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[1,2], Natalia Grabar[3]
(1) Institut libre Marie Haps, Brussels, Belgium
(2) Centre for English Corpus Linguistics, Université catholique de Louvain, Louvain-la-Neuve, Belgium
(3) STL CNRS UMR 8163, Université Lille 1 & 3, Lille, France
marie-aude.lefer@uclouvain.be, natalia.grabar@univ-lille3.fr

Keywords: contrastive morphology, evaluative prefixation, semantics, translation, parallel corpora, automatic alignment, French, English

1 Introduction

This paper deals with French evaluative prefixes and aims to find out whether translation can shed light on the semantics of these prefixes by adopting Noël's (2003) 'translations as evidence for semantics' approach.

Evaluative morphology (esp. augmentatives and diminutives) has been extensively discussed in the field of morphological typology (see e.g. Stump, 1993; Bauer, 1997; Grandi & Montermini, 2005; Körtvélyessy & Stekauer, 2011). Corpus-based descriptions of evaluative morphology, however, are still sorely lacking for many languages, including French. A first attempt at an exhaustive inventory and general discussion of French evaluative prefixation is found in Fradin & Montermini (2009) (in addition to prefixation, it also deals with the -*ET* suffixation). One of the important insights offered by Fradin & Montermini's overview is that in French, like in many other languages, evaluative prefixes can be classified along the following two dimensions (see Wierzbicka, 1991; Grandi, 2002; the sub-categories are taken from Guilbert, 1971: L and Cartoni, 2008: 287-291):

- Quantity dimension with a maximum/minimum axis (so-called 'measurativity') and the two semantic values BIG and SMALL:
  - BIG: increase, abundance
  - SMALL: decrease, attenuation, approximation
- Quality dimension with a positive/negative axis (so-called 'appreciativity') and the two semantic values GOOD and BAD:
  - GOOD: excess (excessive degree), superiority (higher degree)
  - BAD: lack, inferiority (lower degree)

The borderlines between these two dimensions and between the sub-meanings of the BIG/SMALL and GOOD/BAD values, however, are not watertight. In fact, as pointed out by Fradin & Montermini (2009: 241), semantic shifts are commonly observed, both between evaluative prefixes and other semantic categories of prefixes (e.g. location) and within the category of evaluative prefixes itself (e.g. from BIG to GOOD with *méga-* and *maxi-*). Within the group of GOOD prefixes, Guilbert (1971: L) makes a distinction between a set of prefixes conveying HIGHER DEGREE (*archi-*, *extra-*, *super-* and *ultra-*, as in *archibondé*, *archifou*, *extra-fin*, *extra-fort*, *superfin*, *supercarburant*, *ultra-chic*, *ultra-royaliste*) and a set of prefixes conveying EXCESSIVE DEGREE (*hyper-* and *sur-*, as in *hyperémotivité*, *hypernerveux*, *suralimentation, surpeuplé*, *surestimer*) but the question can be asked whether this is a sharp distinction or whether some of the evaluative prefixes can convey both sub-meanings.

2 Objectives

In this study, we would like to go beyond the current descriptions of French evaluative prefixes by providing corpus-based insights into the use and semantics of these word-forming elements, especially with regard to the semantic sub-categorization of evaluation in French.

To do so, we study French evaluative prefixation in translation, using translations derived from a parallel corpus as evidence for semantics. This new approach is inspired by Noël (2003), who states that "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" (ibid., 767). Noël's (2003) hypothesis is that 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 under investigation (ibid., 770). Similar approaches have been adopted in computational semantics for monolingual word sense disambiguation tasks (cf. Banea & Mihalcea, 2011 and the references cited therein) and in dictionary-based morphological research in Cartoni & Namer's (2012) study of Fr. *-iste* and It. *-ista*. In this paper, we wish to assess the potential benefits of using Noël's (2003) corpus-based approach in the field of word-formation. To do so, we analyze French evaluative prefixes alongside their English translation equivalents in a parallel corpus aligned at word level, paying particular attention to incongruent, non-morphological translations, such as periphrastic translations (as opposed to congruent translations, i.e. translations into prefixes), as they are likely to 'spell out' the meaning of the source language prefixes. One of the ultimate objectives of our project is to present a corpus-based semantic classification of French evaluative prefixes that would account for the subtle differences in meaning between the prefixes that belong to the same semantic (sub-)category.

Our study takes stock of insights from theoretical and empirical linguistics (morphology, lexical semantics and corpus linguistics), Natural Language Processing (NLP) and translation studies. The following sections present the empirical data on which the study is based and the extraction and alignment method we used. We then propose some preliminary observations based on the results of the study.

