Collection of a Large Database of French-English SMT Output Corrections

Marion Potet[1], Emmanuelle Esperança-Rodier[1], Laurent Besacier[1] and Hervé Blanchon[2]

[1]UJF-Grenoble1, [2]UPMF-Grenoble2
LIG UMR 5217, Grenoble, F-38041, France
FirstName.LastName@imag.fr

Abstract

Corpus-based approaches to machine translation (MT) rely on the availability of parallel corpora. To produce user-acceptable translation outputs, such systems need high quality data to be efficiently trained, optimized and evaluated. However, building high quality dataset is a relatively expensive task. In this paper, we describe the data collection and analysis of a large database of 10.881 SMT translation output hypotheses manually corrected. These post-editions were collected using Amazon's Mechanical Turk, following some ethical guidelines. A complete analysis of the collected data pointed out a high quality of the corrections with more than 87 % of the collected post-editions that improve hypotheses and more than 94 % of the crowdsourced post-editions which are at least of professional quality. We also post-edited 1,500 gold-standard reference translations (of bilingual parallel corpora generated by professional) and noticed that 72 % of these translations needed to be corrected during post-edition. We computed a proximity measure between the different kind of translations and pointed out that reference translations are as far from the hypotheses as from the corrected hypotheses (i.e. the post-editions). In light of these last findings, we discuss the adequation of text-based generated reference translations to train setence-to-sentence based SMT systems.

Keywords: Machine translation, post-editions, Amazon Mechanical Turk, evaluation

1. Context

In recent years, statistical machine translation systems (SMT) have made significant improvements in term of translation quality. However, MT outputs are far from being perfect and a most common practice is to manually correct them to produce high-quality translations. This correction task is called post-edition. Among the works aimed at collecting datasets of post-edited translations, we can cite (Specia et al., 2010), where 4,000 sentences and their translation hypotheses generated by four English-Spanish SMT have been manually post-edited and evaluated by professional translators. This work shows that such small-sized corpora can be useful to improve existing MT evaluation metrics. In another similar work described in (Specia, 2011), 2,525 French-English translations and 1,000 English-Spanish translations have been post-edited by professionals. In this work, we focused on building a bigger corpus and we hope these data will help research in the machine translation area.

2. Collection of post-edited translations

2.1. Baseline SMT

We created a baseline SMT system to translate news data from French into English. It is a phrase-based system using a log-linear model of 14 weighted feature functions. Our baseline system can be considered as state-of-the-art since it was presented to the international evaluation campaign WMT 2010[1] (described in (Potet et al., 2010) as *system 3*). The translations produced by this system will be later named *translation hypotheses*. In the meantime, for each source utterance a given professional translation reference named *gold-standard* is available.

2.2. Post-edition task

In our scenario, the task consists in correcting our SMT translation hypotheses. Given the French source sentence and its English translation hypothesis, a bilingual annotator has to verify the hypothesis quality and correct it if needed. We have chosen to post-edit 10,881 translations taken from several news corpora of the WMT evaluation campaign (from 2006 to 2010).
After post-edition, for each of the 10,881 source sentences we have:

- the translation *hypothesis* of our SMT system;
- the *corrected hypothesis* (i.e. the human post-edition);
- the *reference* translation (i.e. gold-standard) given in the initial parallel corpus.

In addition to post-editing these hypotheses and in order to control and evaluate the post-edition process itself, we also post-edited 1,500 gold-standard reference translations from the same corpus. A second corpus of 1,500 sentences has been created, containing in addition, the post-edition of the *reference* translation (named *corrected reference*).

