

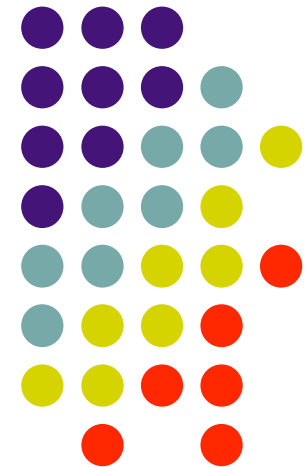
Acquiring Lexical and Ontological Information from Texts

Alessandro Lenci

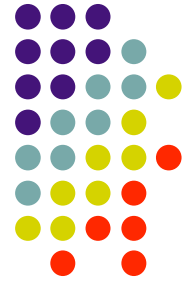
Dipartimento di Linguistica “T. Bolelli”
Università di Pisa



SIABO PhD School: Ontology & Lexicon
Copenhagen, December 1-4 2008

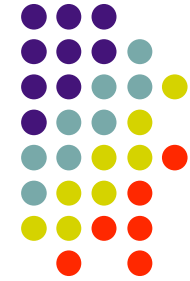


Outline



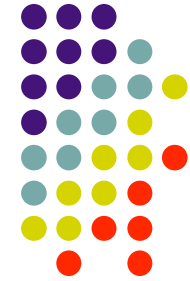
- Learning ontological and lexical information from texts
 - key issues and methods
- Case studies
 - association measures
 - extracting terms
 - Word Space Models

From language to knowledge



- The human expert (either the linguist or a domain expert) is supposed to be the most reliable knowledge source to build an ontology
- **Natural language** represents another not less crucial knowledge source for ontology building
 - documents - from Wikipedia to scientific papers and technical reports - are the primary repository of the knowledge of a certain community
- Texts can be **mined** to identify the knowledge items most relevant to feed the ontology creation process

From language to knowledge



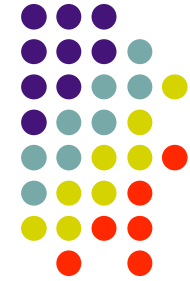
- The challenge is how to carve the formal structure of the conceptual system out of the **implicit ways** in which knowledge is expressed in natural language structures
- Ambiguity of (lexical and grammatical) linguistic structures
 - **N modifiers**
 - *senate member* *rice paper*
 - *book page* *summer vacation*
 - *football player* *cold virus*
 - **N V N**
 - *The man broke the vase.*
 - *The man saw the vase.*

Onto(lexical) learning



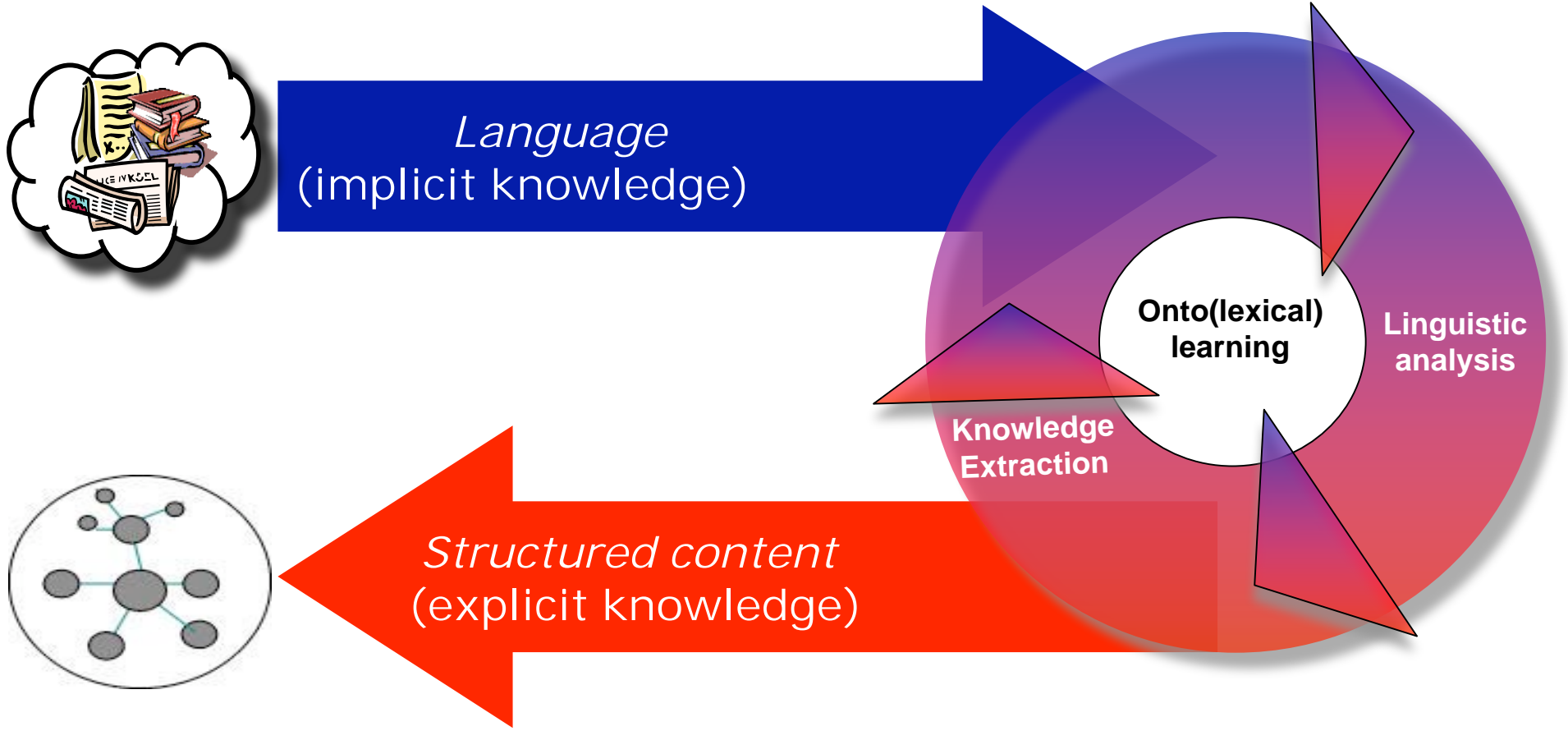
- **Ontology learning**
 - the use of NLP, machine learning, and AI-derived methods to acquire knowledge from texts in support to ontology development
- **Lexical acquisition**
 - (semi-)automatic extraction of lexical information from texts in support to the development of computational lexical resources
 - subcategorization frames, selectional preferences, semantic relations, etc.
- More a difference of emphasis and applications, rather than of methods
 - lexical acquisition is more oriented towards the text-driven acquisition of specific **linguistic properties** of lexical items
 - e.g. collocations, terms, subcategorization information, etc.

On the verge of a paradox

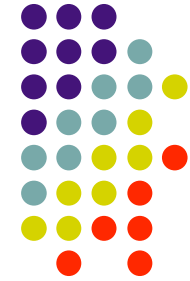


- Linguistic structures can be interpreted in virtue of the meaning of their lexical items
 - *book page vs. football player*
 - *the letter to John vs. the train to London*
- The circle of onto(lexical) knowledge extraction
 - the hypothesis is that **linguistic structures are conducive to the semantic content of the lexical items they contain**, but...
 - ...linguistic structures also depend on the semantic content of their components
- The challenge of onto(lexical) extraction from natural language documents is how to turn this potentially vicious circle into a **virtuous process**

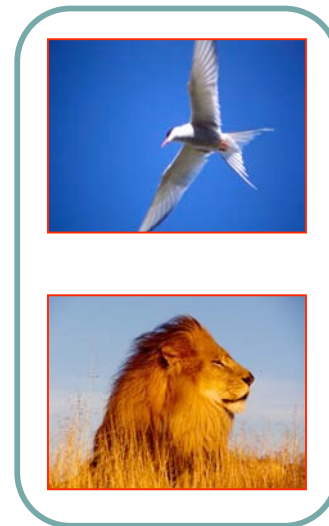
From language to knowledge



The sources of meaning






- Semantic knowledge is grounded on conceptual representations depending on **salient features of the world**
 - **similarity** and **dissimilarity** play a critical role in many cognitive tasks
 - categorization, object recognition, memory organization, etc.
 - perceived similarity between two entities depends on **common features (properties)**
 - **shape, dimension, function, parts, locations**, etc.



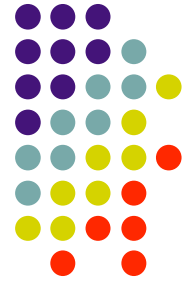
Semantic similarity



- **Semantic similarity** and **dissimilarity** deal with the meaning of words
 - *bird* → *lion*, *car*
 - *airplane* → *lion*, *car*
- *What are the sources of semantic similarity?*
 - semantic similarity may depend on **referential similarity**
 - *bird* is more similar to *lion* than to *airplane*, because  is more similar to  than to 
 - semantic similarity may depend on **word distributional similarity**
 - *bird* is more similar to *lion* than to *airplane*, because *bird* and *lion* are **used in similar (linguistic) contexts**

Meaning, use and context

Distributional evidence



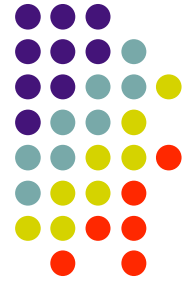
“What people know when they know a word is not how to recite its dictionary definition – they know how to **use** it (when to produce it and how to understand it) in everyday discourse (Miller, 1986). Knowing how to use words is a basic component of knowing a language, and how that component is acquired is a central question for linguists and cognitive psychologists alike. The search for an answer can begin with the cogent assumption that people learn how to use words by observing how words are used. And because words are used together in phrases and sentences, this starting assumption directs attention immediately to the importance of **context**”

Miller & Charles (1991: 4)

- Word meaning as **Contextual Representation**
 - encounters with a word in various contexts lead eventually to the construction of a **contextual representation of that word**
 - “An abstraction of information in the set of natural linguistic contexts in which a word occurs” [Charles 2000]

Distributional evidence

Statistical semantics



“**Statistical Semantics** is the study of “how the statistical patterns of human word usage can be used to figure out what people mean, at least to a level sufficient for information access” (Furnas 2006). How can we figure out what words mean, simply by looking at patterns of words in huge collections of text? What are the limits to this approach to understanding words?”

