HAL open science

Identification of Fertile Translations in Medical Comparable Corpora: a Morpho-Compositional Approach

Estelle Delpech, Béatrice Daille, Emmanuel Morin, Claire Lemaire

▶ **To cite this version:**

Estelle Delpech, Béatrice Daille, Emmanuel Morin, Claire Lemaire. Identification of Fertile Translations in Medical Comparable Corpora: a Morpho-Compositional Approach. AMTA, Oct 2012, San Diego, CA, United States. pp.10. hal-00730325

HAL Id: hal-00730325

https://hal.science/hal-00730325v1

Submitted on 11 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Identification of Fertile Translations in Medical Comparable Corpora : a Morpho-Compositional Approach

Estelle Delpech, Béatrice Daille, Emmanuel Morin
Université de Nantes - LINA FRE CNRS 2729
2 rue de la Houssinière BP 92208
44322 Nantes Cedex 3, France
{name.surname}@univ-nantes.fr

Claire Lemaire
Lingua et Machina
c/o Inria Rocquencourt BP 105
Le Chesnay Cedex 78153, France
cl@lingua-et-machina.com

Abstract

This paper defines a method for lexicon in the biomedical domain from comparable corpora. The method is based on compositional translation and exploits morpheme-level translation equivalences. It can generate translations for a large variety of morphologically constructed words and can also generate 'fertile' translations. We show that fertile translations increase the overall quality of the extracted lexicon for English to French translation.

1 Introduction

Comparable corpora are composed of texts in different languages which are not translations but deal with the same subject matter and were produced in similar situations of communication so that there is a possibility to find translation pairs in the texts. Comparable corpora have been used mainly in the field of Cross-Language Information Retrieval and Computer-Aided Translation (CAT). In CAT, which is our field of application, comparable corpora have been used to extract domain-specific bilingual lexicons for language pairs or subject domains for which no parallel corpora is available. Another advantage of comparable corpora is that they contain more idiomatic expressions than parallel corpora do. Indeed, the target texts of parallel corpora are translations and bear the influence of the source language whereas the target texts of comparable corpora are original, spontaneous productions. The main drawback of comparable corpora is that much fewer translation pairs can be extracted than in parallel corpora because (i) not all source language terms do have a translation in the target texts and (ii) when there is a translation, it may not be present in its canonical form, precisely because the target texts are not translations. As observed by Baker (1996), translated texts tend to bear features like explication, simplification, normalization and leveling out. For instance, an English-French comparable corpus may contain the English term *post-menopausal* but not its "normalized" or "canonical" translation in French (*post-ménopausique*). However, there might be some morphological or paraphrastic variants in the French texts like *post-ménopause 'post-menopause'* or *après la ménopause 'after the menopause'*. The solution that consists in increasing the size of the corpus in order to find more translation pairs or to extract parallel segments of text (Fung and Cheung, 2004; Rauf and Schwenk, 2009) is only possible when large amounts of texts are available. In the case of the extraction of *domain-specific* lexicons, we quickly face the problem of data scarcity: in order to extract high-quality lexicons, the corpus must contain text dealing with very specific subject domains and the target and source texts must be highly comparable. If one tries to increase the size of the corpus, one takes the risk of decreasing its quality by lowering its comparability or adding out-of-domain texts. Studies support the idea that the quality of the corpora is more important than its size. Morin et al. (2007) show that the discourse categorization of the documents increases the precision of the lexicon despite the data sparsity. Bo and Gaussier (2010) show that they improve the quality of a lexicon if they improve the comparability of the corpus by selecting a smaller - but more comparable - corpus from an

initial set of documents. Consequently, one solution for increasing the number or translation pairs is to focus on identifying translation variants. This paper explores the feasibility of identifying "*fertile*" translations in comparable corpora. In parallel texts processing, the notion of fertility has been defined by Brown et al. (1993). They defined the fertility of a source word e as the number of target words to which e is connected in a randomly selected alignment. Similarly, we call a fertile translation a translation pair in which the target term has more words than the source term. We propose to identify such translations with a method mixing morphological analysis and compositional translation : (i) the source term is decomposed into morphemes: *post-menopausal* is split into *post-* + *menopause*[1] ; (ii) the morphemes are translated as bound morphemes or fully autonomous words: *post-* becomes *post-* or *après* and *menopause* becomes *ménopause* and (iii) the translated elements are recomposed into a target term: *post-ménopause*, *après la ménopause*.

This paper falls into 4 sections. Section 2 outlines recent research in compositionality-based lexicon extraction. Section 3 explains the algorithm of morpho-compositional translation. Experimental data and results and described in sections 4 and 5.

2 Related work

Most of the research work in lexicon extraction from comparable corpora concentrates on same-length term alignment. To our knowledge, only Daille and Morin (2005) and Weller et al. (2011) tried to align terms of different lengths. Daille and Morin (2005) focus on the specific case of multi-word terms whose meanings are non-compositional and tried to align these multi-word terms with either single-word terms or multi-word terms using a context-based approach[2].Weller et al. (2011) concentrate on aligning German NOUN-NOUN compounds to NOUN NOUN and NOUN PREP NOUN structures in French and English.