2.3. Data collection platform

For this post-edition task, we used a crowdsourcing platform: Amazon's Mechanical Turk (MTurk)[2]. MTurk is an online platform that allows a "requester" to propose a paid or unpaid work and a "worker" to perform the proposed tasks. These tasks are called "Human Intelligence Tasks" (or HITs) because they rely on human know-how. Many recent articles have shown the effectiveness of MTurk to

[1]http://www.statmt.org/wmt10/translation-task.html

[2]http://www.mturk.com

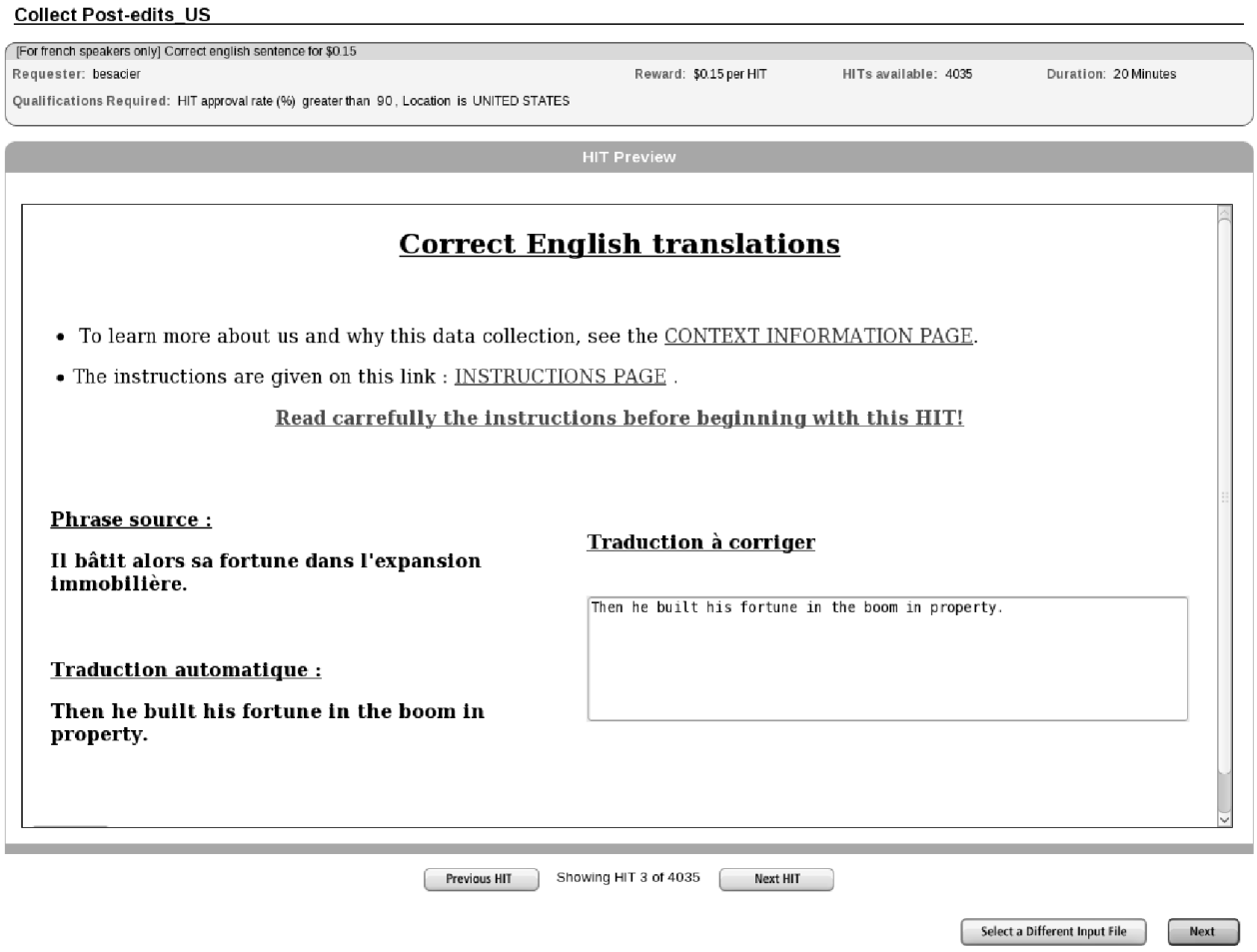

Figure 1: Post-edition task interface

create annotated data in the field of natural language processing but, however, the ethical, social and economic aspects implied by such tools are subject to intense debates (Fort et al., 2011). So, as it was done in (Gelas et al., 2011), we defined some guidelines of "good practice" to use Mturk for research purpose:
- the collected information should be used for non-profit organization and available for free to the community;
- as done traditionally for data collection, the worker should be informed, in HITs descriptions, about the context of the task (Who we are? What are we doing? And why?);
- the worker should be paid a decent amount (with a reasonable hourly rate);
- workers should be filtered by country of residence according to the task, to avoid those who consider MTurk as their major source of income.

