Travaux du 19ème CIL | 19th ICL papers
Congrès International des Linguistes, Genève 20-27 Juillet 2013
International Congress of Linguists, Geneva 20-27 July 2013

19th International
Congress of Linguists
July 21-27 2013
Geneva - Switzerland

Marie-Aude LEFER and Natalia GRABAR

Institut libre Marie Haps, Brussels, Belgium
Université catholique de Louvain, Belgium
Université Lille, France

marie-aude.lefer@uclouvain.be

natalia.grabar@univ-lille3.fr

French evaluative prefixes in translation: From automatic alignment to semantic categorization

oral presentation in workshop: 131 TheoreticAl and Computational MOrphology: New Trends and Synergies [TACMO] (Bruno CARTONI, Delphine BERNHARD & Delphine TRIBOUT)

Published and distributed by: Département de Linguistique de l'Université de Genève, Rue de Candolle 2, CH-1205 Genève, Switzerland
Editor: Département de Linguistique de l'Université de Genève, Switzerland
ISBN:978-2-8399-1580-9

French evaluative prefixes in translation: From automatic alignment to semantic categorization

Marie-Aude Lefer[1], Natalia Grabar[2]

(1) Institut Marie Haps & Université catholique de Louvain, Belgium
(2) STL CNRS UMR 8163, Université Lille 1& 3, France

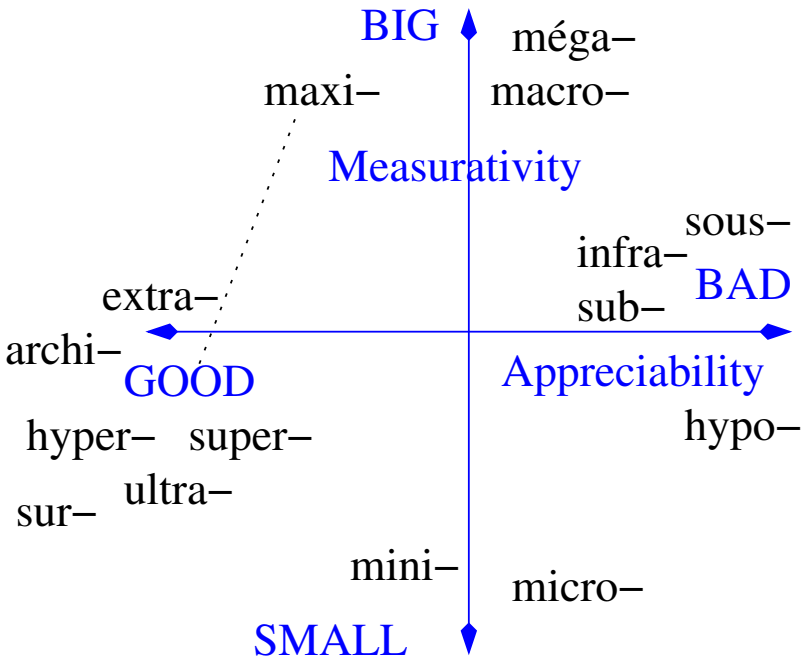
Outline of the presentation

- Context
- Objectives
- Data and Approach
- Results and Discussion
- Conclusion and Perspectives

Context

Evaluative morphology

- Morphological typology:
  - (Stump, 1993; Bauer, 1997; Grandi & Montermini, 2005; Körtvélyessy & Stekauer, 2011)
- First attempt at an exhaustive description of French evaluative morphology:
  - prefixation, -ET suffixation (Fradin & Montermini, 2009)



- Very few corpus-based studies

Context

Existing semantic classification of French evaluative morphology

- (Wierzbicka, 1991; Grandi, 2002; Cartoni, 2008; Fradin & Montermini, 2009):
  1. Quantity dimension with a maximum/minimum axis (so-called *measurativity*):
     - BIG: increase, abundance
     - SMALL: decrease, attenuation, approximation
  2. Quality dimension with a positive/negative axis (so-called *appreciativity*):
     - GOOD: excess (excessive degree), superiority (higher degree)
     - BAD: lack, inferiority (lower degree)
  3. Common semantic shifts (Fradin & Montermini, 2009):
     - between other semantic categories of prefixes (e.g. location) and evaluative prefixes
     - within the category of evaluative prefixes itself (e.g. from BIG to GOOD with *méga-*, *maxi-*)
  4. Other categories within GOOD prefixes (Guilbert, 1971):
     - HIGHER DEGREE (*archi-*, *extra-*, *super-*, *ultra-*) and EXCESSIVE DEGREE (*hyper-*, *sur-*)
     - ...sharp distinction or ambiguity?

