LiLT volume 11, issue 6 December 2014

Evaluative prefixes in translation: From automatic alignment to semantic categorization

Marie-Aude Lefer[1,2] and Natalia Grabar[3]

This article aims to assess to what extent translation can shed light on the semantics of French evaluative prefixation by adopting Noël (2003)'s 'translations as evidence for semantics' approach. In French, evaluative prefixes can be classified along two dimensions (*cf.* (Fradin and Montermini 2009)): (1) a quantity dimension along a maximum/minimum axis and the semantic values BIG and SMALL, and (2) a quality dimension along a positive/negative axis and the values GOOD (EXCESS; HIGHER DEGREE) and BAD (LACK; LOWER DEGREE). In order to provide corpus-based insights into this semantic categorization, we analyze French evaluative prefixes alongside their English translation equivalents in a parallel corpus. To do so, we focus on periphrastic translations, as they are likely to 'spell out' the meaning of the French prefixes. The data used were extracted from the Europarl parallel corpus (Koehn 2005; Cartoni and Meyer 2012). Using a tailor-made program, we first aligned the French prefixed words with the corresponding word(s) in English target sentences, before proceeding to the evaluation of the aligned sequences and the manual analysis of the bilingual data. Results confirm that translation data can be used as evidence for semantics in morphological research and help refine existing semantic descriptions of evaluative prefixes.

Keywords: contrastive morphology, evaluative prefixation, seman-

[1] Institut libre Marie Haps (Translation - Interpreting), 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@ilmh.be, natalia.grabar@univ-lille3.fr

LiLT Volume 11, Issue 6, December 2014.
Theoretical and Computational Morphology: New Trends and Synergies.

tics, translation, parallel corpora, Natural Language Processing, automatic alignment, French, English

1 Evaluative morphology

This article deals with French evaluative prefixation and aims to determine to what extent translation data can shed light on the semantics of evaluative prefixes by adopting Noël (2003)'s 'translations as evidence for semantics' approach. The present section is a brief introduction to evaluative morphology, especially prefixation, both from a monolingual (French) and cross-linguistic (typological and contrastive) perspective.

Evaluative morphology, *i.e.* morphological processes used to express augmentation, diminution, endearment/approval and contempt/pejoration, has been quite extensively discussed in the field of morphological typology (see *e.g.* (Stump 1993; Dressler and Merlini Barbaresi 1994; Bauer 1997; Körtvélyessy and Stekauer (eds)). Topics addressed include, among others, prefix-suffix neutrality in quantitative evaluation (Grandi & Montermini, 2005), the diachronic developments of augmentative and diminutive suffixes (Grandi 2011) and phonetic iconicity in evaluative morphology (*e.g.* (Körtvélyessy 2011)). Typically, these typological studies rely on data collected from grammars or questionnaires submitted to native speakers of the languages under consideration and examine large samples of languages (*e.g.* 55 languages in (Grandi and Montermini 2005)).

Corpus-based descriptions of evaluative morphology, by contrast, are still sorely lacking for many languages and language pairs, including French and the English-French pair.[4] A first attempt at an exhaustive inventory and general discussion of French evaluative prefixation is found in (Fradin and Montermini 2009) (in addition to prefixation, this study also deals with the -ET suffixation). One of the important insights offered by (Fradin and Montermini 2009)'s overview is that in French, like in many other languages, evaluative prefixes can be classified along the following two dimensions (see also (Grandi 2002) and (Amiot 2004); the sub-categories are taken from (Guilbert 1971: p. L) and (Cartoni 2008: p. 287-291)).

- The quantity dimension along a maximum/minimum axis (so-called 'measurativity') and the two semantic values BIG and SMALL:
    - BIG: increase, abundance

[4] In contrastive linguistics, an exception is Andor (2005)'s corpus-based study of English *super-*, *hyper-*, *mega-* and *ultra-* and their Hungarian equivalents. It is based on the Bank of English and the Hungarian National Corpus.

  - SMALL: decrease
- The quality dimension along a positive/negative axis (so-called 'appreciability') and the two semantic values GOOD and BAD:
  - GOOD: excess, higher degree
  - BAD: lack, lower degree

These two dimensions, along with their corresponding French prefixes, are graphically represented in Figure 1.