2.4. Interface

We developed the interface shown in figure 1. The corpus to post-edit was divided into sentences, so, a HIT was corresponding to one translation to be corrected. We proposed a hourly rate of about $15 (or $0.15 per post-edited sentence) and redirected workers to a Web page providing information about the task context.

2.5. Instructions to the Human Post-Editors

Participants were requested to necessarily fairly understand French and be fluent English speakers. In order to better fit in with these constraints, we only authorized USA, Canada and France inhabitants to participate to the task and wrote the instructions in French (the interface was in English). We did not use any particular qualification test to guarantee that the post-editors were English native speakers. The analysis of the post-editions provided by non-necessarily native *vs* native post-editors is analysed in section 3. The post-edition task's instructions were the following:

"The task is to correct English translations of French sentences. Each translation must be checked and, if necessary, corrected into a correct translation of the French sentence. For each sentence (or HIT) to correct, be sure to follow these instructions:
(1) the produced translation must be a syntactically and semantically acceptable English sentence;
(2) make sure that all the French sentence concepts and notions appear in the translation (and vice versa, there should not be concept in translation that are not included in the source sentence);
(3) you must minimize the corrections: make only the necessary changes for compliance with (1) and (2);

Comparison with MT output	Translation judgment			Total
	Perfect	Pretty good	Partially mistaken	
Better	183 sent.	43 sent.	50 sent.	87.1 %
Equivalent	10 sent.	20 sent.	0	12.5 %
Worse	0	1 sent.	0	0.3 %
Total	64.0 %	17.3 %	19.0%	/ 311 sent.

Table 1: Subjective evaluation of post-edition quality

(4) you should keep the translation as close as possible to the French sentence (in terms of vocabulary, word order, etc.);
(5) transcribe the punctuation as it is in the source sentence.
Corrected translations will be checked one by one and any breach of one of these instructions would be a motif of a rejection."
Once a worker submitted a HIT, the requester had the choice to validate it (in that case the worker was paid and the post-edition registered) or reject it. This reviewing has been carried out by ourselves and post-editions which did not respect the given instructions were rejected.

3. Analysis of the data collected

3.1. Collection analysis

We submitted a total of 12,381 translated phrases to be corrected. The data collection spanned about 4 months and its cost was $2,040 ($1,855 for Workers and $185 for MTurk fees). We did some statistics about the participants and the validated/rejected HITs. In total, 553 people took part in the task but only 70 % of them have really helped (i.e. had at least one validated post-edition). This means that 30 % of those who participated have seen all their work rejected for poor quality. On average, a submitted post-edition has 63 % chance (about two chances out of three) of being rejected for poor quality and, on average, a participant submitted around 60 post-editions. The majority, i.e. 62 %, of participants submitted less than 10 post-editions and only six contributors have produced more than 500 validated sentences.

3.2. Evaluation of post-edition quality

3.2.1. Subjective Evaluation of post-edition quality

A former professional post-editor estimated the quality of the collected post-editions evaluating a sample of our post-edited corpus. To select the sample to evaluate, we proceeded as follows: we considered the sentences close to the overall corpus average BLEU score computed between post-editions and corresponding MT outputs (the BLEU score can be interpreted as a measure distance), selected at most 2 sentences for Turkers who post-edited at all less than 10 sentences and 5 sentences for the others. We then, randomly extracted a sample of 311 setences from this sub-corpus.
First, each of these 311 post-edition was compared to its corresponding translation hypothesis and labeled as Better, Equivalent or Worse. Subsequently, regardless of this decision, each post-edition was evaluated as a translation of the source sentence and classified according to its quality. The classification is the following:

Perfect No error in the post-edition;

Pretty good Post-edition contains one or more oversight and unsignificant mistakes which make the translation correct but imprecise. This can be due to wrong or missing uppercase or punctuation, inappropriate term, minor grammar mistake, etc.;