We chose to work in the framework of compositionality-based translation because: (i) compositional terms form more than 60% of the new terms found in techno-scientific domains, and especially in the field of biomedecine (Namer and Baud, 2007) (ii) compositionality-based methods have been shown to clearly outperform context-based ones for the translation of terms with compositional meaning (Morin and Daille, 2010) (iii) we believe that compositionality-based methods offer the opportunity to generate fertile translations if combined with a morphology-based approach.

2.1 Principle of compositional translation

Compositional translation relies on the principle of compositionality which states that "the meaning of the whole is a function of the meaning of the parts" (Keenan and Faltz, 1985, 24-25). Applied to bilingual lexicon extraction, compositional translation ($\mathcal{CT}$) consists in decomposing the source term into atomic components ($\mathcal{D}$), translating these components into the target language ($\mathcal{T}$), recomposing the translated components into target terms ($\mathcal{R}$) and finally filtering the generated translations with a selection function ($\mathcal{S}$):

$$\begin{aligned}
&\mathcal{CT}(\text{“ab”})\\
&\quad = \mathcal{S}(\mathcal{R}(\mathcal{T}(\mathcal{D}(\text{“ab”}))))\\
&\quad = \mathcal{S}(\mathcal{R}(\mathcal{T}(\{\text{a, b}\})))\\
&\quad = \mathcal{S}(\mathcal{R}(\{\mathcal{T}(\text{a}) \times \mathcal{T}(\text{b})\}))\\
&\quad = \mathcal{S}(\mathcal{R}(\{\text{A, B}\}))\\
&\quad = \mathcal{S}(\{\text{A, B}\}, \{\text{B, A}\})\\
&\quad = \text{“BA”}
\end{aligned}$$

where "ab" is a source term composed of *a* and *b*, "BA" is a target term composed of B and A and there exists a bilingual resource linking *a* to A and *b* to B.

2.2 Implementations of compositional translation

Existing implementations differ on the kind of atomic components they use for translation.

Lexical compositional translation (Grefenstette, 1999; Baldwin and Tanaka, 2004; Robitaille et al., 2006; Morin and Daille, 2010) deals with multi-word term to multi-word term alignment and uses lexical words[3] as atomic components : *rate of evap-*

[1]We use the following notations for morphemes: trailing hyphen for prefixes (*a-*), leading hyphen for suffixes (*-a*), both for confixes (*-a-*) and no hyphen for autonomous morphemes (*a*). Morpheme boundaries are represented by a plus sign (+).

[2]Context-based methods were introduced by Rapp (1995) and Fung (1997). They consist in comparing the contexts in which the source and target terms occur. Their drawback is that they need the source and target terms to be very frequent.

[3]as opposed to grammatical words: preposition, determiners, etc.

oration is translated into French *taux d'évaporation* by translating *rate* as *taux* and *evaporation* as *évaporation* using dictionary lookup. Recomposition may be done by permutating the translated components (Morin and Daille, 2010) or with translation patterns (Baldwin and Tanaka, 2004).

Sublexical compositional translation deals with single-word term translation. The atomic components are subparts of the source single-word term. Cartoni (2009) translates neologisms created by prefixation with a special formalism called Bilingual Lexeme Formation Rules. Atomic components are the prefix and the lexical base: Italian neologism *anticonstituzionale 'anticonstitution'* is translated into French *anticonstitution* by translating the prefix *anti-* as *anti-* and the lexical base *constituzionale* as *constitution*. Weller et al. (2011) translate two types of single-word term. German single-word term formed by the concatenation of two neoclassical roots are decomposed into these two roots, then the roots are translated into target language roots and recomposed into an English or French single-word term, e.g. $Kalori_1metrie_2$ is translated as $calori_1metry_2$. German NOUN_1-NOUN_2 compounds are translated into French and English $\text{NOUN}_1\text{NOUN}_2$ or NOUN_1 PREP NOUN_2 multi-word term, e.g. $Elektronen_{N1}$-$mikroskop_{N2}$ is translated as $electron_{N1}$ $microscope_{N2}$.

2.3 Challenges of compositional translation

Compositional translation faces four main challenges which are **(i) morphosyntactic variation:** source and target terms' morphosyntactic structures are different: *anti-*$\underline{cancer}_{\text{NOUN}}$ → *anti-*$\underline{cancéreux}_{\text{ADJ}}$ *'anti-cancerous'* ; **(ii) lexical variation:** source and target terms contain semantically related - but not equivalent - words: *machine translation* → *traduction automatique 'automatic translation'* ; **(iii) fertility:** the target term has more content words than the source term: *isothermal snowpack* → *manteau neigeux isotherme 'isothermal snow mantel'* ; **(iv) terminological variation:** a source term can be translated as different target terms: *oophorectomy* → *ovariectomie 'oophorectomy', ablation des ovaires 'removal of the ovaries'*.