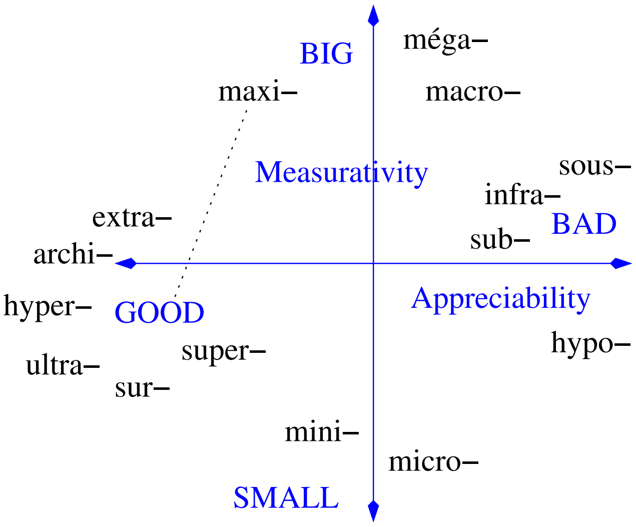

FIGURE 1 French evaluative prefixes along the quantity and quality dimensions (taken from Fradin and Montermini (2009)).

The borders between these two dimensions and between the submeanings of the BIG/SMALL and GOOD/BAD values, however, are not watertight. In fact, as pointed out by (Fradin and Montermini 2009: p.241), semantic shifts are commonly observed, both between evaluative prefixes and other semantic categories of prefixes, such as location (*e.g. extra-territorial, superstructure, surveste*)[5], and within the category of evaluative prefixes itself (*e.g.* from BIG to GOOD with *méga-* and *maxi-*, *cf.* the dotted line in Figure 1)[6].

Within the group of GOOD prefixes, (Guilbert 1971: p. L) makes a distinction between a set of prefixes conveying HIGHER DEGREE (*viz. 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 EXCESS (*viz. hyper-* and *sur-*, as in *hyperémotivité,*

[5] These examples are taken from Amiot (2004).

[6] However, according to (Amiot 2004), these two prefixes rarely convey higher degree (*e.g. maxi-bronzage, méga-fête*). This use seems to be restricted to advertising and teenagers' language.

hypernerveux, suralimentation, surpeuplé, surestimer) but the question is whether this is a sharp distinction or whether some of the GOOD prefixes can express both meanings. For instance, in (Amiot 2004), both *hyper-* and *sur-* are considered to convey HIGHER DEGREE (Fr. *haut degré*; *e.g. hypersensible, surexcité*) alongside EXCESS. In fact, Amiot (2004) distinguishes between EXCESS uses, which are typically found with nominal bases (*hypertension, suralimentation*) and verbal bases (*surévaluer*) and which have *hypo-* and *sous-* counterparts, and HIGHER DEGREE uses, often found with adjectival bases (*hyperraffiné, surexcité*). This is only a general trend, with exceptions (*e.g.* $surexposé_{excess}$). Moreover, Amiot (2004) notes that *hyper-* and *sur-* are more commonly used to express EXCESS.

2 Aims of the present study

In this study, we aim to revisit the current descriptions of French evaluative prefixes by providing corpus-based insights into the use and semantics of these word-forming elements. Special attention will be given to the semantic 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 innovative 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 (2003)'s 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) (see also (Dyvik 1998)).

