Evaluating RBMT output for -ing forms: A study of four target languages

Nora Aranberri-Monasterio
Dublin City University

Sharon O'Brien
Dublin City University

-ing forms in English are reported to be problematic for Machine Translation and are often the focus of rules in Controlled Language rule sets. We investigated how problematic -ing forms are for an RBMT system, translating into four target languages in the IT domain. Constituent-based human evaluation was used and the results showed that, in general, -ing forms do not deserve their bad reputation. A comparison with the results of five automated MT evaluation metrics showed promising correlations. Some issues prevail, however, and can vary from target language to target language. We propose different strategies for dealing with these problems, such as Controlled Language rules, semi-automatic post-editing, source text tagging and "post-editing" the source text.

1. Introduction

The focus of this paper is on evaluating the Machine Translation (MT) output for one linguistic feature, -ing forms, into four target languages (French, Spanish, German and Japanese). Our interest in -ing forms stems from our study of Controlled Language (CL). Controlled Language is defined as "an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style" (Huijsen, 1998, p. 2). CL rules can be implemented to reduce ambiguities in the source text in order to improve the machine translated output (Bernth and Gdaniec, 2001; O'Brien, 2003).

CL rule sets often include one or more rules on -ing forms in English. O'Brien (2003) found that six of the eight CLs she analysed shared a rule which recommended avoiding "gerunds". According to Dervišević and Steensland (2005), AECMA Simplified English does not allow the use of either gerunds or present participles, with the exception of certain technical terms. The Microsoft Manual of Style for Technical Publications (MSTP) (Microsoft Corporation, 1998) cautions against the use of gerunds. There is at least some consensus, then, that -ing forms can be problematic for RBMT. The following example, taken from our research corpus, illustrates the problem:

ST: *Viewing and changing active jobs*
DE: **Anzeigende und ändernde in Arbeit befindliche Programmteile*
ES: **Viendo y modificando tareas activas*
FR: **Les JOBs actifs de visionnement et changeants*
JA: **実行中のジョブを表示し、変更します*

Rules controlling the use of -ing forms are often formulated in very general terms (e.g. “Avoid the use of -ings”) and, consequently, technical writers find them difficult to implement. -ing forms can be categorised into different functional linguistic categories and sometimes the CL rule seeks only to govern “gerunds” (e.g. “Avoid the use of gerunds”).[1] There are two problems with this. First, previous research has not made it clear why gerunds, and not other categories of -ing forms, are specifically targeted. Second, technical writers typically do not have a background in grammar or linguistics and the term “gerund” is therefore difficult for them to comprehend.[2] Given that there is general consensus that -ing forms, or at least gerunds, can create problems in RBMT output, coupled with the vagueness of rules governing this phenomenon, we felt that there was a need for more detailed research on this topic.

Our research was co-funded by Enterprise Ireland and Symantec under the Innovation Partnerships Programme. Symantec implemented the MT system, Systran, in 2006 to meet their increasing translation volumes for security alert information. They also implemented customised Controlled Language checking rules using the acrocheck™ tool.[3] Both Systran (version 5.05) and acrocheck™ were used in this research and the corpus was compiled from Symantec technical documentation.

Our primary research questions were: How problematic for RBMT are -ing forms, and what processes can we implement to reduce those problems for at least four target languages? To answer these questions it was necessary to evaluate the MT output for -ing forms. Automatic evaluation metrics such as BLEU (Papineni et al., 2002) and NIST (NIST report 2002) are commonly used for MT evaluation. Human evaluation of output is also used either in conjunction with automatic metrics or on its own. Much has been written about how these metrics work and on how human evaluation results correlate (or not) with automated metrics (cf. Callison-Burch et al., 2006). However, little has been written specifically on the evaluation of MT output for the -ing form and, to the best of our knowledge, no detailed, contrastive analysis has been published to date on -ing forms and their MT output into multiple target languages.

Section 2 of this paper discusses the methodology used for compiling the corpus, classifying the -ing forms and evaluating them. Section 3 gives details of our results and Section 4 provides the conclusions and an outline of future work.

