It reduces ontology development time and costs
  - It guarantees that extracted concepts are textually grounded
- Cons
  - It entails the typical acquisition bottleneck:
    - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure

Ontology Learning: a dynamic process



Ontology Learning: an incremental process

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

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

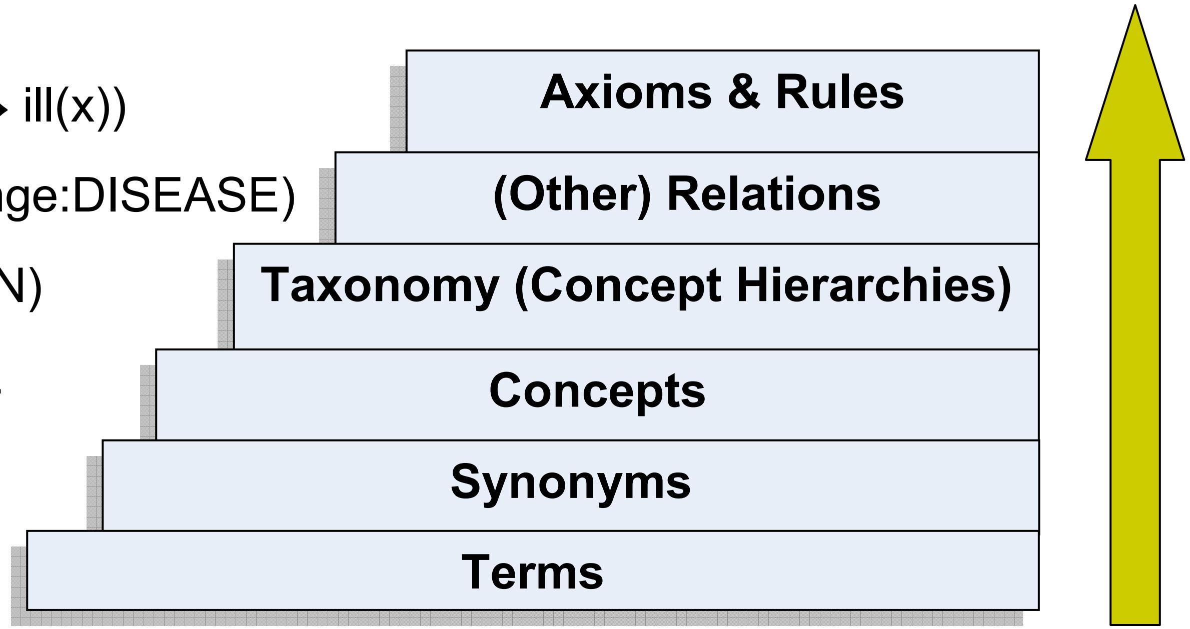
cure (dom:DOCTOR, range:DISEASE)

is_a (DOCTOR, PERSON)

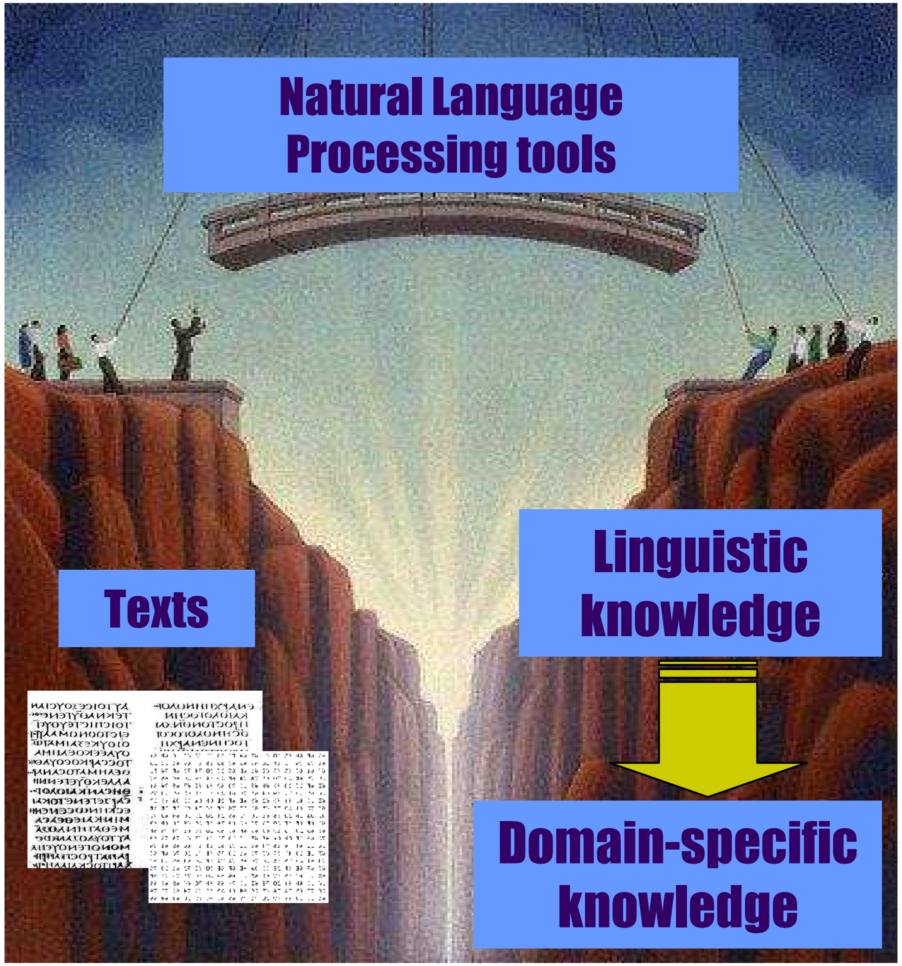
DISEASE:=<Int,Ext,Lex>

{*disease*, *illness*}

disease, *illness*, *hospital*

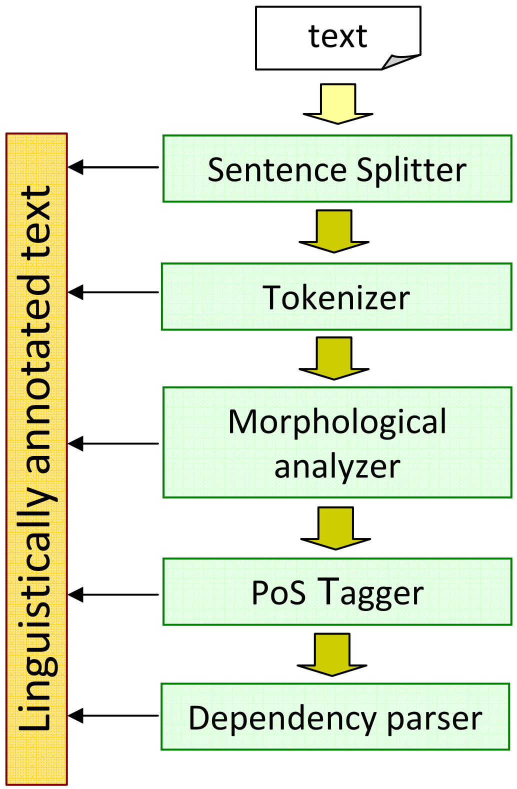


Automatic linguistic annotation: the first step of Ontology Learning



- It is carried out by tools able to access the content of texts by processing the Natural Language in which they are written
- From text to knowledge
  - linguistic
    - morpho-syntactic, syntactic, lexical-semantic
  - domain-specific
    - e.g. legal domain knowledge

Linguistic annotation as an incremental process



Splits the text into sentences

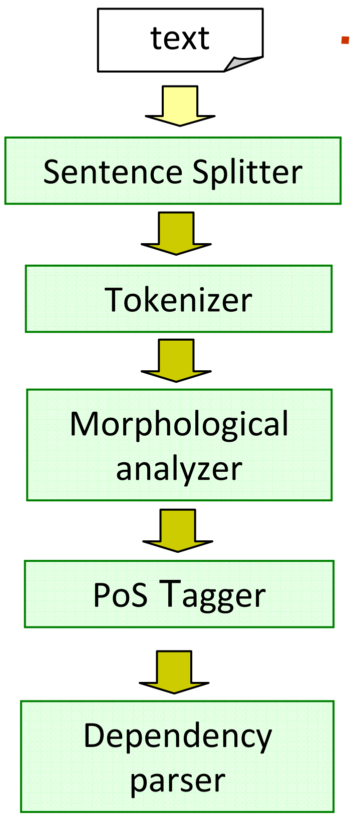
Segments each sentence into orthographic units (tokens)

Assigns the possible morphological analyses to each token

Selects the appropriate morphological interpretation in the specific context

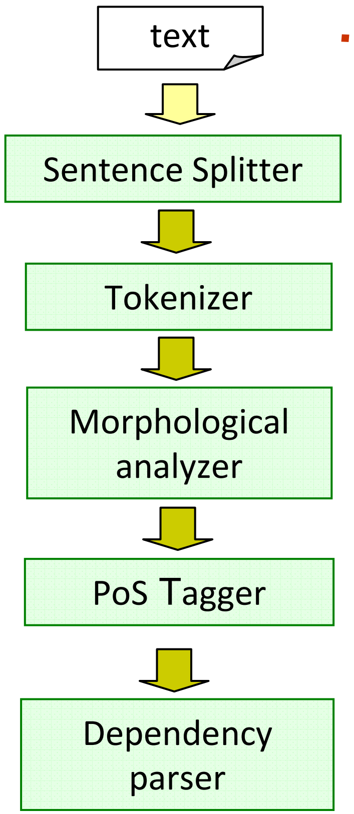
Identifies dependency relations between tokens (e.g. subject, object, etc.)

Linguistic annotation as an incremental process: 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.)

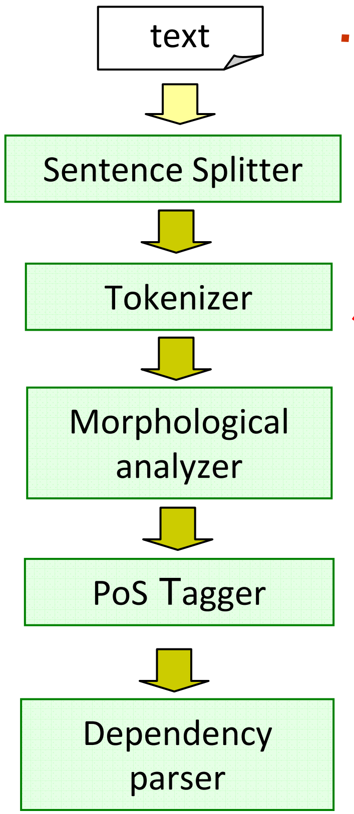
Linguistic annotation as an incremental process: an example



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 as an incremental process: an example



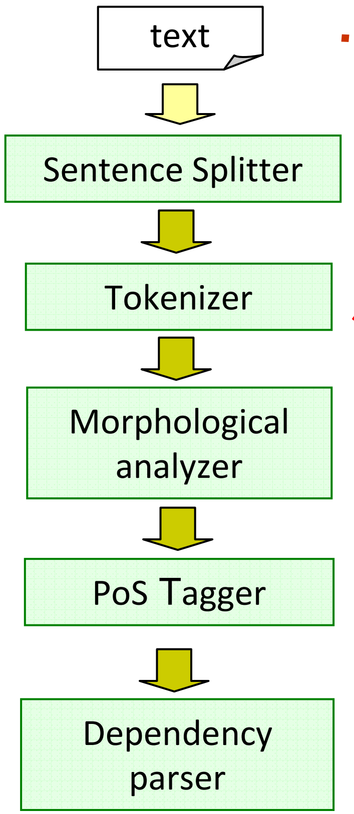
Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

- 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 as an incremental process: an example