Similar approaches have been adopted in computational semantics for monolingual word sense disambiguation tasks (Dagan et al. 1991; Diab and Resnik 2002; Ide et al. 2002; Ng et al. 2003; Tufis et al. 2004; Navigli 2009). This research field has been recently grouped around a specific SemEval competition task on cross-lingual word sense disambiguation (Lefever and Hoste 2010), whose aim is to evaluate the efficiency of automatic natural language processing (NLP) systems. The underlying hypothesis is that word sense ambiguities in a given language can be resolved by using translations in other languages. This hypothesis can be verified thanks to the growing availability of parallel and aligned corpora. Banea and Mihalcea (2011), for example, show that, depending on its context of occurrence, the English noun *plant* can be translated as French *plante* ('living thing in soil') or *usine* ('factory'). Taking into account the French translations of *plant* makes it possible to disambiguate the word in the source text. Banea and Mihalcea (2011)'s multilingual approach improves the overall results of word sense disambiguation by up to 25%. Improving the quality of lexicon bootstrapping in one language using translations in other languages is another application of the method proposed: for instance, Ziering et al. (2013) showed that the results with German and English data were improved by 25%. Similarly, in dictionary-based morphological research, Cartoni and Namer (2012) examine the semantics of Fr. *-iste* and It. *-ista* agent nouns by relying on translation data drawn from an Italian-French bilingual dictionary. To our knowledge, however, Noël (2003)'s approach has not yet been used in corpus-based morphological studies. This is precisely what we aim at here: assessing the potential benefits of using the 'translations as evidence for semantics' approach in the field of word-formation. In doing so, we pay particular attention to non-morphological translations, such as periphrastic translations (as opposed to translations with 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. It is important to note from the outset that we will not be looking into the specific semantics of individual prefixed lexemes. Rather, by focusing on recurrent periphrastic translation patterns, we hope to shed light on the general semantic features of French evaluative prefixes.

Our study takes stock of insights from theoretical and empirical linguistics (morphology, lexical semantics and corpus linguistics), NLP and translation studies. Sections 3 and 4 present the empirical data on which the study is based and the extraction and alignment method we used, respectively. We then propose an overview of the results in Section 5. The article ends with some concluding remarks in Section 6.

3 Data

3.1 French evaluative prefixes

The data used in this study were extracted from a French-to-English parallel corpus (see Section 3.2) on the basis of an inventory of French evaluative prefixes (*cf.* (Fradin and Montermini 2009; Cartoni 2008: p. 240)). Clearly delineating which processes belong to evaluative morphology and which do not is a tricky issue. As rightly pointed out by (Bauer 1997: p. 538), "*although diminutives and augmentatives may provide the core of evaluative morphology, its borders are rather imprecise*". Following (Cartoni 2008: p. 131-135), the attenuation prefixes *demi-*, *mi-* and *semi-* are considered to be part of evaluative prefixation, while in (Fradin and Montermini 2009), they are viewed as quantitative prefixes. The two approximation prefixes *quasi-* and *pseudo-* are also included in our set of evaluative prefixes (*cf.* (Cartoni 2008)). The complete list of prefixes investigated here is given in Table 1.

Semantic value	Prefixes and examples
[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*

TABLE 1 Set of prefixes investigated.

3.2 Corpus used

The study is based on the Europarl parallel corpus of parliamentary debates (Koehn 2005) and more particularly the 'directional' Europarl version made available by Cartoni and Meyer (2012), where the source and target languages, and hence the translation direction, are clearly identified (see Cartoni et al. (2013) for an overview of the use of the corpus in contrastive linguistics and translation studies). The corpus is aligned at sentence level and each pair of aligned sentences has its own identifier. In this study, we relied on a French-to-English subcorpus of Europarl containing 7,878 parallel documents, which corresponds to approximately 10 million running words. The reason for using Europarl is that large and representative French-to-English translation corpora are still sorely lacking. Naturally, a cautionary remark is in order: the results presented in Section 5 below are only valid for the text type under investigation. The analysis of other text types (*e.g.* news items, novels, instruction manuals, research articles, forums) might lead to different, though complementary, trends.

3.3 Bilingual lexicon

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*}, {*sous*, *under*}.

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. analysis of the bilingual data (classification according to the translation strategies used to render the French prefixes).

The first two steps (extraction and alignment) involve both automatic and manual data processing, while the third and fourth steps (evaluation and analysis) are completely manual. Steps 1 and 2 are detailed in the following paragraphs.

The automatic part of the first step consists in projecting the prefixes on the source language sentences and spotting the words that potentially 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. In doing so, 66,398 sentence pairs were extracted. 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*),
- prefixes can be 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 of step 1 consists in filtering the list of the automatically extracted words in order to discard the ones that are clearly not morphologically prefixed, even though formally they contain a prefix-like initial string (*e.g.* *extracteur, maximal, miette*).