Solutions to morphosyntactic, lexical and to some extent terminological variation have been proposed in the form of thesaurus lookup (Robitaille et al., 2006), morphological derivation rules (Morin and Daille, 2010), morphological variant dictionaries (Cartoni, 2009) or morphosyntactic translation patterns (Baldwin and Tanaka, 2004; Weller et al., 2011). Fertility has been addressed by Weller et al. (2011) for the specific case of German NOUN-NOUN compounds.

3 Morpho-compositional translation

3.1 Underlying assumptions

Morpho-compositional translation (morpho-compositional translation) relies on the following assumptions:

Lexical subcompositionality. The lexical items which compose a multi-word term or a single-word term may be split into semantically-atomic components. These components may be either **free** (i.e. they can occur in texts as autonomous lexical items like *toxicity* in *cardiotoxicity*) or **bound** (i.e. they cannot occur as autonomous lexical items, in that case they correspond to bound morphemes like *-cardio-* in *cardiotoxicity*).

Irrelevance of the bound/free feature in translation. Translation occurs regardless of the components' degree of freedom: *-cardio-* may be translated as *cœur 'heart'* as in *cardiotoxicity* → *toxicité pour le cœur 'toxicity to the heart'*.

Irrelevance of the bound/free feature in allomorphy. Allomorphy happens regardless of the components' degree of freedom: *-cardio-, cœur 'heart', cardiaque 'cardiac'* are possible instantiations of the same abstract component and may lead to terminological variation as in *cardiotoxicity* → *cardiotoxicité 'cardiotoxicity', toxicité pour le cœur 'toxicity to the heart', toxicité cardiaque 'cardiac toxicity'*.

Like other sublexical approaches, the main idea behind morpho-compositional translation is to go beyond the word level and work with subword components. In our case, these components are morpheme-like items which either (i) bear referential lexical meaning like confixes[4] (*-cyto-, -bio-, -ectomy-*) and autonomous lexical items (*cancer, toxicity*) or (ii) can substantially change the

[4] we use the term *confix* as a synonym of neoclassical roots (Latin or Ancient Greek root words).

meaning of a word, especially prefixes (*anti-, post-, co-...*) and some suffixes (*-less, -like...*). Unlike other approaches, morpho-compositional translation is not limited to small set of source-to-target structure equivalences. It takes as input a single morphologically constructed word unit which can be the result of prefixation '*pretreatment*', confixation '*densitometry*', suffixation '*childless*', compounding '*anastrozole-associated'* or any combinations of the four. It outputs a list of n words who may or may not be morphologically constructed. For instance, *postoophorectomy* may be translated as *postovariectomie 'postoophorectomy', après l'ovariectomie 'after the oophorectomy'* or *après l'ablation des ovaires 'after the removal of the ovaries'*.

3.2 Algorithm

As an example, we show the translation of the adjective *cytotoxic* into French using a toy dataset. Let $Comp^{l}_{type}$ be a list of components in language l where $type$ equals $pref$ for prefixes, $conf$ for confixes, $suff$ for suffixes and $free$ for free lexical units ; $Trans$ be the translation table which maps source and target components ; Var^{l} be a table mapping related lexical units in language l ; $Stop^{l}$ a list of stopwords in language l ; $Corpus^{l}$ a lemmatized, pos-tagged corpus in language l:

$Comp^{en}_{conf}$ = {-cyto-} ;
$Comp^{en}_{free}$ = {cytotoxic, cytotoxicity, toxic} ;
$Comp^{fr}_{conf}$ = {-cyto-} ;
$Comp^{fr}_{free}$ = {cellule, toxique} ;
$Trans$ = {{-cyto- → -cyto-, cellule},
{toxic → toxique}} ;
Var^{en} = {cytoxic → cytoxicity} ;
$Stop^{fr}$ = {pour, le} ;
$Corpus^{fr}$ = "le/DET cytotoxicité/N être/AUX le/DET propriété/N de/PREP ce/DET qui/PRO être/AUX toxique/A pour/PREP le/DET cellule/N ./PUN" ;
'*The cytotoxicity is the property of what is toxic to the cells.*'

Morpho-compositional translation takes as input a source language single-word term and outputs zero or several target language single-word terms or multi-word terms. It is the result of the sequential application of four functions to the input single-word term: decomposition ($\mathcal{D}$), translation ($\mathcal{T}$), recomposition ($\mathcal{R}$) and selection ($\mathcal{S}$).

3.2.1 Decomposition function

The decomposition function $\mathcal{D}$ works in two steps $\mathcal{D}_1$ and $\mathcal{D}_2$.