3 Data

The data used in this study were extracted from a French-to-English parallel corpus on the basis of an inventory of French evaluative prefixes (cf. Fradin & Montermini, 2009: 240; Cartoni, 2008). Following Cartoni (2008: 131-135), attenuation and approximation prefixes are also considered as being part of evaluative prefixation.[1] The set of prefixes investigated in our study is given below:

[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*

[1] Fradin & Montermini (2009) classify *demi-*, *mi-* and *semi-* as quantitative prefixes.

We used the Europarl6 parallel corpus (Koehn, 2005) and more particularly the 'directional' Europarl6 version made available by Cartoni & Meyer (2012), where the source and target languages are clearly identified. The corpus is aligned at sentence level and each pair of aligned sentences has its own identifier. In this work, we relied on a French-to-English subcorpus containing 7,878 parallel documents (10+ million running words). To test the alignment method, we also built a small set of French and English prefix pairs, such as {*méga, mega*}, {*demi, half*}, {*sur, over*}.

4 Methodology

The main steps of our methodology are the following: (1) detection of the source sentences that contain the evaluative prefixes investigated in the study and extraction of the corresponding target sentences; (2) alignment of French prefixed words with the corresponding word(s) in English target sentences; (3) evaluation of the aligned sequences; (4) manual analysis of the bilingual data. The first two steps (extraction and alignment) involve both automatic and manual data processing, while the third step (evaluation) is completely manual. The automatic part of the first step consists in projecting the prefixes on the source language sentences and spotting the words that contain these prefixes. The sentences containing the potentially prefixed words in French are then collected together with the corresponding aligned sentences in the target subcorpus. Prefixed words can have three types of spelling, which are all catered for by the extraction method: prefixes can be attached to the base word (as in *ultralibéral*), hyphenated to the base (as in *ultra-libéral*) or the prefix and the base can be separated by a space (as in *ultra libéral*). The manual processing phase consists in filtering the automatically extracted words in order to discard the ones that are not morphologically prefixed, even though they contain a prefix-like initial string (e.g. *extra*cteur, *maxi*mal, *mi*ette) and the words that are diachronically analyzable but opaque in synchrony (e.g. *extra*dition, *extra*poler, *hypo*crisie).

During the second step, alignment is performed at word level. The objective here is to detect, in the target sentence, the word (or the segment) that corresponds to the source prefixed word. This task was performed separately with the existing word-alignment tool GIZA++ (Och & Ney, 2000) and with a tailor-made alignment program. GIZA++ applies several alignment models, such as IBM-4, IBM-5 and HMM. As for the tailor-made program, it applies several heuristics (the strings underlined in the following examples make it possible to align prefixed words and their equivalents in the source and target sentences): (a) detection of a word that begins with the same prefix in the target sentence, e.g. {*ultra*libérales, *ultra*liberal}; (b) detection of the equivalent of the source base word in the target sentence (after having removed the prefix in the source word and replaced accented characters by non-accented characters; the source and target words have to at least share their first four letters), e.g. {*une région ultra*sensible, *an extremely* sensitive *region*} or {*cette société ultra-*urbaine, *this predominantly* urban *society*}; (c) detection of a word that begins with a translation of the source prefix in the target sentence, e.g. {*sur*pêche, *over-fishing*}, {*sous-développement, under*development} or {*demi-mesures, half-measures*}; (d) exploration of the neighboring context of the prefixed word in the source sentence and detection of the corresponding words in the target sentence, e.g. {*des* machines *ultraperformantes permettent, since high-performance* machines *permit a higher level*} or {*de la surenchère* systématique, *refuses to* systematically *try to outdo the*}. As shown in the last two examples, the extracted segments in the source and target sentences may be larger than the relevant words or segments. At that stage, we manually adjust and validate the extracted segments and complete the alignment when no alignment could be performed automatically. The aligned data also allows for the filtering of some more irrelevant data, e.g. thanks to the detection of the locative meaning of some prefixes, which is often easier to detect when examining the French prefixed word in context or the English translations (cf. {*extra européens, non-European*} or {*ultra*périphériques, *outer*most}).

The corrected alignments can be evaluated by means of the BLEU precision measure (Papinemi et al., 2002), which is usually adopted for the evaluation of machine translation results. It typically consists in counting the number of words in the original target sequence and in the adjusted target sequence: the number of common words between them corresponds to the BLEU measure. For instance, in the alignment {*des* machines *ultraperformantes permettent, since high-performance* machines *permit a higher level*}, the right target sequence is *high-performance*, which means that only one word is correct among the 7 aligned words in English. The

precision rate is computed as follows: 1/7=0.14 for this alignment.