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 word-alignment tool `GIZA++` (Och and Ney 2000) and with a tailor-made alignment program. We wanted to compare these two tools because, as will be clear below, they rely on two different approaches for the detection of words or segments to be aligned. In this study we used the aligner that offered the best coverage (see section 5.1). `GIZA++` applies several statistical alignment models, such as `IBM-4`, `IBM-5` and `HMM`. The main clues for aligning at word level are provided by the contexts of use of the words. As for the tailor-made program, it relies on lexical information and the presence of cognates. It also applies several heuristics. For instance, 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.* {*ultralibérales*, *ultraliberal*};

(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 ultrasensible*, *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.* {*surpêche*, *over-fishing*}, {*sous-développement*, *underdevelopment*} or {*demi-mesures*, *half-measures*};

(d) exploration of the neighboring context of the prefixed word in the source sentence and detection of the corresponding word(s) 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 adjusted and validated the extracted segments and completed the alignment when no alignment could be performed automatically. The aligned, contextualized data also allow the irrelevant data to be filtered, *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*}, {*ultrapériphériques*, *outermost*}). We also weeded out numerous cases where the prefix-like string, which was separated from the following word by a space, did not function as a prefix but rather as another part of speech (*e.g. demi* in *deux ans et demi* ('two years and half'), *micro* in *intervention hors micro* ('off-microphone intervention')).

The relevant and corrected alignments were then evaluated by means of the BLEU precision measure (Papinemi et al. 2002). This measure is widely used for the evaluation of automatic machine translation results. Quality is considered to be the correspondence between the automatic output and human translation. The measure has often been reported to correlate quite significantly with human judgment (Coughlin 2003). It consists in counting the number of words in the automatically aligned target sequence and in the adjusted target sequence: the percentage of common words between them corresponds to the BLEU measure. The values of the BLEU score range between 1 (perfect alignment) and 0. For example, consider the following cases:

- In {*mini-traité*, *mini-treaty*}, {*microclimat*, *microclimate*}, {*micro-entreprises*, *micro-companies*}, {*ultratechnique*, *ultra-technical*}, all the extracted and aligned words are correct, which gives a precision rate of 1;
- In {*ultrasimple*, *a wholly simple system*}, two words (*viz. wholly simple*) out of four are correct, which results in a precision rate of 0,5 (2/4);
- In {*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, which gives a precision rate of 1/7=0.14;
- If no alignment is proposed, the precision rate is 0.

The scores were first computed for each aligned segment and then averaged on the whole dataset.

5 Results and discussion

5.1 Assessing the automatic alignment method

For the 4,574 prefixed words extracted from the French source sentences, `GIZA++` and the tailor-made program generated 2,268 and 3,566 alignments respectively. The alignment rates show that `GIZA++` generated alignments for almost 50% of the data, while the tailor-made heuristics aligned c. 80% of the data. In view of this difference in alignment rate, we chose to work with the results provided by the tailor-made program. Among them, we found:

(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 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 rate on the target sequences is 0.76, which is satisfactory in view of comparable automatic alignment tasks. For instance, a similar task was undertaken during the Word Alignment challenge held in 2003 (see (Mihalcea and Pedersen 2003)). On English-French data with null alignments allowed, the precision rates varied between 0.37 and 0.72 when evaluated against the set of sure translations from the reference set. When probable translations from the reference set were also considered, the precision rate reached 0.77. As expected, the performance of the tailor-made program in this study decreases when the prefix is separated by a blank from the corresponding base word (in source or target segments), as well as when the translation of the prefix is non-morphological. After a final deduplication phase, we were left with 1,985 validated bilingual segments.

5.2 'Translation as evidence for semantics' in morphology: general trends

Sur-, *sous-*, *quasi-*, *ultra-* and *super-* are found to be the most frequent evaluative prefixes in the Europarl corpus, with more than 200 tokens and at least 55 different types each (see Table 2). The other prefixes, however, appear to be (very) infrequent: *hyper-/archi-* (GOOD), *macro-/méga-* (BIG), *micro-/mini-* (SMALL), *pseudo-* (APPROXIMATION), *demi-/semi-/mi-* (ATTENUATION) (no occurrences of evaluative *hypo-*, *maxi-* and *sub-* were found in Europarl). It is also important to add that some of them occur in a very limited set of prefixed words (*e.g. macro-* in only 13 types, such as *macroéconomique* and *macrofinancier*).