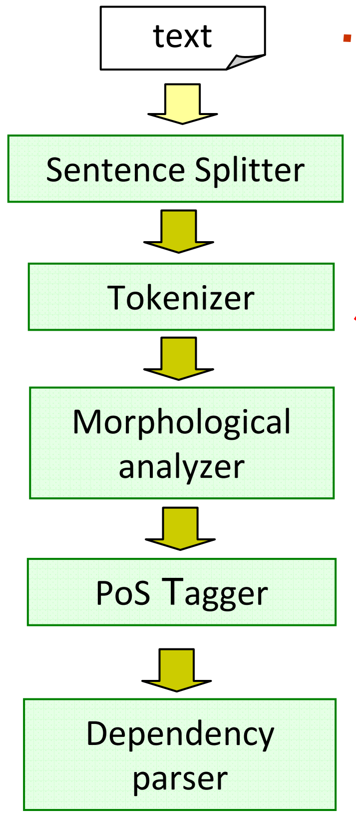
Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

- 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 as an incremental process: an example



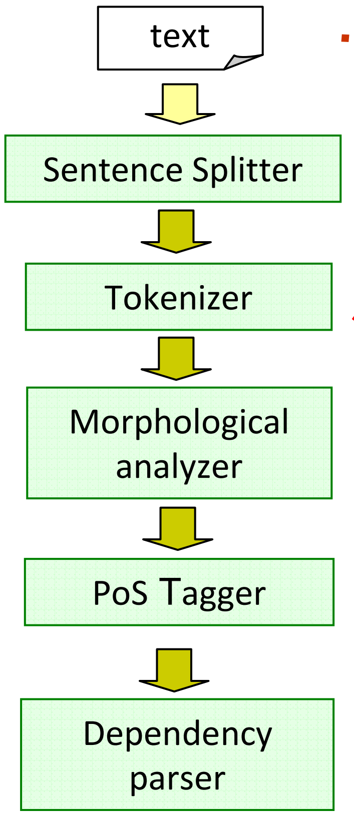
Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l’avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

- 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 as an incremental process: an example



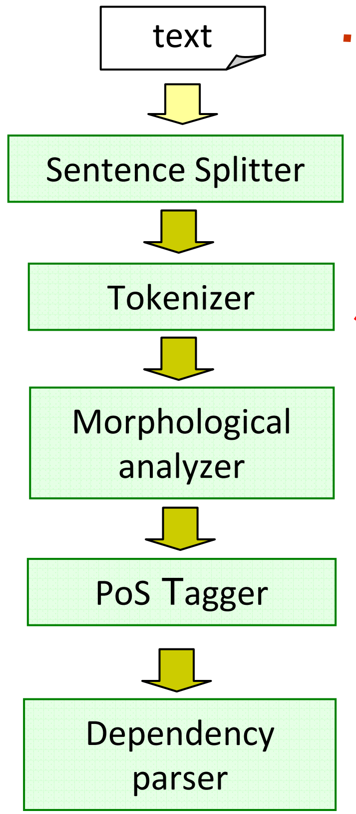
Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

- 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	DEP
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 as an incremental process: an example

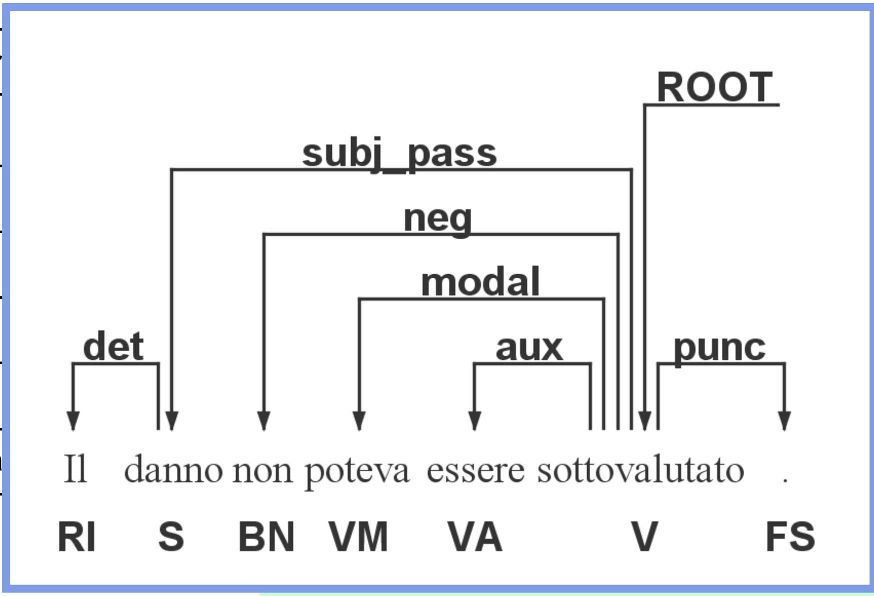


Il danno non poteva essere sottovalutato. Il sig. Rossi decise perciò di chiamare l'avvocato. (The damage could not be understimated. Mr. Rossi decided therefore to call the lawyer.)

- 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



SS

schema

Linguistic annotation "layer cake"

MODIF(*decreto,presente*) SUBJ(*stabilire,decreto*) OBJD(*stabilire,norma*)	**Dependency analysis**
decreto DECRETO#S@MS#	**POS-tagging**
DECRETARE#V@S1IP# DECRETO#S@MS#	**Morphological analysis**
Il | presente | decreto | stabilisce | […] | dell' | inquinamento | da | rumore | .	**Tokenization**
Il presente decreto stabilisce le norme per la prevenzione dell'inquinamento da rumore. (This decree sets the rules for the prevention of noise pollution.)	**Sentence Segmentation**

Ontology Learning systems differentially exploit different levels of linguistic annotation of texts in an incremental fashion

Evaluating Ontology Learning results (1)

- Evaluation approaches
  - comparing the ontology to a “**gold standard**” in terms of
    - Precision: percentage of correctly acquired items with respect to all acquired items
    - Recall: percentage of correctly acquired items with respect to all items in the gold standard
    - Pros: measure of the reliability and coverage of the ontology wrt a resource manually built by a domain-expert
    - Cons: time consuming since the gold standard needs to be built manually

Evaluating Ontology Learning results (2)

- using the ontology in an **application** and evaluating the results (task-based evaluation)
  - Broad range of applications, e.g. Information Retrieval, to support advanced functionalities for semantic annotations of document collections
  - Pros: carried automatically
  - Cons: no information about the reliability of the acquired data
- evaluation is **done by humans** who try to assess how well the ontology meets a set of predefined criteria, requirements, etc.
  - Pros: meets specific needs
  - Cons: time consuming manual evaluation

Part 2

Ontology Learning in the Legal Domain

Ontology learning in the legal domain: the need

- The current situation
  - A number of legal ontologies have been proposed
    - mostly focusing on **upper level concepts**
    - **hand-crafted** by domain experts
    - typically built following a **top-down approach**
- The need
  - Large knowledge-based applications in the legal domain need broad-coverage ontologies
    - whose content is reusable and continuously updated knowledge
    - where concepts are linked with their corresponding linguistic (textual) realization
- A possible solution
  - The “middle-out” approach to ontology construction
    - integrating top-down and buttom-up approaches for automated ontology-learning from texts
  - Ontology Learning
    - combining NLP technologies with Machine Learning techniques

Ontology learning in the legal domain: so far …

- Relatively few attempts made so far to semi-automatically induce legal domain ontologies from texts
- Among them:
  - 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
  - exploitation of a middle-out approach
    - LKIF Core ontology (Hoekstra et al., 2007)
    - LOIS (Peters et al., 2005)
    - OPJK (Casellas, 2008)
    - DALOS (Agnoloni et al., 2009)

Ontology learning in the legal domain: the main challenges

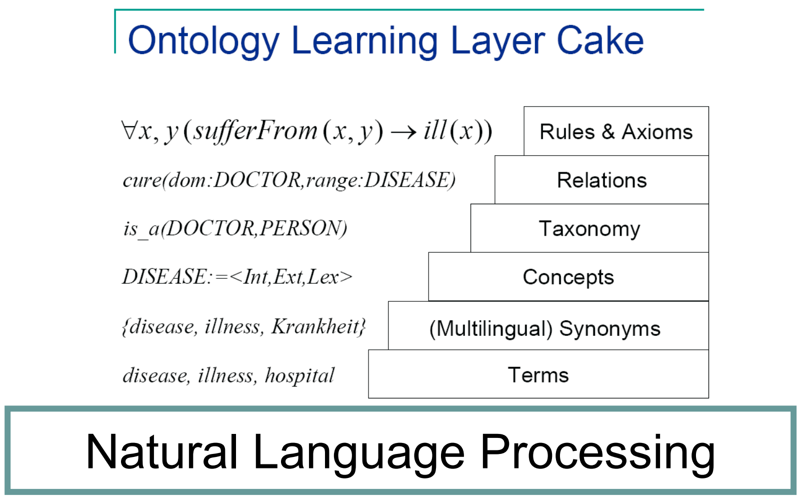
- The typical **ontological acquisition bottleneck**
  - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure
  - "One of the main obstacles to progress in the field of artificial intelligence and law is the natural language barrier" (McCarty, ICAIL 2007)

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

- The **«epistemological promiscuity»** as a common attitude in constructing legal ontologies (Breuker & Hoekstra, 2004)
  - i.e. domain independent concepts of law are tainted with concepts referring to the world the legal domain is about

Ontology learning in the legal domain: the ontological acquisition bottleneck

- Technologies in the area of knowledge management are typically confronted with the problem of processing linguistic structure
  - Particular 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)
- Which domain-specific issues must be taken into account addressing the legal language processing task?
  - To what extent legal language differ from ordinary language?
  - Do these differences have impact on the performances of NLP tools?



Ontology learning in the legal domain: the main challenges

- The typical **ontological acquisition bottleneck**
  - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure
  - "One of the main obstacles to progress in the field of artificial intelligence and law is the natural language barrier" (McCarty, ICAIL 2007)

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

- The **«epistemological promiscuity»** as a common attitude in constructing legal ontologies (Breuker & Hoekstra, 2004)
  - i.e. domain independent concepts of law are tainted with concepts referring to the world the legal domain is about

Ontology learning in the legal domain: the peculiarity of legal language (1)

- Which are the linguistic peculiarities of legal texts?
  - Focus on legislative texts

- Two main different perspectives of analysis carried out by (Venturi 2008, 2012; Dell'Orletta et al., 2012):
  - comparative
    - Whether and to what extent a legal text is something different from ordinary language?
  - contrastive
    - Are there features which are shared by the "legalese" of different languages (e.g. legal Italian vs legal English)?