Partially mistaken The post-edition contains one or more significant or serious error which make the translation incomplete or wrong (in this case, the translation does not transmit the whole meaning of the source sentence). This can be due to a grammar error, adjective or concept omission or mistranslation;

We should note that this evaluation was carried out impartially: post-editions containing just one mistake or error on a word are considered as "bad" even if the rest of the sentence is good ; in the same way, a post-edition containing several mistakes is classified according to its biggest error.
The distribution of the sentences according to errors and post-edition labels is presented in table 1. The results showed that 87.1 % of post-editions improved the hypothesis, while 12.5 % are considered as equivalent (they do not improve the translation hypotheses). Even if the corrections improve translation quality, some post-edited sentences still contain errors that should have been corrected. Indeed, 64 % of the post-editions are perfectly good translations of the source sentence but 19 % of them are partially mistaken.

3.2.2. Comparison with professional post-editions

Finally, we manually evaluated our post-editions collected with MTurk comparing some of them with professional post-editions. In a previous similar work, presented in (Specia, 2011), a sub-corpus of 2,525 sentences included in our 10,881 post-edited sentences on MTurk had been translated by another SMT and the obtained translations were post-edited by a professional translator (a bilingual native speaker of French and English with a first degree in Translation studies). It is important to note that the instructions and post-edition context are equivalent in both cases: from the translation hypothesis, post-editors were asked to perform the minimum number of corrections needed to obtain a "publishable" translation of the source sentence, with no time constraints and the job was also paid. Moreover, the SMT systems used in both experiments are very similar (both are phrase-based SMT trained with the Moses toolkit and default settings; our SMT has a BLEU score of 20.20 while the other one has a BLEU score of 20.76).

To confirm the similarity of the two systems, we computed for each of them, the distribution of sentence-based TER scores (Translation Error Rate) and those are very close which clearly illustrates very similar SMT.
First, as seen in table 2, we noticed that MTurk post-editions are farther from our translation hypotheses than the professional post-editions from the corresponding SMT. In the same way, only 5.5 % of the translation hypotheses have not been corrected by the Turkers against 13 % with the professional translators. This means that, in a very similar experiment, professional post-editors correct fewer sentences than MTurk post-editors. Except for that, TER score (at sentence level) distributions between translation hypotheses and post-editions are very similar in both experiments.

	Professional	Mturk
Proximity (=BLEU score)	65,30	50,70
Non-post-edited hypotheses	13 %	5.5 %

Table 2: Comparison of professional and MTurk post-editions (total of 2,525 sentences)

We then drew our attention to sentences for which the translation hypothesis was the same for both systems. About 6 % (146 sentences) generated the same translation hypothesis regardless of the translation system. Among these, 146 common translation hypotheses, 35 (or 24 %) also have the same post-edition. We thus obtained a set of 111 sentences for which we can directly compare the professional post-editions with our MTurk post-editions. Results of the subjective analysis of these sentences are presented in table 3. The human preference evaluation shows that 26.1 % of the MTurk post-editions are considered to be better than the professional ones. We can notice that, in a large majority of cases (i.e. for 67.6 % of sentences) both post-editions are judged as equivalent, but surprisingly, the professional post-editions are judged better than the Mturk ones in only 6.3 % of cases. In other words, as 93.7 % of the Mturk post-editions are considered at least of professional quality, we can infer about the high quality of our collected corpus.

Preference	♯ sentences (%)
Better Mturk post-edition	29 (26.1 %)
Better professional post-edition	7 (6.3 %)
Equivalent post-edition	75 (67.6 %)

Table 3: Quality comparison between professional and MTurk post-editions (total of 111 sentences)

3.3. Overall post-edition corpus caracteristics

3.3.1. Translation hypotheses post-edition

The post-edition task gives us a corpus of 10,881 sentences (corpus-10881) to analyze translation hypothesis corrections. Some examples of hypothesis corrections are given in figure 6. The ratio between hypothesis size and post-edition size is 1.02; this means that the post-editing task adds little or no words to the translation hypothesis. On the other hand, as seen in table 4, 9 % of the hypotheses did not require any correction during the post-edition. That is to say, 9 % of MT outputs have been considered as perfect translations of the source sentence.