Wikipedia – “Statistical Semantics”

- **Statistical semantics**
 - *how can word usage be turned into semantic representations?*
 - what is the design of contextual representations?
 - how does the abstraction process work?
 - *to what extent distributional similarity can really model semantic knowledge?*

Contextual representations

some references



- **Philosophy of language**
 - “For a large class of cases - though not for all - in which we employ the word 'meaning' it can be defined thus: *the meaning of a word is its use in the language*”
(Ludwig Wittgenstein, *Philosophical Investigations*, 43)
- **Structural linguistics**
 - “If we consider words or morphemes A and B to be more different in meaning than A and C. then we will often find that the [contextual] distributions of A and B are more different than the [contextual] distributions of A and C. In other words, *difference in meaning correlates with difference of [contextual] distribution*”
(Zellig Harris, *Papers in Structural and Transformational Linguistics*, 1970)
- **Corpus linguistics**
 - “You shall know a word by *the company* it keeps”
(John R. Firth, *Selected Papers*, 1957)

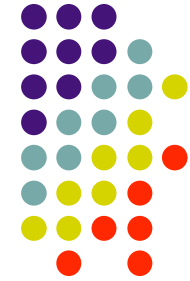
Meaning, use, and context



- Context
 - communication extralinguistic “setting”
 - entities, events, participants, etc.
 - linguistic context (cotext)
 - words and construction with which a word co-occurs
- In contextual semantic representations, context is approximated with co-text
 - we learn the meaning of many terms simply from language (often before having any experience with the corresponding entities)
 - abstract terms, scientific terms, ecc.
 - cf. *idiosyncrasy*, *apotropaic*, *love*, *jealousy*, *synchrotron*, etc.
 - the semantic content of a word corresponds to its use distribution in the various linguistic contexts
 - meanings (semantic type) are inductive, usage-based generalizations

Onto(lexical) learning

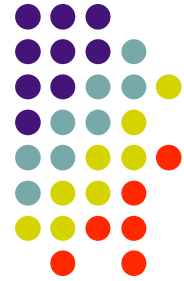
the core questions



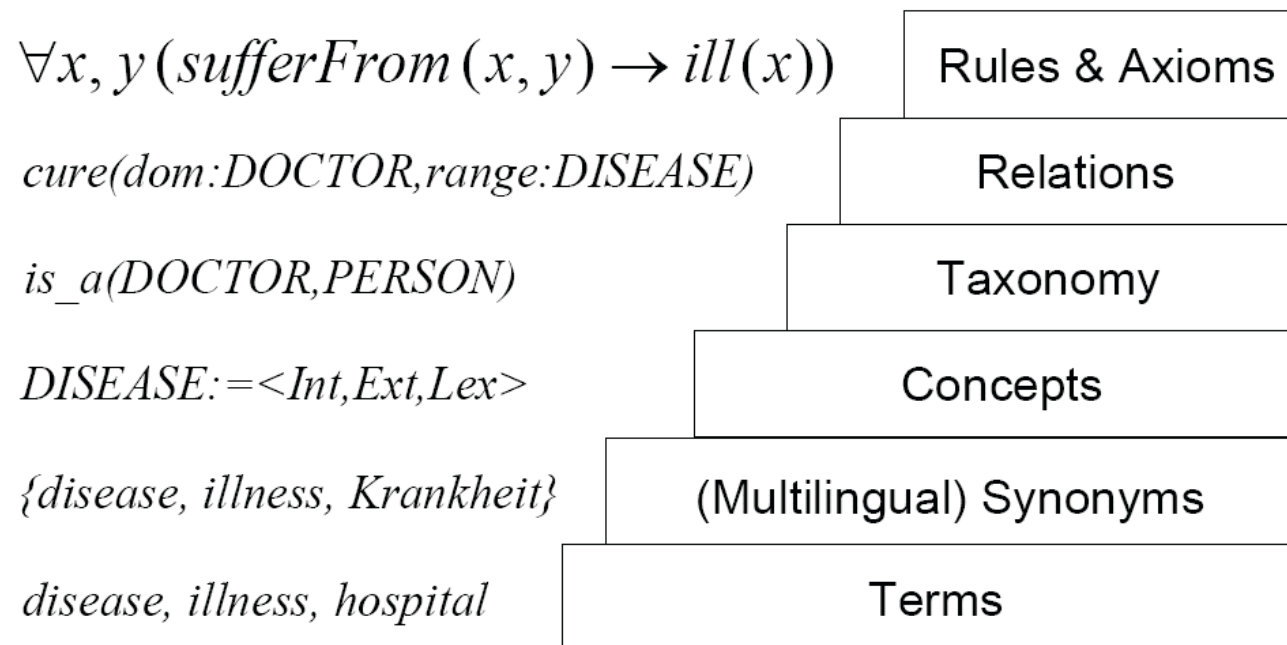
- *What?*
 - what pieces of semantic knowledge can we extract with computational linguistic methods?
- *Where?*
 - which language sources (i.e. text types) can be used for onto(lexical) knowledge extraction? How the source text type affects the target semantic information?
- *How?*
 - which methods can be applied, and how they relate to the type of knowledge we want to extract?
- *Why?*
 - what are the advantages of using NLP and learning methods for onto(lexical) building?

What?

the targets of ontology learning

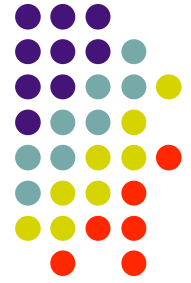


Ontology Learning Layer Cake



Where?

the sources for ontology learning



- **Fully-structured knowledge resources**
 - cf. thesauri, glossaries, web taxonomies, etc.
- **Semi-structured knowledge resources**
 - cf. machine readable dictionaries
 - “*idiosyncrasy* |₁ idēə¹ sɪŋ krəsē | noun (pl. **-sies**) (usu. **idiosyncrasies**) a mode of behavior or way of thought peculiar to an individual”
- **Text corpora**
 - currently the standard for onto(lexical) learning
 - domain corpora, general corpora (e.g. BNC), the Web

How?

methods for ontology learning



- Some mixture of NLP techniques and statistical (or machine learning) analysis
- NLP (e.g. PoS tagging, parsing, etc.) is used to define a **more abstract and linguistically grounded notion of context**
- Statistical analysis is used to select **significant word co-occurrence patterns**
 - i.e. not all contexts are equally informative

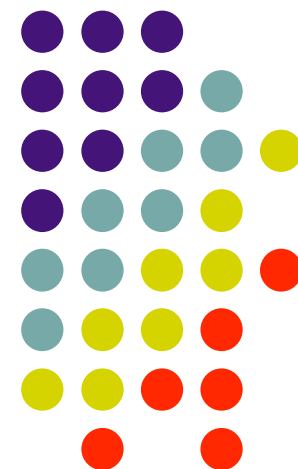
Onto(lexical) learning

some examples



- General methodological assumptions
 - syntagmatic properties of words can be used to investigate their paradigmatic properties
- Case studies
 - association measures
 - term extraction
 - word space models
 - semantic similarity

Association measures and term extraction



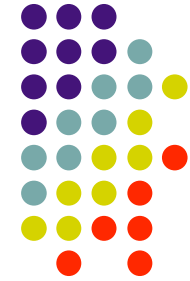
Word associations



- There are word sequences that have a strong degree of reciprocal association (cf. Manning & Schütze 1999, Evert 2007)
 - **collocations**
 - *strong tea, high season, best practice, break the rules,*
 - **support verb constructions**
 - *have a coffee, take a shower, etc.*
 - **terms**
 - *credit card, operating system, prime minister, etc.*
 - **phrasal verbs**
 - *break up, get off, etc.*
 - **idioms**
 - *eat the dust, bite the bullet, kick the bucket*
- Some common properties
 - **high degree of conventionalization**
 - **reduced compositionality**
 - **high structural rigidity**

Association measures

www.collocations.de



- Co-occurrence computation
 - **token bigrams** = pairs of words
 - “the dog chases the dog”
 - 5 tokens
 - <the dog>, <dog chases>, <chases the>, <the dog>
 - 4 bigrammi tokens
 - **type bigrams** = types of word pairs
 - “the dog chases the dog”
 - 3 word types
 - <the dog>, <dog chases>, <chases the>
 - 3 type bigrams
- **Contingency table** to classy word co-occurrences

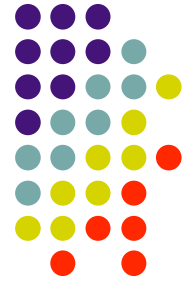
	$V = v$	$V \neq v$
$U = u$	O_{11}	O_{12}
$U \neq u$	O_{21}	O_{22}

$O_{11} + O_{12} + O_{21} + O_{22} = N$

contingency table for the
bigram <u v>

Mutual Information (MI)

(Church & Hanks 1990)



- Mutual Information

$$MI(w_1, w_2) \equiv \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

- It compares the probability to observe the bigram $\langle w_1, w_2 \rangle$ with the probability of observing w_1 and w_2 independently one from the other
 - $MI(w_1, w_2) \approx 0 \rightarrow$ weak association between the words
 - $MI(w_1, w_2) \gg 0 \rightarrow$ significant associations between the words

How to compute MI



Probabilities can be estimated with the **relative frequency** in a corpus C

$$\log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} = \log_2 \frac{\frac{f(w_1, w_2)}{|C|}}{\frac{f(w_1)}{|C|} \cdot \frac{f(w_2)}{|C|}}$$

bigram relative frequency

With some transformations, we get:

$$\log_2 \frac{\frac{f(w_1, w_2)}{|C|}}{\frac{f(w_1)}{|C|} \cdot \frac{f(w_2)}{|C|}} = \log_2 \frac{f(w_1, w_2)}{|C|} \cdot \frac{|C|^2}{f(w_1)f(w_2)} = \log_2 \frac{f(w_1, w_2) \cdot |C|}{f(w_1)f(w_2)}$$



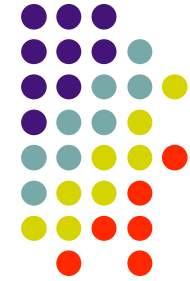
How to compute MI

- Four quantities are necessary to compute the MI of the bigram *honorary doctor*.
 - **bigram frequency**
 - $f(\textit{honorary}, \textit{doctor}) = 12$
 - **frequency of w_1**
 - $f(\textit{honorary}) = 111$
 - **frequency of w_2**
 - $f(\textit{doctor}) = 621$
 - **corpus length**
 - $|C| = 15$ millions tokens

Church & Hanks (1990)

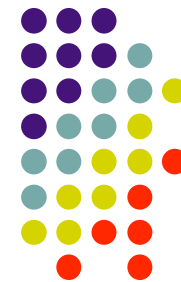
$I(x, y)$	$f(x, y)$	$f(x)$	x	$f(y)$	y
11.3	12	111	honorary	621	doctor
11.3	8	1105	doctors	44	dentists
10.7	30	1105	doctors	241	nurses
9.4	8	1105	doctors	154	treating
9.0	6	275	examined	621	doctor
8.9	11	1105	doctors	317	treat
8.7	25	621	doctor	1407	bills
8.7	6	621	doctor	350	visits
8.6	19	1105	doctors	676	hospitals
8.4	6	241	nurses	1105	doctors