- Monitoring the distribution of a number of linguistic characteristics between
  - legal texts vs newspapers (i.e. ordinary language)
  - different subvarieties of legal language (i.e. EU, national and regional laws)
  - Italian vs English EU laws

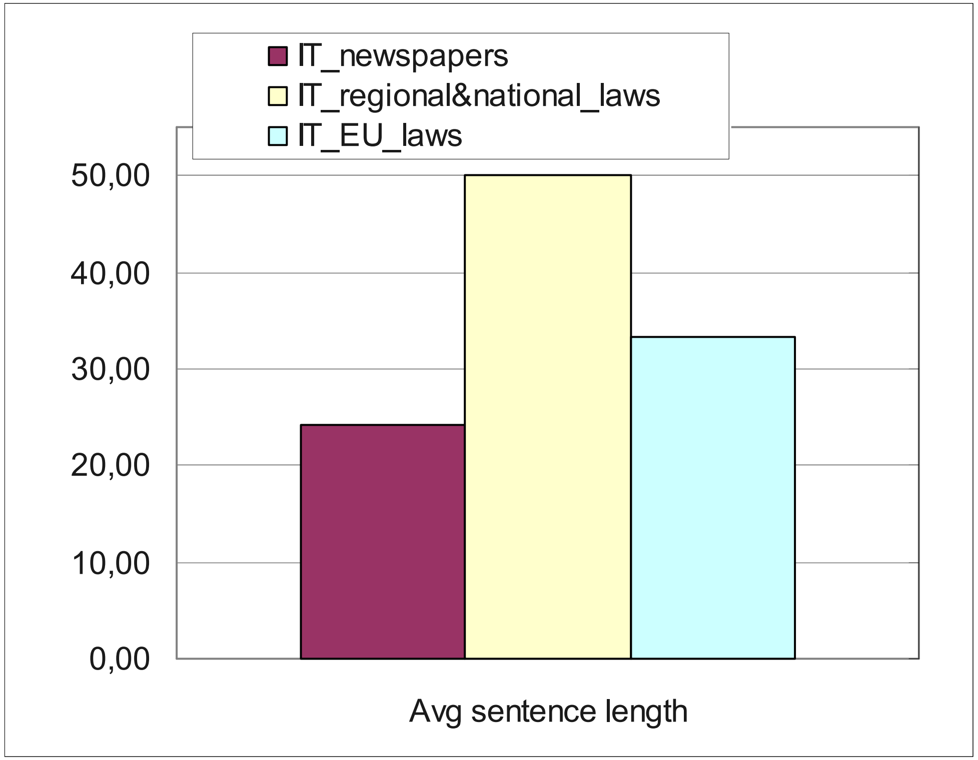
Ontology learning in the legal domain: the peculiarity of legal language (2)

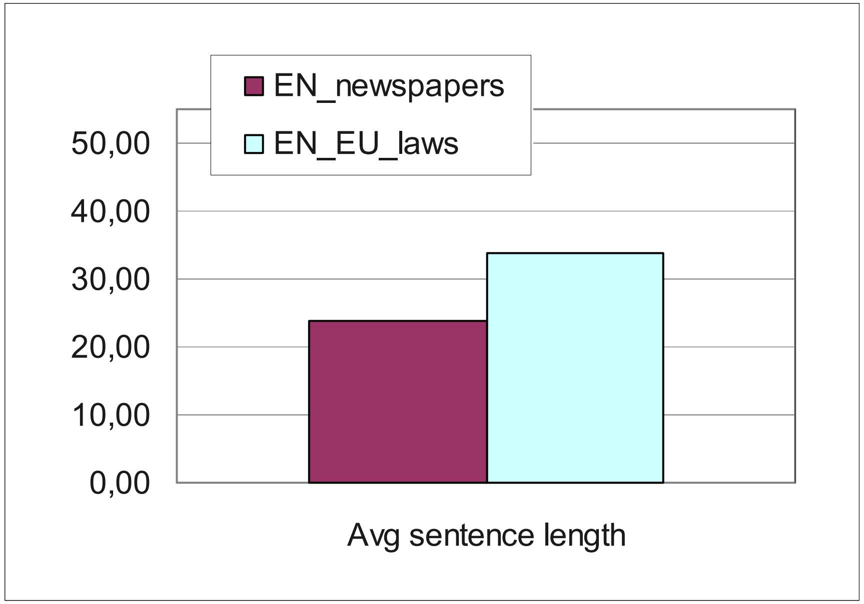
- The monitored corpora
  - Italian:
    - a corpus of newspapers representative of the ordinary Italian
    - a collection of laws enacted by the European Commission
    - a collection of laws enacted by Italian State and Regions
  - English:
    - a corpus of newspapers representative of the ordinary English
    - a collection of laws enacted by the European Commission
  - The two corpora of Italian and English European laws contain aligned sentences (i.e. translations of the same text)

- Wide range of monitored linguistic characteristics:
  - *raw* features (e.g. sentence length, word length)
  - morpho-syntactic features (e.g. the distribution of verbs, nouns, etc.)
  - syntactic (concerning the structure of the sentence)

Ontology learning in the legal domain: the peculiarity of legal language (3)

- Distribution of *raw* textual features

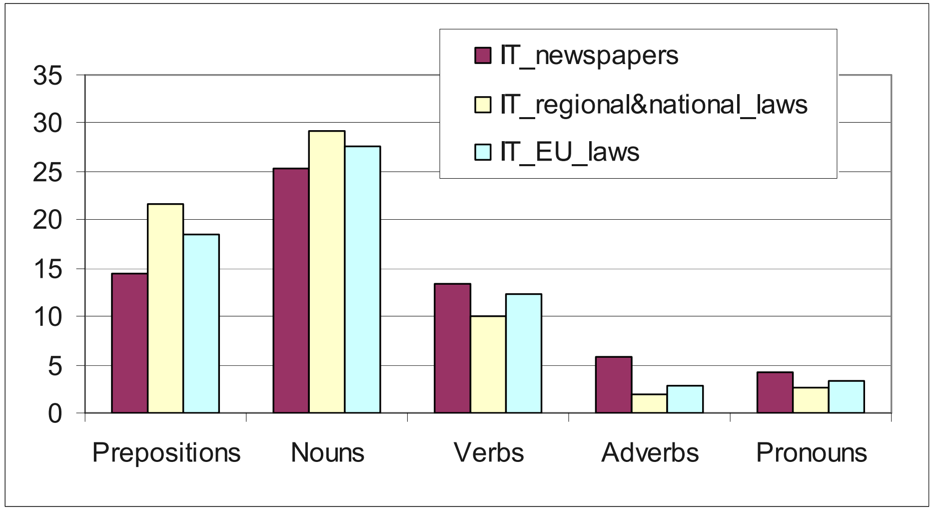


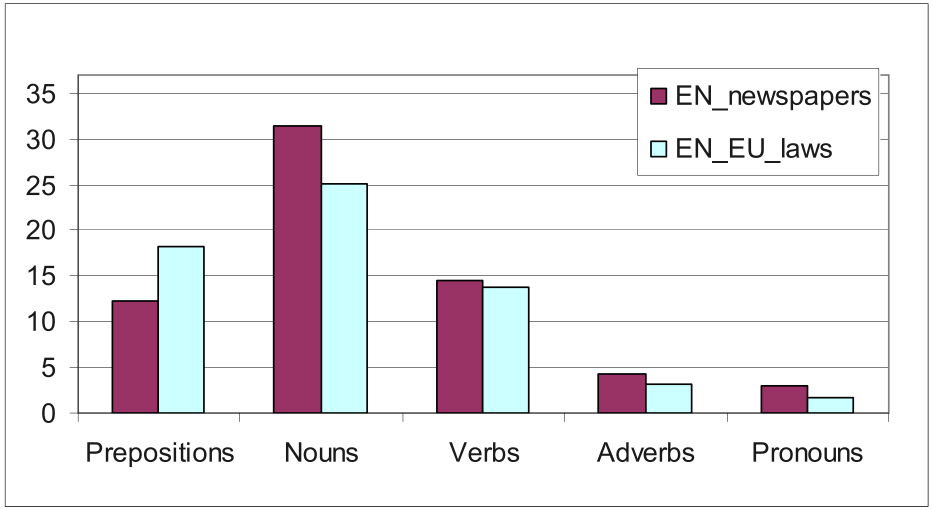


- IT and EN legal texts contain longer sentences wrt newswire texts
  - Italian Regional and national texts contain the longest sentences

Ontology learning in the legal domain: the peculiarity of legal language (4)

- Distribution of morpho-syntactic features





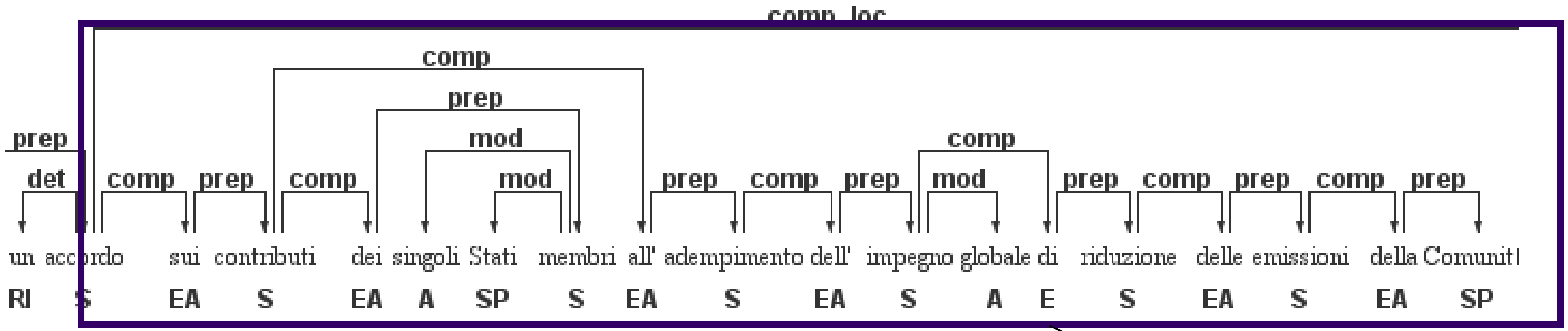
- For both IT and EN, legal texts are characterized by:
  - higher % of prepositions
    - connected with long “chains” of complements
  - lower % of verbs, adverbs, pronouns

Ontology learning in the legal domain: the peculiarity of legal language (5)

- Distribution of syntactic features

Ontology learning in the legal domain: the peculiarity of legal language (5)

- Distribution of syntactic features
  - Sequence of consecutive prepositional complement (embedded complement ‘chains’)

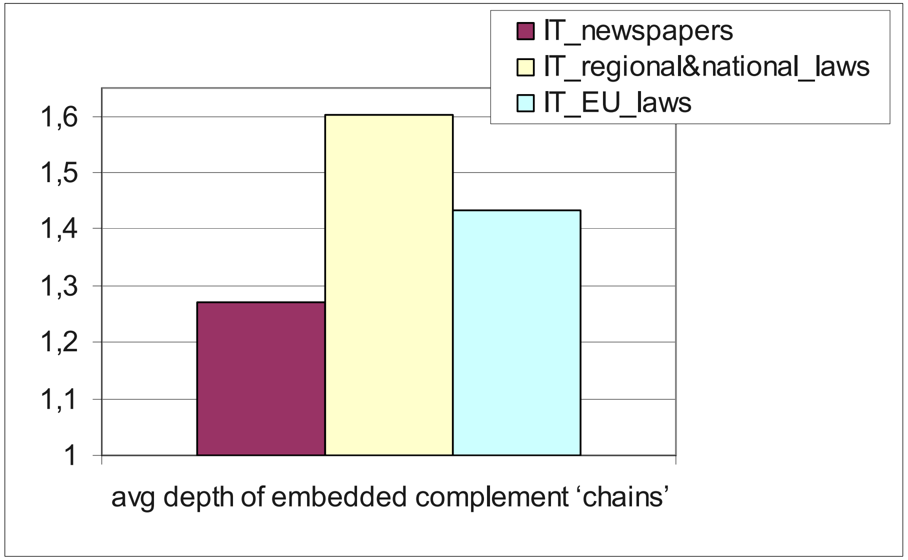


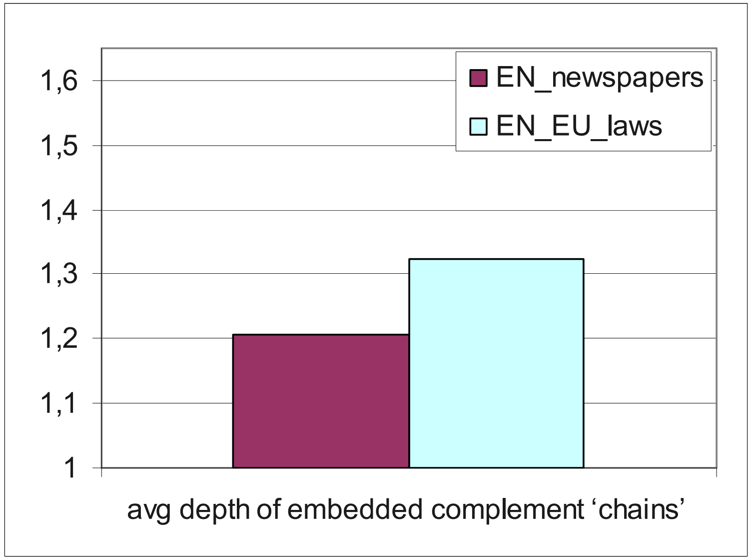
Embedded complement ‘chain’ (length=6)

