An appraisal of recent breakthroughs in machine translation: the case of past participle-based compound adjectives in ESP

Évaluation des récentes avancées de la traduction automatique : le cas des adjectifs composés formés à partir d’un participe passé en anglais de spécialité

François Maniez

Electronic version
URL: http://journals.openedition.org/asp/5059
DOI: 10.4000/asp.5059
ISSN: 2108-6354

Publisher
Groupe d’étude et de recherche en anglais de spécialité

Printed version
Date of publication: 1 November 2017
Number of pages: 29-48
ISSN: 1246-8185

Electronic reference

This text was automatically generated on 20 April 2019.

Tous droits réservés
An appraisal of recent breakthroughs in machine translation: the case of past participle-based compound adjectives in ESP

Évaluation des récentes avancées de la traduction automatique : le cas des adjectifs composés formés à partir d’un participe passé en anglais de spécialité

François Maniez

1. Introduction

While the problems posed by the translation of denominal adjectives from French into English and vice-versa are well documented (Maniez 2009), the translation into Romance languages of compound adjectives consisting of a past participle preceded by a noun phrase (e.g. decay-corrected, receptor-mediated or technetium-99m-labeled) has received little attention so far, in spite of the very thorough treatment of participial compound pre-modifiers (henceforth PCPs) by Mestivier (2015). As stressed by Maniez and Thoiron (2004), such compound adjectives are more frequently used in scientific English than in general language and very often present a challenge when translating into Romance languages, where the equivalent of head modification in a noun phrase cannot be expressed with the same economy as that afforded by the ‘cognitive packing’ that such PCPs allow (Gledhill & Pecman 2015). As pointed out by Biber (1988: 105), PCPs, insofar as they “pack information into fewer words and structures” than predicative adjectives or relative clauses, are “a more integrated form of nominal elaboration” whose frequency has constantly increased in scientific writing in the past few decades. Figure 1, for instance, shows how the use of the PCP evidence-based has markedly increased since the
beginning of the 1990s in the Google Books corpus in relation to the occurrence of the phrase (is) based on evidence.¹

Figure 1. Frequency of use of evidence-based and (is) based on evidence between 1980 and 2008 in the Google Books corpus

² For translators from English into Romance languages, a major difficulty lies in the fact that some very frequently used adjectives of this type are absent from terminological databases (whose entries consist almost exclusively of nouns or complex noun phrases), as the productivity of English compounding mechanisms makes it impossible to record all existing forms.

2. Identification of the most productive past participle forms used in PCPs ending in -ed in the Corpus of Contemporary American English

³ We used a list of the 500 most frequently used compound adjectives ending in -ed in the Corpus of Contemporary American English (henceforth COCA) as a basis for our study (see Davies 2008, for an introduction to the design and use of this well-known 520-million-word corpus). We selected the 464 forms in that list which were used five times or more in the corpus. Table 1 shows the thirty most frequently used bigrams consisting of a PCP followed by a noun, with their respective frequencies.
Table 1. Most frequently used bigrams consisting of a PCP ending in -ed followed by a noun and their frequency in the COCA Corpus (ACAD:MED)

<table>
<thead>
<tr>
<th>Bigram</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>evidence-based practice</td>
<td>499</td>
</tr>
<tr>
<td>coal-fired power</td>
<td>341</td>
</tr>
<tr>
<td>evidence-based practices</td>
<td>340</td>
</tr>
<tr>
<td>sun-dried tomatoes</td>
<td>331</td>
</tr>
<tr>
<td>star-spangled banner</td>
<td>276</td>
</tr>
<tr>
<td>curriculum-based measurement</td>
<td>275</td>
</tr>
<tr>
<td>state-owned enterprises</td>
<td>231</td>
</tr>
<tr>
<td>mortgage-backed securities</td>
<td>225</td>
</tr>
<tr>
<td>physician-assisted suicide</td>
<td>222</td>
</tr>
<tr>
<td>rocket-propelled grenades</td>
<td>208</td>
</tr>
<tr>
<td>computer-aided design</td>
<td>205</td>
</tr>
<tr>
<td>health-related fitness</td>
<td>200</td>
</tr>
<tr>
<td>norm-referenced tests</td>
<td>175</td>
</tr>
<tr>
<td>rocket-propelled grenade</td>
<td>165</td>
</tr>
<tr>
<td>self-regulated learning</td>
<td>162</td>
</tr>
<tr>
<td>U.S.-led coalition</td>
<td>157</td>
</tr>
<tr>
<td>health-related quality</td>
<td>154</td>
</tr>
<tr>
<td>value-added tax</td>
<td>152</td>
</tr>
<tr>
<td>self-directed learning</td>
<td>146</td>
</tr>
<tr>
<td>faith-based organizations</td>
<td>144</td>
</tr>
<tr>
<td>grass-fed beef</td>
<td>137</td>
</tr>
<tr>
<td>ground-based telescopes</td>
<td>124</td>
</tr>
<tr>
<td>problem-based learning</td>
<td>124</td>
</tr>
<tr>
<td>republican-controlled congress</td>
<td>124</td>
</tr>
<tr>
<td>school-based interventions</td>
<td>122</td>
</tr>
<tr>
<td>faith-based initiative</td>
<td>119</td>
</tr>
<tr>
<td>home-cooked meal</td>
<td>119</td>
</tr>
<tr>
<td>community-based organizations</td>
<td>118</td>
</tr>
<tr>
<td>community-based services</td>
<td>111</td>
</tr>
<tr>
<td>peer-reviewed journals</td>
<td>100</td>
</tr>
</tbody>
</table>

