A contrastive Approach to Multi-word Term Extraction from Domain-specific Corpora

Francesca Bonin, Felice Dell' Orletta, Giulia Venturi and Simonetta Montemagni

LREC Malta 2010

Outline

- Introduction and aims
- Multiword extraction process
  - Multi-word candidates extraction
  - Contrastive Re-ranking of extracted terms
- Case studies
- Evaluation
- Conclusions

The aim

- **Extraction of domain specific terminology**
  - **Focus on multiwords.**
  - Filtering terms from noise:
    - Open-domain terms, e.g.
      - *anno successivo* "following year"
  - For multi-domain terminology: Singling out terms which belong to different domains.
    - *This is the case in the legal domain*
    - e.g. environmental terms from legal terms

      Rifiuto pericoloso "dangerous waste"

      singled out from

      Diritto nazionale "national law"

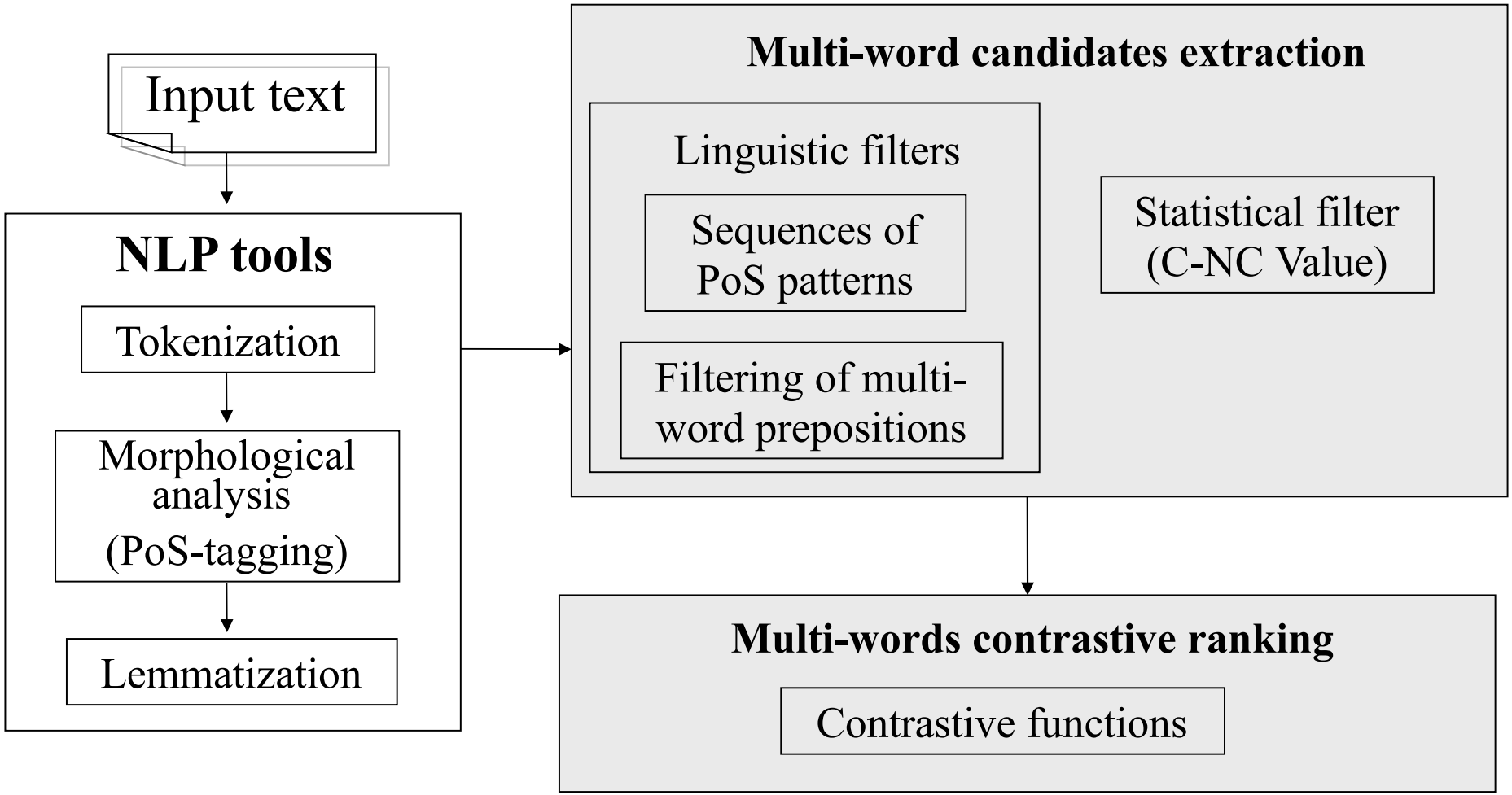
Approaches to Terminology extraction

- The state of the art of TE proposes a wide variety of approaches:
  - Linguistic
  - Statistical
  - **Mixed (Linguistic and statistical)**
  - **Contrastive**
- Statistical approaches based on e.g. term frequency/inverse document frequency, log likelihood, mutual information, up to more sophisticated approaches such as C-NC Value
- Contrastive approaches: usually applied on single terms extraction
  - They face multiwords extraction expanding the single terms heads
    - e.g. [Basili et. al 2001]: contrastive selection via Heads (CsvH)
    - Single candidate terms selection using a **contrastive function:** distribution in the target and contrastive corpora;
    - Multi-words ranking.

Our approach: main features

- Multiwords based: we consider multiwords as unique elements, independent from single terms
- **Combines** different approaches: linguistic + statistical + contrastive
- **Multi-layered approach**
  - We split the multiword extraction process in two steps:
    - Extraction of well-formed multi-word candidates' shortlist
    - Multi-word re-ranking.
  - Benefits of two-step approach:
    - Overcomes the multi-word term sparseness problem