2. Methodology

2.1. Classifying -ing forms

In order to analyse what effect -ing forms have on MT output in different languages, we needed a useful classification system. Traditionally, words ending in -ing have been divided into two categories: gerunds and participles (Quirk et al., 1985). Huddleston and Pullum (2002) claim that the current usage of the English language shows no systematic correlation of differences in form, function and aspect between the traditional gerund and present participle. They propose that words with a verb base and the -ing suffix be classified as gerundial nouns (genuine nouns); gerund-participles (forms with a strong verbal flavour); and participial adjectives (genuine adjectives).

In grammar books, -ing forms are described under the sections of different types of word classes, phrases or clauses in which they can appear, that is, a syntactic description of the -ing form is spread throughout the grammar description. However, no classification has focused on -ing forms as a main topic or in a detailed manner. Izquierdo (2006) faced this deficiency when carrying out a contrastive study of the -ing form and its translation into Spanish. She compiled a general language parallel English-Spanish corpus, mainly consisting of texts extracted from fiction, and analysed the -ing forms, comparing the theoretical framework set out in grammar books and the actual uses found in her corpus. She established a functional classification of -ing forms (see Table 1).

Table 1: Izquierdo's (2006) functional classification of -ing forms

Functions	Adverbial	Progressive	Characterisation		Referential
Grammatical Structures	time	past	Pre-modifiers	Post-modifiers	catenative
	process	present	participial adjective	reduced relative clause	prepositional clause
	purpose	future		nominal adjunct	subject
	contrast	conditional		adjectival adjunct	direct object
	place	etc.			attribute complement
	condition				comparative subordinate
	etc.				

Izquierdo's classification was considered suitable for our study for several reasons. Firstly, we focus on RBMT systems, that is, the analysis, transfer and generation modules are built upon grammatical rules. Therefore, a classification that could describe fixed grammatical patterns was considered appropriate.

Secondly, by using the syntax-based tagger of our CL checker (acrocheck™) during a search of our corpus, the behaviour of the checker for these particular forms would be better understood.

Thirdly, -ing forms cannot be classified in isolation; contextual information must be considered in order to distinguish a gerundial noun from a participial adjective or a gerund-participle. The functional classification would provide boundaries for this context.

Additionally, it would allow us to test whether the same classification used for general language would be suitable for a specialised domain such as IT.

2.2. Corpus Compilation

One of the first questions we had to answer regarding the design of our research was whether to use a test suite or a corpus in order to study the -ing form. Test suites allow the researcher to isolate the linguistic structures under study and to perform an exhaustive analysis of all the possible combinations of a specific linguistic phenomenon, with the certainty that each variation will only appear once (Balkan et al., 1994). On the other hand, a corpus allows the researcher to focus on authentic and real texts, on language as it is used (McEnnery et al., 2006, pp. 6-7). Given that this research focuses on text produced and machine-translated in an industrial context, we felt it was important to use a corpus that represented -ing forms as they are produced by technical writers. However, the corpus approach is not without its problems, which we discuss below.

It is essential to ensure the validity of a corpus, i.e. its suitability for studying the selected linguistic phenomenon. Literature on corpus design highlights the difficulties in guaranteeing ecological and sample validity. Yet, authors concur that the decisions made must depend on the purpose of each study (Bowker & Pearson, 2002, pp. 45-57; Kennedy, 1998, pp. 60-85; Olohan, 2004, pp. 45-61).

Kennedy (1998, pp. 60-70) highlights three design issues to be taken into consideration when building a corpus: stasis and dynamism; representativeness and balance; and size. A dynamic corpus is one that is constantly upgraded whereas a static corpus includes a fixed set of texts gathered in a specific moment in time. The aim of the present research is to study the current performance of our RBMT system when dealing with -ing forms given the current level of MT system development and source text quality. Dynamic corpora are mainly used when trying to capture the latest uses of language or when studying linguistic changes over time. Since we did not expect the use of -ing forms to change, we opted for a static corpus.