Table 3 provides a summary of the translation patterns observed in English target texts. We have made a distinction between the following translation strategies for the French prefixed words found in the

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

TABLE 2 Breakdown per evaluative prefix in Europarl (tokens and types).

Europarl dataset:

- translation with a derivative containing an evaluative prefix, *e.g.* {*sous-estimer*, *underestimate*};
- translation with a derivative containing a non-evaluative prefix, *e.g.* {*sous-utilisé*, *unused*};
- translation with a non-prefixed word (which can be a simplex word, a suffixed word or a compound), *e.g.* {*sous-alimenté*, *starving*}, {*sous-équipé*, *ill-equipped*}, {*surpoids*, *obesity*},
- translation with a periphrasis, *e.g.*{*ultra-concurrence*, *competition taken to extremes*}, {*hyper-fédéraliste*, *extremely federalist*};
- zero translation, when the prefixed word is not translated in the target text.

As can be seen in Table 3, the vast majority of words containing an evaluative prefix in French are translated with morphologically and semantically similar words in English (1,459 instances; 73.5%), which is rather unsurprising in view of the fact that the two languages share a large number of evaluative prefixes. Periphrastic translations account for c. one fourth of the data (453 instances; 22.8%). The other translation strategies, by contrast, are rather infrequent. Naturally, these are only overall trends. Noticeable differences can be observed between prefixes. For instance, while *super-* displays a mere 2.9% of periphrastic translations (it is mainly translated with *super-*

Translation strategy	Total # tokens	%
Translation with an evaluative prefix (same semantic category)	1,459	73.5%
Translation with a periphrasis	453	22.8%
Translation with 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 with a non evaluative prefix (another semantic category)	5	0.3%

TABLE 3 Overview of translation strategies in Europarl.

in English, *e.g.* {*superpuissance*, *superpower*}), *quasi-* is translated by means of a periphrasis in almost 90% of instances despite the existence of the English prefix *quasi-* (*e.g.* {*quasi-épave*, *virtual wreck*}, {*quasi-identique*, *almost identical*}, {*quasi-général*, *more or less general*}, {*quasi-unanime*, *practically unanimous*}). *Sous-*, like *super-*, is also rarely paraphrased (3.9%), with *ultra-* and *sur-* in between the two extremes (12.6% and 27.1% of periphrases, respectively). The exact breakdown for the top 5 most frequent prefixes is given in Table 4.

Translation strategy	*sur-*		*sous-*		*quasi-*		*ultra-*		*super-*	
	#	%	#	%	#	%	#	%	#	%
Derivative containing an evaluative prefix	321	64.8	286	93.2	25	9.5	201	87.4	201	95.7
Derivative containing a non-evaluative prefix	1	0.2	1	0.3	0	0	0	0	0	0
Non-prefixed word	37	7.5	7	2.3	3	1.1	0	0	2	0.9
Periphrasis	**134**	**27.1**	**12**	**3.9**	**233**	**88.9**	**29**	**12.6**	**6**	**2.9**
Zero translation	2	0.4	1	0.3	1	0.4	0	0	1	0.5

TABLE 4 Translation strategies for the prefixed words containing the top 5 most frequent evaluative prefixes in Europarl.

In the remainder of this article, we will restrict our discussion to (recurrent) periphrastic translations and leave aside the other translation patterns where the meaning of the source language prefix is not explicitly spelled out in target texts. It turns out that *sur-* is the best candidate to assess the full potential of the 'translations as evidence for semantics' approach because it is the most frequent evaluative prefix in Europarl, with a substantial amount of periphrastic translations (495 validated entries, with 134 instances of periphrastic translations). Section 5.3 is therefore devoted to *sur-* and, more generally, to the set of GOOD prefixes and the HIGHER DEGREE vs. EXCESS distinction. Before zooming in on *sur-*, Table 5 offers an overview of the periphrastic translation patterns uncovered for the other semantic categories of evaluative prefixes (note that no periphrastic patterns were found for the BIG prefixes), together with their frequency of occurrence.