Il Consiglio è giunto ad un **accordo** <u>sui contributi dei singoli Stati membri all’adempimento dell’impegno globale di riduzione delle emissioni della Comunità</u> nelle conclusioni del Consiglio del 16 giugno 1998. (*The Council agreed <u>upon the contributions of each Member State to the overall Community reduction commitment</u> in the Council conclusions of 16 June 1998*)

Ontology learning in the legal domain: the peculiarity of legal language (5)

- Distribution of syntactic features

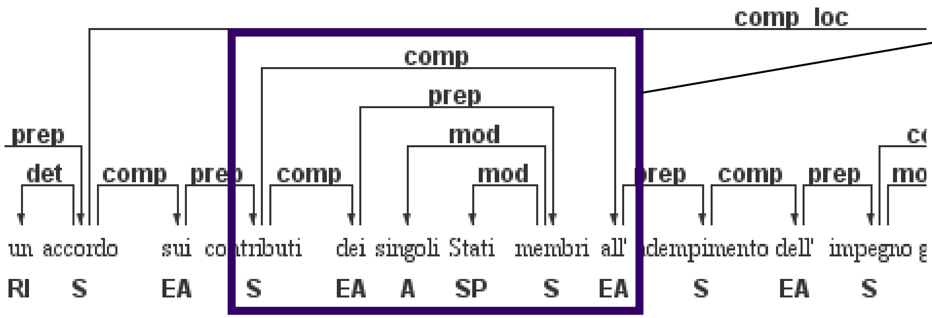




- Both IT and EN legal texts contain the deeper **embedded complement “chains”** wrt newswire texts
  - Regional and national IT texts contain the deepest “chains”

Ontology learning in the legal domain: the peculiarity of legal language (6)

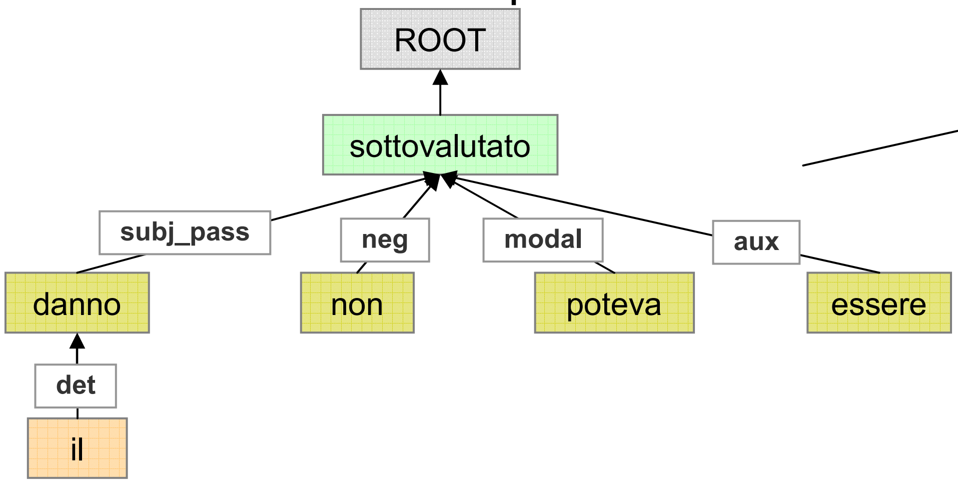
- Distribution of syntactic features
  - Dependency link length



Length of dependency link (length=5)

Calculated in terms of the number of words occurring between the syntactic head (i.e. *contributi*) and the dependent (i.e. *all'*)

  - Parse tree depth

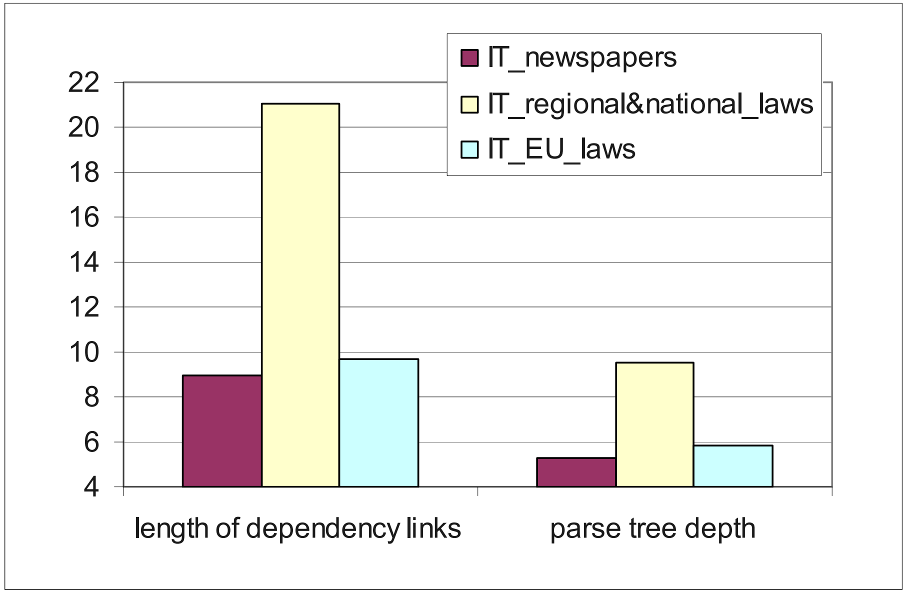


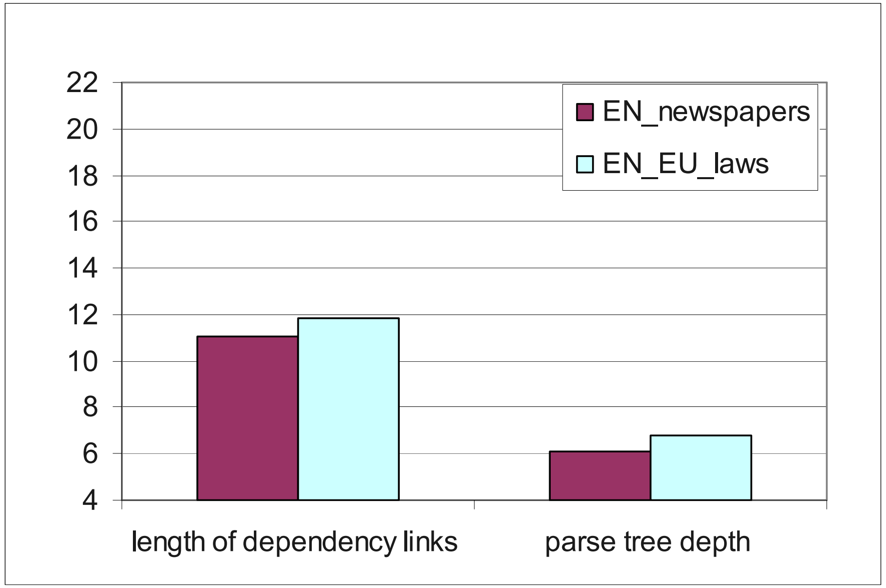
Parse tree depth (depth=2)

Calculated in terms of the longest path from the ROOT (i.e. *sottovalutato*) of the dependency tree to some leaf (i.e. *il*)

Ontology learning in the legal domain: the peculiarity of legal language (6)

- Distribution of syntactic features





- Both IT and EN legal texts are characterized by
  - longer **dependency links** (on average)
  - higher **parse trees** (on average)
- Regional and national IT texts contain the longest links and the highest trees

- Note that:
  - McDonald and Nivre (2007) report that statistical parsers have a drop in accuracy when analyzing long distance dependencies
  - Parse tree depth is another feature reflecting sentence complexity

Ontology learning in the legal domain: the impact of legal language on NLP tools (1)

- What is the performance of state-of-the-art NLP tools on legal texts?
  - A challenge for all NLP tasks
  - Dramatic drop of accuracy when syntactic parsers are tested on domains outside of the data from which they are trained or developed on (Gildea, 2001)
- Main focus on **dependency parsing** since it represents a prerequisite for any advanced text processing tasks
- So far, few answers
  - Scarcity of *gold* corpora of legal texts manually annotated with syntactic information with respect to which this performance could be evaluated

Ontology learning in the legal domain: the impact of legal language on NLP tools (2)

- The answers so far:
  - McCarty (2007) - English
    - "How accurate is Collins' parser on sentences from judicial opinions?
    - Impossible to answer due to the lack of a gold standard legal corpus
    - From a qualitative analysis, "the parser is very good on the internal structure of sentences, but it is weaker on prepositional phrase attachments and coordinated conjunctions"
  - Walter (2009) - German
    - PReDS parser, trained on general language (newspaper) texts
    - Gold standard legal (case law) corpus of 100 sentences
    - PReDS performance decreases from 86.74% to 64%
    - This corpus is currently encoded following the PReDS parser native annotation format: its exploitation would require the conversion into some kind of standard representation format (e.g. CoNLL)

Ontology learning in the legal domain: the impact of legal language on NLP tools (3)

- More recently, two initiatives
  - *Domain Adaptation Track* at Evalita 2011 – Italian
  - SPLeT-2012 *Shared Task on Dependency Parsing of Legal Texts* – Italian and English

- Each track was organized into different subtasks devoted to
  - 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
    - i.e. a domain outside of the data from which they were trained or developed

Ontology learning in the legal domain: the impact of legal language on NLP tools (3)

- 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

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

- Accuracy results for English:

Participant System	**Newspaper test**	**EU legal test**
1	88.81	78.90

- Significant drop of accuracy on the legal test wrt newspaper test

Ontology learning in the legal domain: the impact of legal language on NLP tools (4)

- Lower performance of parsing systems on legal texts wrt newspapers
  - For both Italian and English
  - Across different subvarietes of legal language
  - In all cases the accuracy of analysis results MUST be improved

- The need for a strategy for **adapting** general purpose dependency parsers to the legal domain
  - Ongoing efforts in this direction
  - Benefit for a number of real word applications for the legal domain which rely on the automatic linguistic annotation of text
    - Such as Ontology Learning

Ontology learning in the legal domain: the main challenges

- The typical **ontological acquisition bottleneck**
  - as knowledge is mostly conveyed through text, content access requires understanding the linguistic structure
  - “One of the main obstacles to progress in the field of artificial intelligence and law is the natural language barrier” (McCarty, ICAIL 2007)
- The **peculiarity of legal language and its impact on NLP tools**
  - Legal syntax is “convoluted and unnatural” (McCarty, NaLEA 2009) wrt ordinary language
  - What is the performance of state-of-the-art NLP tools on legal texts?
- The **«epistemological promiscuity»** as a common attitude in constructing legal ontologies (Breuker & Hoekstra, 2004)
  - i.e. domain independent concepts of law are tainted with concepts referring to the world the legal domain is about