4 The regular expression used in our query (\*-*ed) did not allow for the selection of specific syntactic patterns, since the components of compound adjectives are not individually part-of-speech-tagged (we should also add that while this regular expression may occasionally return a compound adjective formed with the past participle of an irregular verb -- such as grass-fed -- it will fail to return those that are formed with most other irregular verbs). Consequently, a wide variety of word formation patterns was to be observed in our original list: while it seems possible to paraphrase the meaning of most adjectives by developing an underlying prepositional phrase attached to the head noun (community-based = based in the community, health-related = related to health) as remarked by Mestivier (op. cit.: 3), some of the items listed in Table 1 clearly do not fit that pattern. Here are some examples of forms that were originally returned by our query and the reason why they were discarded:

- the first component of the compound adjective is a prefix or an adverb: pre-formatted, so-called, well-defined.
- the compound adjective is not formed with a past participle, but by adding the -ed suffix to a whole complex noun phrase: left-sided, open-ended, middle-aged, able-bodied.

5 Since we assume that such forms are lexicalized and are generally assigned stable translation equivalents in Romance languages, we will consider them to be outside the scope of our study, and we will focus instead on compound adjectives that are formed with the past participle of a regular verb, whether they express an agentine relation (computer-aided = aided by a computer, physician-assisted = assisted by a physician) or any other relation that can be expressed through the use of a prepositional phrase.
Close examination of the data in Table 1 reveals that while some of the most frequent forms (such as sun-dried tomatoes or star-spangled banner) clearly belong to general English (GE), the overwhelming majority are related to English for Specific Purposes (ESP) and more specifically to scientific English. On looking more specifically at the past participle forms involved, we also notice that many of the adjectives shown in Table 1 are formed with the same past participle (as many as ten are formed on based alone). Bearing that fact in mind, we identified the most productive past participles in the formation of such adjectives in the medical part (ACAD:MED) of the Academic sub-section of the COCA by computing their frequencies separately.

Table 2. Twenty most productive past participles in the formation of PCPs in the medical part of the Academic sub-section of the COCA Corpus (ACAD:MED)

<table>
<thead>
<tr>
<th>Past participle</th>
<th>X-Ved types</th>
<th>X-Ved tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>based</td>
<td>383</td>
<td>2238</td>
</tr>
<tr>
<td>related</td>
<td>371</td>
<td>1694</td>
</tr>
<tr>
<td>induced</td>
<td>157</td>
<td>350</td>
</tr>
<tr>
<td>associated</td>
<td>104</td>
<td>259</td>
</tr>
<tr>
<td>oriented</td>
<td>68</td>
<td>140</td>
</tr>
<tr>
<td>treated</td>
<td>56</td>
<td>167</td>
</tr>
<tr>
<td>mediated</td>
<td>49</td>
<td>143</td>
</tr>
<tr>
<td>shaped</td>
<td>46</td>
<td>100</td>
</tr>
<tr>
<td>controlled</td>
<td>37</td>
<td>197</td>
</tr>
<tr>
<td>infected</td>
<td>33</td>
<td>169</td>
</tr>
<tr>
<td>adjusted</td>
<td>25</td>
<td>136</td>
</tr>
<tr>
<td>assisted</td>
<td>24</td>
<td>164</td>
</tr>
<tr>
<td>cent(e)red</td>
<td>23</td>
<td>204</td>
</tr>
<tr>
<td>weighted</td>
<td>21</td>
<td>172</td>
</tr>
<tr>
<td>linked</td>
<td>18</td>
<td>63</td>
</tr>
<tr>
<td>reported</td>
<td>16</td>
<td>416</td>
</tr>
<tr>
<td>tested</td>
<td>13</td>
<td>38</td>
</tr>
<tr>
<td>defined</td>
<td>11</td>
<td>79</td>
</tr>
</tbody>
</table>
Table 2 shows the total number of distinct compound adjectives formed with the twenty most productive past participles and the aggregate frequency for all of those forms combined. The figures shown in the table confirm that a great number of the adjectives under study are formed with only a handful of past participle forms which are very productive. Indeed, the first four participles in the list (based, related, induced and associated) are used for more than two thirds of all types and tokens of such PCPs (based alone accounts for more than a quarter of all the types of those adjectives and almost a third of all the forms listed in the table).

### 3. Machine Translation and the translation of PCPs in scientific texts

Machine translation (henceforth MT) has made extraordinary progress in the past two decades. Early methods based on the use of dictionaries and parsing algorithms obtained fairly good results when applied to scientific texts for two reasons:

- Their content generally presents fewer polysemous words than general language, leading to a lesser degree of semantic ambiguity.
- Complex terms, which most often are present in the form of long noun phrases, represent an important part of such texts, and can easily be translated if their translations equivalents (TEs) are included in the machine-readable glossaries or translation memories that underlie such MT programs.