    - Multi-word contrastive ranking: independent from single terms ranking.

General workflow



Step 1: Multi-word terms (MWT) candidates extraction

MWT Candidate extraction process:

- Linguistic filters
  - Based on automatic POS tagged/lemmatized text
  - We identify sequences of allowed POS patterns in order to cover most of the Italian morphosyntactic multi–words structures:
  - Noun+(Prep+(Noun|ADJ)+ |Noun|ADJ)+
    - *Diritto nazionale – "national law"*
    - *Presidente della Repubblica – "President of the Republic"*
  - Filtering of domain specific multi-word preposition, automatically extracted with a first run of the same process using the patterns
  - Noun-Prep-Noun
    - *ai sensi di – "by law"*
- Statistical filters: C-NC Value (Frantzi & Ananiadou 1999).

Step 2: MWT contrastive ranking

- Candidates multi-word terms are re-ranked using a contrastive method against a reference corpus.

  i. Single domain – contrast against open domain corpus [for filtering noisy general terms]

  ii. Double domain – contrast corpus sharing only one of the domains [for singling out different term types in multi-domain corpora]

  – In case i) - TFITF contrastive function

    - Basili et. al 2001 approach: contrastive selection via Heads (CsvH).
    - Our approach: Basili et. al 2001 function directly applied to multiwords.

  – In case ii) – CSmw contrastive function

    - based on arctan.
    - Particularly suitable for dealing with low frequency events

Step 2: MWT contrastive ranking - *TF-ITF contrastive function*

- **TFITF: Term frequency Inverse Term frequency**
  - Variant of Basili et al. 2001
  - applied to multi-word terms without passing through single head terms
  - A list *L* of candidate multi-words is extracted with C-NC Value
  - *L* toplist is ranked on TF-ITF score

$$TFITF = \log(f_i(t)) * IWF(t)$$

  - Where $f_i(t)$ is the frequency of the candidate term (multi-word) *t*, and *IWF* is the inverse word frequency.

$$IWF(t) = \log(N / F(t))$$

    - N: sum of all F*(t)* for each *t* in *L*
    - *F(t)*: *t* frequency in the contrastive corpus