Semantic value	French (# tokens)	Periphrastic translation patterns (# occurrences)
ATTENUATION	*demi-N (30)*	*half a N (2), little N (1), partial N (2), partly ADJ (1)*
	semi-N/ADJ (16)	*partially ADJ (1), virtually ADJ (1)*
APPROXIMATION	*pseudo-N (35)*	*so-called N (2), would-be N (1), imaginary N (1)*
	quasi-N/ADJ/ADV (262)	*almost ADJ/ADV (135), more or less ADJ (5), practically ADJ (11), virtually ADJ/ADV (49), near N (7), virtual N (13)*
SMALL	*micro-N (131)*	*small N (2)*
BAD	*sous-N/ADJ (307)*	*inadequate N (1), lack of N (1), badly ADJ (1), insufficiently ADJ (1), less than ADJ (1)*

TABLE 5 Periphrastic translation patterns of the ATTENUATION, APPROXIMATION, SMALL and BAD prefixes in Europarl.

As appears from Table 5, the periphrastic translations found in Europarl – even though they are relatively infrequent in some cases – quite accurately reflect the evaluative meaning of the prefixes. Consider, for example, $semi_{attenuation}$ and $demi_{attenuation}$: {*en régime de semi-liberté*, *partially free*}, {*demi-solution*, *partial solution*}, {*demi-échec*, *partial failure*}, {*demi-satisfaction*, *partly satisfied*}, where the attenuative meaning of the prefixes is clearly spelled out in the translated data by means of the adjective *partial* and the adverbs *partially* and *partly* (*semi-* and *demi-* can also function as quantitative prefixes in French). Larger translation corpora and/or bilingual lexicographic data (*cf.* (Cartoni and Namer 2012)) will be needed in future studies to examine 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.

5.3 The prefix *sur-*: a window onto the EXCESS vs. HIGHER DEGREE distinction

Zooming in on the recurrent periphrastic translations of *sur-* (*i.e.* leaving aside cases where *sur-* is translated with *over-*, *e.g.* {*suradministré*, *over-administered*}, and one-off periphrastic translations), 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*}, {*surconsommation*, *excess consumption*}, {*suremballage*, *excess packaging*}, {*surpression*, *excess pressure*}, {*surréglementation*, *excessive regulation*}, {*surexposition*, *excessive exposure*}, {*surendettement*, *excessive debts*} (28 occurrences in our dataset);
- *too much/too many*, *e.g.* {*surendettement*, *too much debt*}, {*suremploi*, *too many jobs*} (6 occurrences);
- *overly* 'too', *e.g.* {*sururbanisé*, *overly built-up*}, {*surpuissant*, *overly powerful*}, {*surintensif*, *overly intensive*} (5 occurrences).

As can be seen in the examples mentioned above, *sur*$_{excess}$ mainly applies to nominal bases, with the corresponding *excess(ive) N* periphrasis in English, but it is also used on adjectival bases, where it is paraphrased as *overly ADJ* (*cf. surexposé* mentioned in Amiot (2004)).

Interestingly, the typical EXCESS periphrases uncovered for *sur-* make it possible to disambiguate the two meanings of the GOOD value (EXCESS and HIGHER DEGREE) for other, less frequent prefixes. A case in point is *ultra-*: *ultra*$_{excess}$ is paraphrased as *excessive(ly)* (*e.g.* {*ultra-échangisme*, *excessively free market*}[7], {*ultraconcurrence*, *excessive competition*}) while *ultra*$_{higher-degree}$ 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, for example, {*propositions hyper dirigistes*, *highly authoritarian proposals*} with {*hyperréglementation*, *excessive regulation*} or {*hyperconcentration*, *excessive concentration*}. Once again, the corpus data show that the EXCESS interpretation, which is commonly found with nominal bases, is also found in a few derivatives with adjectival bases: {*ultrasécuritaire*, *excessively security-conscious*}, {*des formes de travail hyperflexibilisées*, *excessively*