Association measures



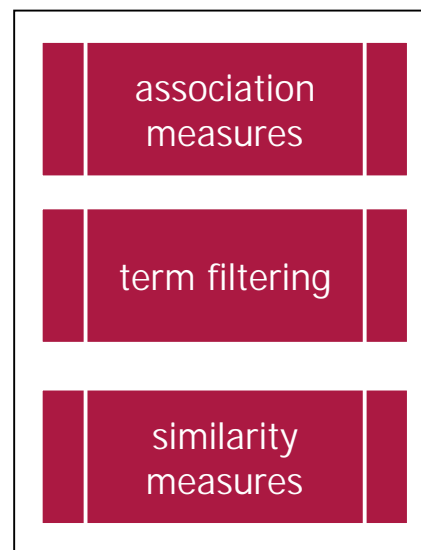
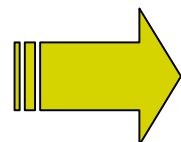
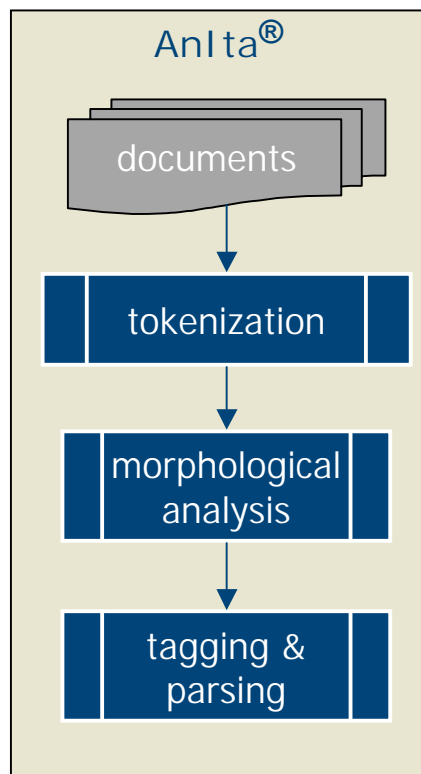
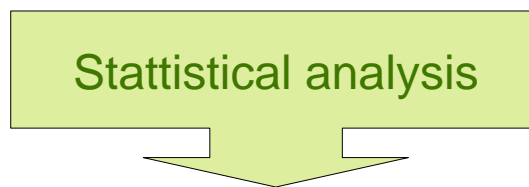
- The absolute frequency of a bigram is not enough to quantify the association strength
- Besides considering the number of times we see two words together, we have also to consider **the number of times we observe them independently one from the other**
 - $f(a, doctor) = 41\dots$ but $f(a) = 284690$
 - *a doctor* is not a strongly associated bigram because the number of times in which *a* occurs with a word different from *doctor* is much higher than the frequency of the bigram *a doctor*

Text-2-Knowledge (T2K[®])

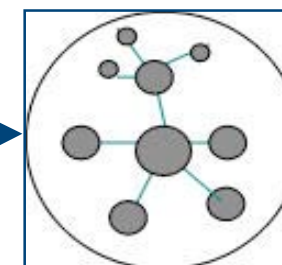


- Automatic extration of domain terminology for Italian
 - Hybrid architecture
 - NLP modules + statistical filtering
 - Department of Linguistics, Univ. Pisa & ILC-CNR
 - Goals
 - term extraction
 - thesaurus creation
 - semantically related terms
 - Domains of applications
 - law and administrative texts
 - medical reports

The architecture of T2K[®]



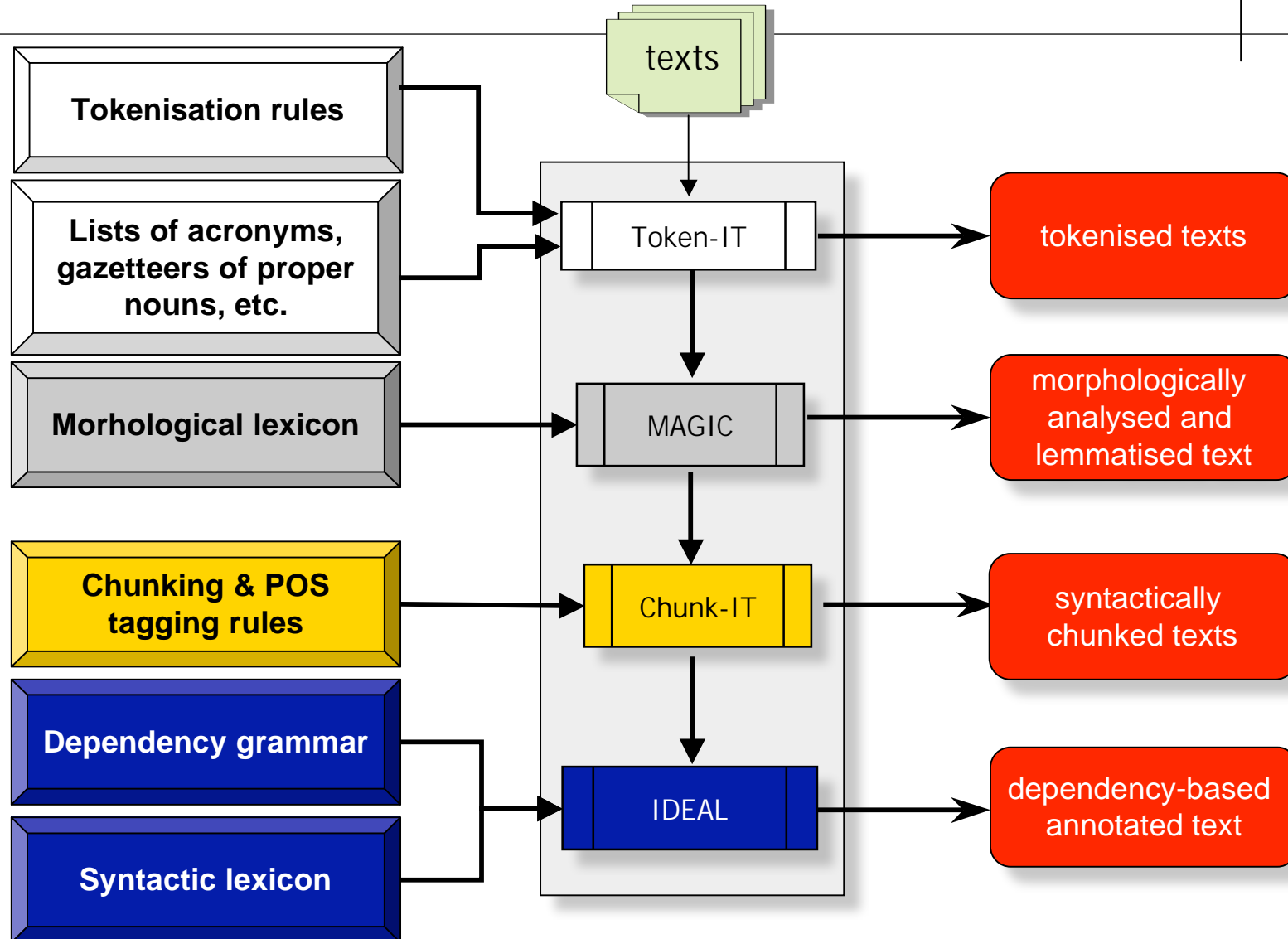
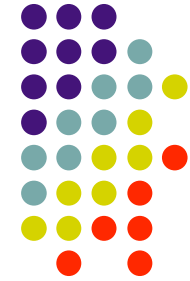
term-bank & thesaurus



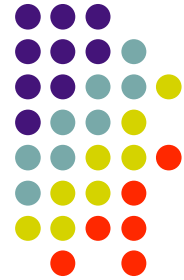
```
Oggi a <word sem="location"> Bruxelles </word> vertice della  
<word sem="institution"> Ue </word> sull' <word  
sem="location"> IRAQ </word>. <word  
sem="person"> Blair </word> <word lemma="scrivere"  
pos="V"> scrive </word> agli altri per chiedere una posizione  
<word lemma="comune" pos="A"> comune </word>  
La <word sem="institution"> NATO </word> in <word  
lemma="difesa" pos="S"> difesa </word> della <word  
sem="location"> Turchia </word>
```

indexed texts

Anlta: a suite of NLP tools



Anlta: a suite of NLP tools



tokenization

L'
obiettivo
centrale
della
Linguistica
Computazionale

lemmatization and
morphological
analysis

LO#RD@FS@MS# L'#PQ@FS3@MS3# L'#SP@NN#
OBIETTIVO#A@MS# OBIETTIVO#S@MS# OBIETTIVARE#V@S1IP#
CENTRARE#V@S2MP<+LE#PQ@FN3#># CENTRALE#A@FS@MS# CENTRALE#S@FS#
DI#E@FS#
LINGUISTICO#A@FS# LINGUISTICA#S@FS# LINGUISTICA#SP@NN#
COMPUTAZIONALE#A@FS@MS# COMPUTAZIONALE#SP@NN#