Step 2: MWT contrastive ranking - *Csmw contrastive function*

Csmw : Contrastive Selection of multi-word terms

- Specifically designed for dealing with low-frequency events
- **Arctan function's mathematical features suites the low frequency events extraction**
- The statistical weight is calculated directly on multi-word terms

$$CSmw(t) = \arctan(\log(f_i(t)) * K)$$

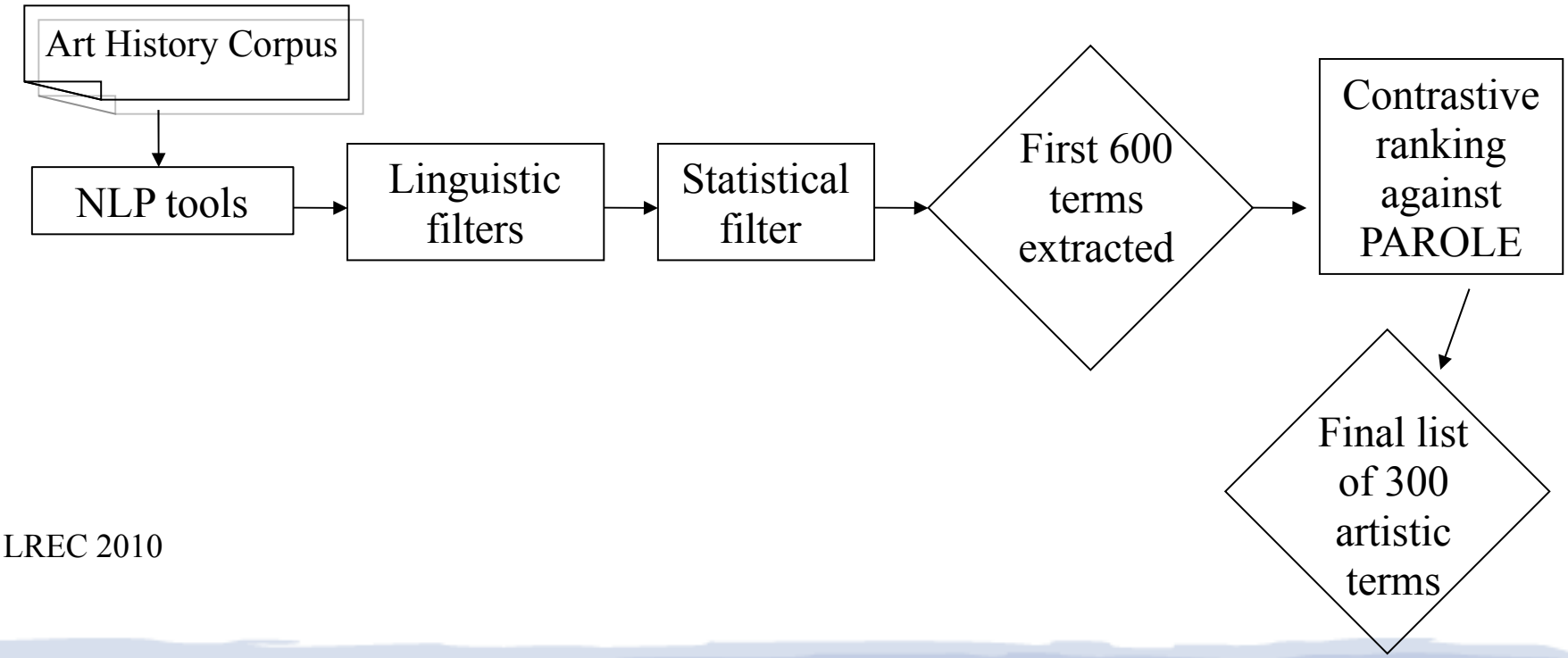
- Where: $K(t) = 1/F_c(t)/N_c$
- C is the set of contrastive domains, $F_c(t)$ is the frequency of *t* in all contrastive domains of C normalized on N_c that is the sum of all frequencies in *C* for each *t* in *L*.