The type of progress accomplished by MT programs in the previous decade can be seen in example 1, which shows the translation of a sentence (quoted in Maniez, 2008) by the Systran software (version 6) in 2007 and a recent translation of the same sentence by the Google Translate MT program:

**Example 1:**

**Systran 6 (2007):** Les algorithmes de gestion seraient plus utiles que des figures montrant des réponses anecdotiques aux traitements hypolipidémiant ou aux photographies des patients faisant des poussées.

**GT (08/2017):** Les algorithmes de gestion seraient plus utiles que les chiffres montrant des réponses anecdotiques aux médicaments abaissant le cholestérol ou des photographies de patients faisant des flexions.

The Systran software that was used in Maniez (2008) came with a medical dictionary, which explains why it returned a terminologically more specialized adjective (hypolipidémiant) for the English PCP cholesterol-lowering than the ‘unpacked’ translation provided by GT (abaissant le cholestérol), which – while it may not be the most usual translation – is quite understandable and suffices to convey the meaning of the original PCP. The GT translation also avoids several mistakes present in the segments in bold characters above, i.e. literal rendering of the polysemous word figure, an incorrect
translation of the noun push-ups, and faulty syntactic analysis of coordination in the final noun phrase, since the translation shows that the segment photographs of patients doing push-ups was interpreted as coordinated with cholesterol-lowering medications instead of figures.

Indeed, however much progress has recently been made in the field of artificial intelligence, human translators still consistently outperform MT in three important tasks:

- identification of referents in anaphora (especially of the interphrastic kind);
- resolution of semantic ambiguity (such as that created by polysemy or homonymy);
- resolution of syntactic ambiguity that cannot be solved using statistical methods.

Translation difficulties involving the latter aspect essentially concern structures in which both pre-modification and coordination are used as well as polysemy and part-of-speech ambiguity, as can be seen in the examples below (all examples quoted henceforth were collected on the Google Translate web site in March and April 2017, and the faulty segments which are commented are in bold characters).

In example 2, the ellipsis that occurs in the word (play)pen results in the MT program selecting by default the commonest meaning of the word pen, that of a writing tool:

(2) Leaving the third door open, he started back down the hallway, eyes locked on Julia playing in her pen.
Laissant la troisième porte ouverte, il recula dans le couloir, les yeux fixés sur Julia jouant dans son stylo.

In example 3, the MT program assumes limited scope for the premodification by forest, which should also apply to burning:

(3) Mr Sharp, the Eagle High spokesman, blamed local villagers for the forest clearing and burning on its sites.
M. Sharp, le porte-parole de Eagle High, a blâmé les villageois locaux pour la déforestation et la combustion sur ses sites.

In example 4, the elliptic causative structure (equivalent to I don’t want a precedent to be set) is not correctly identified by the MT program, and the word set is assigned the wrong part-of-speech (i.e. it is interpreted as a noun):

(4) “I don’t want a precedent set which says a Prime Minister and a handful of MPs or ministers can take away or grant rights.”
“Je ne veux pas d’un ensemble de précédents qui dit qu’un Premier ministre et une poignée de députés ou de ministres peuvent prendre ou accorder des droits.”

The translation of PCPs into French usually involves ‘unpacking’ the premodifying information through the use of a reduced relative clause in the passive voice, as in example 5:

(5) Although this is common practice in clinical trials, preclinical studies rarely use imaging-based criteria.
Bien que ce soit une pratique courante dans les essais cliniques, les études précliniques utilisent rarement des critères basés sur l’imagerie.

However, ‘repacking’ does not necessarily occur in translations from French into English:

(6) For the final diagnosis of malignancy or benignity, imaging (follow-up scan, MRI, follow-up scintigraphy) and clinical follow-up were used:
Since a Google search gives forty-one results for the sequence *criteria based on imaging*, as opposed to almost 10,000 for *imaging-based criteria*, it is reasonable to assume that the latter translation would be selected by a human translator who is moderately familiar with scientific writing. When the English PCP is part of a very frequently used term, however, such repacking will occasionally be performed by MT translation programs:

(7) Le rôle exact de la tomographie par émission de positons dans la prise en charge des malades avec nodule pulmonaire solitaire reste à définir en l’absence de directives basées sur les preuves.

The exact role of positron emission tomography in the management of patients with solitary pulmonary nodule remains to be defined in the absence of evidence-based guidelines.

The recent changes brought about by the use of neuronal networks (Lewis-Kraus 2016) are clearly instrumental in obtaining such convincing results, even though a number of terms featuring nouns preceded by the PCP *evidence-based* are included in terminological data bases and translation memories (the complex terms in which the compound adjective is followed by the nouns *approach, care, crime prevention, decision-making, epidemiology, healthcare, medical practice* and *medicine* are all recorded in either IATE, Termium or the Grand Dictionnaire Terminologique).

Another problem which seems to be correctly solved following those recent changes is the treatment of PCPs in which the noun is a multiple word unit (MWU), as in example 8:

(8a) The combination of both interpolators is determined by a fuzzy logic-based system.

(8b) La combinaison des deux interpolateurs est déterminée par un système basé sur la logique floue.

It is worth mentioning that ‘repacking’ (i.e. the reconstruction of the PCP when reverting to English) does not occur in this instance, since processing of the French translation gives us the following output:

(8b) The combination of the two interpolators is determined by a system based on fuzzy logic.