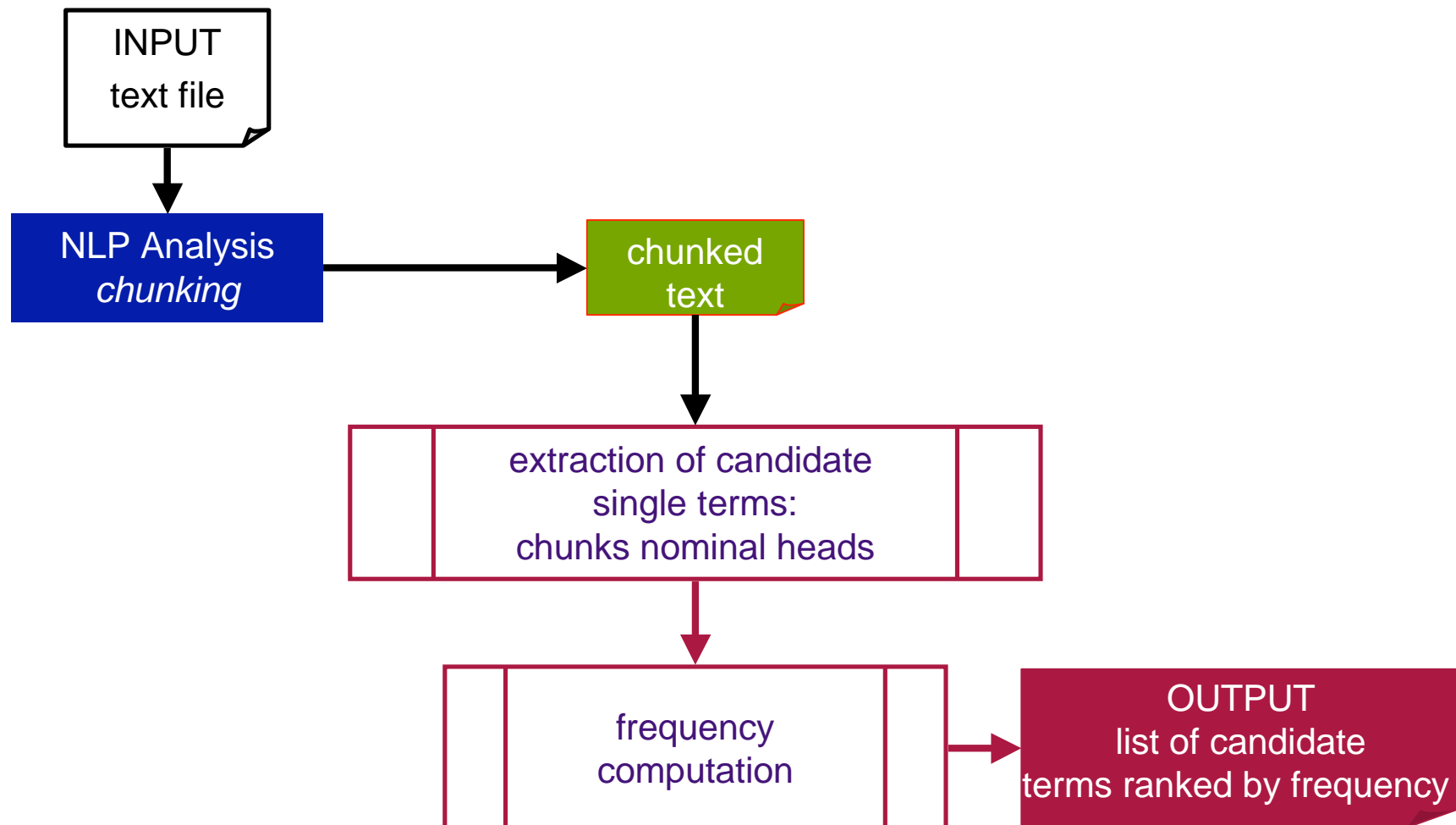
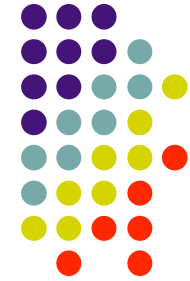
chunking

[[CC: N_C] [DET: LO#RD@MS] [AGR: @MS] [POTGOV: OBIETTIVO#S@MS]]
[[CC: ADJ_C] [AGR: @FS@MS-@FS] [POTGOV: CENTRALE#A@FS@MS]]
[[CC: di_C] [DET: LO#RD@FS] [AGR: @FS] [POTGOV: LINGUISTICA#S@FS]]
[[CC: ADJ_C] [AGR: @FS@MS] [POTGOV: COMPUTAZIONALE#A@FS@MS]]

dependency parsing

MODIF(OBIETTIVO, CENTRALE<role = RESTR>)
ARG(OBIETTIVO, LINGUISTICA<intro = DI>)
MODIF(LINGUISTICA, COMPUTAZIONALE<role = RESTR>)

Single term extraction in T2K

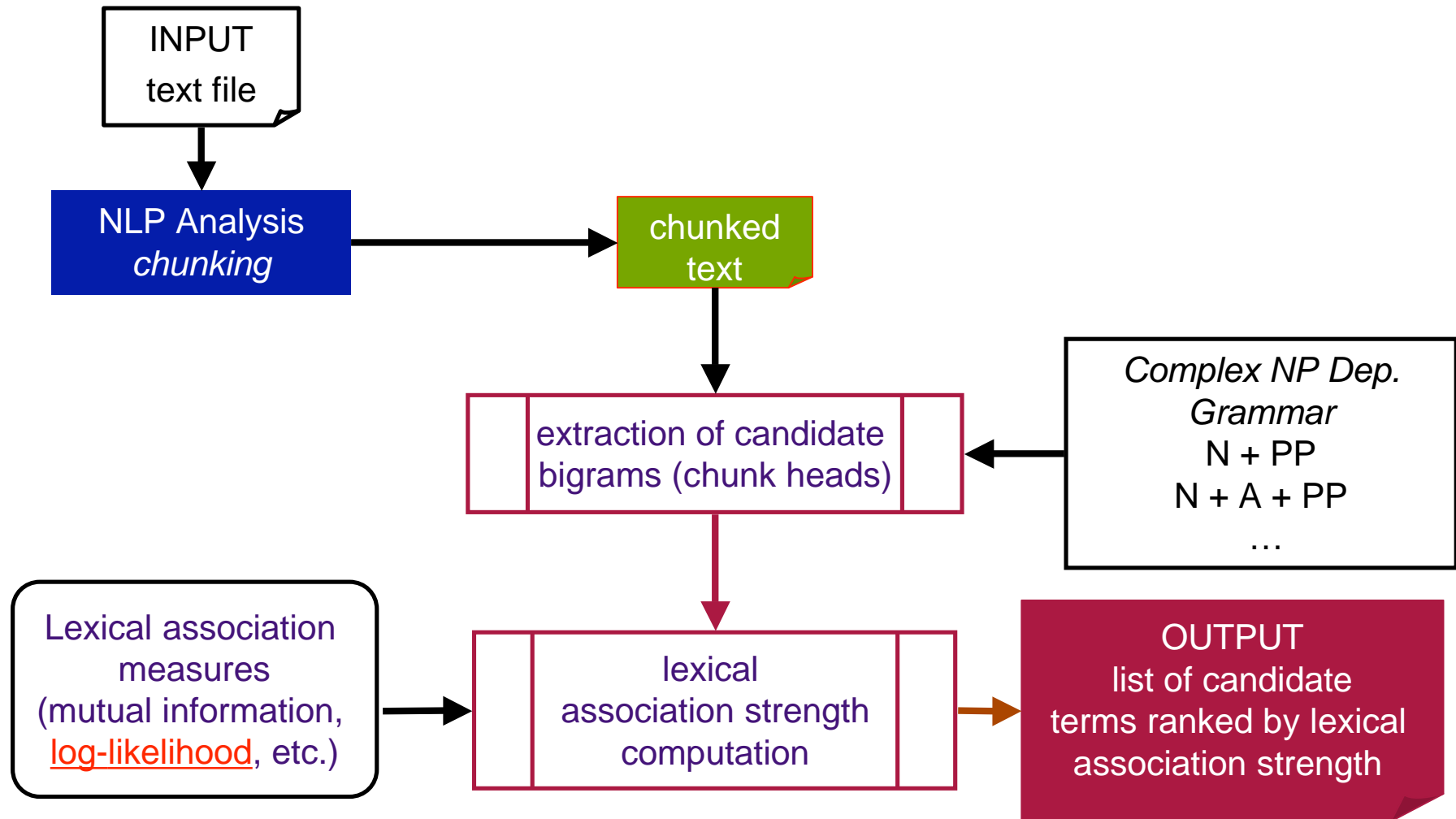
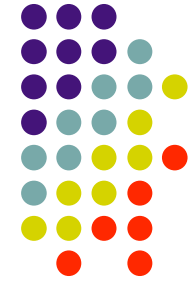


La **Commissione** e gli **Stati** membri collaborano e creano **sinergie** con altri **organismi** internazionali o paneuropei, nel **contesto** degli **obiettivi** di cui all'**articolo 1**, per promuovere la **conservazione** e la **protezione** delle **foreste** ai fini dello **sviluppo** sostenibile.



```
[ [ CC: N_C ] [ DET: LO#RD@FS ] [ AGR: @FS ] [ POTGOV: COMMISSIONE#S@FS ] ]
[ [ CC: COORD_C ] [ CONJTYPE: E#CC ] ]
[ [ CC: N_C ] [ DET: LO#RD@MP ] [ AGR: @MP@MP ] [ POTGOV: STATO#S@MP ] ]
[ [ CC: ADJ_C ] [ AGR: @MP ] [ POTGOV: MEMBRO#A@MP ] ]
[ [ CC: FV_C ] [ AGR: @P3 ] [ POTGOV: COLLABORARE#V@P3IP ] ]
[ [ CC: COORD_C ] [ CONJTYPE: E#CC ] ]
[ [ CC: FV_C ] [ AGR: @P3 ] [ POTGOV: CREARE#V@P3IP ] ]
[ [ CC: N_C ] [ AGR: @FP ] [ POTGOV: SINERGIA#S@FP ] ]
[ [ CC: P_C ] [ PREP: CON#E ] [ AGR: @MP ] [ PREMODIF: ALTRO#A@MP ] [ POTGOV: ORGANISMO#S@MP ] ]
[ [ CC: NA_C ] [ AGR: @FP@MP-@FP ] [ POTGOV: INTERNAZIONALE#A@FP@MP INTERNAZIONALE#S@FP ] ]
[ [ CC: COORD_C ] [ CONJTYPE: O#CC ] ]
[ [ CC: ADJ_C ] [ AGR: @MP ] [ POTGOV: PANEUROPEO#A@MP ] ]
[ [ CC: PUNC_C ] [ PUNCTYPE: ,#@ ] ]
[ [ CC: P_C ] [ PREP: IN#E ] [ DET: IL#RD@MS ] [ AGR: @MS ] [ POTGOV: CONTESTO#S@MS ] ]
[ [ CC: di_C ] [ DET: LO#RD@MP ] [ AGR: @MP ] [ POTGOV: OBIETTIVO#S@MP ] ]
[ [ CC: P_C ] [ PREP: DI#E ] [ AGR: @FP@FS@MP@MS ] [ POTGOV: CUI#P@FP@FS@MP@MS ] ]
[ [ CC: P_C ] [ PREP: A#E ] [ DET: LO#RD@MS ] [ AGR: @MS ] [ POTGOV: ARTICOLO#S@MS ] ]
[ [ CC: ADJ_C ] [ POTGOV: 1#N ] ]
[ [ CC: PUNC_C ] [ PUNCTYPE: ,#@ ] ]
[ [ CC: I_C ] [ PREP: PER#E ] [ AGR: @F ] [ POTGOV: PROMUOVERE#V@F ] ]
[ [ CC: N_C ] [ DET: LO#RD@FS ] [ AGR: @FS ] [ POTGOV: CONSERVAZIONE#S@FS ] ]
[ [ CC: COORD_C ] [ CONJTYPE: E#CC ] ]
[ [ CC: N_C ] [ DET: LO#RD@FS ] [ AGR: @FS ] [ POTGOV: PROTEZIONE#S@FS ] ]
[ [ CC: di_C ] [ DET: LO#RD@FP ] [ AGR: @FP ] [ POTGOV: FORESTA#S@FP ] ]
[ [ CC: P_C ] [ PREP: A#E ] [ DET: IL#RD@MP ] [ AGR: @MP ] [ POTGOV: FINE#S@MP ] ]
[ [ CC: di_C ] [ DET: LO#RD@MS ] [ AGR: @MS ] [ POTGOV: SVILUPPO#S@MS ] ]
[ [ CC: ADJ_C ] [ AGR: @FS@MS ] [ POTGOV: SOSTENIBILE#A@FS@MS ] ]
```