Case Studies

- Art History case study:
  - Aim: domain specific term extraction
  - Corpus of Art history websites, 326,066 tokens.
    - Manually collected by a domain expert
  - Open domain contrastive Corpus: PAROLE.
    - Italian texts of different types, 3 millions tokens.
- Legislative-environmental case study
  - Aim: "double" domain terminology extraction and classification.
  - Collection of Italian European Legal Texts concerning the environmental domain, 394,088 tokens
  - Contrastive corpora used:
    - PAROLE.corpus.(open-domain)
    - Collection of European Legal Texts concerning the consumer protection domain, 72,210 tokens (Domain specific)

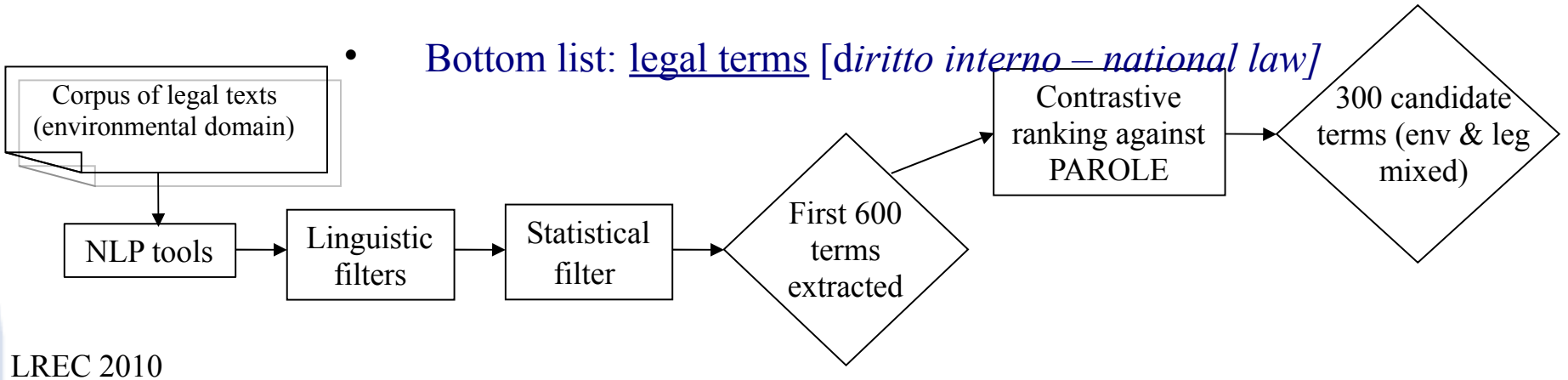
Art History case study

- Extraction of MWT candidates [C-NC-Value]
  - Selection of a top list of C-NC-Value ranked candidates (threshold empirically set at 600 terms).
- Contrast : against the open domain corpus PAROLE.
  - Final list *L* of 300 domain specific terms.