Step 1 of decomposition ($\mathcal{D}_1$) splits the input single-word term into minimal components by matching substrings of the single-word term with the resources $Comp^{src}$, $Comp^{src}_{conf}$, $Comp^{src}_{suff}$, $Comp^{src}_{free}$ and respecting some length constraints on the substrings. For example, one may split a single-word term $SWT_{1,n}$ of n characters into prefix $Pref_{1,i}$ and lexical base $LexBase_{i+1,n}$ provided that $SWT_{1,i} \in Comp^{src}_{pref}$ and $SWT_{i+1,n} \in Comp^{src}_{free}$ and $n - i > \mathcal{L}0$; $\mathcal{L}0$ being empirically set to 5. A single-word term is first split into an optional prefixe + base$_1$, then base$_1$ is split into base$_2$ + optional suffix, then base$_2$ is split into one or several confixes or lexical items. When several splittings are possible, only the ones with the highest number of minimal components are retained.

$$\mathcal{S}(\mathcal{R}(\mathcal{T}(\mathcal{D}_2(\mathcal{D}_1(\text{“cytotoxic”}))))) \\ = \mathcal{S}(\mathcal{R}(\mathcal{T}(\mathcal{D}_2(\{\text{cyto, toxic}\}))))$$

Step 2 of decomposition ($\mathcal{D}_2$) gives out all possible decompositions of the single-word term by enumerating the different concatenations of its minimal components. For example, if single-word term "abc" has been split into minimal components {a,b,c}, then it has 4 possible decompositions: {abc}, {a,bc}, {ab,c}, {a,b,c}. For a single-word term having n minimal components, there exists 2^{n-1} possible decompositions.

$$\mathcal{S}(\mathcal{R}(\mathcal{T}(\mathcal{D}_2(\{\text{cyto, toxic}\})))) \\ = \mathcal{S}(\mathcal{R}(\mathcal{T}(\{\text{cyto, toxic}\}, \{\text{cytotoxic}\})))$$

The concatenation of the minimal components into bigger components increases the chances of finding translations. For example, consider the single-word term *non-cytotoxic* and a dictionary having translations for *non*, *cyto* and *cytotoxic* but no translation for *toxic*. If we stick to the sole output of $\mathcal{D}_1$ {*non-,-cyto-,toxic*}, the translation of *non-cytotoxic* will fail because there is no translation for *toxic*. Whereas if we also consider the output of $\mathcal{D}_2$ which contains the decomposition {*non-,cytotoxic*}, we will be able to translate *non-cytotoxic* because the dictionary has an entry for *cytotoxic*.

3.2.2 Translation function

The translation function $\mathcal{T}$ provides translations for each decomposition output by $\mathcal{D}$. Applying the compositionality principle to translation, we consider that the translation of the whole is a function of the translation of the parts: $\mathcal{T}(a,b) \cong \mathcal{T}(a) \times \mathcal{T}(b)$. For a given decomposition $\{c_1, ...c_n\}$ having n components, there exists $\prod_{i=1}^{n} |\mathcal{T}(c_i)|$ possible translations. Components' translations are obtained using the $Trans$ and Var resources: $\mathcal{T}(c) = Trans(c) \cup Trans(Var^{src}(c)) \cup Var^{tgt}(Trans(c))$. If one of the component cannot be translated, the translation of the whole decomposition fails.

$\mathcal{S}(\mathcal{R}(\mathcal{T}(\{\text{cyto, toxic}\}, \{\text{cytotoxic}\})))$
$= \mathcal{S}(\mathcal{R}(\mathcal{T}(\text{cyto}) \times \mathcal{T}(\text{toxic}), \mathcal{T}(\text{cytotoxic})))$
$= \mathcal{S}(\mathcal{R}(\{\text{cyto, toxique}\},\{\text{cellule, toxique}\},$
$\{\text{cytotoxicité}\}))$

3.2.3 Recomposition function

The recomposition function $\mathcal{R}$ takes as input the translations outputted by $\mathcal{T}$ and recomposes them into sequences of one or several lexical items. It takes place in two steps.

Step 1 of recomposition ($\mathcal{R}_1$) generates, for a given translation of n items, all of the $n!$ possible permutations of these items. As a general rule, $O(n!)$ procedures should be avoided but we are permuting small sets (up to 4 items). This captures the fact that components' order may be different in the source and target language (distortion). Once the components have been permutated, we generate, for each permutation, all the different concatenations of its components into lexical items (like it is done in step 2 of decomposition).

$\mathcal{S}(\mathcal{R}_2(\mathcal{R}_1(\{\text{cyto,toxique}\},\{\text{cellule,toxique}\},$
$\{\text{cytotoxicité}\})))$
$= \mathcal{S}(\mathcal{R}_2(\{\text{cyto,toxique}\}, \{\text{cytotoxique}\},$
$\{\text{toxique,cyto}\}, \{\text{toxiquecyto}\}, \{\text{cellule, toxique}\},$
$\{\text{celluletoxique}\}, \{\text{toxique, cellule}\},$
$\{\text{toxiquecellule}\}, \{\text{cytotoxicité}\}))$

Step 2 of recomposition ($\mathcal{R}_2$) filters out the ouput of $\mathcal{R}_1$ using heuristic rules. For example, a sequence of lexical items $L = \{l_1, ...l_n\}$ would be filtered out provided that $\exists\ l \in L \mid l \in Comp^{tgt}_{pref} \cup Comp^{tgt}_{conf} \cup Comp^{tgt}_{suff}$, i.e. recomposition $\{\text{cytotoxique}\}$ would be accepted but not $\{\text{-cyto-}, \text{toxique}\}$ because *-cyto-* is a bound component (it should not appear as an autonomous lexical item).