Multi-word term extraction in T2K

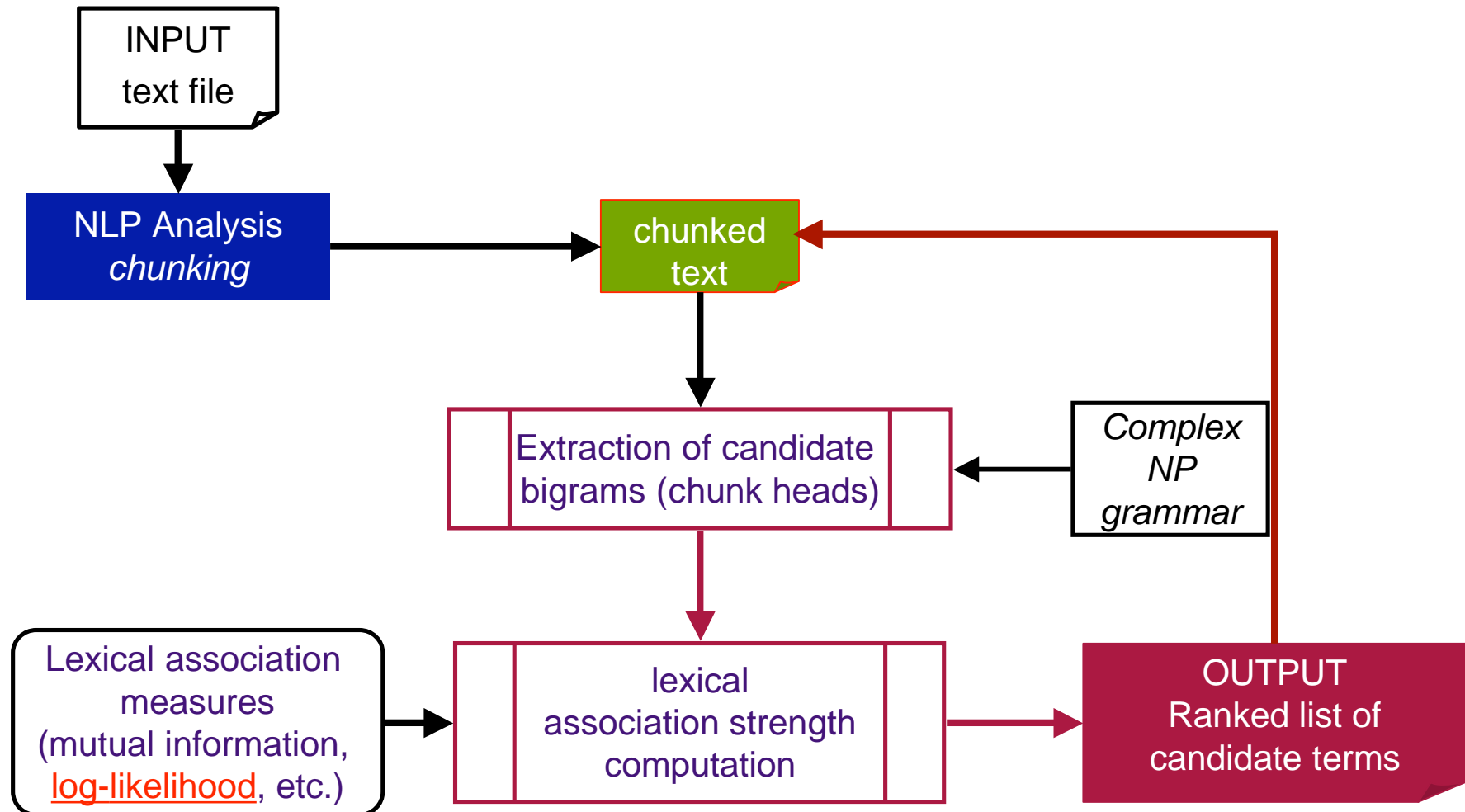


La Commissione e gli **Stati membri** collaborano e creano sinergie con altri **organismi internazionali** o paneuropei, nel **contesto degli obiettivi** di cui all'articolo 1, per promuovere la conservazione e la **protezione delle foreste** ai fini dello **sviluppo sostenibile**.



```
[ [ CC: N_C] [ DET: LO#RD@FS] [ AGR: @FS] [ POTGOV: COMMISSIONE#S@FS]]
[ [ CC: COORD_C] [ CONJTYPE: E#CC]]
[ [ CC: N_C] [ DET: LO#RD@MP] [ AGR: @MP@MP] [ POTGOV: STATO#S@MP]]
[ [ CC: ADJ_C] [ AGR: @MP] [ POTGOV: MEMBRO#A@MP]]
[ [ CC: FV_C] [ AGR: @P3] [ POTGOV: COLLABORARE#V@P3IP]]
[ [ CC: COORD_C] [ CONJTYPE: E#CC]]
[ [ CC: FV_C] [ AGR: @P3] [ POTGOV: CREARE#V@P3IP]]
[ [ CC: N_C] [ AGR: @FP] [ POTGOV: SINERGIA#S@FP]]
[ [ CC: P_C] [ PREP: CON#E] [ AGR: @MP] [ POTGOV: ORGANISMO#S@MP]]
[ [ CC: NA_C] [ AGR: @FP@MP-@FP] [ POTGOV: INTERNAZIONALE#A@FP@MP INTERNAZIONALE#S@FP]]
[ [ CC: COORD_C] [ CONJTYPE: O#CC]]
[ [ CC: ADJ_C] [ AGR: @MP] [ POTGOV: PANEUROPEO#A@MP]]
[ [ CC: PUNC_C] [ PUNCTYPE: ,#@]]
[ [ CC: P_C] [ PREP: IN#E] [ DET: IL#RD@MS] [ AGR: @MS] [ POTGOV: CONTESTO#S@MS]]
[ [ CC: di_C] [ DET: LO#RD@MP] [ AGR: @MP] [ POTGOV: OBIETTIVO#S@MP]]
[ [ CC: P_C] [ PREP: DI#E] [ AGR: @FP@FS@MP@MS] [ POTGOV: CUI#P@FP@FS@MP@MS]]
[ [ CC: P_C] [ PREP: A#E] [ DET: LO#RD@MS] [ AGR: @MS] [ POTGOV: ARTICOLO#S@MS]]
[ [ CC: ADJ_C] [ POTGOV: 1#N]]
[ [ CC: PUNC_C] [ PUNCTYPE: ,#@]]
[ [ CC: I_C] [ PREP: PER#E] [ AGR: @F] [ POTGOV: PROMUOVERE#V@F]]
[ [ CC: N_C] [ DET: LO#RD@FS] [ AGR: @FS] [ POTGOV: CONSERVAZIONE#S@FS]]
[ [ CC: COORD_C] [ CONJTYPE: E#CC]]
[ [ CC: N_C] [ DET: LO#RD@FS] [ AGR: @FS] [ POTGOV: PROTEZIONE#S@FS]]
[ [ CC: di_C] [ DET: LO#RD@FP] [ AGR: @FP] [ POTGOV: FORESTA#S@FP]]
[ [ CC: P_C] [ PREP: A#E] [ DET: IL#RD@MP] [ AGR: @MP] [ POTGOV: FINE#S@MP]]
[ [ CC: di_C] [ DET: LO#RD@MS] [ AGR: @MS] [ POTGOV: SVILUPPO#S@MS]]
[ [ CC: ADJ_C] [ AGR: @FS@MS] [ POTGOV: SOSTENIBILE#A@FS@MS]]
```

Multi-word term extraction in T2K *incremental approach*



Il presente decreto stabilisce le misure e le procedure finalizzate a prevenire e ridurre per quanto possibile gli effetti negativi dell'**incenerimento dei rifiuti pericolosi** sull'ambiente, in particolare l'inquinamento atmosferico, del suolo, delle acque superficiali e sotterranee, nonche' i rischi per la salute umana che ne risultino, in attuazione della direttiva 94/ 67/CE ed ai sensi dell'articolo 3, comma 2, del decreto del Presidente della Repubblica 24 maggio 1988, n. 203 e dell'articolo 18, comma 2, lettera a), del decreto legislativo 5 febbraio 1997, n. 22, come modificato ed integrato dal decreto legislativo 8 novembre 1997, n. 389 e dalla legge 9 dicembre 1998, n. 426.



```
[ [ CC: N_C] [ DET: IL#RD@MS] [ AGR: @MS] [ PREMODIF: PRESENTE#A@MS] [ POTGOV:
DECRETO#S@MS]]
[ [ CC: FV_C] [ AGR: @S3] [ POTGOV: STABILIRE#V@S3IP]]
[ [ CC: N_C] [ DET: LO#RD@FP] [ AGR: @FP] [ POTGOV: MISURA#S@FP]]
[ [ CC: COORD_C] [ CONJTYPE: E#CC]]
[ [ CC: N_C] [ DET: LO#RD@FP] [ AGR: @FP] [ POTGOV: PROCEDURA#S@FP]]
[ [ CC: ADJPART_C] [ AGR: @FP-@FP] [ POTGOV: FINALIZZARE#V@FPPR FINALIZZATO#A@FP]]
[ [ CC: I_C] [ PREP: A#E] [ AGR: @F] [ POTGOV: PREVENIRE#V@F]]
[ [ CC: COORD_C] [ CONJTYPE: E#CC]]
[ [ CC: I_C] [ AGR: @F] [ POTGOV: RIDURRE#V@F]]
[ [ CC: P_C] [ PREP: PER#E] [ DET: QUANTO#D@MS] [ AGR: @MS] [ POTGOV: POSSIBILE#S@MS]]
[ [ CC: N_C] [ DET: LO#RD@MP] [ AGR: @MP] [ POTGOV: EFFETTO#S@MP]]
[ [ CC: NA_C] [ AGR: @MP-@MP] [ POTGOV: NEGATIVO#A@MP NEGATIVO#S@MP]]
[ [ CC: di_C] [ DET: LO#RD@MS] [ AGR: @MS] [ POTGOV: INCENERIMENTO#S@MS]]
[ [ CC: di_C] [ DET: IL#RD@MP] [ AGR: @MP] [ POTGOV: RIFIUTO_PERICOLOSO#S@MP]]
[ [ CC: P_C] [ PREP: SU#E] [ DET: LO#RD@MS] [ AGR: @MS] [ POTGOV: AMBIENTE#S@MS]]
[ [ CC: PUNC_C] [ PUNCTYPE: ,#@]]
[ [ CC: P_C] [ PREP: IN#E] [ AGR: @MS] [ POTGOV: PARTICOLARE#S@MS]]
[ [ CC: N_C] [ DET: LO#RD@MS] [ AGR: @MS] [ POTGOV: INQUINAMENTO#S@MS]]
[ [ CC: ADJ_C] [ AGR: @MS] [ POTGOV: ATMOSFERICO#A@MS]]
[ [ CC: PUNC_C] [ PUNCTYPE: ,#@]]
[ [ CC: di_C] [ DET: IL#RD@MS] [ AGR: @MS] [ POTGOV: SUOLO#S@MS]]
[ [ CC: PUNC_C] [ PUNCTYPE: ,#@]]
```