Objectives

- Corpus-based insights into French evaluative prefixes
- *Translations as evidence for semantics* (Noël, 2003: 767, 770)
  - *translators are language users whose linguistic choices are not only informative about the language they are producing [the target language], they are also highly indicative of their interpretation of the language they are receiving [the source language], and this interpretation is revelatory of the nature of the language that is received*
- Hypothesis:
  - in a parallel corpus, *the semantic nature of the matches in the other language [i.e. the target language]* can shed light on the semantics of the source items
- Similar approaches:
  - word sense disambiguation (Banea & Mihalcea, 2011)
  - dictionary-based morphological study on Fr. *-iste* and It. *-ista* (Cartoni & Namer, 2012)

Rationale

- Analyze French evaluative prefixes
  - alongside their English translation equivalents
  - in a parallel corpus
  - aligned at word level
- Distinction between:
  - congruent translations: translations into prefixes
  - incongruent translations: such as periphrastic translations

  $\Longrightarrow$ likely to *spell out* the meaning of the source language prefixes
- Multidisciplininary framework:
  - Theoretical and empirical linguistics (morphology, lexical semantics and corpus linguistics)
  - Natural Language Processing, computer sciences
  - Translation studies

Data

- Prefixes
- Corpus
- Lexicon

Data

Prefixes

- BIG: *macro-*, *maxi-*, *méga-*
  macromolécule, maxi-bouteille, méga-stade
- SMALL: *micro-*, *mini-*
  micro-ordinateur, minisatellite
- GOOD: *archi-*, *extra-*, *hyper-*, *maxi-*, *méga-*, *super-*, *sur-*, *ultra-*
  archifaux, extra-chouette, hypernerveux, maxi-sale, méga-beau, superbon, surdoué, ultramoderne
- BAD: *hypo-*, *sous-*, *sub-*
  hypotension, sous-alimentation, subaigu
- ATTENUATION: *demi-*, *mi-*, *semi*
  demi-sommeil, mi-sérieux, semi-liberté
- APPROXIMATION: *quasi-*, *pseudo-*
  quasi-mûr, pseudo-scientifique

Data

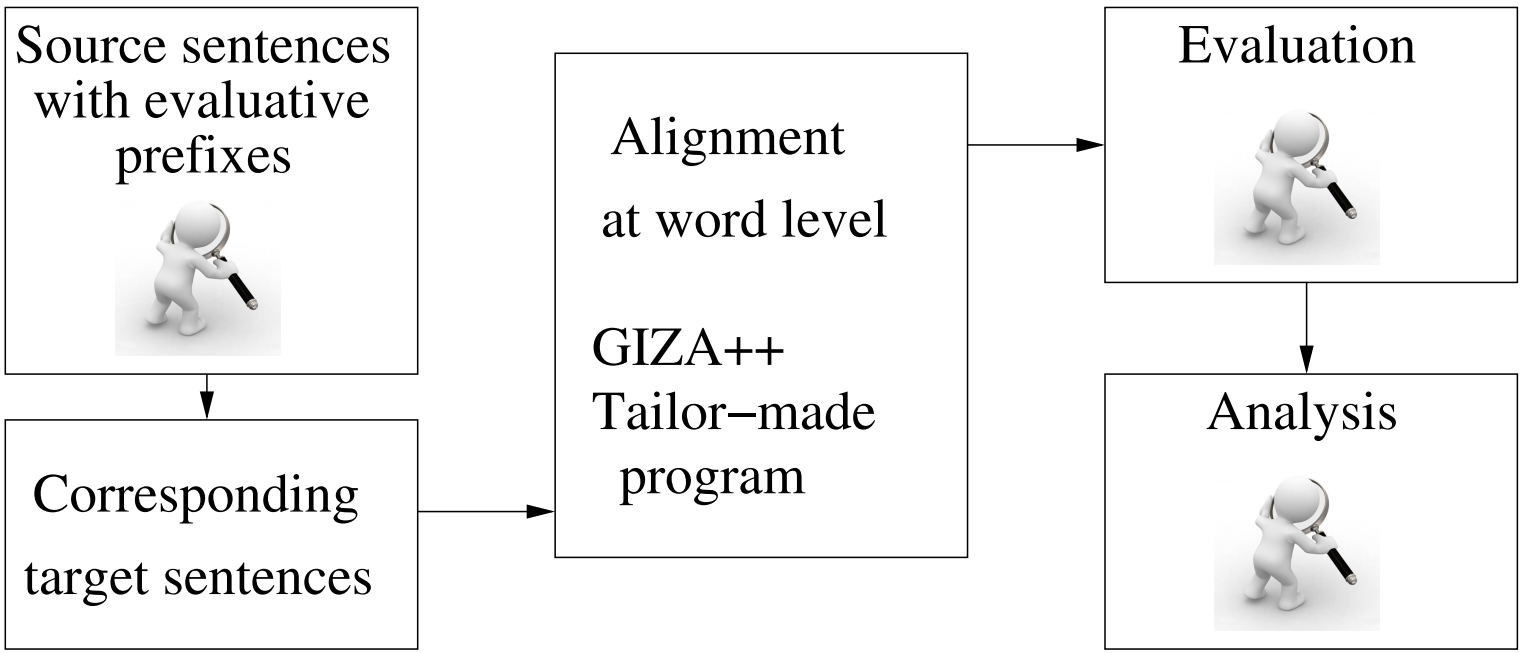
Corpus:

- Europarl6 parallel corpus (Koehn, 2005):
  - aligned at sentence level
- *Directional* Europarl6 (Cartoni & Meyer, 2012):
  - identification of source and target languages
- French-to-English subcorpus: 7,878 parallel documents
  10+ million running words

Lexicon:

- Small set of French and English prefix pairs:
  - {*méga*, *mega*}, {*demi*, *half*}, {*sur*, *over*}...

Approach


Source sentences with evaluative prefixes
Corresponding target sentences
Alignment at word level
GIZA++ Tailor–made program
Evaluation
Analysis

Approach

1. Detection of the source sentences that contain the evaluative prefixes
   - *ultralibéral*, *ultra-libéral*, *ultra libéral*
   - weeding out words such as: *extracteur, maximal, miette*, *extradition, extrapoler, hypocrisie*
2. Extraction of the corresponding target sentences

Approach

3. Alignment of French prefixed words with the corresponding word(s) in English target sentences:
   - GIZA++ (Och & Ney, 2000): several alignment models, such as IBM-4, IBM-5 and HMM
   - tailor-made program with heuristics:
     - same prefix in the target sentence: {*ultra*libérales, *ultra*liberal};
     - removing the prefix in the source word and replacing accented characters (at least first four letters): {*une région ultra*sensible, *an extremely* sensitive *region*};
     - translation of the source prefix in the target sentence: {*sur*pêche, *over*-fishing}, {*sous*-développement, *under*development} or {*demi*-mesures, *half*-measures};
     - neighboring context of the prefixed word in the source sentence: {*de la surenchère systématique*, *refuses to systematically try to outdo the*}.

Approach

4. Evaluation of the aligned sequences:
   - adjustment of the aligned segments
   - deleting locative instances
   - if empty, proposal of new segments
   - evaluation with the BLEU precision measure (Papinemi et al., 2002):
     - the number of common words between extracted and corrected segments
5. Manual analysis of the bilingual data
   - classification according to the strategies used to translate the prefixes in English

Results and Discussion

- 4,574 prefixed words extracted from the French source sentences
- GIZA++: 2,268 alignments (50%)
- Tailor-made program: 3,566 alignments (80%):
  - 1,862 alignments with direct equivalents in English;
  - 214 alignments thanks to the base word;
  - 1,168 alignments thanks to the translations of prefixes;
  - 322 alignments thanks to the neighboring words.
- No alignments for 1,008 words.
- Evaluation by two evaluators working independently
- 2,938 alignments kept after the validation phase
- 1,985 alignments once deduplicated
- Average BLEU precision on the target sequences: 0.76

Results and Discussion

prefix		tokens	types
sur	Appr/GOOD	495	146
sous	Appr/BAD	307	72
quasi	Aprox	262	124
ultra	Appr/GOOD	230	55
super	Appr/GOOD	210	57
micro	Meas/SMALL	142	36
macro	Meas/BIG	140	13
hyper	Appr/GOOD	46	34
mini	Meas/SMALL	44	21
pseudo	Approx	43	41
demi	Atten	31	17
semi	Atten	16	13
méga	Meas/BIG	10	9
mi	Atten	7	7
archi	Appr/GOOD	2	2