In former stages of their development, MT systems would have provided translations based on parsing (syntactic analysis) in which two separate adjectives (*fuzzy* and *logic-based*) qualified the noun *system*. This is indeed a well-known pitfall for human translators as well, which is probably brought about by the fact that the hyphen is interpreted, both visually and semantically, as the stronger link. Absence of hyphens — which happens in fewer than 10% of all cases according to data provided by Mestivier (op. cit.) — for the adjective *stress-induced* may create syntactic ambiguity, as is shown by the following example quoted by the same author:

(9a) A putative relationship between hyperglycemic hormone degradation and stress induced protease activity in the hemolymph of the American cockroach.

Example 9, which is actually the title of a scientific article, should be understood as the coordination of two noun phrases, which precludes understanding *induced* as a conjugated verb form, even though the absence of hyphenation makes such an interpretation syntactically valid. MT actually avoids that pitfall, probably because of the overall frequency of the form *stress-induced* in the training data used by the program:

(9b) Une relation putative entre la dégradation de l’hormone hyperglycémique et l’activité de la protéase induite par le stress dans l’hémolympie du cafard américain.
The mistake in the above translation actually concerns another item: the MT system failed to find a match for the relatively rare word hemolymph (even though the correct TE hémolymphe is quite close) and subsequently coined an unknown French word based on its statistical algorithms for word formation.

Another clear advantage of human translators over MT systems is their grasp of the cohesiveness of argumentation. Example 10 clearly involves a comparison between two models (tumor diameter-based vs. tumor volume-based), a fact that is missed by the MT system (which probably favors the use of the French modèles tumoraux on statistical grounds in the same way as croissance tumoraire rightly translates tumor growth), thus obscuring the comparison by not mentioning volume:

(10) Over the last few years, taking advantage of the linear growth of diameter kinetics, tumor diameter-based rather than tumor volume-based models have been developed for the phenomenological modeling of tumor growth. Au cours des dernières années, en profitant de la croissance linéaire de la cinétique du diamètre, des modèles basés sur le diamètre de la tumeur plutôt que des modèles tumoraux ont été développés pour la modélisation phénoménologique de la croissance tumoraire.

Syntactic variation is also a pitfall for MT systems when combined with polysemy, as shown by the erroneous translation of staging in example 11:


While the ‘packed’ version of the first part of the above example (early prostate cancer staging) does not yield the same error (even though stadification précoce would have been closer to the mark, the suggested translation (dépistage précoce du cancer de la prostate) is reasonably close), selection of the theatrical meaning of stage underscores the lack of cohesiveness in MT output. It is worth mentioning, however, that while proposé is wrongly attached to volume, the PCP tumor volume-based is correctly translated by basé sur le volume tumoral.

4. Study of PCPs formed with the past participle based

We chose to study compound adjectives formed with the most productive past participle (i.e. based, see Table 2). Figure 2 (below) shows absolute and relative frequencies of PCPs formed with based in the five sections of the COCA, with a twofold increase in frequency in the Academic section as opposed to the Magazine and Newspaper sections, and very rare use in the Spoken and Fiction sections.
For each of the three adjectives formed with based that were examined below (community based, evidence-based and population-based), we checked for inclusion in the Grand Dictionnaire Terminologique (henceforth GDT), since its search module allows for queries using wild cards (e.g. *-based), unlike some other such terminology databases. We then checked for the presence of French translation equivalents of expressions using those three compound adjectives on Linguee (a search engine which provides online translations by emulating concordancing software used on a virtual parallel corpus) and submitted those sentences to the Google Translate MT system. In each of the examples below, the corresponding translations are either preceded by L (for Linguee) or MT (for the Google Translate MT system).

The GDT lists 33 items corresponding to the query «*-based». They cover a variety of domains (17 in total), but mostly concern items listed under the heading informatique (computer science/information technology). The past participle based is itself an entry, in which the translation architecturé autour de is provided in relation to the informatique domain. The most commonly used compound adjectives formed with based in the medical part of the Academic section of COCA are community-based (266 occurrences), evidence-based (224), population-based (204) and school-based (168). None of them are present in the GDT as adjectival forms, but a query using several wildcards (*-based *) reveals no fewer than 577 items, including 11 terms starting with the adjective community-based, 4 with evidence-based, 1 with population-based (population-based approach) and 3 with school-based. In some cases, alternative spellings of the same adjective — in its hyphenated (school-based) and non-hyphenated (school based) forms — are listed as separate items, even though they eventually link to the same terminological record.

4.1. Community-based

If we turn to the 100 most frequently used PCP-noun expressions in the ACAD:MED section of COCA, we find that three of them start with the adjective community-based: community-based services (26 occurrences), community-based health (12) and community-based programs (8), the sequence community-based health being actually part of longer noun phrases such as community-based health programs or community-based health services. It turns out that none of those expressions are part of the items listed as a result of our GDT query, which are the following:
- community based distribution of contraceptives
- community-based family planning project
- community-based grocery store
- community-based literacy program
- community based primary health care
- community-based program
- community-based social marketing
- community based tourism
- community based tree plantation

34 Suggested translations range from the use of the relational adjective communautaire (when community-based modifies such noun phrases as grocery store, literacy program, mentoring, program, social marketing or tourism) to the use of prepositional phrases built around the noun collectivité (community based family planning project: projet de planification de la famille à l’échelon de la collectivité, community based primary health care: soins de santé primaires axés sur la collectivité).