The legal-environmental TermBank



KWID	Term	Freq	Lemmatised headword
67	DECISIONE	1344	DECISIONE
674	DECISIONE DEL CONSIGLIO	76	DECISIONE CONSIGLIO
3193	DECISIONE DELL' AUTORITA' COMPETENTE	12	DECISIONE AUTORITA' COMPETENTE
489	DECISIONE DELLA COMMISSIONE	148	DECISIONE COMMISSIONE
3870	DECISIONE DELLA COMMISSIONE CONCERNENTE	10	DECISIONE COMMISSIONE CONCERNENTE
4772	DECISIONE DELLA COMMISSIONE RELATIVA	8	DECISIONE COMMISSIONE RELATIVO
2256	DECISIONE DEL PARLAMENTO	17	DECISIONE PARLAMENTO
8801	DECISIONE QUADRO	4	DECISIONE QUADRO
4602	DECORRENZA DELL' OBBLIGO	8	DECORRENZA OBBLIGO
2	DECRETO	7533	DECRETO
3335	DECRETO DEL DIRETTORE	11	DECRETO DIRETTORE
637	DECRETO DEL MINISTERO	83	DECRETO MINISTERO
145	DECRETO DEL MINISTRO	706	DECRETO MINISTRO
93	DECRETO DEL PRESIDENTE	1043	DECRETO PRESIDENTE
4468	DECRETO DIRETTORIALE	8	DECRETO DIRETTORIALE
1054	DECRETO INTERMINISTERIALE	41	DECRETO INTERMINISTERIALE
279	DECRETO-LEGGE	377	DECRETO-LEGGE
7932	DECRETO-LEGGE COORDINATO	4	DECRETO-LEGGE COORDINATO
20	DECRETO LEGISLATIVO	2632	DECRETO LEGISLATIVO
1560	DECRETO MEDESIMO	26	DECRETO MEDESIMO
197	DECRETO MINISTERIALE	526	DECRETO MINISTERIALE
4704	DECRETO PER GLI IMPIANTI	8	DECRETO IMPIANTO
6339	DECRETO PREVISTO	6	DECRETO PREVISTO
3271	ECOSISTEMI ACQUATICI	12	ECOSISTEMA ACQUATICO
4077	ECOSISTEMI FORESTALI	9	ECOSISTEMA FORESTALE
10758	ECOSISTEMI MARINI	3	ECOSISTEMA MARINO
8459	ECOTOSSICITÀ DI SOSTANZE	4	ECOTOSSICITA' SOSTANZA
9340	EDIFICI DI INTERESSE CULTURALE	3	EDIFICIO INTERESSE CULTURALE
7652	EDIFICI DI NUOVA COSTRUZIONE	4	EDIFICIO COSTRUZIONE
8629	EDIFICI ESISTENTI	4	EDIFICIO ESISTENTE

Fragments of taxonomical chains



APPLICAZIONE

- APPLICAZIONE DEI PARAGRAFI
- APPLICAZIONE DELL' ARTICOLO
- APPLICAZIONE DELLA DIRETTIVA
- APPLICAZIONE DELLA LEGGE
- APPLICAZIONE DELLA TARIFFA
- APPLICAZIONE DELLE DISPOSIZIONI
- APPLICAZIONE DELLE SANZIONI
 - APPLICAZIONE DELLE SANZIONI AMMINISTRATIVE
 - APPLICAZIONE DELLE SANZIONI PREVISTE
- APPLICAZIONE DEL PRESENTE DECRETO
- APPLICAZIONE DEL REGOLAMENTO

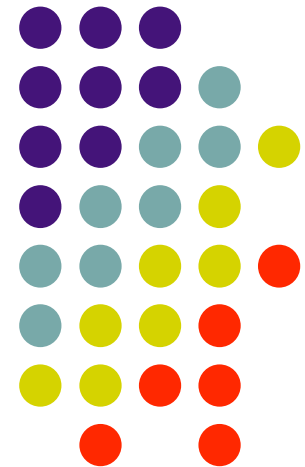
RIDUZIONE

- RIDUZIONE DEI CONSUMI
- RIDUZIONE DELL' INQUINAMENTO
 - RIDUZIONE DELL' INQUINAMENTO ACUSTICO
- RIDUZIONE DELLA PRODUZIONE
- RIDUZIONE DELLE EMISSIONI
 - RIDUZIONE DELLE EMISSIONI INQUINANTI
- RIDUZIONE DEL LIVELLO
- RIDUZIONE DELLO STANZIAMENTO
- RIDUZIONE DEL TENORE

TUTELA

- TUTELA AMBIENTALE
- TUTELA DEI CONSUMATORI
- TUTELA DELL' AMBIENTE
- TUTELA DELL' OZONO STRATOSFERICO
- TUTELA DELLA QUALITÀ
- TUTELA DELLA SALUTE
 - TUTELA DELLA SALUTE PUBBLICA
- TUTELA DELLE ACQUE
 - TUTELA DELLE ACQUE INTERNE
- TUTELA DEL PAESAGGIO
- TUTELA DEL TERRITORIO
 - TUTELA DEL TERRITORIO RURALE

Word space models and semantic similarity



Word Space Models



Distributional Hypothesis

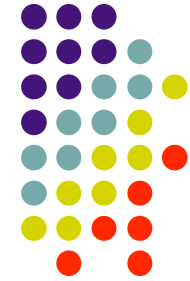
(Harris 1968; Miller & Charles 1991)

Words that tend to occur in similar contexts tend to have similar meanings

Semantic similarity can be defined as distributional similarity

- **Word Space Models** (WSMs) induce distributed semantic representations from textual input
- The meaning of a word is represented as a **vector** recording its co-occurrence with other words

Word Space Models



- **Models of the mental lexicon**
 - lexical priming (Lund *et al.*, 1995), analogical reasoning (Ramscar and Yarlett, 2003), semantic similarity judgements (MacDonald & Ramscar, 2001), semantic deficits (Vigliocco, *et al.* 2004), etc.
- **Natural Language Processing (NLP)**
 - word sense disambiguation, Information Retrieval, ontology and thesaurus learning (Lin 1998, Maedche & Staab 2004, Widdows 2003), etc.

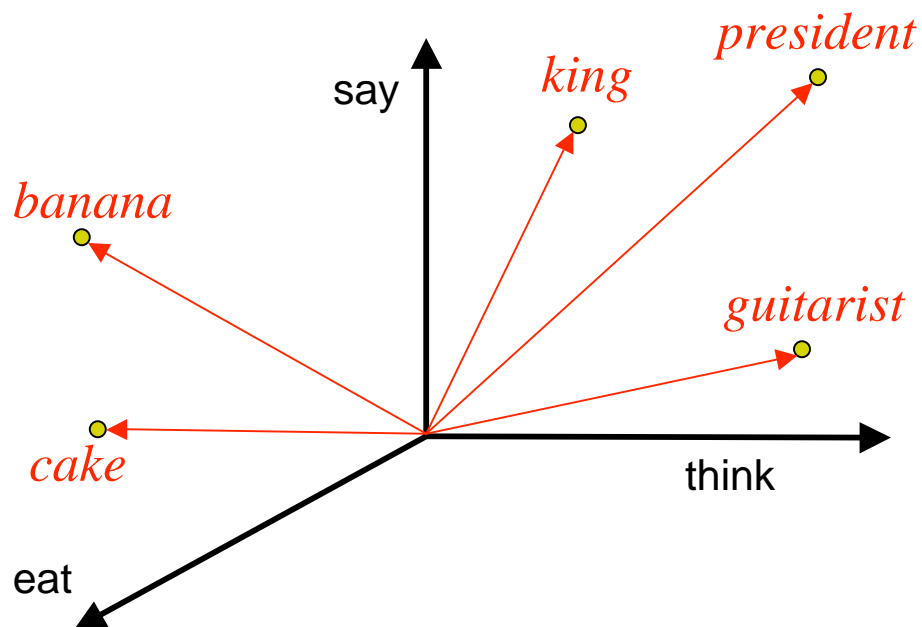
Word space models

- **Latent Semantic Analysis (LSA)** (Landauer & Dumais 1997)
- **Hyperspace Analogue to Language (HAL)** (Burgess & Lund 1997)
- **Random Indexing (RI)** (Karlsgren & Sahlgren 2001)
- **Incremental Semantic Analysis (ISA)** (Baroni, Lenci & Onnis 2007)
- **Tupleware** (Baroni & Lenci in preparation)

Semantic word spaces



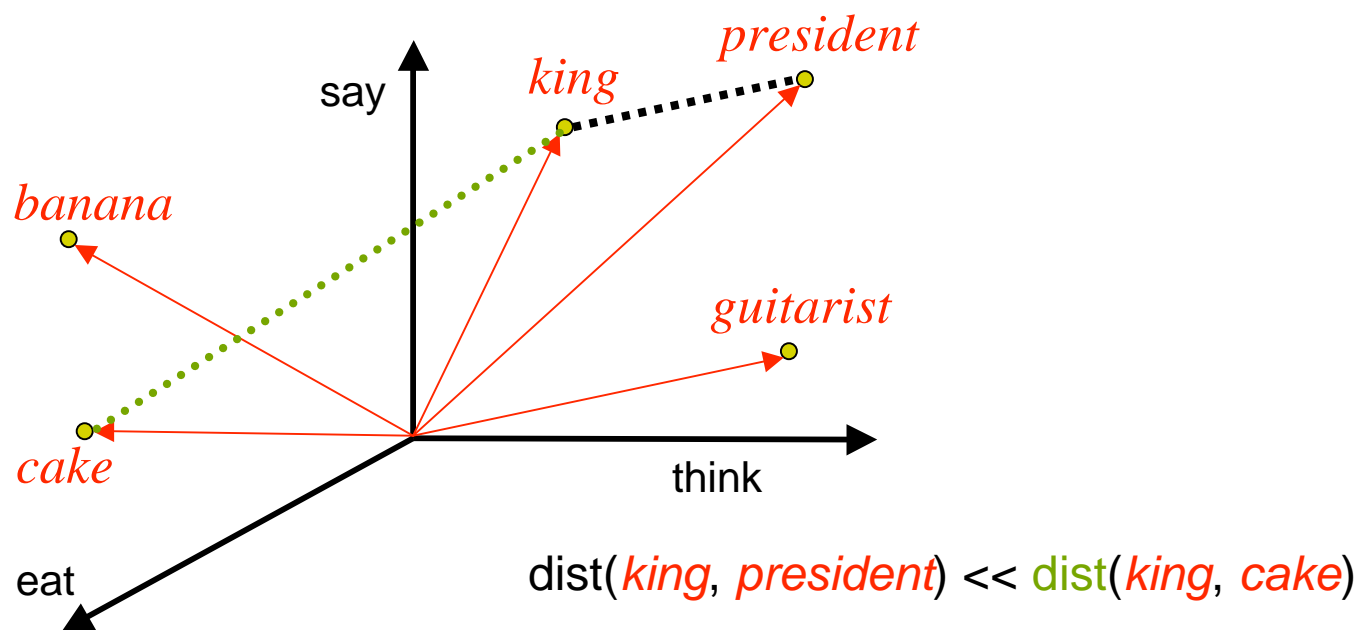
- Each word w is represented as a point in a **distributional space**
 - **linguistic contexts** provide the space dimensions
 - the coordinates of w are determined by the **statistical distribution** of w in the linguistic contexts