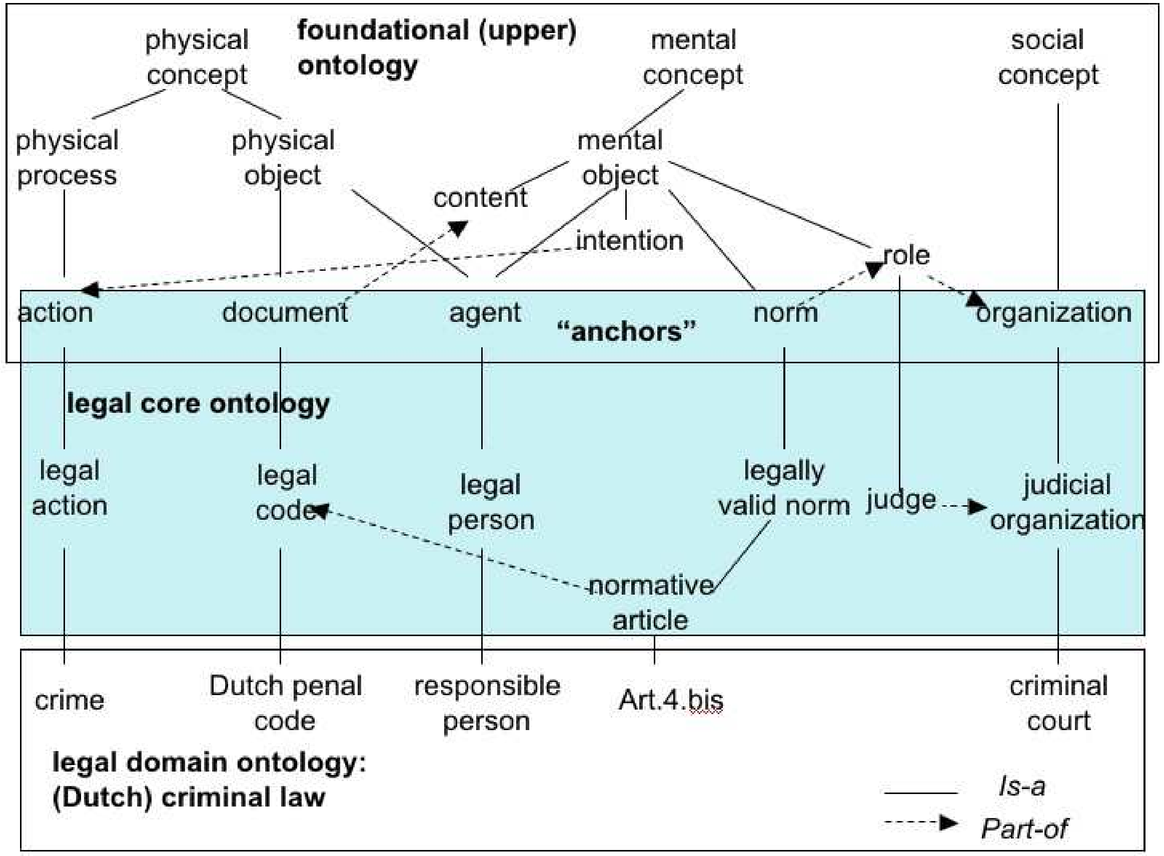
Ontology learning in the legal domain: the challenge of the «epistemological promiscuity» (1)

- 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)

- However, the «epistemological promiscuity» is a common attitude in constructing legal ontologies (Breuker & Hoekstra 2004)
  - i.e. domain independent concepts of law are mixed with concepts referring to the world the legal domain is about
    - «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'.»

Ontology learning in the legal domain: the challenge of the «epistemological promiscuity» (2)

- Discriminating between legal and regulated domain terms and/or concepts is key in a legal ontology learning process
  - It is closely related to the reusability and interoperability issue
  - «ontologies mixed with epistemological frameworks have a far more limited re-use and may pose more interoperability problems than clean ontologies.» (Breuker & Hoekstra 2004)



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

Case studies carried out in the legal domain with T2K

- T2K (*Text–to–Knowledge*) ontology learning system
  - Institute of Computational Linguistics “Antonio Zampolli” (ILC-CNR) and Department of Linguistics of the University of Pisa (Dell’Orletta et al. 2006)
  - It offers a battery of tools for NLP, statistical text analysis and machine language learning, dynamically integrated to induce ontological knowledge from texts
  - Text interpretation ranges from the acquisition of terminology to advanced syntax and ontological/conceptual mapping

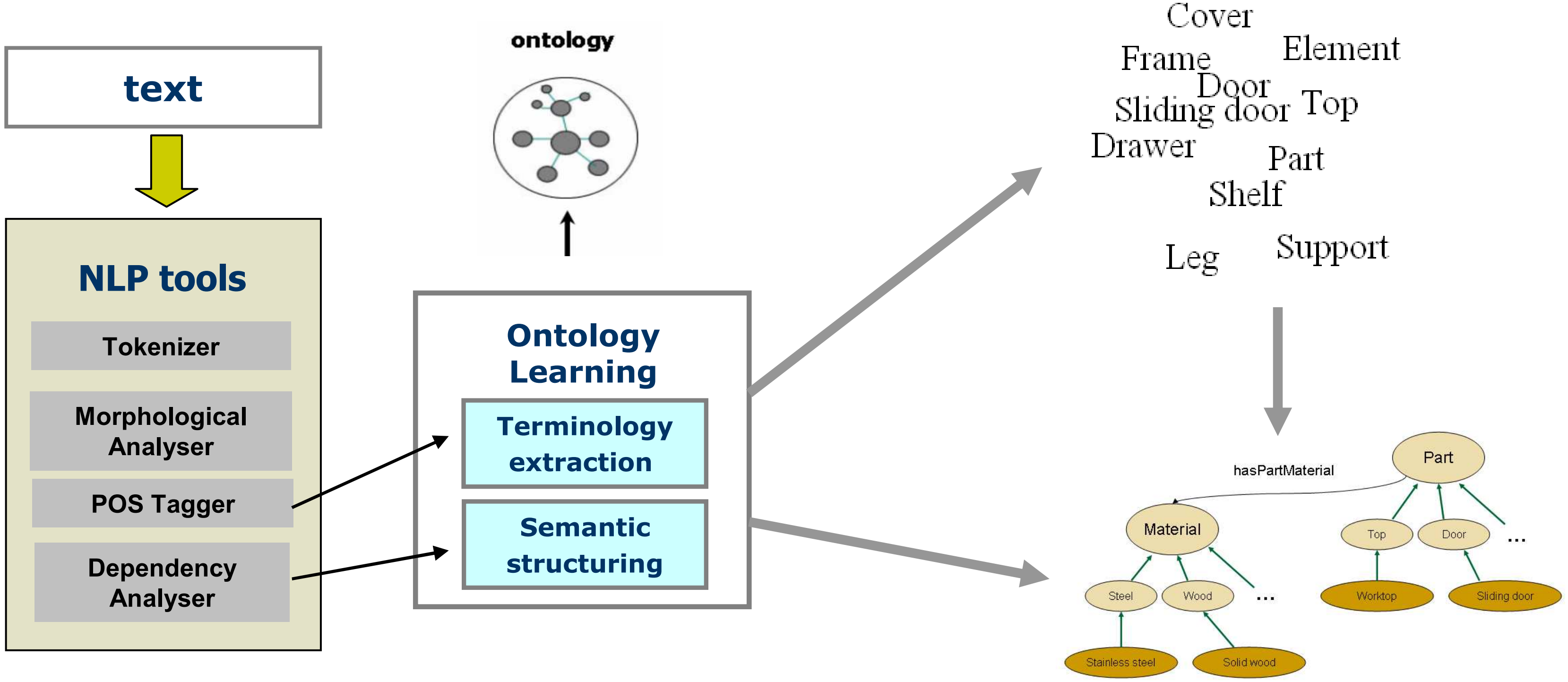
Case studies carried out in the legal domain with T2K

- T2K (*Text–to–Knowledge*) ontology learning system

- Case studies performed so far in the legal domain:
  - T2K_v1
    - Corpus of environmental laws (Venturi, 2006)
      - Institutional and administrative acts by EU, State and Piedmont Region (1,399,617 tokens)
    - Consumer Law corpus (European DALOS project) (Agnoloni et al., 2009)
      - including EU Directives, Regulations and case law on protection of consumers' economic and legal interests (292,609 tokens)
  - T2K_v2
    - 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)

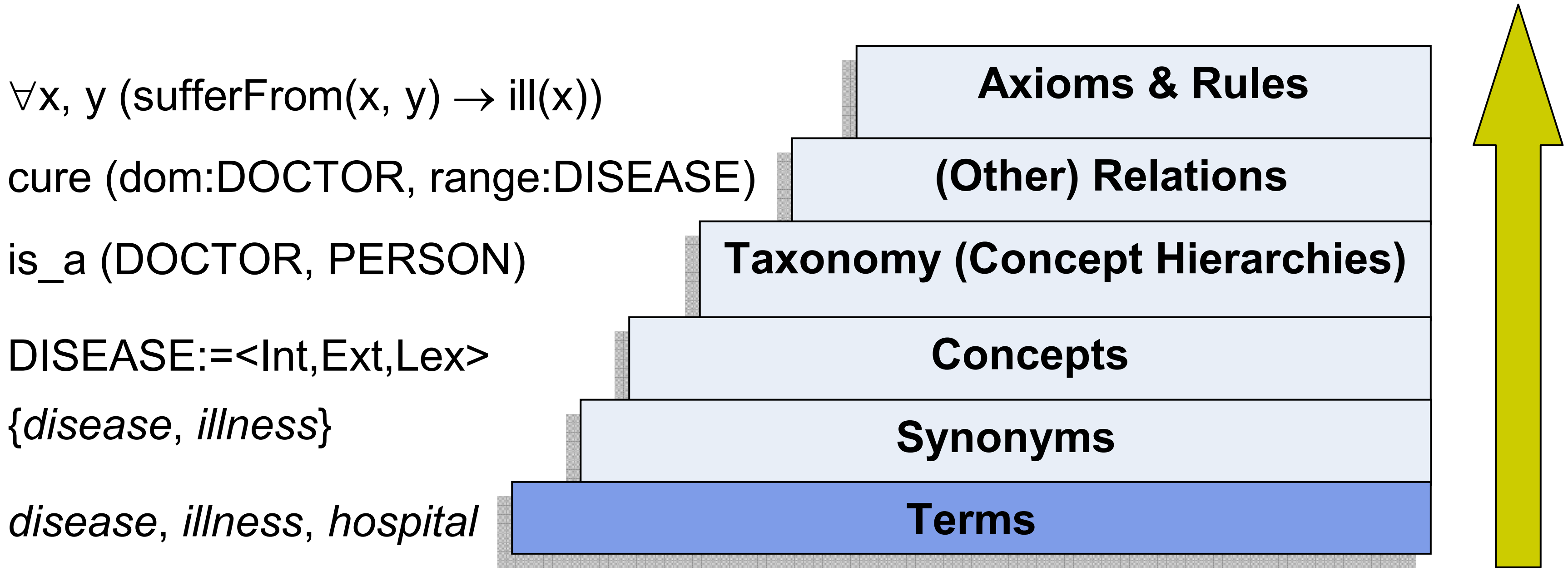
Ontology Learning with T2K

- 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: terminology extraction (1)