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

flexible types of work}. Another recurrent periphrastic translation pattern, which involves the adverb *extremely*, is found for the HIGHER DEGREE meaning of the GOOD value with *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, help refine (Guilbert 1971: p. L)'s distinction between the set of HIGHER DEGREE prefixes (*archi-*, *extra-*, *super-* and *ultra-*) and the two EXCESS prefixes *hyper-* and *sur-*, or Amiot (2004)'s overview of GOOD prefixes (where no mention is made of the EXCESS meaning of *ultra-*). In our dataset periphrastic translations show that *ultra-* and *hyper-* are used to convey both HIGHER DEGREE and EXCESS, while *sur-* is mainly used to convey EXCESS. It should be added that *sur-* seems to denote some kind of HIGHER DEGREE only in a few rare cases, *viz.* when its base refers to money (*e.g.* {*surcoûts*, *additional costs/higher costs*}, {*surpéage*, *additional toll*}, {*surprime*, *additional premium*}, {*surtaxation*, *higher taxes*}). In some of these cases, however, an EXCESS periphrasis is also found for the same prefixed word (*e.g.* {*surcoût*, *excess cost*}). Interestingly, for *sur*-derivatives such as *surtaxe*, *surloyer* and *surpaie*, Amiot (2004) proposes another (yet marginal) sub-meaning of the GOOD value, *viz.* ACCUMULATION. The periphrastic translation pattern with the adjective *additional* uncovered in our study brings support to this semantic interpretation.

In view of the lack of any systematic correspondence between nominal bases and the EXCESS meaning on the one hand and adjectival bases and the HIGHER DEGREE meaning on the other, relying on translation data and adopting a multilingual approach to morphology turns out to be a promising way to explore the semantics of evaluative prefixes conveying the GOOD value.

6 Conclusion and perspectives

Our study has relied on the cross-fertilization of different research fields, *i.e.* morphology, contrastive linguistics, translation studies and NLP. We hope to have shown that these fields can each contribute to common objectives, thanks to their various approaches, methods and tools.

In addition to providing some basic, exploratory corpus-based insights into French evaluative prefixation (such as the low/high frequency of prefixes in Europarl), 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 has also shown 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 contrastive findings presented above will undoubtedly help us improve the alignment rate of our tailor-made program. In the present study, we experimentally relied on a small set of known translation equivalents for some prefixes (a total of six French-English prefix pairs were used). This can be enriched in future work using more prefix pairs or other types of recurring translation equivalents (*e.g.* adverbs found in frequent periphrastic patterns). We believe that this will improve the coverage of the results generated here. In addition, the study demonstrates that prefixes are useful anchor points for automatic alignment at word level (*cf.* (Simard et al. 1992; Kondrak et al. 2003)). However, it is important to note that if, unlike French and English, the languages investigated are morphologically distant (*e.g.* French and Japanese, English and Hungarian), anchor points are less useful as no or little common morphological and semantic regularities can be found. For such pairs of languages, statistical approaches, such as those implemented in `GIZA++`, would probably perform better. As regards the filtering of words that are not morphologically prefixed, although they formally contain a prefix-like initial string, the manual filtering step will be sped up in future work by further controlling the extraction methodology: if the prefix-like string is not followed by a word-like string listed in available lexicons, it should be ruled out.

In addition to improving our alignment program, an obvious follow-up study would consist in complementing the analyses presented here by:

(i) looking at cases where, despite the lack of any evaluative prefix in French source texts, an evaluative prefix is found in English target texts, and

(ii) reversing the approach, *i.e.* examining the French periphrastic translations of English evaluative prefixes.

This would undoubtedly help us refine the exploratory results presented in this study. Needless to say, other text types would ideally need to be taken into account, which currently proves to be difficult in view of the lack of large parallel corpora other than Europarl (or similar corpora emanating from international or European institutions).

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. For example, the dataset analyzed in this study (recurrent periphrastic translation patterns, authentic corpus examples) have been used to inform an online bilingual dictionary of French and English affixes, the MuLeXFoR prototype[8] (see (Cartoni and Lefer 2010) for a general presentation of the tool). In addition, the bilingual dataset could be used to enrich existing electronic dictionaries, which have often been described as being non-exhaustive, and could be used for machine translation. This would help show the impact of the generated results on the performance of machine translation systems.

Acknowledgments

The present study was carried out within the WOG Contragram framework. The two authors thank the anonymous reviewers and the guest-editors of the special issue for their helpful comments and suggestions.