Semantic similarity

- **Distributional (semantic) similarity** between two words is measured by their distance in space



Word Space Models

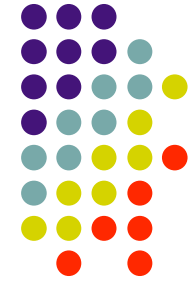
Lowe (2001), Padó & Lapata (2007)



- WSMs are formally defined by the quadruple $\langle T, B, M, S \rangle$
 - **T** is the set of “target” items
 - the elements to which the space provides a semantic representation
 - **B** is the **basis** that defines the space dimensions
 - **linguistic contexts** used to compute the distributional similarity
 - **M** is a co-occurrence matrix
 - cells contain some function of the **co-occurrence frequency** of targets with the basis contexts
 - **S** is some **measure of the distance** between points in space
 - e.g. cosine, euclidean distance, etc. (cf. Mohammad & Hirst 2005 for a survey)

Word Space Models

the co-occurrence matrix

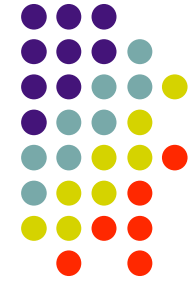


- **Rows** correspond to the target words
- **Columns** correspond to the contexts defined by the basis B
 - typically contexts are represented by a window of n adjacent words
- **Cell values** correspond to (some function of) the **co-occurrence frequency** of a word in a certain context
 - e.g. log-frequency, entropy, association measure (MI, log-likelihood, etc.)

	say	eat	open	think	country	sweet	...
king	6	2	5	4	1	0	
president	10	3	2	3	7	0	
cake	0	4	2	0	0	3	
banana	0	7	0	0	0	1	
...							

Word Space Models

defining the basis



- Models differ for the type of context defining the basis
- Typically, vectors record word co-occurrences within a **sliding context window**
 - each vector dimension d_{ij} records the number of times w_i occurs within a window of n words before and after w_j , where n is an empirically fixed parameter
 - the number of dimensions is a subset of the word types in the corpus
 - minus high frequency stop words and very low frequency items
- Other definitions of contexts are possible
 - syntactic dependencies (Padó & Lapata 2007), paragraphs, documents (Landauer & Dumais 1997)

Word Space Models

measuring the distance in space



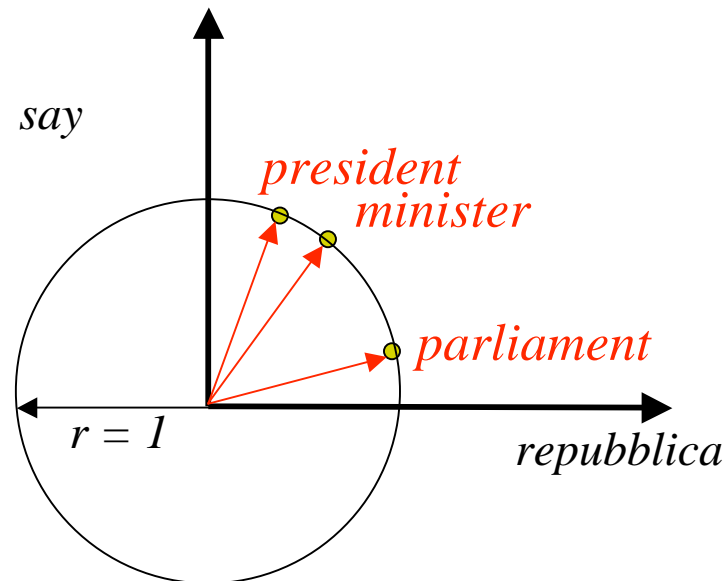
- **Vector normalization**

- each vector w is **normalized**, so that its **length** (norm) is 1

- **vector length:**

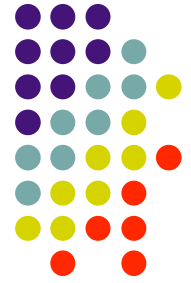
$$|\vec{x}| = \sqrt{\sum_{i=1}^n x_i^2}$$

- each dimension is divided by the vector norm



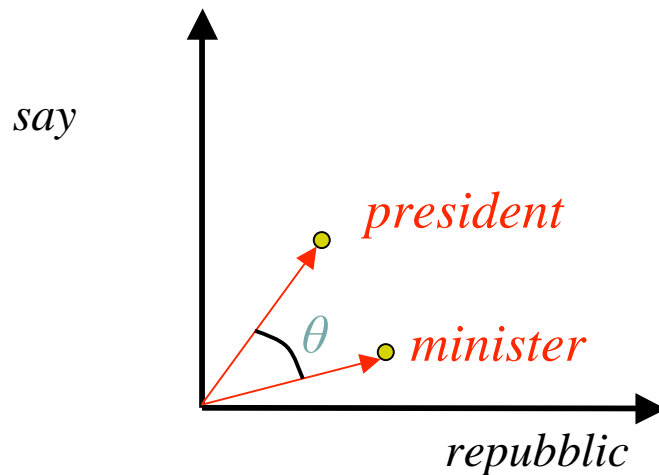
Word Space Models

measuring the distance in space



- **Cosine**

- the higher the **cosine of the angle between two vectors**, the higher the similarity of the corresponding target words



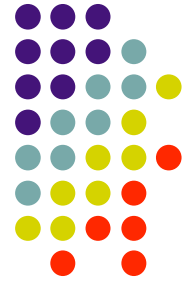
cosine as a similarity measure

$$d(\vec{w}_1, \vec{w}_2) = \cos \theta = \vec{w}_1 \cdot \vec{w}_2 = \sum_{i=1}^n w_{1_i} w_{2_i}$$

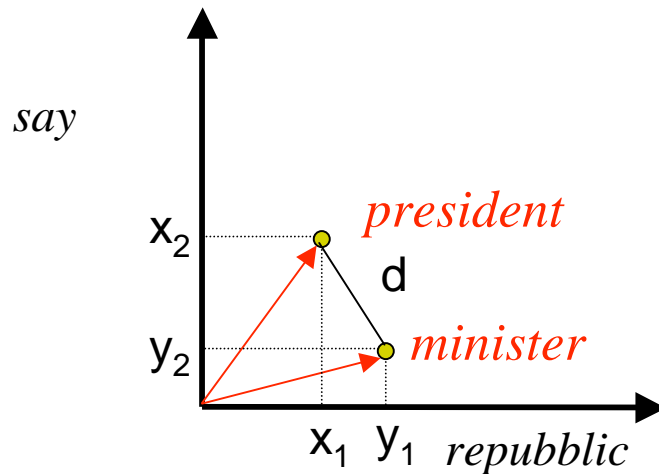
the highest similarity value is **1** ($\cos 0^\circ = 1$)
if two words are semantically independent
(orthogonal) $\cos \theta$ is close to **0** ($\cos 90^\circ = 0$)

Word Space Models

measuring the distance in space



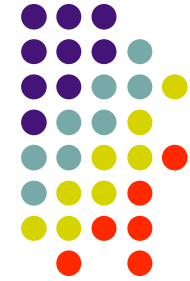
- **Euclidean distance**
 - the shorter is the **euclidean distance** between two **vectors**, the higher the similarity of the corresponding target words



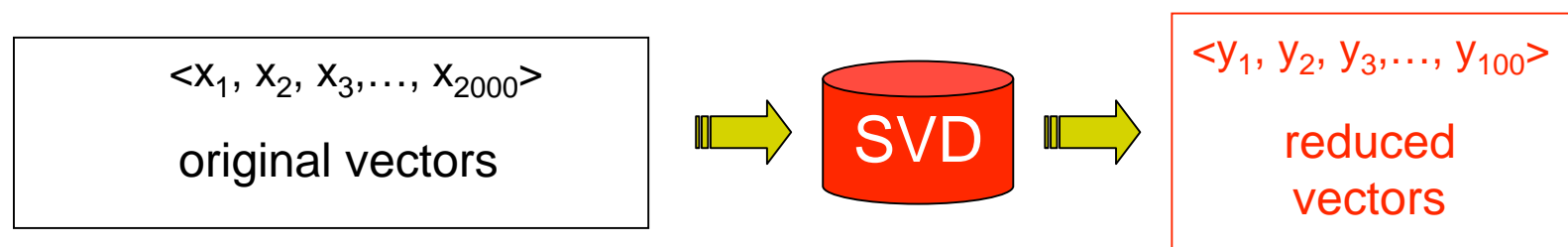
euclidean distance

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

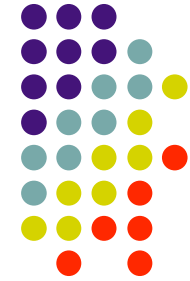
Vector dimensionality



- Reducing vector dimensionality
 - typically, word vectors have a very high number of dimensions and are very sparse
 - matrix transformation methods from linear algebra are often applied to reduce the number of vector dimensions
 - **Singular Value Decomposition (SVD)**