$\mathcal{S}(\mathcal{R}_2(\{\text{cyto,toxique}\}, \{\text{cytotoxique}\},$
$\{\text{toxique,cyto}\}, \{\text{toxiquecyto}\}, \{\text{cellule, toxique}\},$
$\{\text{celluletoxique}\}, \{\text{toxique, cellule}\},$
$\{\text{toxiquecellule}\}, \{\text{cytotoxicité}\}))$
$= \mathcal{S}(\{\text{cytotoxique}\}, \{\text{toxiquecyto}\},$
$\{\text{cellule, toxique}\}, \{\text{celluletoxique}\},$
$\{\text{toxique, cellule}\}, \{\text{toxiquecellule}\}, \{\text{cytotoxicité}\})$

These concatenations correspond to the final lexical units which will be matched against the target corpus with the selection function. For example, the concatenation $\{\text{toxique}_{\text{A}}, \text{cellule}_{\text{B}}\}$ corresponds to a translation made of two distinct lexical items: *toxique* followed by *cellule*. The concatenation $\{\text{cytotoxique}_{\text{AB}}\}$ corresponds to only one lexical item: *cytotoxique*.

3.2.4 Selection function

The selection function $\mathcal{S}$ tries to match the sequences of lexical items outputted by $\mathcal{R}$ with the lemmas of the tokens of the target corpus. We call $T = \{t_1, ...t_m\}$ a sequence of tokens from the target corpus, $l(t_k)$ the lemma of token t_k and $p(t_k)$ the part-of-speech of token t_k. We call $L = \{l_1, ...l_n\}$ a sequence of lexical items outputted by $\mathcal{R}$. L matches T if there exists a strictly increasing sequence of indices $I = \{i_1, ...i_n\}$ such as $l(t_{i_j}) = l_j$ and $\forall j, 1 \leq j \leq n$ and $\forall i, 1 \leq |i_{j-1} - i_j| \leq \mathcal{L}1$ and $\forall t_k |\ k \notin I,\ l(t_k) \in Stop^{tgt}$; $\mathcal{L}1$ being empirically set to 3.

$= \mathcal{S}(\{\text{cytotoxique}\}, \{\text{toxiquecyto}\},$
$\{\text{cellule, toxique}\}, \{\text{celluletoxique}\},$
$\{\text{toxique, cellule}\}, \{\text{toxiquecellule}\}, \{\text{cytotoxicité}\})$
= "cytotoxicité/N", "toxique/A pour/PREP le/DET cellule/N"
'*cytotoxicity*', '*toxic to the cells*'

In other words, L is a subsequence of the lemmas of T and we allow at maximum $\mathcal{L}1$ closed-class words between two tokens which match the lemmas of L.

For a given sequence of lexical items L, we collect from the target corpus all sequences of tokens $T_1, T_2, ...T_p$ which match L according to our abovementioned definition. We consider two sequences $T1$ and $T2$ to be equivalent candidate translations if $|T1| = |T2|$ and $\forall (t1_i, t2_j)$ such that $t1 \in T1, t2 \in$

$T2, i = j$ then $l(t1_i) = l(t2_j)$ and $p(t1_i) = p(t2_j)$, i.e. if two sequences of tokens correspond to the same sequence of (lemma, pos) pairs, these two sequences are considered as the same candidate translation.

4 Experimental data

We worked with three languages: English as source language and French and German as target languages.

4.1 Corpora

Our corpus is composed of specialized texts from the medical domain dealing with breast cancer. We define specialized texts as texts being produced by domain experts and directed towards either an expert or a non-expert readership (Bowker and Pearson, 2002). The texts were manually collected from scientific papers portals and from information websites targeted to breast cancer patients and their relatives. Each corpus has approximately 400k words (cf. table 1). All the texts were pos-tagged and lemmatized using the linguistic analysis suite XELDA[5]. We also computed the comparability of the corpora. We used the comparability measure defined by (Bo and Gaussier, 2010) which indicates, given a bilingual dictionary, the expectation of finding for each source word of the source corpus its translation in the target corpus and *vice-versa*. The English-French corpus' comparability is 0.71 and the English-German corpus' comparability is 0.45. The difference in comparability can be explained by the fact that German texts on breast cancer were hard to find (especially scientific papers): we had to collect texts in which breast cancer was not the main topic.

Readership	EN	FR	DE
experts	218.3k	267.2k	197.2k
non-experts	198.2k	184.5k	201.7k
TOTAL	416.5k	451.75k	398.9k

Table 1: Composition and size of corpora in words

4.2 Source terms

We tested our algorithm on a set of source terms extracted from the English texts. The extraction was done in a semi-supervised manner. **Step 1:** We wrote a short seed list of English bound morphemes. We automatically extracted from the English texts all the words that contained these morphemes. For example, we extracted the words *postchemotherapy* and *poster* because they contained the string *post-* which corresponds to a bound morpheme of English. **Step 2:** The extracted words were sorted : those which were not morphologically constructed were eliminated (like *poster*), and those which were morphologically constructed were kept (like *postchemotherapy*). The morphologically constructed words were manually split into morphemes. For example, *postchemotherapy* was split into *post-*, *-chemo-* and *therapy*. **Step 3:** If some bound morphemes which were not in the initial seed list were found when we split the words during step 2, we started the whole process again, using the new bound morphemes to extract new morphologically constructed words. We also added hyphenated terms like *ER-positive* to our list of source terms.