35 A query on the Linguee search engine provides a translation for the expression community-based health services that makes use of the adjective communautaire:

(12) Twelve per cent of the groups surveyed were involved in the arts; organizations that serve ethnocultural communities and community-based health services each made up 10.0% of respondents.

L: Douze pour cent des organismes participants sont engagés dans le domaine des arts, 10 % sont des organisations au service de communautés ethnoculturelles et 10 % des organismes de santé communautaire.

MT: Douze pour cent des groupes interrogés étaient impliqués dans les arts; Les organisations qui servent les communautés ethnoculturelles et les services de santé communautaires représentent chacune 10,0% des répondants.

36 MT treatment of the PCP (for which the adjective communautaire is also used) reflects human usage, and the human translation provided here is only marginally better (since literal translation is avoided for the nouns groups and services) and is improved for the prepositional phrase in the arts by the addition of the noun domaine (dans le domaine des arts). The other expressions for which translations are available on Linguee (where the PCP community-based modifies health advocacy groups, health care, health and social service organizations, health clinics and health insurance schemes) do not match those that are most frequently used in ACAD:MED. Using the Linguee search engine is nevertheless a useful option for translators, even though the various meanings of the word community and the wide variety of head nouns modified by the adjective community-based make it difficult to readily select one particular translation that will fit all of its uses. We should also add that while the adjective communautaire is widely used, a few translations resort to French expressions such as au niveau local or de proximité, which convey the notion of a geographically restricted area.

4.2. Evidence-based

37 The GDT lists two terminological records whose main entry starts with the adjective evidence-based: evidence-based medicine and evidence-based care. As both components of the compound adjective evidence-based can be translated in several ways, a comprehensive list of all the variations for the French equivalent is given in the terminological record that is devoted to evidence-based medicine, which can be synthesized in the following manner (parentheses indicate words that can be dispensed with; slashes indicate alternatives):
médecine basée/fondée sur les/des [preuves/données probantes/faits (prouvés)]. A much shorter translation, médecine factuelle is mentioned as the term that has been most currently used in the past few years. The same type of variation is mentioned in the terminological record that is devoted to evidence-based care.

38 A query on the Linguee search engine provides examples of translations for evidence-based medicine, as well as the other expression used in ACAD:MED, evidence-based practice. The result of our query also shows translations for expressions in which evidence-based is followed by the nouns approach, decision-making, information, policy, policy-making and research. The examples provided by Linguee reveal that while médecine factuelle and its longer synonyms are often used as TEs, some other expressions (médecine fondée sur l’expérience clinique, médecine scientifique) can occasionally be found. As far as the expression evidence-based practice is concerned, the two most commonly found translation equivalents are médecine fondée sur l’expérience clinique and médecine fondée sur des données probantes.

39 In example 13, the Linguee translation uses one of the many possibilities listed in the GDT (fondées sur des données probantes), while MT provides another (fondées sur des preuves). Translations also vary for the noun guidelines, for which both lignes directrices and directives are a fitting TE.

(13) Several evidence-based guidelines for the prevention and treatment of osteoporosis are currently available.
L: Il existe actuellement plusieurs lignes directrices (fondées sur des données probantes) pour la prévention et le traitement de l’ostéoporose.
MT: Plusieurs directives fondées sur des preuves pour la prévention et le traitement de l’ostéoporose sont actuellement disponibles.

4.3. Population-based

40 Only one terminological record (population-based approach) concerns a term which includes this adjective. The French translation is approche en population générale, and the English definition, coined by the World Health Organization in 1995, suggests that use of the term is restricted to the context of prevention measures against alcohol-related problems, although web evidence suggests otherwise (see below). Judging by the variety of expressions in which the adjective is used, some of them, notably population-based study (see below), probably deserve to be added to the GDT.

41 A query on the Linguee search engine provides a translation for expressions in which the adjective population-based precedes a great number of nouns, including approach, data, measures, prevention, surveillance, strategy, (cohort) study and survey. Various French translations are recorded for the most commonly used expression (population-based study): étude en population (générale) and étude de population are the most frequently encountered, and there are two instances of the relational adjective populationnel. Example 14 shows that use of the term population-based approach does not necessarily follow the sense restrictions mentioned in the GDT. In this case, MT translation, although more literal, is devoid of obvious mistakes.

(14) […] offer evidence-based cancer screening through a systematic population-based approach with quality assurance at all appropriate levels.
L: […] offrir un dépistage du cancer reposant sur des données probantes, grâce à une approche systématique s’adressant à la population et avec une assurance de qualité à tous les niveaux appropriés.
5. Study of PCPs formed with the past participle *induced*

Access to a terminological database such as the Grand Dictionnaire is a precious tool for translators, and a wide variety of terms starting with compound adjectives are included in terminological records (for instance, 577 hits were obtained for those formed with the past participle *based* alone). But while using online search engines such as Linguee may help translators find TEs that are not mentioned in terminological databases (or check their currency if they are indeed listed in those databases), such searches do not always yield satisfying results. We will now focus on PCPs that are formed with the past participle *induced*, which is actually more productive than *based* in the medical field: out of the 577 records returned by the query «*-based *» in the GDT, only 20 belong to the Medicine field, as opposed to 38 of the 211 records returned by the query «*-induced *». Table 3 shows the 12 complex terms listed in the GDT that include PCPs formed with the past participle *induced*, (an entry is also devoted to the adjective *drug-induced*, for which the translation *d’origine médicamenteuse* is given).