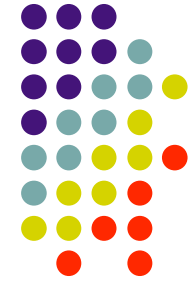
Singular Value Decomposition



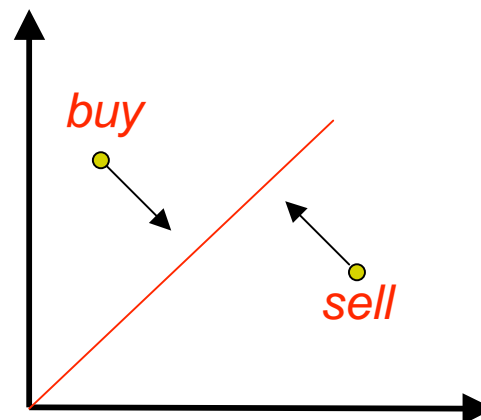
- Similar to **Principal Component Analysis**
 - maps the original vector space onto a reduced space, formed by the “essential” dimensions
 - semantic dimensions “hidden” in the original distributional matrix
 - the dimensions are those that best explain the variance in the data

$$\begin{matrix} & c & & m & & m & & c \\ & \boxed{} & = & \boxed{} & \times & \boxed{} & \times & \boxed{} \\ w & & & w & & & & \end{matrix}$$

Singular Value Decomposition



- SVD is claimed to capture contextual second order similarity relations
 - **first order similarity**
 - two words are similar because they occur in the same contexts
 - eg. *buy* a book, *buy* a newspaper
 - **second order similarity**
 - two words are similar because they occur in contexts that are in turn distributionally similar
 - eg. *buy* a book, *sell* a newspaper



Semantic neighbors



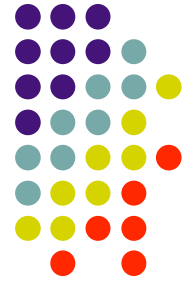
table:1.000000
sideboard:0.686550
chair:0.658235
sit:0.633039
sip:0.631396
armchair:0.620863
stool:0.591989
sofa:0.586884
tray:0.573620
mug:0.554641
coffee:0.541843
napkin:0.540975
seat:0.539092
saucer:0.531629
tablecloth:0.527330
tea:0.518063
trestle:0.517287

cat:1.000000
dog:0.745510
pet:0.699355
animal:0.606473
prey:0.601927
human:0.591936
fox:0.583327
spider:0.573000
predator:0.557195
monkey:0.555635
snake:0.553745
bird:0.542290
rodent:0.540798
parrot:0.537788
pheasant:0.533675
wolf:0.533063
fowl:0.527398

car:1.000000
van:0.755437
driver:0.728991
vehicle:0.708992
park:0.699351
motorist:0.692586
motor:0.686240
lorry:0.661483
scooter:0.660576
getaway:0.642833
drive:0.642715
bike:0.616735
truck:0.614012
steal:0.595915
carriageway:0.594908
hatchback:0.594367
jaguar:0.591013

Testing distributional similarity

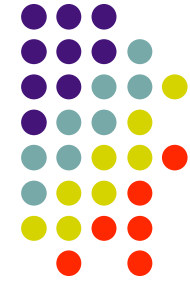
synonym identification



- Standard test set
 - synonym section in TOEFL (Test of English as a Foreign Language)
 - 80 items for which subjects must select the correct synonym among 4 candidates
 - target = furnish {supply, impress, protect, advise}
 - target = physician {chemist, pharmacist, nurse, doctor}
- Landauer & Dumais (1997)
 - LSA: 64.4% TOEFL participants: 64.5%
- Rapp (2004)
 - LSA: 92% Macquarie University (non-native): 86.75%
Macquarie University (native): 97.75%

Synonym identification

state of the art (ACL wiki)



Algorithm	Reference for algorithm	Reference for experiment	Type	Correct	95% confidence
RES	Resnik (1995)	Jarmasz and Szpakowicz (2003)	Hybrid	20.31%	12.89–31.83%
LC	Leacock and Chodrow (1998)	Jarmasz and Szpakowicz (2003)	Lexicon-based	21.88%	13.91–33.21%
LIN	Lin (1998)	Jarmasz and Szpakowicz (2003)	Hybrid	24.06%	15.99–35.94%
Random	Random guessing	1 / 4 = 25.00%	Random	25.00%	15.99–35.94%
JC	Jiang and Conrath (1997)	Jarmasz and Szpakowicz (2003)	Hybrid	25.00%	15.99–35.94%
LSA	Landauer and Dumais (1997)	Landauer and Dumais (1997)	Corpus-based	64.38%	52.90–74.80%
Human	Average non-English US college applicant	Landauer and Dumais (1997)	Human	64.50%	53.01–74.88%
DS	Pado and Lapata (2007)	Pado and Lapata (2007)	Corpus-based	73.00%	62.72–82.96%
PMI-IR	Turney (2001)	Turney (2001)	Corpus-based	73.75%	62.72–82.96%
HSO	Hirst and St.-Onge (1998)	Jarmasz and Szpakowicz (2003)	Lexicon-based	77.91%	68.17–87.11%
JS	Jarmasz and Szpakowicz (2003)	Jarmasz and Szpakowicz (2003)	Lexicon-based	78.75%	68.17–87.11%
PMI-IR	Terra and Clarke (2003)	Terra and Clarke (2003)	Corpus-based	81.25%	70.97–89.11%
CWO	Ruiz-Casado et al. (2005)	Ruiz-Casado et al. (2005)	Web-based	82.55%	72.38–90.09%
PPMIC	Bullinaria and Levy (2006)	Bullinaria and Levy (2006)	Corpus-based	85.00%	75.26–92.00%
GLSA	Matveeva et al. (2005)	Matveeva et al. (2005)	Corpus-based	86.25%	76.73–92.93%
LSA	Rapp (2003)	Rapp (2003)	Corpus-based	92.50%	84.39–97.20%
PR	Turney et al. (2003)	Turney et al. (2003)	Hybrid	97.50%	91.26–99.70%

Conclusions



- Computational corpus-based methods can support onto(lexical) development
 - still, lots of open issues
 - various aspects of meaning are extremely challenging
 - cf. relation extraction, argument structures, etc.
 - moving from extracted terms to full-blown ontologies lies beyond current system capacities
- Word space models are promising in providing **a new bridge towards cognitive research**
 - appealing for psycho-computational modelling
 - possibility to define semantic representations **without the need to stipulate any type of semantic *a priori***

Conclusions



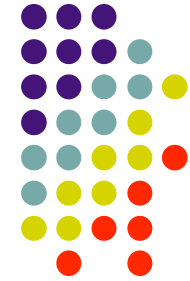
- Text-driven onto(lexical) learning methods can be used to obtain systems of semantic types more “**attuned**” with language use
 - the semantic space emerges from the way linguistic expressions distribute and interact in context
- We can explore and test the **contribution of syntagmatic information in shaping ontologies** of semantic types relevant for linguistic explanation and semantic representation

Possible topics for research essays



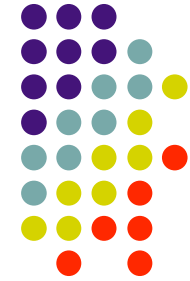
- Compositionality in word-space models
- Corpus-based approaches to polysemy
- Extracting semantic relations from text: the limits of the state of the art
- Beyond the Generative Lexicon: testing the limits of the QS
- Exploring co-compositional process
- Feeding the QS with corpus-based information

References



- Charles, W.G. (2000), "Contextual correlates of meaning", *Applied Psycholinguistics*, 21: 505-524
- Church, K.W. & P. Hanks (1990), "Word association norms, mutual information, and lexicography", *Computational Linguistics*, 16(1): 22-29
- Evert, S. (2004), *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, PhD dissertation
- Firth, J. R. (1957), *Papers in Linguistics*, London, Oxford University Press
- Harris Z. (1968), *Mathematical Structures of Language*, New York, Wiley
- Harris, Z.S. (1970), *Papers in Structural and Transformational Linguistics*, D. Reidel Publishing Company, Dordrecht-Holland
- Lund K., Burgess C., & R.A. Atchley (1995), "Semantic and associative priming in high-dimensional semantic space", *Proceedings of the Cognitive Science Society*, Hillsdale, N.J., Erlbaum Publishers: 660-665
- McDonald S. & M. Ramsar (2001). "Testing the distributional hypothesis: The influence of context on judgements of semantic similarity", *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Edinburgh, LEA: 611-616
- Manning, C.D. & H. Schütze (1999), *Foundations of Statistical Natural Language Processing*, Cambridge MA, MIT Press
- Miller, G.A. & W.G. Charles (1991), "Contextual correlates of semantic similarity", *Language and Cognitive Processes*, 6: 1-28
- Ramsar M. & D. Yarlett (2003), "Semantic grounding in models of analogy: An environmental approach", *Cognitive Science*, 27(1): 41-71
- Vigliocco, G., Vinson, D.P, Lewis, W. & M.F. Garrett, (2004), "Representing the meanings of object and action words: The featural and unitary semantic space hypothesis", *Cognitive Psychology*, 48: 422-488

References



- Baroni M., Lenci A., & L. Onnis (2007), “ISA meets Lara: A fully incremental word space model for cognitively plausible simulations of semantic learning”, in *Proceedings of the ACL Workshop on Cognitive Aspects of Language Acquisition*, Praha: 49-56
- Burgess, C. & K. Lund (1997), “Modelling parsing constraints with high-dimensional context space”, *Language and Cognitive Processes*, 12: 1-34.
- Karlgren, J. & M. Sahlgren (2001), “From words to understanding”, in Uesaka Y., Kanerva P. & H. Asoh (eds.), *Foundations of real-world intelligence*, Stanford, CSLI: 294-308
- Landauer, Th.K. & S.T. Dumais (1997), “A Solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge”, *Psychological Review*, 104(2): 211-240
- Lin, D. (1998), “An Information-Theoretic Definition of Similarity”, in *Proceedings of the 15th International Conference on Machine Learning*: 296-304
- Lowe, W. (2001), “Towards a theory of semantic space”, Proceedings of the 23rd Annual Conference of the Cognitive Science Society, Philadelphia, PA, LEA: 576-581
- Padó S. & M. Lapata (2007), “Dependency-based construction of semantic space models”, *Computational Linguistics*, 33(2): 161-199
- Maedche, A. & S. Staab (2004), “Ontology Learning”, in Staab, S. & R. Studer (eds.), *Handbook on Ontologies*, Springer
- Widdows, D. (2003), *Geometry and Meaning*, Stanford CA, CSLI

References



- Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A. & L. Prevot (eds.) (2009), *Ontologies and the Lexicon. A Natural Language Processing Perspective*, Cambridge, Cambridge University Press
- Lenci, A., Montemagni, S., Pirrelli, P. & G. Venturi (2007), “NLP-based ontology learning from legal texts. A case study”, *Proceedings of LOAIT 2007*
- Lenci, A., (2009), “The Life Cycle of Knowledge” in Huang, C.-R. *et al.* (eds.) (2009)
- Rapp, R (2004), “A Freely Available Automatically Generated Thesaurus of Related Words”, *Proceedings of LREC 2004*, Lisbona, ELRA: 395-398
- Saif M. & G. Hirst (2005), “Distributional measures as proxies for semantic relatedness”, in press