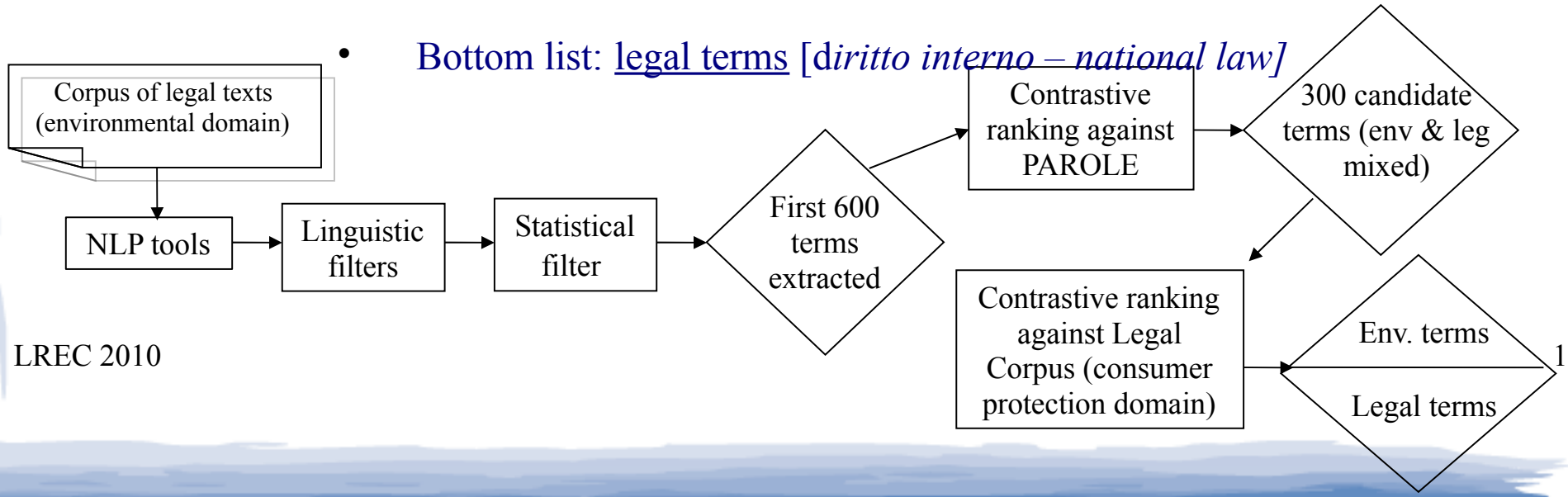
Legislative Case study

- Extraction of MWT candidates [C-NC-Value]
  - Selection of a top list of C-NC-Value ranked candidates (threshold empirically set at 600 terms).
- 1th contrast : against the open domain corpus PAROLE.
  - List *L* of 300 legal and environmental terms.
- 2th contrast: against Legal Corpus on consumer protection
  - Final list *L* new ranking:
    - Top list: environmental terms [*rifiuto pericoloso – dangerous waste]*
    - Bottom list: legal terms [d*iritto interno – national law]*



Legislative Case study

- Extraction of MWT candidates [C-NC-Value]
  - Selection of a top list of C-NC-Value ranked candidates (threshold empirically set at 600 terms).
- 1th contrast : against the open domain corpus PAROLE.
  - List *L* of 300 legal and environmental terms.
- 2th contrast: against Legal Corpus on consumer protection
  - Final list *L* new ranking:
    - Top list: environmental terms [*rifiuto pericoloso – dangerous waste]*
    - Bottom list: legal terms [d*iritto interno – national law]*



Evaluation methodology

- Automatic evaluation using gold standard resources
  - Term list provided by domain experts.
  - EartH, Environmental Applications Reference Thesaurus
- Manual evaluation, of unmatched terms, carried out by a domain expert
  - Gold standard resources do not have proper coverage of complex terms.
  - Art domain - Art History Department, University of Pisa.
  - Environmental – Institute of Atmospheric pollution (CNR).
  - Legal – Scuola Superiore Sant Anna, Pisa (Ossevatorio sul danno alla persona)

Evaluation has been carried on wrt the results obtained with:

- *NC-Value*
- *Csmw*
- *CsvH*
- *TF-ITF*

Evaluation – Art history domain

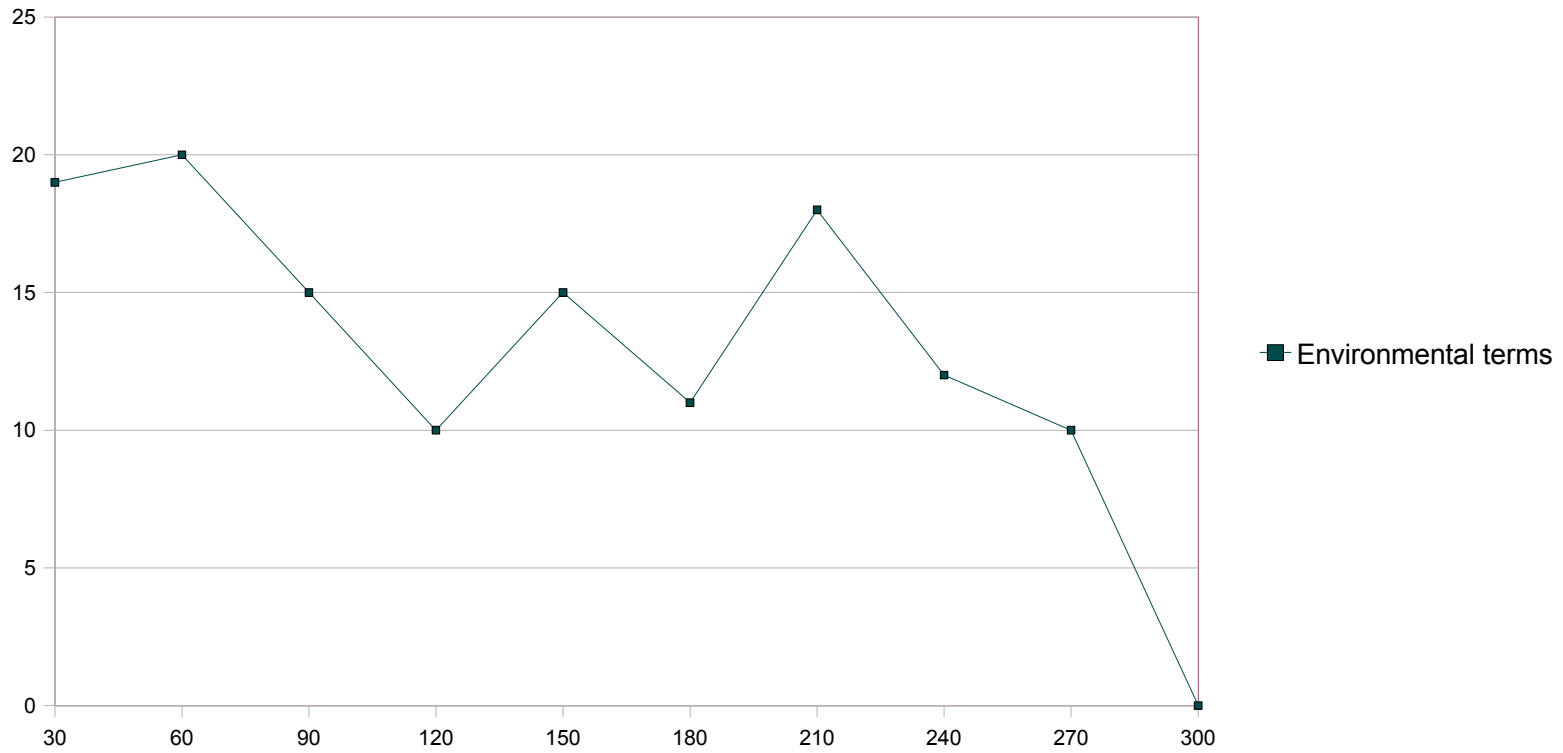
- List of 300 extracted artistic terms
- Extracted MWT distributed into 10 groups of 30 terms each.
- Out of the first 300 terms, CsvH method retrieved the largest amount of Artistic terms.
- TFITF and Csmw have

more domain-specific terms in the top list .

Group	NC-Value	CS-vH	TFITF	Csmw
0-30	24	28	25	25
30-60	20	21	25	24
60-90	20	23	26	25
90-120	18	20	21	24
120-150	20	24	22	26
Tot	102	116	119	124