Table 3. Terms including a compound adjective formed with *induced* in the GDT

<table>
<thead>
<tr>
<th>Term</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>acetylcholine-induced spasm</td>
<td>spasme acétylcholinique</td>
</tr>
<tr>
<td>digitalis-induced bradycardia</td>
<td>bradycardie digitalique</td>
</tr>
<tr>
<td>drug-induced hepatitis</td>
<td>hépatite médicamenteuse</td>
</tr>
<tr>
<td>drug-induced parkinsonism</td>
<td>syndrome parkinsonien neuroleptique</td>
</tr>
<tr>
<td>exercise-induced asthma</td>
<td>asthme d’effort</td>
</tr>
<tr>
<td>heparin-induced thrombocytopenia</td>
<td>thrombocytopénie induite par l’héparine</td>
</tr>
<tr>
<td>leukocytopenia-induced</td>
<td>leucopéniant</td>
</tr>
<tr>
<td>medication-induced headache</td>
<td>céphalée d’origine médicamenteuse</td>
</tr>
<tr>
<td>platelet-induced cerebral thrombosis</td>
<td>thrombose cérébrale par agrégation plaquettaire</td>
</tr>
<tr>
<td>pregnancy-induced hypertension</td>
<td>hypertension gravidique</td>
</tr>
<tr>
<td>radiation induced malignancy</td>
<td>altération maligne causée par les rayonnements</td>
</tr>
<tr>
<td>trans urethral laser induced prostatectomy</td>
<td>TULIP</td>
</tr>
</tbody>
</table>

It is worth noting that the causal relation expressed by the verb *induce* is translated with a verb form only twice, through use of the verbs *causer* (altération maligne causée par les rayonnements) and *induire* (thrombocytopénie induite par l’héparine). Other ways in...
which this relation is expressed include use of prepositions (asthme d’effort, trombose cérébrale par agrégation plaquettaire) or modification by a relational adjective (spasme acétylcholinique, bradycardie digitalique, hépatite médicamenteuse, hypertension gravidique). In the case of the term syndrome parkinsonien neuroleptique, the class of drugs is actually specified, so that the adjective neuroleptique actually means neuroleptic-induced (both neuroleptic-induced and antipsychotic-induced are widely used adjectives in the literature).

Example 15 indicates that use of the relational adjective acétylcholinique does not necessarily apply to other nouns than spasme, since the translation equivalent déclenché par l’acétylcholine is used for acetylcholine-induced on Linguee, while MT provides a more literal translation including the past participle induit:

(15) In anesthetized guinea pigs, salbutamol completely prevents acetylcholine-induced bronchospasm at the dose of 100 mcg/kg intravenously.

L: Chez des cobayes anesthésiés, le salbutamol administré par voie intraveineuse à une dose de 100 µg/kg prévient complètement le bronchospasme déclenché par l’acétylcholine.

MT: Dans les cochons d’Inde anesthésiés, le salbutamol empêche complètement le bronchospasme induit par l’acétylcholine à la dose de 100 mcg / kg par voie intraveineuse.

Other frequently used adjectives such as drug-induced will often be translated by the relational adjective médicamenteux, whose use implies that the causal relationship is obvious, as suggested by the translation for drug-induced hepatitis in Table 3. The translation given on Linguee in example 16 follows this usage, while MT once again provides a literal ‘unpacked’ translation, in which polysemy triggers an improper rendering of drug as drogue:

(16) The monograph also states that drug-induced hepatitis and hepatic necrosis have been reported.

L: La monographie indique aussi qu’on a signalé des cas d’hépatite médicamenteuse et de nécrose hépatique.

MT: La monographie indique également que l’hépatite induite par la drogue et la nécrose hépatique ont été rapportées.

In example 17, the same type of literal translation is at work in MT (asthme induit par l’exercice), whereas the term recommended by the GDT (asthme d’effort) is used in the translation provided by Linguee. It is to be noted that while use of combattre as a TE for take care of may seem slightly inaccurate, the MT alternative prendre en compte is totally off the mark. Such pitfalls should definitely be pointed out to students by instructors who train them in post-editing skills, as the extent to which recent neural networks translation systems have minimized the need for grammatical correction in the target language definitely makes the risk of overlooking such erroneous statistical matchings much greater than it once was.

(17) I did some research myself and happened to find out that exercise-induced asthma can be taken care of by increasing the oxygen capacity in your lungs.

L: Je me suis intéressé de plus près à cela et j’ai découvert qu’on peut combattre l’asthme d’effort en augmentant sa capacité pulmonaire.

MT: J’ai fait des recherches moi-même et j’ai découvert que l’asthme induit par l’exercice peut être pris en compte en augmentant la capacité d’oxygène dans vos poumons.