Representativeness is the second design issue highlighted by Kennedy (ibid, pp. 62-65). He points out that it is difficult to ensure that the conclusions drawn from the analysis of a particular corpus can be extrapolated to the language or genre studied (ibid, pp. 60). We focus on the -ing words which appear in IT manuals (user guides, installation guides, administrator's guides, etc.). These documents have in common that they are made up of descriptive and procedural text-types. Text-types are "groupings of texts that are similar with respect to their linguistic form" (Biber, 1988, pp. 70), which means that the syntactic patterns tend to be stable. This increases the representativeness of our data. Yet, additional controls suggested by Bowker and Pearson (2002, pp. 49-52) were taken into account in selecting the texts: text length, number, medium, subject, type, authorship, language and publication date. Complete texts were used so as to ensure that any variation in -ing form use from one text section to the next would be represented. Bowker and Pearson also recommend that studies of linguistic features include a series of texts written by a series of authors so as to avoid idiosyncratic uses affecting the results. In order to address this issue, texts describing different products and written by different writing teams were included. It was decided to use texts which had not undergone any language control, that is, the selected texts should not have been written following the Controlled Language rules. This would make it possible to measure the extent to which -ing forms cause problems prior to implementation of CL rules and also allow us to develop procedures for fixing any problems encountered (see "Conclusions").

Bowker and Pearson (2002, p. 48) state that in studies related to language for specialized purposes (LSP), corpus sizes ranging from between ten thousand to several hundred thousand words have proven "exceptionally useful". Following this, the initial corpus created for this project amounted to 494,618 words.

We feel that the ecological validity of the corpus was ensured by using real texts that meet both the relevant number, authorship and date variation, and stability in subject, type and medium as required for the population for which we intend to draw conclusions.

With the classification system in place and the corpus compiled, we then had to extract all occurrences of -ing forms in the corpus and classify them before sending them to the RBMT system and having their translations evaluated. Using acrocheck™ to extract as many instances as possible of -ing forms, 8,316 instances were classified from a total of 10,417 in the corpus, i.e. 79.83%.[4] Such high correlation with a classification presented from general language further reflects the suitability and coverage of our corpus. The classification of the 8,316 instances is shown in the appendix. One modification was made to Izquierdo's classification, i.e. we introduced the category of "Titles" starting with -ing forms. Titles have a high level of occurrence in instruction manuals and they are not always handled correctly by RBMT systems. Titles which start with -ing forms often require a different translation from identical -ing forms in running text. For example:

ST [Title]: *<u>Using</u> default su or query credentials*
ES [Title]: *<u>Usar</u> configuración predeterminada su o credenciales de consulta*
Vs.
ST: *Creating a policy for true image restore by <u>using</u> the Policy Wizard*
ES: *Crear una política para el establecimiento de la imagen verdadera <u>usando</u> el asistente de políticas*

This difference causes difficulties for RBMT systems, which are not yet designed to distinguish between running text and titles. It was therefore considered essential to study the performance of the MT system when dealing with this particular structure, especially given their frequency of occurrence (25% of -ing forms in the corpus).

2.3. Human Evaluation

Since our research focuses on evaluating the RBMT output for -ing forms and little work has to date been done using automated metrics for specific sub-sentential linguistic constituents (with the exception of constituents such as subjects, NPs and CNPs evaluated by Callison-Burch et al. (2007)), we opted for a human evaluation. While large-scale machine translation evaluation initiatives such as the NIST Open Evaluation or the Shared Translation Task in the ACL Workshop on Statistical MT use unlimited numbers of human judges to evaluate as many examples as they choose, smaller experiments report results based on 3 to 7 judges and this is expected to allow for enough variation, particularly if the evaluation is performed by experts (Pierce, 1966; Elliott et al., 2004; Estrella et al., 2007). In keeping with this trend, we hired four professional translators per target language to judge whether the translations of the -ing forms were "correct" or "incorrect". By "correct" we mean grammatical and accurate. MT evaluation experiments have recently converged into reporting the parameters "fluency" and "adequacy" (or "accuracy") (Cieri et al., 2007; etc). Whereas the meaning of adequacy seems generally agreed upon, the meaning of fluency is less clear. As Mutton et al. (2007) discuss, authors have defined fluency differently, ranging from closer to grammaticality (Pan and Shaw, 2004) to an intuitive reaction (NIST MT Evaluation Plan Guidelines, 2005), as well as including attributes such as rhythm and flow, among others (Coch, 1996). The LDC (2003) defines fluency as the "degree to which the translation is well-formed according to the grammar of the target language", thereby bringing it close to the definition of grammaticality. Since the research was being conducted in an industrial setting, our aim was to learn whether the translation generated by the RBMT system was ready for publication. The standard acceptable for publication depends on the expected function of the translated text. For our context, the minimum quality required for publication was set as a grammatical text which transferred the same information as the original text.