We obtained a set 2025 English terms with this procedure. For our experiments, we excluded from this set the source terms which had a translation in the general language dictionary and whose translation was present in the target texts. The final test set for English-to-French experiments contains 1839 morphologically constructed source terms. The test set for English-to-German contains 1824 source terms.

4.3 Resources used in the translation step $\mathcal{T}$

Tables 2 and 3 show the size of the resources we used for translation.

General language dictionary We used the general language dictionary which is part of the linguistic analysis suite XELDA.

Domain-specific dictionary We built this resource automatically by extracting pairs of cognates from the comparable corpora. We used the same technique as (Hauer and Kondrak, 2011): a SVM classifier trained on examples taken from online dictionaries[6].

Morpheme translation table To our knowledge, there exists no publicly available morphology-based bilingual dictionary. Consequently, we asked trans-

[5] http://www.temis.com

[6] http://www.dicts.info/uddl.php

lators to create an *ad hoc* morpheme translation table for our experiment. This morpheme translation table links the English bound morphemes contained in the source terms to their French or German equivalents. The equivalents can be bound morphemes or lexical items.

In order to handle the variation phenomena described in section 2.3, we used a **dictionary of synonyms** and lists of **morphologically related words**. The dictionary of synonyms is the one part of the XELDA linguistic analyzer. The lists of morphologically related words were built by stemming the words of the comparable corpora and the entries of the bilingual dictionary with a simple stemming algorithm (Porter, 1980).

	EN→FR	EN→DE
General language	38k→60k	38k→70k
Domain-specific	6.7k→6.7k	6.4k→6.4k
Morphemes (TOTAL)	242→729	242→761
prefixes	50→134	50→166
confixes	185→574	185→563
suffixes	7→21	7→32

Table 2: Nb. of entries in the multilingual resources

	EN→EN	FR→FR	DE→DE
Synonyms	5.1k→7.6k	2.4k→3.2k	4.2k→4.9k
Morphol.	5.9k→15k	7.1k→18k	7.4k→16k

Table 3: Nb. of entries in the monolingual resources

4.4 Resources used in the decomposition step ($\mathcal{D}$)

The decomposition function uses the entries of the bound morphemes translation table (242 entries) and a list of 85k lexical items composed of the entries of the general language dictionary and English words extracted from the Leipzig Corpus (Quasthoff et al., 2006) which is a general language corpus.

5 Evaluation

5.1 Evaluation metrics

As explained in section 2.2, compositional translation consists in *generating* candidate translations. These candidate translations can be filtered out with a classifier (Baldwin and Tanaka, 2004), by keeping only the translations which occur in the target texts of the corpus (Weller et al., 2011; Morin and Daille, 2010) or by using a search engine (Robitaille et al., 2006). Unlike alignment evaluation in parallel texts, there is no reference alignmens to which the selected translations can be compared and we cannot use standard evaluation metrics like AER (Och and Ney, 2000). It is also difficult to find reference lexicons in specific domains since the goal of the extraction process is to create such lexicons. Furthermore, we also wish to evaluate if the algorithm can identify non-canonical translations which, by definition, can not be found in a reference lexicon. Usually, the candidate translations are annotated manually as *correct* or *incorrect* by native speakers. Baldwin and Takana (2004) use two standards for evaluation: *gold-standard*, *silver-standard*. Gold-standard is the set of candidate translations which correspond to canonical, reference translations. Silver-standard corresponds to the gold-standard translations plus the translations which "capture the basic semantics of the source language expression and allow the source language expression to be recovered with reasonable confidence" (op. cit.).

The first evaluation metric is the *precision* P which is the number of correct candidate translations $|Corr|$ over the total number of generated candidate translations $|A|$: $P = \frac{|Corr|}{|A|}$. In addition to precision, we propose to indicate the *coverage* C of the lexicon, i.e. the proportion of source terms (ST) which obtained at least one candidate translation regardless of its accuracy:

$$C = \frac{\sum_{i=1}^{|\text{ST}|} \alpha(\text{ST}_i)}{|\text{ST}|}$$

were $\alpha(\text{ST}_i)$ returns 1 if $|A(\text{ST}_i)| \geq 1$ else 0. As augmenting coverage tends to lower precision, we also compute OQ, the *overall quality* of the lexicon, to get an idea of the coverage/precision tradeoff: $OQ = P \times C$.