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



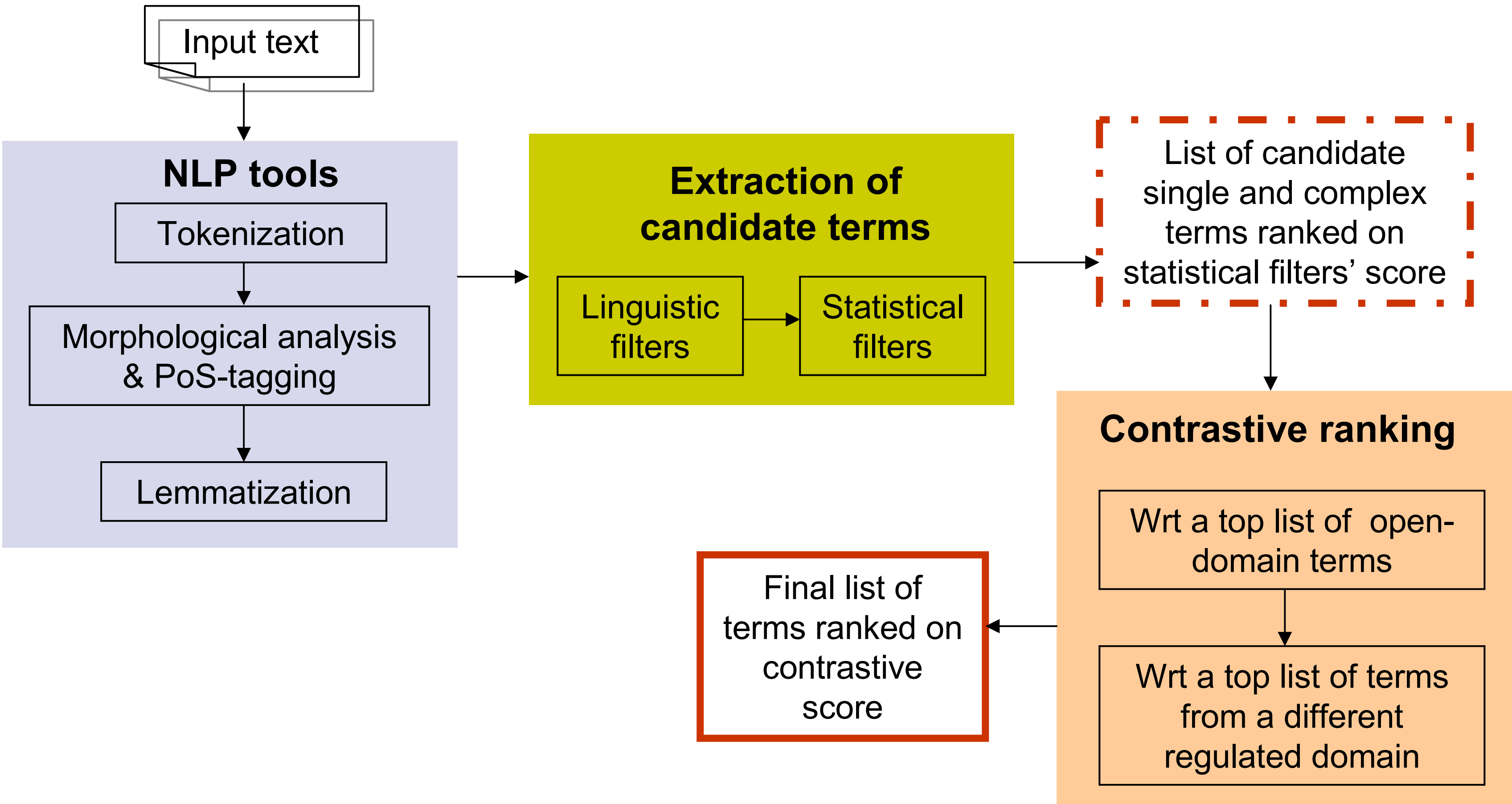
Ontology Learning: terminology extraction (2)

- 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

- Dealing with legal texts, we need to take into account:
  - the extraction of terms corresponding to legal relevant concepts
    - e.g. *law, legislative decree*
  - the identification of the specific domain of the collection of legal documents (i.e. the 'world knowledge')
    - e.g. *consumer, hazardous substance*
- It can be a helpful starting point for any further construction of legal ontologies where **legal** and **world knowledge** is kept separate

Ontology Learning: terminology extraction (3)

- The multi-layered term extraction architecture of T2K_v2



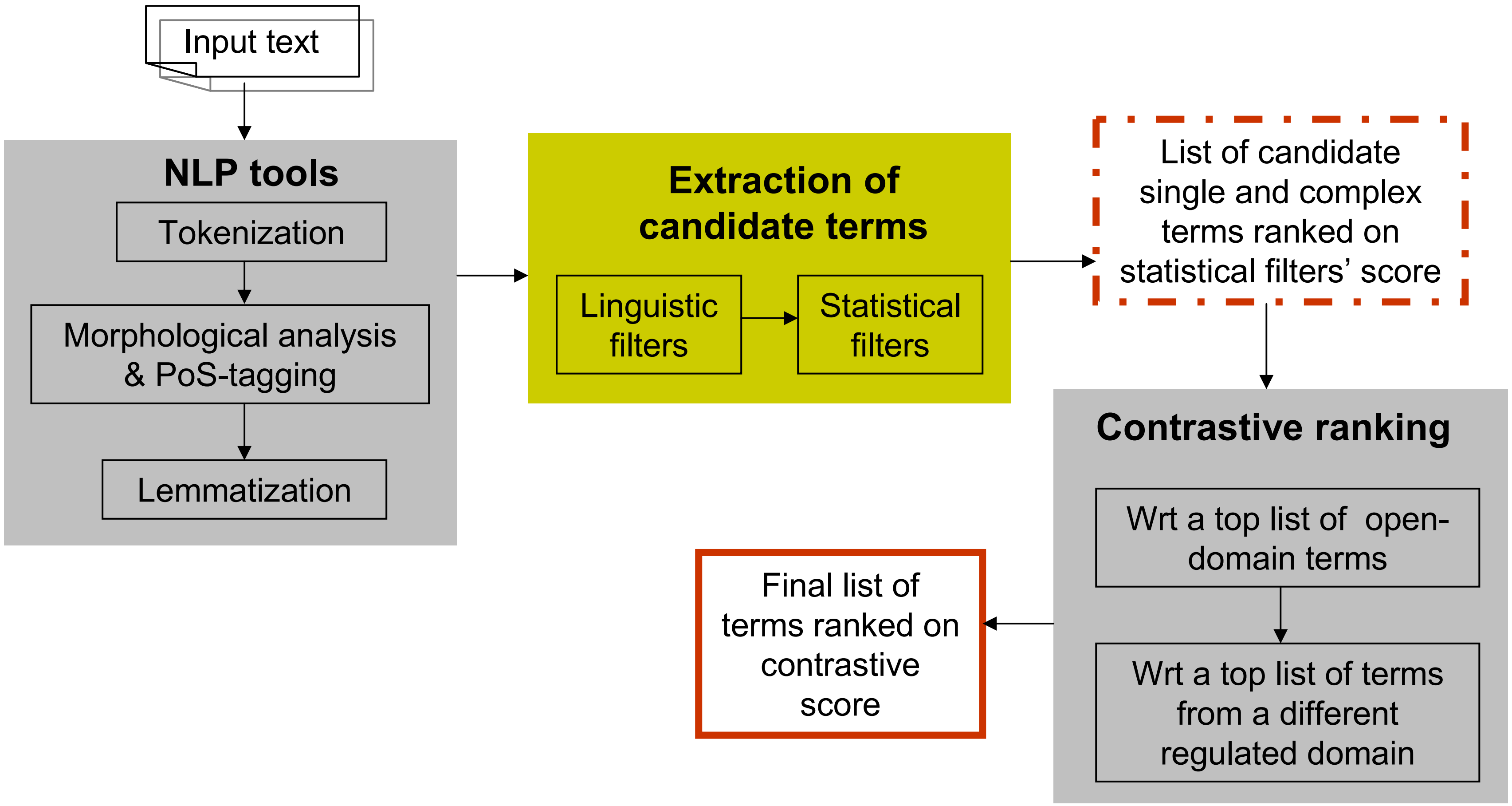
Ontology Learning: terminology extraction (4)

- 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: terminology extraction (5)

- The multi-layered term extraction architecture of T2K_v2



Ontology Learning: terminology extraction (6)

- Extraction of candidate terms
  - 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: terminology extraction (6)

- Linguistic filters
  - **Single terms**: nouns (S), e.g. *impianto* (*plant*), *direttiva (directive)*

Forma	Lemma	CPoSTag	PosTag	Tratti morfologici
Il	il	R	RD	num=s|gen=m
piano	piano	S	S	num=s|gen=m
nazionale	nazionale	A	A	num=s|gen=n
di	di	E	E	_
riduzione	riduzione	S	S	num=s|gen=f
delle	di	E	EA	num=p|gen=f
emissioni	emissione	S	S	num=p|gen=f
in	in	E	E	_
nessun	nessun	D	DI	num=s|gen=m
caso	caso	S	S	num=s|gen=m
può	potere	V	VM	num=s|per=3|mod=i|ten=p
esonerare	esonerare	V	V	mod=f

Forma	Lemma	CPoSTag	PosTag	Tratti morfologici
un	un	R	RI	num=s|gen=m
impianto	impianto	S	S	num=s|gen=m
dal	da	E	EA	num=s|gen=m
rispetto	rispetto	S	S	num=s|gen=m
della	di	E	EA	num=s|gen=f
pertinente	pertinente	A	A	num=s|gen=n
normativa	normativa	S	S	num=s|gen=f
comunitaria	comunitario	A	A	num=s|gen=f
,	,	F	FF	_
compresa	comprendere	V	V	num=s|mod=p|gen=f
la	il	R	RD	num=s|gen=f
direttiva	direttiva	S	S	num=s|gen=f
96/61/CE.	96/61/CE.	S	SP	_

Ontology Learning: terminology extraction (6)

- Linguistic filters
  - **Single terms**: nouns (S), e.g. *impianto* (*plant*), *direttiva (directive)*
  - **Multi-word terms**: part-of-speech patterns
    - noun+preposition+noun (S+E+S), e.g. *riduzione di emissione (emission reduction)*
    - noun+adjective (S+A), e.g. *piano nazionale (national plan), normativa comunitaria (Community legislation)*

Forma	Lemma	CPoSTag	PosTag	Tratti morfologici
Il	il	R	RD	num=s|gen=m
piano	piano	S	S	num=s|gen=m
nazionale	nazionale	A	A	num=s|gen=n
di	di	E	E	_
riduzione	riduzione	S	S	num=s|gen=f
delle	di	E	EA	num=p|gen=f
emissioni	emissione	S	S	num=p|gen=f
in	in	E	E	_
nessun	nessun	D	DI	num=s|gen=m
caso	caso	S	S	num=s|gen=m
può	potere	V	VM	num=s|per=3|mod=i|ten=p
esonerare	esonerare	V	V	mod=f

Forma	Lemma	CPoSTag	PosTag	Tratti morfologici
un	un	R	RI	num=s|gen=m
impianto	impianto	S	S	num=s|gen=m
dal	da	E	EA	num=s|gen=m
rispetto	rispetto	S	S	num=s|gen=m
della	di	E	EA	num=s|gen=f
pertinente	pertinente	A	A	num=s|gen=n
normativa	normativa	S	S	num=s|gen=f
comunitaria	comunitario	A	A	num=s|gen=f
,	,	F	FF	_
compresa	comprendere	V	V	num=s|mod=p|gen=f
la	il	R	RD	num=s|gen=f
direttiva	direttiva	S	S	num=s|gen=f
96/61/CE.	96/61/CE.	S	SP	_

Ontology Learning: terminology extraction (7)

- Statistical filters
  - **Single terms**: ranked on the basis of their 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

Ontology Learning: terminology extraction (7)