3.3.2. Reference translations post-edition

For the sub-corpus of 1,500 sentences (corpus-1500), we also (in addition to SMT hypothesis corrections) collected and analyzed *gold standard* translation corrections. Some examples of reference corrections are given in figure 7. The rates of non-post-edited translations on this sub-corpus of 1,500 sentences are presented in table 4. We observe the same trend as for the whole corpus of 10,881 sentences: 10 % of the hypotheses derived from the SMT did not require corrections. But surprisingly, only 28 % of the *gold standard* translations were considered as correct, i.e. 72 % of the reference translations were corrected during post-edition. This can be explained by the fact that reference professional translations have been produced in a context of a whole text translation, that is to say, they have been produced considering the entire story. However, as developing automatic translation systems requires news stories texts splitted into sentences, we can easily imagine that a missing concept in a sentence translation appears, in fact, in a previous sentence. So, these professional translations are not really appropriate for a sentence-based training or evaluation.

Rate of non-post-edited sentences	Overall corpus (10,881 sent.)	Sub-corpus (1,500 sent.)
Translation hypotheses	9 %	10 %
Gold standard translations	/	28 %

Table 4: Rate of translations considered as perfect according to the corpus

3.3.3. Distance between the different translations

We computed the BLEU score between different corpora: the obtained score can be interpreted as a proximity measure (between 0 and 100) accounting for the n-gram intersection of two texts. Figure 2 shows a distance calculated by $d = 100 - BLEUscore$. In the same way, table 5 represents the BLEU scores calculated between the different translations for corpus-1500 and corpus-10881.
We observe that reference translations are farther from system translation hypotheses than corrected translations. Similarly, corrected translations and *gold standards* are dramatically different while both are supposed to be a correct translation of the same source sentence.

4. Conclusion

We described a large database of 10,881 SMT outputs manually post-edited. The corpus was collected using a crowdsourcing Web platform (MTurk), and the data collection method required us to define clear guidelines and strict control reviewing to guarantee correction quality. Post-editions quality was evaluated and carefully compared with a professional work. Analysis results shows that even untrained-human post-editor (native of the target language

Corpus	Type of compared translations	BLEU Score
corpus-10881	Hypothesis & Reference	27,03
	Hypothesis & Corrected Hypothesis	61,41
	Corrected Hypothesis & Reference	32,20
corpus-1500	Hypothesis & Reference	29,48
	Hypothesis & Corrected Hypothesis	65,10
	Corrected Hypothesis & Reference	36,96
	Reference & Corrected Reference	71,90
	Corrected Hypothesis & Corrected Reference	43,40

Table 5: Proximity, in terms of BLEU score, between the different corpora

Source Sentence	Translation Hypothesis	Corrected Hypothesis
• La police anti-émeutes les ont aussitôt encerclés et sont intervenus sans ménagement, jetant plusieurs d'entre eux à terre.	• The anti-riot police were immediately surrounded and spoke blunty, several of them on land.	• ~~The~~ **A**nti-riot police**men** ~~were~~ immediately surrounded **them** and ~~spoke blunty~~ **stepped in ruthlessly**, **throwing** several of them ~~on land~~ **to the ground**.
• Le cinquième candidat affirme ne soutenir ni le pouvoir, ni l'opposition.	• The fifth candidate says it support nor the current leadership, nor the opposition.	• The fifth candidate says ~~it~~ **he** support ~~nor~~ **neither** the current leadership, nor the opposition.
• Forte mobilisation à Copenhague et à travers le mode, pour le climat.	• Strong involvement in Copenhague and in the world climate.	• Strong ~~involvement~~ **mobilization** in Copenhague and ~~in~~ **across** the world **for the** climate.
•Il y a des rivières qui s'assèchent en Afrique, des cours d'eau où l'on peut marcher comme on ne l'avait jamais fait avant.	• There are rivers are drying up in Africa, rivers where you can walk as it had never done before.	• There are rivers ~~are~~ drying up in Africa, ~~rivers~~ **watercourses** where ~~you~~ **one** can walk as it had never done before.