In example 18, the compound adjective radio-induit, an information-packing TE which has gained currency in the past few decades in French medical articles in the field of nuclear
medicine (Maniez op. cit.) is used instead of the ‘unpacked’ equivalent suggested by the GDT (causé par les rayonnements), while MT once again uses a literal unpacked translation:

(18) Different rates of tumors, including radiation-induced secondary malignancies, have been reported in males and females.

L: On a observé des taux différents de tumeurs, y compris les malignités secondaires radio-induites, chez les hommes et les femmes.

MT: Des taux différents de tumeurs, y compris des tumeurs secondaires induites par rayonnement, ont été signalés chez les hommes et les femmes.

Examples below confirm the frequent use of radio-induit (which accounts for about half of all translations found on Linguee) and show the difficulties that may arise when the PCP is preceded by another adjective. In Example 19, even though MT attributes a correct syntactic dependency to the adjective known, the resulting expression (cancer connu par rayonnement) seems too elliptic in French and does not achieve the same degree of accuracy as human translation, while chez l'enfance should be replaced by chez l'enfant.

(19) Further, there was no evidence that childhood leukemia mortality in particular, a known radiation-induced cancer, was increased above expectation.

L: Rien n'indiquait non plus que la mortalité due à la leucémie infantile en particulier (cancer connu pour être radio-induit) dépassait les prévisions.

MT: En outre, il n’y avait aucune preuve que la mortalité liée à la leucémie chez l’enfant en particulier, un cancer connu par rayonnement, était supérieure aux attentes.

Example 20 provides an example of a potential syntactic ambiguity created by MT, since the French adjective aigu may be construed as a modifier for the preceding noun (rayonnement).

(20) Outside that zone, no acute radiation-induced effects have been reported.

L: Au-delà de cette zone, aucun effet radio-induit aigu n’a été signalé.

MT: À l’extérieur de cette zone, aucun effet induit par rayonnement aigu n’a été signalé.

6. A concrete case of professional translation: MT

Treatment of the PCP image-guided

In order to illustrate the amount of time potentially gained by using MT systems, we will now briefly examine the various steps which a translator might follow when confronted with example 21a, which includes the adjective image-guided, the PCP with the highest frequency in the English part of an English-French comparable corpus of over 700 scientific articles in the field of nuclear medicine that was compiled by our research center, the Centre de Recherche en Terminologie et Traduction (CRTT, EA 4162).

(21a) We present the topic of image-guided radiotherapy (IGRT) on the basis of a selective review of the literature.

As was mentioned previously, PCPs rarely constitute separate entries in terminological databases, and image-guided is no exception to this general rule. Here are the results provided by the three online data bases which are most often used by professional translators:

IATE: the term image-guided minimally invasive therapy is mentioned along with its associated initialism (IGMIT), but no translation is provided in any of the 25 European languages covered by the data base.

GDT: the term image-guided [brain surgery / neurosurgery] is mentioned, for which the translation neurochirurgie assistée par imagerie médicale is provided.
Termium: the term image-guided surgery is mentioned, for which the translation [chirurgie / intervention chirurgicale] guidée par l’image is provided.

Confronted with such a degree of variation, the translator is left to wonder which TE to use (there are potentially four ways in which the PCP might be translated, since two solutions are given for both image and guided).

The Linguee data base provides three different translations for image-guided radiotherapy. Checking frequency for each of them on the Google search engine gives us 32 occurrences of radiothérapie assistée par l’image, 21 of radiothérapie assistée par imagerie and 48 of radiothérapie guidée par imagerie.

Relying on statistical information consisting of billions of words’ worth of data, MT translation of example 20a actually provides a solution which is very close to what is needed:

(21b) Nous présentons le sujet de la radiothérapie guidée par image (IGRT) sur la base d’une revue sélective de la littérature.

Another query conducted on any search engine with the character string radiothérapie guidée par image without quotes will quickly reveal that radiothérapie guidée par l’image is actually by far the most commonly used term, with several thousands of results (strangely enough, it is to be found on Linguee only as a translation equivalent for radiation therapy, a synonym for radiotherapy).

7. Conclusion

In this study of the lexicalization and translation into French of compound adjectives ending with a past participle in scientific English, we have seen that the Grand Dictionnaire Terminologique covers a large number of such adjectives (over 500 records for such productive past participles as based and controlled, for instance) and that while their translation by the online translating service Google Translate has notably improved, it still represents a difficulty for MT systems, especially as regards the production of such adjectives when English is the target language. Very often, wide variation in the translation of these adjectives provides an explanation for the fact that no equivalent is provided for the adjective alone (e.g. evidence-based or radiation-induced) but instead only for complex noun phrases that include them. The relative scarcity of coverage of such forms in well-established terminology databases (Termium Plus, IATE) certainly warrants further investigation in this area.

The respective merits of the other two lexical resources that were consulted might also be researched further. The Linguee search engine is a powerful tool that can be put to good use by confirmed translators who can accurately rate the reliability of its translations. A similar tool is provided by Reverso (which also has a dictionary based on the 2005 edition of the Collins dictionary) and seems to give even more promising results when used on some of the PCPs mentioned here. All these recent developments clearly show that use of the Word Wide Web as a parallel corpus, which was heralded by Kilgarriff (2001) and Resnick & Smith (2003), will continue in years to come. With the increasing availability of reliable parallel corpora, the need to use comparable corpora, which afford access to examples of language use unaltered by the translation process (Baker 1998) may even be obviated. While such corpora remain a very useful tool for tracking variation of translation equivalents in specialized language, their compilation
and use are time-consuming, and the new direction taken by MT systems actually fulfills one of their functions, since target language-related collocational statistics obviously play a key role in generating their output.