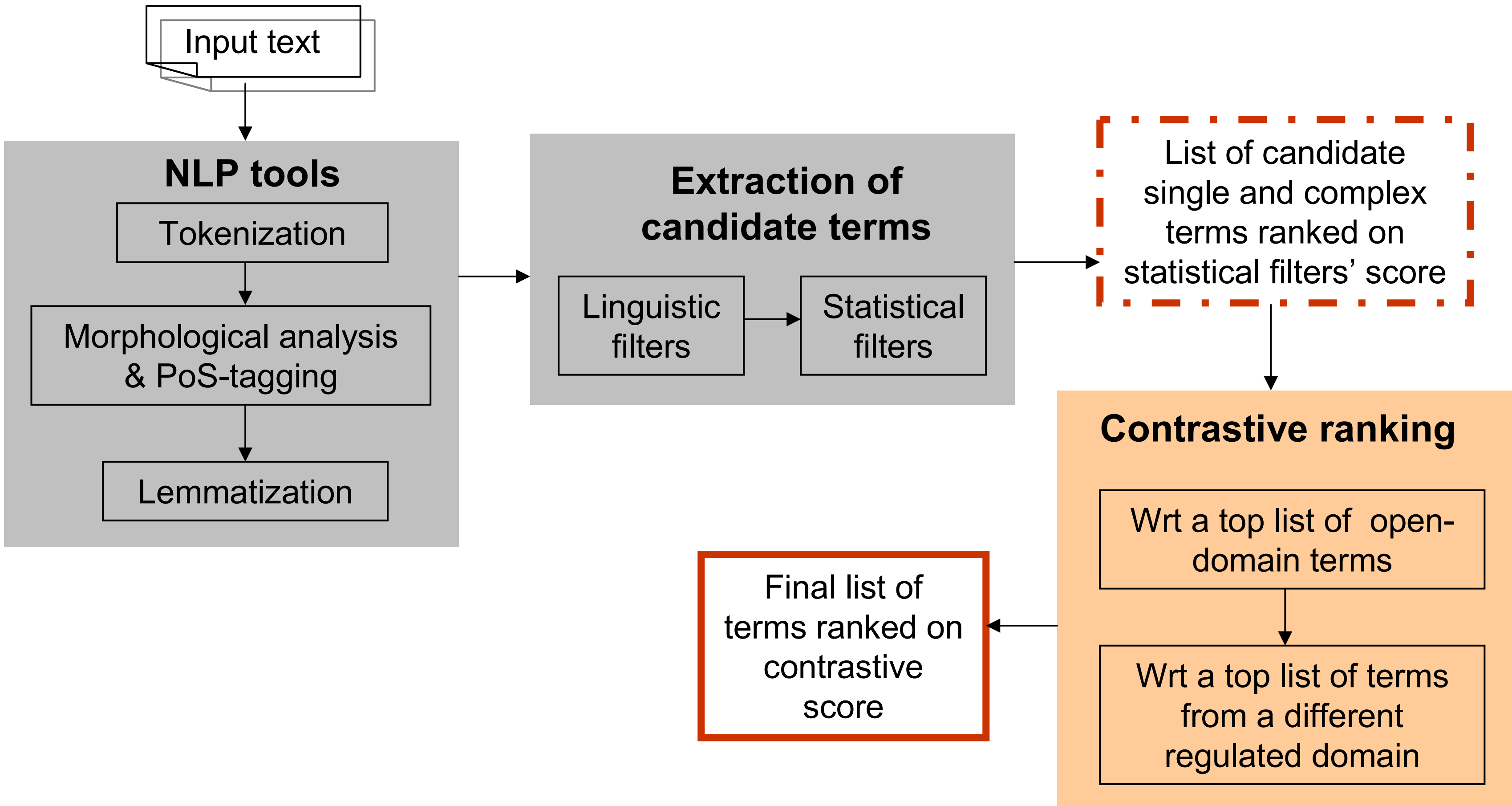
- Statistical filters
  - **Single terms**: ranked on the basis of their frequency of occurrence in the input text
  - **Multi-word terms**: ranked on the score of the C-NC Value (Frantzi & Ananiadou 1999), assessing the likelihood for a term of being a well-formed and relevant multi-word term

impianto	1,570796318
amministratore	1,570796316
emission	
gas	
sostanza	
energia	
serra	
produzio	
deposito	
tabella	
riduzione	
stoccagg	
veicolo	
quota	
protocoll	
fonte	
costrutto	
elettricità	
inquinam	
autovettu	
aria	
strategia	
unità	
carbonio	
quantità	
acqua	
gestore	
misurazio	
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: terminology extraction (8)

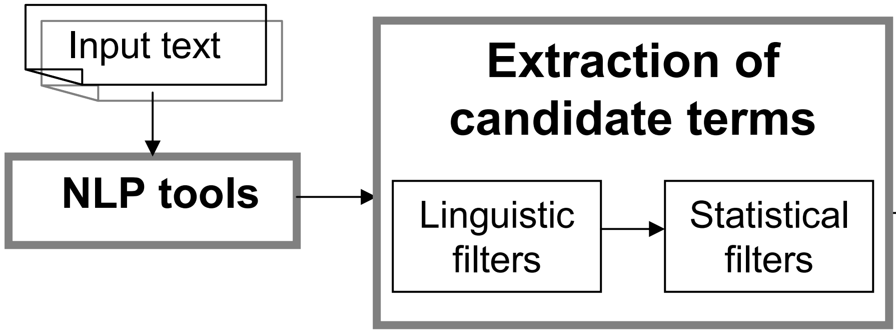
- The multi-layered term extraction architecture of T2K_v2



Ontology Learning: terminology extraction (9)

- Contrastive ranking
  - Aimed at assessing the domain relevance of terms on the basis of the contrastive distribution of relevant candidate terms across an input corpus and a different corpus
  - The ranking of identified terms is revised on the basis of a contrastive score calculated for the same terms with respect to a different corpus
- An example
  - Input corpus: a collection of EU Italian Directives concerning the environmental domain
  - Two contrasted corpora:
    - a corpus of Italian texts of different types (newspapers, books, etc.) testifying general language usage
      - to prune common words
    - a corpus of containing EU Italian Directives on consumer protection
      - to discriminate legal and regulated-domain terminology

Ontology Learning: terminology extraction (10)



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

Output of the statistical filters:

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

Ontology Learning: terminology extraction (10)

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 the general language corpus

Ontology Learning: terminology extraction (10)

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 the general language corpus

Contrast against a top list of terms from the corpus of Directives on 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: terminology extraction (11)

- Evaluation of results
  - Accuracy of the extracted list of terms
    - An example: Consumer Law corpus (European DALOS project) (Agnoloni et al., 2009)
    - Reference resources selected as a gold standard
      - the thesaurus of DOGI Archive
      - JurWordNet
    - Precision
      - calculated as the percentage of correctly acquired terms with respect to all acquired terms
      - 85.38%
    - Recall wrt relevant 56 European Union Legal Concepts (EULG)
      - calculated as the percentage of correctly acquired terms with respect to all terms in the gold standard
      - 80.69%

Ontology Learning: terminology extraction (12)

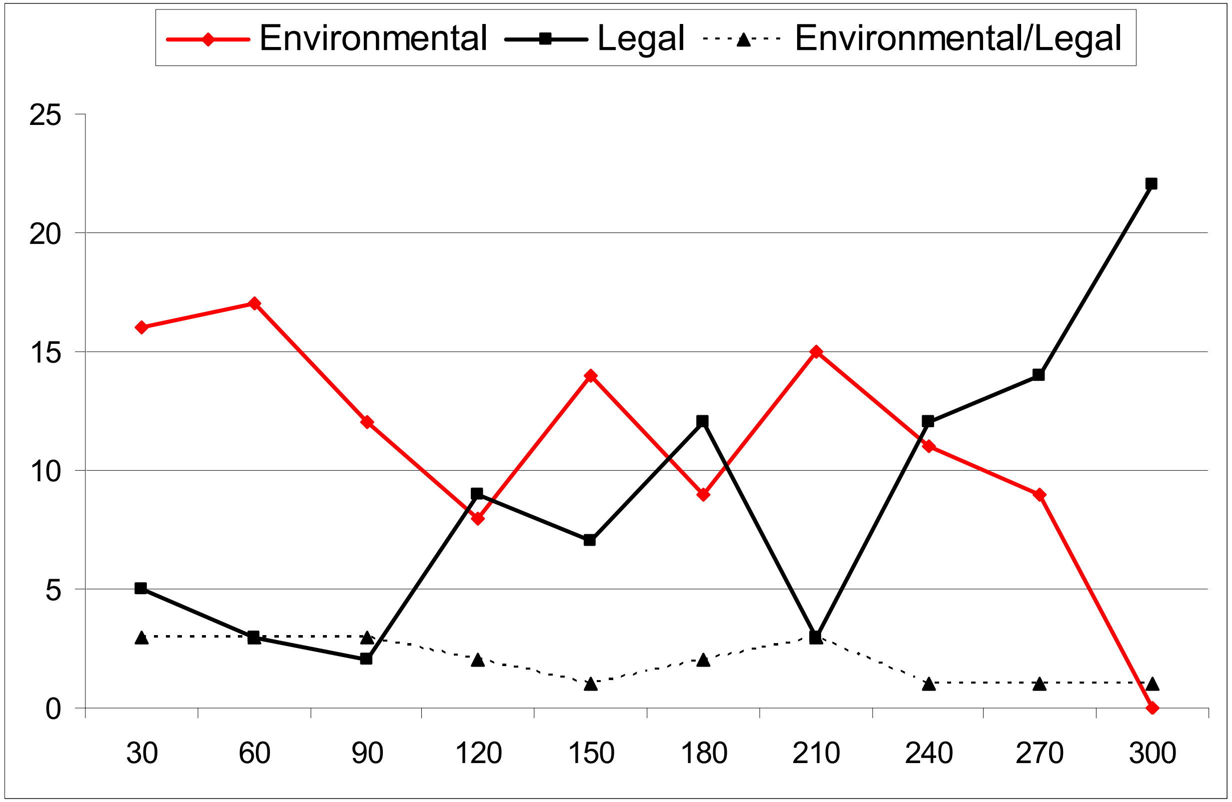
- Evaluation of results
  - Ability to discriminate legal and domain-specific (world) terms
    - An example: Corpus of environmental laws (Bonin et al., 2010)

- Reference resources selected as gold standard
  - The thesaurus EARTh (Environmental Applications Reference Thesaurus)
    - To assess the accuracy of the terms acquired as belonging to the environmental
  - The Dizionario giuridico, Edizioni Simone
    - To assess the accuracy of the terms acquired as belonging to the legal

Ontology Learning: terminology extraction (12)