5 Results: general overview

For the 4,574 prefixed words extracted from the French source sentences, GIZA++ and the tailor-made program generated respectively 2,268 and 3,566 alignments, among which we find: (a) 1,862 alignments with direct equivalents in English; (b) 214 alignments thanks to the base word; (c) 1,168 alignments thanks to the translations of prefixes; (d) 322 alignments thanks to the neighboring words. For 1,008 words the corresponding segments could not be extracted automatically. The alignment rates show that GIZA++ generated the alignments for nearly 50% of the data, while the tailor-made heuristics aligned c. 80% of the data. In view of this difference in alignment rate, we have chosen to work with the results provided by the tailor-made program. The alignments were divided in two subsets and validated by two evaluators working independently and applying the same validation criteria. 2,938 alignments were kept after the validation phase (several words were discarded during this manual filtering, cf. above). The validation (a, b, c, and d types) reveals that the mean BLEU precision on the target sequences is 0.76. After a final deduplication phase, we were left with 1,985 validated bilingual segments.

Some of the prefixes appear to be (very) infrequent in the Europarl corpus (*archi-*: 2 tokens, *demi-*: 31 tokens, *hyper-*: 46 tokens, *méga-*: 10 tokens, *mi-*: 7 tokens, *mini-*: 44 tokens, *pseudo-*: 43 tokens, *semi-*: 16 tokens, *sub-*: no occurrence) or occur in a very limited set of prefixed words (e.g. *macro-* in *macroéconomie/iste/ique*). Generalizations cannot be formulated for these prefixes at this stage because of the small amount of data available in the corpus analyzed for this study. Nevertheless, it is interesting to note that the periphrastic translations found in Europarl reflect the evaluative meaning of these prefixes quite accurately. Consider, for example, $semi_{\text{ATTENUATION}}$ and $demi_{\text{ATTENUATION}}$: *en régime de semi-liberté - partially free*, *demi-solution - partial solution*, *demi-échec - partial failure*. Larger translation corpora would be needed to study these prefixes in more detail but the data we have at our disposal already point to the usefulness of the 'translations as evidence for semantics' approach in morphology.

It turns out that the prefix *sur-* is the best candidate to test the potential of the 'translations as evidence for semantics' approach because it is the most frequent evaluative prefix in Europarl (495 validated entries). Zooming in on the non-morphological, periphrastic translations of *sur-* (i.e. leaving aside cases where *sur-* is translated into *over-*, e.g. *suradministré - over-administered*), which represent 134 occurrences out of 495 (27%), we find that it is possible to identify a range of typical periphrases of the EXCESS meaning. In Europarl sur_{EXCESS} is paraphrased as *excess(ive)* (e.g. *sur-bureaucratisation - excess of bureaucracy*, *suremballage - excess packaging*, *surpression - excess pressure*, *surréglementation - excessive regulation*), *overly* 'too' (e.g. *sururbanisé - overly built-up*, *surfiscalité - overly high taxation*, *surpuissant - overly powerful*) and *too much/too many* (e.g. *surendettement - too much debt*, *suremploi - too many jobs*). Interestingly, the typical EXCESS periphrases uncovered for *sur-* make it possible to disambiguate the two sub-meanings of the GOOD value (EXCESS and SUPERIORITY) for other, less frequent prefixes. A case in point is *ultra-*: $ultra_{\text{EXCESS}}$ is paraphrased as *excessively* (e.g. *ultra-échangisme - excessively free market*[2]) while $ultra_{\text{SUPERIORITY}}$ is rather rendered as *highly* 'very/to a high level' (e.g. *domaine ultrasensible - highly sensitive area*; *centres ultraspécialisés - highly specialized centers*). The same observation also holds for *hyper-* (compare *propositions hyper dirigistes - highly authoritarian proposals* with *hyperflexibilité - excessive flexibility* or *hyperconcentration - excessive concentration*). Another recurrent periphrastic translation pattern, which involves the adverb *extremely*, is found for the SUPERIORITY sub-meaning of the GOOD value with the prefixes *hyper-*, *super-* and *ultra-* (e.g. *hyper dangereux - extremely dangerous*, *superqualifié – extremely qualified*, *ultrasensible - extremely sensitive*). These corpus findings, provided they are confirmed in larger-scale studies relying on other text types, could help refine Guilbert's (1971: L) distinction mentioned above between the set of SUPERIORITY (i.e. higher degree) prefixes (*archi-*, *extra-*, *super-* and *ultra-*) and the two EXCESS prefixes *hyper-* and *sur-*, as in our dataset periphrastic translations show that *ultra-* and *hyper-* are used to convey both SUPERIORITY and EXCESS, while we find that *sur-* is mainly