**BIBLIOGRAPHY**


GLEDHILL, Christopher & Mojca PECMAN. 2015. “Cognitive packing and rhetorical packaging in English science writing: a comparative analysis of alternating pre-modified versus post-modified nominals”. In Colloque International de Linguistique GReG PLS IV; “Marqueurs et structures, quelles articulations dans la (re) construction du sens?” 6-7 November.


Webography (corpora and terminological data bases)

Corpus of Contemporary American English <http://corpus.byu.edu/coca/>

Corpus de médecine nucléaire du Centre de Recherche en Terminologie et Traduction <http://crtt.univ-lyon2.fr/les-corpus-medicaux-du-crtt-613310.kjsp>

Google Books Corpus (Ngram Viewer Interface <https://books.google.com/ngrams>

Grand Dictionnaire Terminologique <http://gdt.oqlf.gouv.qc.ca/>

IATE (InterActive Terminology for Europe) <http://iate.europa.eu/>

Linguee, dictionnaire et moteur de recherche de traductions <http://www.linguee.fr/>

Reverso Context, traducteur de mots et d’expressions <http://context.reverso.net/translation/>

Termium Plus <http://www.btb.termiumplus.gc.ca/>

NOTES

1. While Pechenick et alii (2015) have proved a certain degree of overrepresentation of scientific data in the Google Books corpus, it is reasonable to assume that the contrast would be even more pronounced if scientific English alone was taken into account, since the phrase based on evidence is more likely to occur in general English than its PCP version.

2. For a detailed description of MT performance for the <N N and N> or <ADJ N and N> structures, see Maniez 2008.

3. It is worth pointing out that Linguee, even though it is an interesting resource for experienced translators, has its limitations and should be used with caution, in combination with other resources.

4. In such cases, using a comparable corpus for the translation of PCPs often proves useful, as demonstrated for the adjective radiation-induced in the specialized field of nuclear medicine by Maniez (2013).

5. Let us add that the translation of leukocytes-induced by leucopéniant seems erroneous, since the causal relation is apparently inverted (leucopéniant actually means leukocytopenia-inducing).

6. The total number of words is over 3 million words for the English section and over 2.5 million words for the French section; the corpus is freely available for research at the URL listed below in the Webography section.

ABSTRACTS

In this article, we study the automatic translation into French of compound adjectives in English for Specific Purposes, with a special focus on adjectives consisting of a past participle preceded by a noun or a noun phrase. Such adjectives are ubiquitous in scientific English, and while many are part of terms for which a stable translation equivalent exists, some remain a challenge for translators, since head modification in English noun phrases often corresponds to less economic syntactic structures in Romance languages. After a brief review of the most productive past participle forms used in English compound adjectives, we evaluate the range of coverage of
compound adjectives formed with the past participles ‘based’ and ‘induced’ in a well-known terminology database (the Grand Dictionnaire Terminologique) and attempt to assess the quality of the output of the recently improved Google Translation online translation service for sentences containing such compound adjectives within expressions which are not yet lexicalized.

Dans cet article, nous étudions la traduction automatique en français d’adjectifs composés (AC) en anglais de spécialité, en mettant l’accent sur les adjectifs constitués d’un participe passé précédé d’un nom ou d’un groupe nominal. Ces adjectifs sont omniprésents dans l’anglais scientifique ; si nombre d’entre eux font partie de termes pour lesquels il existe un équivalent de traduction stable, certains constituent encore un défi pour les traducteurs, dans la mesure où la prémodification dans le groupe nominal en anglais correspond souvent à des structures syntaxiques moins économiques dans les diverses langues romanes. Après un bref passage en revue des formes de participe passé les plus productives utilisées dans les AC, nous évaluons l’étendue de la couverture des AC formés à partir des participes passés based et induced dans une base de données terminologique connue et nous tentons d’évaluer la qualité des résultats obtenus par Google Translation à la suite des récentes améliorations qu’il a subies, notamment dans le cas de phrases où ces AC sont utilisés dans des expressions qui ne sont pas encore lexicalisées.

INDEX

Keywords: compound adjective, corpus linguistics, machine translation, parallel corpus, terminology database, specialized translation

Mots-clés: adjectif composé, base de données terminologique, corpus parallèle, linguistique de corpus, traduction spécialisée, traduction automatique

AUTHOR

FRANÇOIS MANIEZ

François Maniez is a Professor of English linguistics and has been teaching English for Specific Purposes at Lumière Lyon 2 University (France) since 1990. He has directed the Centre for Research in Terminology and Translation since 2007, and was the coordinator of the translation of the bilingual (English-French) version of Dorland’s pocket medical dictionary in 2008. His main research interests are corpus linguistics, lexicology, lexicography and ESP, with a particular focus on the syntactic and lexical aspects of medical translation. francois.maniez@univ-lyon2.fr