A morphologically driven reference semantic lexicon for French

Fiammetta NAMER / University of Nancy - ATILF - CNRS

Pierrette BOUILLON / University of Geneva / ISSCO

Evelyne JACQUEY / ATILF - CNRS Nancy

Why ?

- Lexical semantic information is needed
  - NLP in man-machine dialog
  - Information retrieval, Electronic documents, …
  - Major categories : nouns, verbs, adjectives
    - More than 90% of the French lexicon (more than 91.000 lexemes)
- Main properties of any reference lexical resource
  - Free access and using
  - Normalization and robustness
  - Coherence and coverage

How ?

- Learning from morphology
  - Automatic decomposition of constructed lexemes
  - First semantic skeletons
- Learning from corpora
  - A lexicographic corpus : TFLi definitions
  - Journalistic and technical corpora
  - Database FRANTEXT
- Crossing the methodologies
- A same representation format : GLT

Advantages and originality

- Reuse of existing resources and devices
- Morphology
  - Rough coherent semantic for a quantitatively significant part
- Corpora learning
  - Usage specifications for constructed lexemes
  - Simple lexemes
- Merging the methodologies and sharing the same representation format
  - Coherence and coverage of the whole resource

Lexical semantics from morphology

- DériF results
  - Pairs (Lexeme L, Base B) + Lexical Formation Rule (LFR)
  - L-to-B pairs are annotated with constraints LFRs prototypically impose on B and/or L

- LFR for *dé-* prefixed deadjectival verbs
  1. DESSOULER/VERB (sober up) ==> SOUL,ADJ (drunk) / dé:prefix
  2. "(suppress - deprive from) SOUL character"
  3. SOUL/ADJ:(predicative, _, temporary)
  4. DESSOULER/VERB:(dynamic, trans, [cause,theme], causative) || (dynamic, intrans, [theme], resultative)

Representation within GLT formalism

- L and B entries are built
- Entries specifications depend on DériF annotations provided
  - *DESSOULER*

FORM	not 1 (soul'(e1:temporary_state,y:ind)	
AGENT	FORM	FORM 2 dessouler-act(e2:act,y)
		AGENT 1
	AGENT	CAUSE(x:agent, 2)

  - *SOUL*

FORM	soul' (e:temporary_state,y:ind)

Why another method and input ?

- All simple lexemes cannot be annoted by DériF
- DériF semantic annotations are only governed by LFRs → no usage specifications
  - Semantic content selection
    - DETERRER (dig up SMTH) / TERRE (earth) [(initial) ground]
    - DENEIGER (clear snow from SMWHRE) / NEIGE (snow) [figure]
  - Detection of restrictions to a specific LFR
    - ABREUVOIR (watering place) / ABREUVER

Lexical semantics from corpora

- Lexicographic corpus
  - Part-of-speech and XML-tagged machine readable TLFi
  - Corpus of the definitions for nouns, verbs, adjectives (1.421.530 lemmatized occurrences)
  - Regular expressions based learning
- GLT specifications for quale and corpus
  - Hypothesis
    - Telic, formal, agentive and constitutive predicates can be automatically detected in definitions because of their regularity

Lexical semantics from corpora

- Nouns referring to prototypically functional entities
  - GLT : FORMAL and TELIC quale are predicated
  - TLFi definitions corpus
    - TELIC : some characteristic expressions
      - "*servir à*" (useful/allows to) / "*utilisé pour*" (intended/used to)
    - FORMAL : nominal phrases before TELIC expressions
- *BALAI (broom, brush)*
  - **Ustensile de ménage**$_{\text{[FORM]}}$ servant au **nettoyage**$_{\text{[TELIC]}}$ …
    (housework ustensil used for sweeping)

    FORM x:**housework_ustensil**

    TELIC **sweep-act**(e1:act, y:ind, w:room, x)

Crossing morphology and corpora

- Semantic content selection
- Underspecified semantics with DériF results
  1. DÉTERRER/VERB (dig up) ==> TERRE,NOUN (earth) / dé:prefix
  2. "Remove smth from TERRE || Remove TERRE from smth"
  3. DÉTERRER/VERB:(dynamic, trans, [cause,theme], causative)
- TLFi corpus : selection
  - Only the meaning "Remove smth from TERRE" is recensed
    - A.- Tirer de terre (to dig up SMTH from earth) [(initial) ground]

Crossing morphology and corpora

- Restrictions detection: deverbal -oir nouns
- DériF results
  - X- oir/NOUN ==> X, VERB/oir:suffix
  - "Instrument of X || Place of X"
  - X/VERB: (dynamic,-,-,-)
- Usages in corpora
  - Instrument and place : ABREUVOIR (watering place)
  - Instrument or place : HACHOIR (cleaver) or (chopping board)
  - Place only : FUMOIR (smoking room)
  - Instrument only : RASOIR (razor)

Representation within GLT formalism

- LFR prediction → instrument•localisation type
- Usage restrictions
  - Endocentric dotted type: ABREUVOIR
  - Exocentric dotted type: HACHOIR
  - Exocentric dotted type and shadowing of instrument : FUMOIR
  - Exocentric dotted type and shadowing of localisation : RASOIR

Evaluation

- DériF results
  - 35,5% of TLFi (35.500 lexemes), using 85 LFRs
  - 45.500 (Lexeme, Base) relations
- Explanation
  - Not all LFRs implemented
  - Large amount of TLFi lexemes are simple
- But advantage of Dérif
  - It can be applied to any lexicon
  - Biomedical lexicon = 59% of constructed lexemes

Conclusion

- An original methodology to derive a reference semantic lexicon for French
  - It relies on the morphological properties of the lexicon
    - Coherence of the produced lexicon
    - Collaboration of morphology with other knowledges and device (lexicographic and textual corpora) to derive deep representations of the meaning
- Expected results
  - A GL lexicon
  - Tools for extracting new entries