[2] Fr. *libre-échangisme* is translated into En. *free market* and *free trade* in the Europarl corpus.

used to convey EXCESS.[3]

6 Concluding remarks

In addition to providing exploratory corpus-based insights into French evaluative prefixation (such as the low/high frequency of prefixes), our study has made it possible to confirm the usefulness of translations derived from parallel corpora as semantic evidence in morphology. However, it should be borne in mind that in cases where French evaluative prefixes are not paraphrased in the English target texts, semantic categorization cannot easily be performed on the basis of bilingual translation data.

Our study also shows how NLP can contribute to the 'translations as evidence for semantics' approach by making it possible to automatically extract and align French evaluative prefixes and their English translation equivalents. Taking into account the findings of this pilot study will undoubtedly help us improve the alignment rate of our tailor-made program. In addition, the study demonstrates that prefixes are useful anchor points for automatic alignment at word level.

The aligned bilingual dataset analyzed here can be further explored in contrastive or translation studies to offer new, corpus-based insights into French-English word-formation. These, in turn, can then be used in applied fields such as machine or computer-assisted translation, bilingual e-lexicography and second/foreign language learning/teaching. In follow-up studies, we will address the use of the bilingual data obtained in this study in some of these fields.

References

Banea, C. and Mihalcea, R. (2011). Word Sense Disambiguation with Multilingual Features. In *Proceedings of the International Conference on Computational Semantics* (ICCS 2011), Oxford, UK, January 2011, 25-34.

Bauer, L. (1997). Evaluative Morphology: In Search of Universals. *Studies in Language*, 21(3): 533-575.

Cartoni, B. (2008). *De l'incomplétude lexicale en traduction automatique : Vers une approche morphosémantique multilingue*. PhD thesis. Université de Genève: Genève.

Cartoni, B. and Meyer, T. (2012). Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), May 2012, Istanbul, Turkey.

Cartoni, B. and Namer, F. (2012). Linguistique contrastive et morphologie : Les noms en *-iste* dans une approche onomasiologique. In *Actes du Congrès Mondial de Linguistique Française* (CMLF 2012), July 2012, Lyon, France, 1245-1259.

Fradin, B. and Montermini, F. (2009). La morphologie évaluative. In Fradin, B., Kerleroux, F. and Plénat, M. (eds.), *Aperçus de morphologie du français*. Saint Denis: PUV, 231-266.

Grandi, N. (2002). *Morfologie in contatto. Le costruzioni valutative nelle lingue del Mediterraneo*. Milan: FrancoAngeli.

Grandi, N. and Montermini, F. (2005). Prefix-Suffix Neutrality In Evaluative Morphology. In Booij, G., Guevara, E., Ralli, A., Sgroi, S. and Scalise, S. (eds.), *Morphology and Linguistic Typology*. *On-line*

[3] It seems that *sur-* denotes a HIGHER DEGREE only in a few rare cases, when its base refers to money (e.g. *surcoût* - *additional cost*, *higher cost*). In some of these cases, however, an EXCESS periphrase is also found for the same prefixed word (e.g. *surcoût* - *excess cost*).

Proceedings of the Fourth Mediterranean Morphology Meeting (MMM4), September 2003, Catania, University of Bologna, 2005, 143-156.

Guilbert, L. (1971). De la formation des unités lexicales. *Grand Larousse de la langue française*. Paris: Larousse. I: IX-LXXXI.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, 79-86.

Körtvélyessy, L. and Stekauer, P. (eds.) (2011). *Diminutives and Augmentatives in the Languages of the World*. *Lexis*, 6.

Noël, D. (2003). Translations as evidence for semantics: An illustration. *Linguistics*, 41(4): 757-785.

Och, F.J. and Ney, H. (2000). Improved Statistical Alignment Models. In *Proceedings of ACL*, 440-447.

Papinemi, K., Roukos, S., Ward, T., Henderson, J. and Reeder, F. (2002). BLEU: a method for automatic evaluation of Machine Translation. In *Proceedings of ACL*, 311-318.

Stump, G. T. (1993). How peculiar is evaluative morphology? *Journal of Linguistics*, 29: 1-36.

Wierzbicka, A. (1991). *Cross-cultural pragmatics: The semantics of human interaction*. Berlin: Mouton de Gruyter.