Results and Discussion

Translation into an evaluative prefix (same semantic category)	1,459	73.5%
Translation into a periphrase	453	22.8%
Translation into a non-prefixed word (simplex word or compound)	60	3%
Zero translation (the prefixed word is not translated in the targed segment)	8	0.4%
Translation into a non evaluative prefix (another semantic category)	5	0.3%

- Different trends for the analyzed prefixes:
  - 90% of periphrastic translations for *quasi-*
  - 3% of periphrastic translations for *super-*

Results and Discussion

- (Very) infrequent prefixes in the Europarl corpus:
  - *archi-*, *hyper-*, *méga-*, *mini-*, *pseudo-*, *semi-*, *sub-*
  - or occurring in a very limited set of prefixed words (e.g. *demi-* in *demi-mesure* and *macro-* in *macroéconomie/iste/ique*)
- Periphrastic translations reflect the evaluative meaning of the prefixes quite accurately:
  - *semi-*$_{ATTENUATION}$ and *demi-*$_{ATTENUATION}$:
    - {*en régime de semi-liberté*, *partially free*}, {*demi-solution*, *partial solution*}, {*demi-échec*, *partial failure*}
  - *quasi-*
    - {*quasi-nudité*, *near-nakedness*}, {*quasi-épave*, *virtual wreck*}, {*quasi-identique*, *almost identical*}, {*quasi-général*, *more or less general*}, {*quasi-unanime*, *practically unanimous*}
- Usefulness of the *translations as evidence for semantics* approach in morphology

Results and Discussion

Zoom on *sur-*: 495 validated entries

- leaving aside cases where *sur-* is translated into *over-*
- *sur-*$_{EXCESS}$:
  - excess(ive): {*surbureaucratisation*, *excess of bureaucracy*}, {*suremballage*, *excess packaging*}, {*surpression*, *excess pressure*}, {*surréglementation*, *excessive regulation*}
  - overly 'too': {*sururbanisé*, *overly built-up*}, {*surfiscalité*, *overly high taxation*}, {*surpuissant*, *overly powerful*}
  - too much/too many: {*surendettement*, *too much debt*}, {*suremploi*, *too many jobs*}

Results and Discussion

Disambiguation of sub-meanings

EXCESS and SUPERIORITY (in GOOD value):

- *ultra-*
  - *ultra-*$_{EXCESS}$: {*ultra-échangisme*, *excessively free market*}
  - *ultra-*$_{SUPERIORITY}$: {*domaine ultrasensible*, *highly sensitive area*}, {*centres ultraspécialisés*, *highly specialized centers*}
- *hyper-*
  - *hyper-*$_{EXCESS}$: {*hyperflexibilité*, *excessive flexibility*}, {*hyperconcentration*, *excessive concentration*}
  - *hyper-*$_{SUPERIORITY}$: {*propositions hyper dirigistes*, *highly authoritarian proposals*}
- Need to be confirmed in larger-scale studies
- Refine Guilbert's (1971) distinction between:
  - SUPERIORITY (higher degree) prefixes *archi-*, *extra-*, *super-*, *ultra-*
  - EXCESS (excessive degree) prefixes *hyper-*, *sur-*,
- In our dataset *ultra-* and *hyper-* convey both SUPERIORITY and EXCESS, while *sur-* is only used to convey EXCESS.

Conclusion

- Corpus-based insights into French evaluative prefixation: low/high frequency of prefixes
- Valid for parliamentary debates
- Usefulness of translations derived from parallel corpora as semantic evidence in morphology
- NLP contributes to the 'translations as evidence for semantics'
- Prefixes are useful anchor points for automatic alignment at word level

Perspectives

- Testing the methodology on:
  - other corpora
  - other languages
  - other translation directions (English-to-French) and language pairs
  - other morphological phenomena
- Assess the generalisability of the method
- Future exploitation:
  - machine or computer-assisted translation
  - bilingual elexicography
  - second/foreign language learning/teaching
- Application: Mulexfor database (MULtilingual LEXeme-FOrmation Rules) (Cartoni & Lefer, 2010), https://sites.google.com/site/mulexfor/

Conclusion and Perspectives