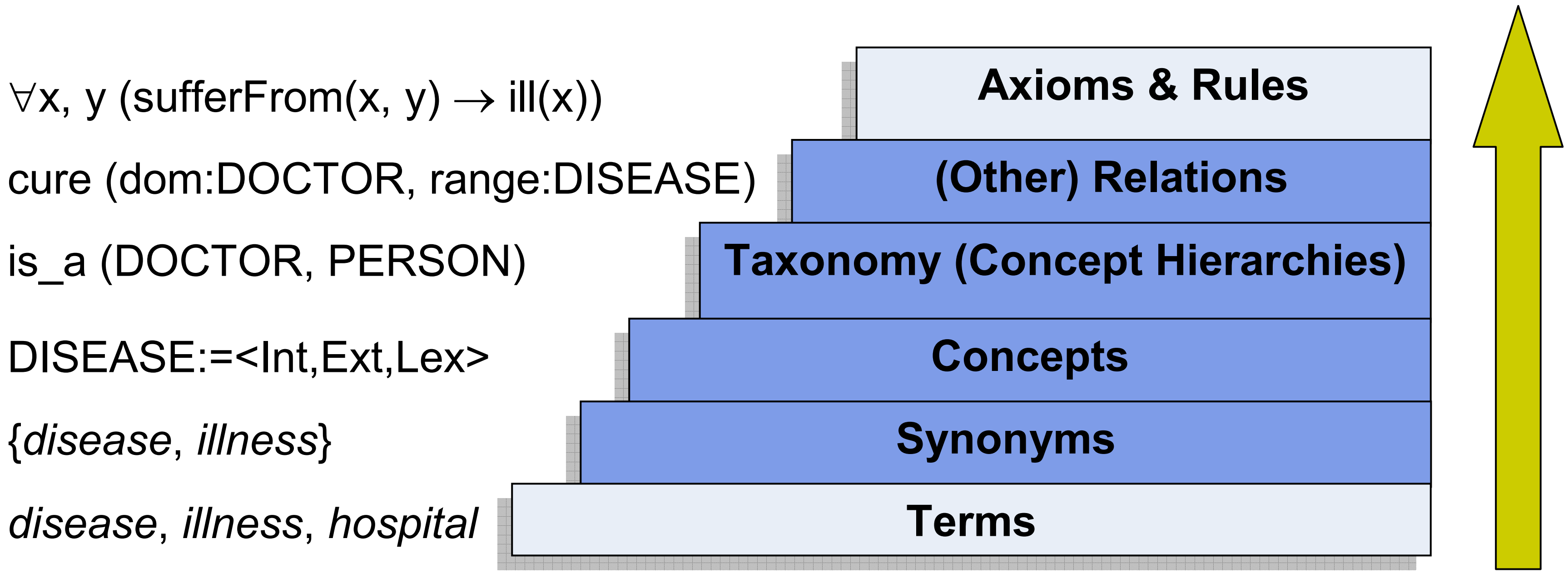
- Evaluation of results
  - Ability to discriminate legal and domain-specific (world) terms
    - Corpus of environmental laws (Bonin et al., 2010)

- The extracted terms were classified as
  - environmental terms, e.g. *sostanza pericolosa* (hazardous substance)
  - legal terms, e.g. *disposizione nazionale* (national provision)
  - terms which can refer to both domains, e.g. *politica ambientale* (environmental policy)



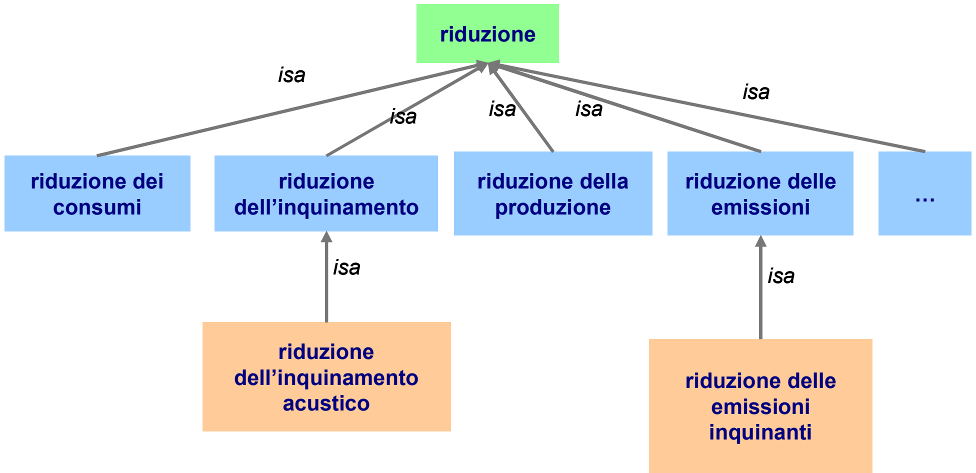
Ontology Learning: semantic structuring (1)

- Extracted terms are organised by T2K into:
  - fragments of taxonomical chains
  - clusters of semantically related terms
- Both are useful inputs which can be used to support a semi-automatic construction of an ontology



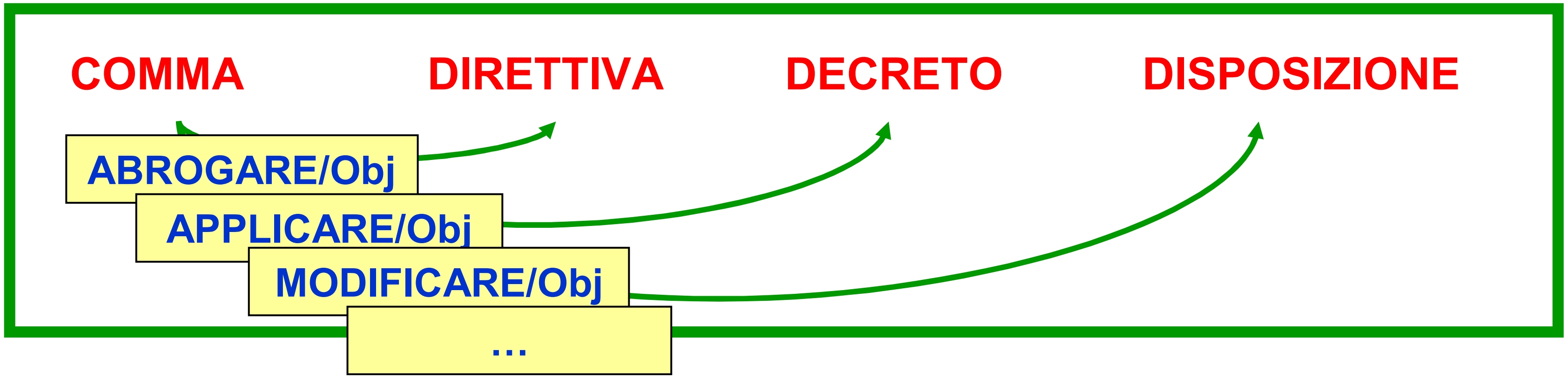
Ontology Learning: semantic structuring (2)

- Fragments of taxonomical chains
  - Example extracted from the Corpus of environmental laws (Lenci et al., 2008)
- simple and multi-word terms structured in a vertical hierarchy
  - On the basis of their internal linguistic structure (head sharing)



Ontology Learning: semantic structuring (3)

- Clusters of semantically related terms
  - Example extracted from the Corpus of environmental laws (Lenci et al., 2008)
  - inferred through dynamic distributionally-based similarity measures
  - using a contex-sensitive notion of semantic similarity
  - computing the most relevant co-occurring verb/subject and verb/object pairs (dependency-annotated text)
  - terms which share the property of being interchangeable in a syntactic context are clustered



Ontology Learning: semantic structuring (3)

- Clusters of semantically related terms
    - Example extracted from the Corpus of environmental laws (Lenci et al., 2008)
  - inferred through dynamic distributionally-based similarity measures
  - using a contex-sensitive notion of semantic similarity
  - computing the most relevant co-occurring verb/subject and verb/object pairs (dependency-annotated text)
  - terms which share the property of being interchangeable in a syntactic context are clustered

INQUINAMENTO (pollution)
DANNO AMBIENTALE (environmental damage)
INQUINAMENTO MARINO (sea pollution)
EFFETTI NOCIVI (harmful effect)
CONSEGUENZA (consequence)
INQUINAMENTO ATMOSFERICO (air pollution)

- Different kinds of relations are collapsed
  - e.g. *inquinamento* (pollution)/ *danno ambientale* (environmental damage): effect-cause relation
  - e.g. *inquinamento* (pollution)/ *inquinamento atmosferico*: taxonomical relation

From bricks of knowledge to a domain ontology

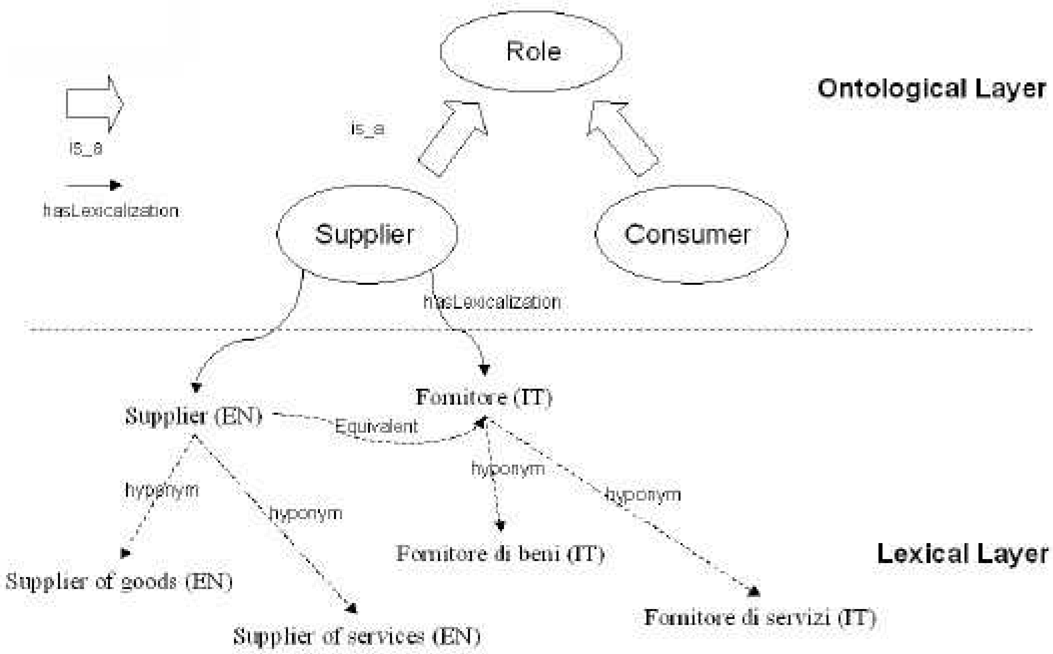
T2K at work: an example of a "middle-out" approach to ontology construction

- 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 **Ontologycal layer**, containing the conceptual modelling at a language independent level
  - the **Lexical layer**, containing multi-lingual terminology conveying the concepts represented at the Ontologycal layer

T2K at work: an example of a "middle-out" approach to ontology construction

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

- **Lexical layer**
  - Bottom-up approach
  - Terms are automatically extracted from a corpus of Consumer Protection laws
  - They are automatically organized into taxonomical relations
  - They are linked to their translation equivalent
- **Ontologycal layer**
  - Top-down approach
  - Domain-specific concepts and their relationships are manually defined by domain expert



T2K current directions of research

- Semi-automatic induction and labelling of basic ontological classes from acquired proto-conceptual structures
  - e.g. making explicit the kind of semantic relation linking semantically related terms which were clustered
    - e.g. *inquinamento* (pollution)/ *danno ambientale* (environmental damage): effect-cause relation

- Extension of the acquired domain-ontology with concept-linking relations
  - e.g. events (typically expressed by verbs) as connecting elements between concepts
    - e.g. L'*autorità amministrativa* competente **accerta** la *compatibilità paesaggistica* (The relevant administrative authority verifies the landscape compatibility)

To conclude

- McCarty, ICAIL 2007
  - "One of the main obstacles to progress in the field of artificial intelligence and law is the natural language barrier"
    - raw materials of the law are embodied in natural language (cases, statutes, regulations, etc.)
    - knowledge-based legal information systems need to access the content embedded in legal texts

Ontology Learning (NLP+ML)
is a feasible solution which can help to overcome 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)**

  - 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.
  - Walter, S. and M. Pinkal. Automatic extraction of definitions from german court decisions. In Proceedings of the International Conference on Computational Linguistics (COLING-2006) “Workshop on Information Extraction Beyond The Document”: 20–28, Sidney, 2006.

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.