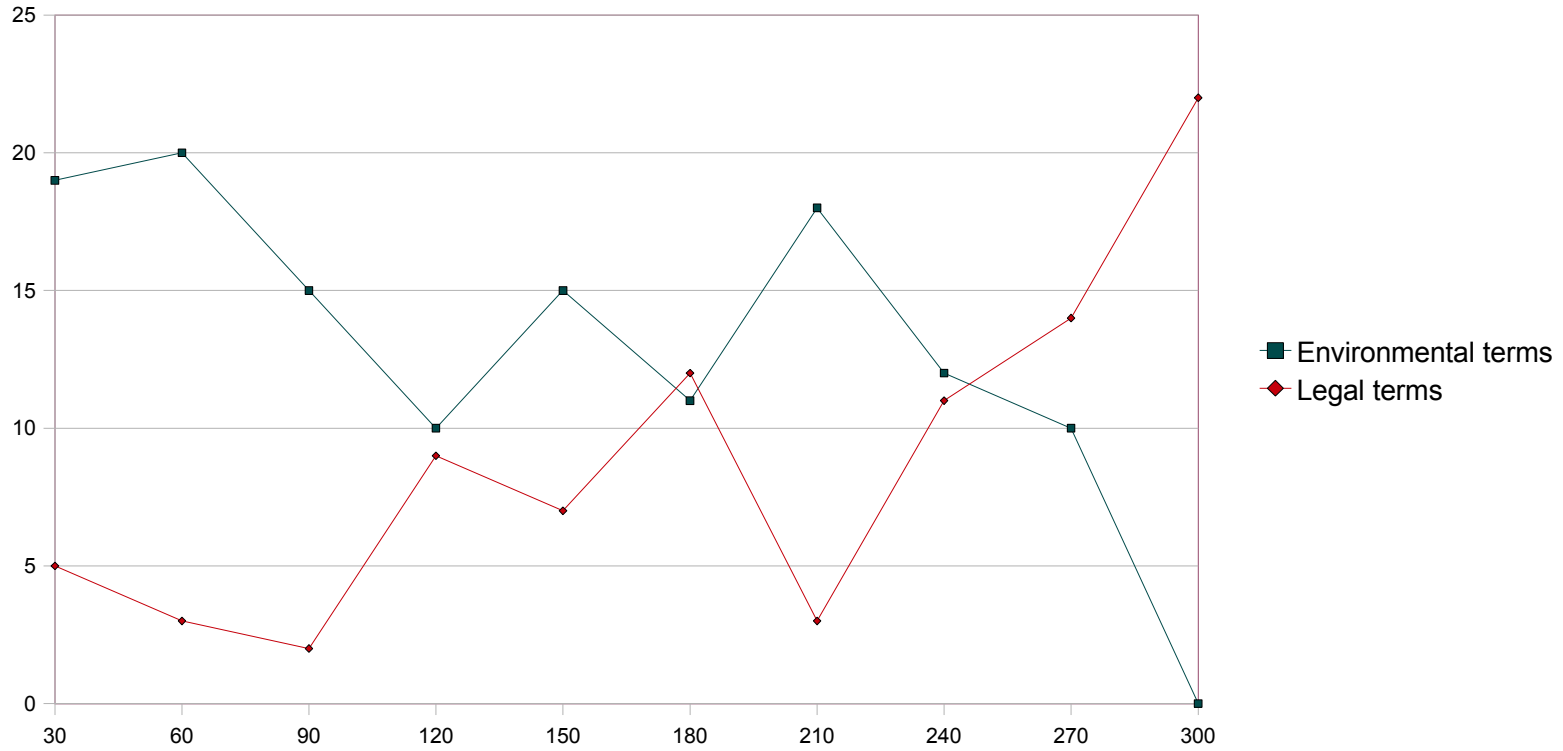
Evaluation – Legal domain

- List of 300 extracted artistic terms
- Extracted MWT distributed into 10 groups of 30 terms each.
- Top list: mainly environmental terms
- Bottom list: mainly legal terms



Evaluation – Legal domain

- List of 300 extracted artistic terms
- Extracted MWT distributed into 10 groups of 30 terms each.
- Top list: mainly environmental terms
- Bottom list: mainly legal terms



Conclusions and future developments

Novel approach to MWT extraction combining the C–NC value method with a contrastive ranking technique, aimed at:

- Reducing noise deriving from common words
- Discriminating semantically different types of terms within heterogeneous terminologies (as in the legal domain)

- **Current directions of research include:**

  Improvements to the MWT extraction algorithm

  Improvements of the multi-domain terminology extraction task

  Application of the proposed approach to identify neologisms from diachronic corpora of newspapers texts.

Thanks for your attention!