References

Amiot, D. 2004. Haut degré et préfixation. intensité, comparaison, degré. *Travaux linguistiques du Cerlico* 17:91–104.

Andor, J. 2005. A lexical semantic-pragmatic analysis of the meaning potentials of amplifying prefixes in english and hungarian. A corpus-based case study of near synonymy. In *The Corpus Linguistics Conference Series 1(1)*. Available online: *http://www.corpus.bham.ac.uk/PCLC/*.

Banea, C and R Mihalcea. 2011. Word sense disambiguation with multilingual features. In *International Conference on Computational Semantics (ICCS 2011)*, pages 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 MA Lefer. 2010. Improving the representation of word-formation in multilingual lexicographic tools: the MuLeXFoR database. In *XIV Euralex International Congress*, pages 581–591.

Cartoni, B and T Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *8th International Conference on Language Resources and Evaluation (LREC)*.

[8] See MuLeXFor-Marie Haps: *https://sites.google.com/site/mulexfor/*

Cartoni, B and F Namer. 2012. Linguistique contrastive et morphologie : les noms en -iste dans une approche onomasiologique. In *CMLF*, pages 1245–1259.

Cartoni, B, S Zufferey, and T Meyer. 2013. Using the europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics* 27:23–42.

Coughlin, D. 2003. Correlating automated and human assessments of machine translation quality. In *MT Summit IX*, pages 23–27.

Dagan, Ido, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *ACL*, pages 130–137.

Diab, M and P Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *ACL*, pages 255–262.

Dressler, WU and L Merlini Barbaresi. 1994. *Morphopragmatics: Diminutives and Intensifiers in Italian, German and Other Languages*. Berlin: Mouton de Gruyter.

Dyvik, H. 1998. A translational basis for semantics. In S. Johansson and S. Okselfjell, eds., *Corpora and Crosslinguistic Research: Theory, Method and Case Studies*, pages 51–86.

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

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

Grandi, N. 2011. Renewal and innovation in the emergence of indo-european evaluative morphology. In L. Körtvélyessy and P. Stekauer, eds., *Diminutives and Augmentatives in the Languages of the World*, pages 5–26.

Grandi, N and F Montermini. 2005. Prefix-suffix neutrality in evaluative morphology. In G. Booij, E. Guevara, A. Ralli, S. Sgroi, and S. Scalise, eds., *Fourth Mediterranean Morphology Meeting (MMM4)*, pages 143–156.

Guilbert, L. 1971. De la formation des unités lexicales. In P. Larousse, ed., *Grand Larousse de la langue franaise*, pages IX–LXXXI.

Ide, N, T Erjavec, and D Tufis. 2002. Sense discrimination with parallel corpora. In *ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, pages 79–86.

Kondrak, G, D Marcu, and K Knight. 2003. Cognates can improve statistical translation models. In *HLT-NAACL 2003*, pages 46–48.

Körtvélyessy, L. 2011. A cross-linguistic research into phonetic iconicity. In L. Körtvélyessy and P. Stekauer, eds., *Diminutives and Augmentatives in the Languages of the World*, pages 27–40.

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

Lefever, Els and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *SemEval 2010*, pages 82–87.

Mihalcea, R and T Pedersen. 2003. An evaluation exercise for word alignment. In *Workshop Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10.

Navigli, R. 2009. Word sense disambiguation: a survey. *ACM Computing Surveys* 42:1–69.

Ng, HT, B Wang, and YS Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *ACL*, pages 455–462.

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

Och, FJ and H Ney. 2000. Improved statistical alignment models. In *ACL*, pages 440–447.

Papinemi, K, S Roukos, T Ward, J Henderson, and F Reeder. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Simard, Michel, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*. Montréal, Canada. Available online *http://www-rali.iro.umontreal.ca/Publications/sfiTMI92.ps/*.

Stump, GT. 1993. How peculiar is evaluative morphology? *Journal of Linguistics* 29:1–36.

Tufis, Dan, Radu Ion, and Nancy Ide. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *COLING*, pages 1312–1318.

Ziering, Patrick, Lonneke van der Plas, and Hinrich Schütze. 2013. Multilingual lexicon bootstrapping. Improving a lexicon induction system using a parallel corpus. In *International Joint Conference on Natural Language Processing*, pages 844–848.