5.2 Results

Compositional-translation methods give better results when they are applied to general language texts rather than domain-specific texts. This is due to the fact that the translations of the components can be easily found in dictionaries since they belong to the general language and it is also easier to collect large

corpora. **Working with general language texts**, Baldwin and Takana (2004) were able to generate candidate translations for 92% of their source terms and they report 43% (gold-standard) to 84% (silver standard) of correct translations. The size of their corpus exceeds 80M words for each language. Cartoni (2009) works on the translation of prefixed Italian neologisms into French. He considers that the generated neologisms have a "confirmed existence" if they occur more than five times on Internet. He finds that between 42% and 94% of the generated neologisms fall into that category. **Regarding domain-specific translation**, Robitaille et al. (2006) use a search engine to build corpus from the web and incrementally collect translation pairs. They start with a list of 9.6 pairs (on average) with a precision of 92% and end up with a final output of 19.6 pairs on average with a precision of 81%. Morin and Daille (2009) could generate candidate translations for 15% of their source terms and they report 88% of correct alignments. The size of their corpus is 700k words per language. Weller et al. (2011) were able to generate 8% of correct French translations and 18% of correct English translations for their 2000 German compounds. Their corpus contains approximately 1.5M words per language.

We ran the morpho-compositional translation prototype on the set of source terms described in section 4.2. The output candidate translations were manually annotated by two translators. Like Baldwin and Takana (2004), we used three annotation values: canonical translation, recoverable translation and incorrect. In our case, recoverable translations correspond paraphrastic and morphological translation variants. For example, the canonical translation for *post-menauposal* is *post-ménopausique*. Recoverable translations are *post-ménopause 'post-menopause'* and *après la ménopause 'after the menopause'*. Fertile translations can be canonical translations if a non-fertile translation would have been more awkward. For example, the canonical translation for *oestrogen-sensitive* is *sensible aux œstrogènes 'sensitive to oestrogens'*. A non-fertile translation would sound very unnatural. We computed inter-annotator agreement on a set of 100 randomly selected candidate translations. We used the Kappa statistics (Carletta, 1996) and obtained a high agreement (0.77 for English to German translations and 0.71 for English to French).

First, we tested the impact of the linguistic resources described in section 4.3 (B for Baseline dictionaries, D for Domain-specific dictionary, S for Synonyms, M for Morphologically related words). We also tested a simple Prefix+lemma translation (Pref) in similar vein to the work of Cartoni (2009) to serve as a line of comparison with our method. The results are given in tables 4 and 5. The best results in terms of overall quality are obtained with the combination of the baseline and domain-specific dictionaries (BD). Morphologically related words and synonyms increase coverage to the cost of precision. Regarding English-to-French translations, we were able the generate translations for 26% of the source terms. The gold-standard precision is 60% and the silver standard precision is 67%. Regarding English-to-German translations, we were able the generate translations for 26% of the source terms. The gold-standard precision is 39% and the silver-standard precision is 43%. The prefix+lemma translation method has a very high precision (between 84% and 76%) but produces very few translations (between 1% and 2%). Coverage and precision scores compare well with other approaches knowing that we have very small domain-specific corpora (400k words per language) and that our approach deals with a large number of morphological constructions. The lower quality of the German translations can be explained by the fact that the English-German corpus is much less comparable than the English-French corpus (0.45 vs. 0.71).

	C	P		OQ	
		GOLD	SILVER	GOLD	SILVER
Pref	.01	.84	.9	.01	.01
B	.12	.57	.60	.07	.07
BS	.15	.50	.53	.08	.08
BM	.23	.28	.37	.06	.09
BD	.26	**.60**	**.67**	**.16**	**.17**
BSMD	**.39**	.33	.44	.13	**.17**

Table 4: Scores for the EN→FR lexicon

We also tested the impact of the fertile translations on the quality of the lexicon. Tables 6 and 7 show the evaluation scores with and without fertile translations. As expected, fertile translations

	C	P		OQ	
		GOLD	SILVER	GOLD	SILVER
Pref	.02	.76	.86	.02	.02
B	.13	.35	.39	.05	.05
BS	.16	.31	.35	.05	.05
BM	.22	.23	.29	.05	.06
BC	.26	**.39**	**.43**	**.10**	.11
BCSM	**.36**	.27	.34	**.10**	**.12**

Table 5: Scores for the EN→DE lexicon

	C		P		OQ	
	-f	+f	-f	+f	-f	+f
B	.04	.12	**.81**	.57	.03	.07
BS	.05	.15	.69	.50	.03	.08
BM	.11	.23	.20	.28	.02	.06
BD	.16	.26	.70	**.60**	**.11**	**.16**
BSMD	**.24**	**.39**	.31	.33	.07	.13
avg. gain	+11		-8.6		+4.8	

Table 6: Scores without (-f) and with (+f) fertile translations (EN→FR)

	C		P		OQ	
	-f	+f	-f	+f	-f	+f
B	.06	.13	.80	.35	.05	.05
BS	.08	.16	.69	.31	.05	.05
BM	.12	.22	.40	.23	.05	.05
BC	.17	.26	**.65**	**.39**	**.11**	**.10**
BCSM	**.24**	**.36**	.43	.27	.10	**.10**
avg. gain	+9.2		-28.4		-0.2	