Table 6: Examples of translation hypothesis corrections

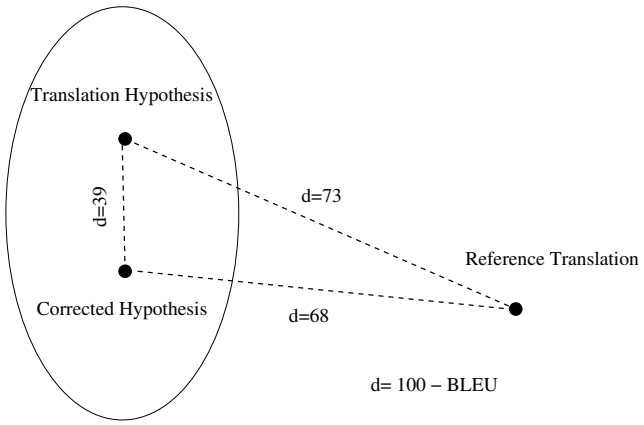

Figure 2: Distances between the different translations

or not) were able to generate high quality translations in a moderately controlled context. Strict instructions in background have certainly contributed of the task success. The large database described in this paper is freely available to the community and can be downloaded from:
http://www-clips.imag.fr/geod/User/marion.potet/index.php?page=download.

As suggested in (Potet et al., 2011), the collected data can be part of an iterative translation workflow. So, as they are supposed to be the correct translations closest to our baseline general-domain standard phrase-based SMT outputs, such corpora could be used, for example, to improve machine translation evaluation metrics (Specia et al., 2010), for MT confidence estimation (Bach et al., 2011) or for PBMT automatic post-edition purpose (Kuhn et al., 2010).

Another finding of our work concerns the adequacy of the professional reference sentences usually used to train, tune and test models and metrics. As these translations have been produced considering a whole context they contain, at the sentence level, errors like source concept omissions. In addition to that, they are very far from machine outputs in terms of syntax and style. These two facts question their relevance for SMT training and evaluation.

5. References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Hu-*

Source Sentence	Reference Translation	Corrected Reference
• Mais une fiscalité insuffisante peut également produire les mêmes effets.	• Too little taxation can do the same.	• **But** ~~Too little~~ **an insufficient** taxation can **also** do ~~have~~ the same **effects**.
• Le malaise français n'a certainement pas été induit par ces réformes.	• The French malaise has nothing to do with any of them.	• The French malaise ~~has nothing to do with~~ **was certainly not induced by** ~~any of them~~ **these reforms**.
• Mais quelle est la signification réelle de ces deux principes ?	• But what do solidarity and subsidiarity really mean?	• But what ~~do solidarity and subsidiarity really mean~~ **is the real meaning of these two principles** ?
• Les traités européens expriment clairement cette subsidiarité verticale.	• In the European Treaties, we find a clear expression of vertical subsidiarity.	• ~~In~~ **T**he **e**uropean **t**reaties ~~we find a clear expression of~~ **clearly express this** vertical subsidiarity.

Table 7: Examples of reference translation corrections

man Language Technologies (ACL-HLT), Portland, Oregon, June.

Karn Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Journal of Computational Linguistics*, 37:413–420, June.

Hadrien Gelas, Solomon Teferra Abate, Laurent Besacier, and Francois Pellegrino. 2011. Evaluation of crowdsourcing transcriptions for african languages. In *proceedings of Human Language Technologies for Development (HLTD)*, Alexandrie, Egypt, May.

Roland Kuhn, Pierre Isabelle, Cyril Goutte, Jean Senellart, Michel Simard, and Nicola Ueffing. 2010. Recent advances in automatic post-editing. *Journal of Multilingual computing and technology*, 21(1):43–46.

Marion Potet, Laurent Besacier, and Hervé Blanchon. 2010. The lig machine translation system for wmt 2010. In *proceedings of Workshop on Machine Translation (WMT'10)*, Uppsala, Sweden.

Marion Potet, Emmanuelle Esperança Rodier, Hervé Blanchon, and Laurent Besacier. 2011. Preliminary experiments on using users post-editions to enhance a smt system. In *proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT)*, Leuven, Belgium, May.

Lucia Specia, Nicolas Cancedda, and Marc Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In *proccedings of the 7th Conference on International Language Resources and Evaluation (LREC)*, Valletta, Malta, May.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 73–80, Leuven, Belgium, May.