Table 7: Scores without (-f) and with (+f) fertile translations (EN→DE)

enables us to increase the size of the lexicon but they are less accurate than non-fertile translations. Fertile translations increase the overall quality of the English-French lexicon by 4% to 5%. This is not the case for English-German translations: fertile translations result in a big drop in precision. The overall quality does not significantly change. This might be partly due to the low comparability of the corpus but we think that the main reason lies in the morphological type of the languages involved in the translation. It is worth noticing that, if we consider only the non-fertile translations, the English-German lexicon has generally better scores than the English-French one. In fact, fertile variants are more natural and frequent in French than in German. English and German are Germanic languages with a tendency to build new words by agglutinating words or morphemes into one single word. Noun compounds such as *oestrogen-independent* or *Östrogen-unabhängige* are common in these two languages. Conversely, French is a Romance language which prefers to use phrases composed of two nouns and a preposition rather than a single-noun compound (*oestrogen-independent* would be translated as *indépendant des œstrogènes 'independent to oestrogens'*). It is the same with the bound morpheme/single word alternance. The term *cytoprotection* will be translated into German as *Zellschutz* whereas in French it can be translated as *cytoprotection* or *protection de la cellule 'protection of the cell'*.

6 Conclusion and future work

We have proposed a method based on the compositionality principle which can extract translations pairs from comparable corpora. It is capable of dealing with a largely variety of morphologically constructed terms and can generate *fertile* translations. The added value of the fertile translations is clear-cut for English to French translation but not for English to German translation. The English-German lexicon is better without the fertile translations. It seems that the added-value of fertile translations depends on the morphological type of the languages involved in translation. Future work includes the improvement of the identification of morphological variants. The morphological families extracted by the stemming algorithm are too broad for the purpose of translation. For example, the words *desirability* and *desiring* have the same stem but they are too distant semantically to be used to generate translation variants. We need to restrict the morphological families to a small set of morphological relations (e.g. noun ↔ relational adjective links). We will also work out a way to rank the candidate translations. Several lines of research are possible : go beyond the target corpora and learn a language model from a larger target corpus, mix compositional translation with a context-based approach, learn part-of-speech patterns translation probabilities from a parallel corpora (e.g. learning that it is more probable that a noun is translated as another noun or as a noun phrase rather than an adverb). A last improvement

could be to gather morpheme correspondences from parallel data.

References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Somers H., Amsterdam & Philadelphia, john benjamins edition.

Baldwin, T. and Tanaka, T. (2004). Translation by machine of complex nominals. In *Proceedings of the ACL 2004 Workshop on Multiword expressions: Integrating Processing*, pages 24–31, Barcelona, Spain.

Bo, L. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *23me International Conference on Computational Linguistics*, pages 23–27, Beijing, Chine.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge, London/New York.

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Cartoni, B. (2009). Lexical morphology in machine translation: A feasibility study. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 130–138, Athens, Greece.

Daille, B. and Morin, E. (2005). French-English terminology extraction from comparable corpora. In *Proceedings, 2nd International Joint Conference on Natural Language Processing*, volume 3651 of *Lecture Notes in Computer Sciences*, page 707718, Jeju Island, Korea. Springer.

Fung, P. (1997). Finding terminology translations from non-parallel corpora. pages 192–202, Hong Kong.

Fung, P. and Cheung, P. (2004). Mining Very-Non-Parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of EMNLP 2004*, pages 57–63, Barcelona, Spain.

Grefenstette, G. (1999). The world wide web as a resource for example-based machine translation tasks. *ASLIB'99 Translating and the computer*, 21.

Hauer, B. and Kondrak, G. (2011). Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 865–873, Chiang Mai, Thailand.

Keenan, E. L. and Faltz, L. M. (1985). *Boolean semantics for natural language*. D. Reidel, Dordrecht, Holland.

Morin, E. and Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. In Rayson, P., Piao, S., Sharoff, S., Evert, S., and B., V. M., editors, *Language Resources and Evaluation (LRE)*, volume 44 of *Multiword expression: hard going or plain sailing*, pages 79–95. Springer Netherlands.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Namer, F. and Baud, R. (2007). Defining and relating biomedical terms: Towards a cross-language morphosemantics-based system. *International Journal of Medical Informatics*, 76(2-3):226–33.

Och, F. and Ney, H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th Conference on Computational Linguistics*, volume 2, pages 1086–1090.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Quasthoff, U., Richter, M., and Biemann, C. (2006). Corpus portal for search in monolingual corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 1799–1802, Genoa, Italy.

Rapp, R. (1995). Identifying word translations in Non-Parallel texts. pages 320–322, Boston, Massachussets, USA.

Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 16–23, Athens, Greece.

Robitaille, X., Sasaki, X., Tonoike, M., Sato, S., and Utsuro, S. (2006). Compiling French-Japanese terminologies from the web. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 225–232, Trento, Italy.

Weller, M., Gojun, A., Heid, U., Daille, B., and Harastani, R. (2011). Simple methods for dealing with term variation and term alignment. In *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, pages 87–93, Paris, France.