The use of human evaluators limited how many examples could be judged. We used a stratified systematic sampling technique to extract an evaluation set of 1,800 examples. Evaluators were asked to judge the translation of the -ing words only. They were presented with a source segment in which the -ing form to be judged was highlighted (so that it could be easily identified), together with the machine translation of the segment and a post-edited version, which they were told to use for reference purposes as an example of what could be accepted for publication. Due to the novelty of the constituent-based approach, evaluators were provided with some guidelines. This allowed for better understanding of our aim and, we hoped, a higher level of consistency.

The analysis of the results was performed by a native-speaker linguist for each target language. They were asked to examine the examples that the evaluators had judged as incorrect. They were provided with guidelines to ensure that the results from all four target languages could be compared.

2.4. Testing for Correlations between Human and Automated Metrics

Human evaluation is time-consuming and expensive, and its reliability has been a hot topic in recent years (Vilar et al., 2007). As an alternative, automated metrics have been proposed to measure the quality of MT output. Most of these metrics compare the machine translation output against one (or more) reference translations and report a score based on their similarity. The most widespread within MT evaluation experiments are string-based metrics, such as BLEU and NIST. These metrics, however, report the results for the text or sentence level, and their usefulness for calculating scores for a sub-sentential linguistic feature remains largely unexplored. Therefore, we decided to test correlations between the constituent-based human evaluation and a constituent-based automatic evaluation.[5]

We chose 5 different metrics that could be run using short constituents, namely, n-gram-based NIST (NIST Report, 2002), word-based GTM (Turian et al., 2003), TER (Snover & Dorr, 2006; Przybocki et al., 2006) and METEOR (Banerjee & Lavie, 2005), and character-based edit-distance (NLTK). The most widespread BLEU metric did not allow us to work on short constituents, as it uses a geometric mean to average the n-gram overlap (therefore, if one of the values of n produces a zero score, the total score is nullified). NIST, however, combines the scores for 1 to 5 n-grams using the arithmetic average and can be used with short segements. The GTM metric, based on precision and recall and the composite f measure instead of n-grams, pays less attention to word order. Thus there is no penalty for short segments and it can be used with constituents. We chose TER because it also calculates the distance between the MT-generated output and the reference translation, but does so by counting the number of insertions, deletions and substitutions required, based on edit-distance techniques. The last word-based metric used was METEOR. This metric diverges from the previous ones in that it uses a stemmer to calculate the scores. Finally, we included a character-based edit-distance metric, to examine whether better correlations could be found by using a character-based metric instead of a word-based metric in short constituents.

In order to obtain the -ing constituents, we asked the native-speaker linguists to read the 1,800-sentence evaluation set provided to the human evaluators and to highlight the translation of the -ing words in the MT output and the post-edited versions. They were required to do so according to the same criteria given to the human evaluators in the guidelines. The highlighted segments were extracted and treated as sentence units for input to the metrics. The segments obtained from the post-edited version were used as reference segments.

3. Results

The results of the human evaluation showed that for German, Japanese and Spanish, 72%-73% of the -ing forms were grammatically and accurately translated (see Figure 1). The average for French was lower, with 52% of the examples classified as correct. This lower score was mainly due to two frequently occurring -ing constituents, which were consistently translated incorrectly by the RBMT system for French. The human evaluation outcome, although impossible to compare with other problematic structures due to lack of similar exhaustive research, demonstrates that this RBMT system handles -ing words quite well. Yet there is clearly room for improvement.

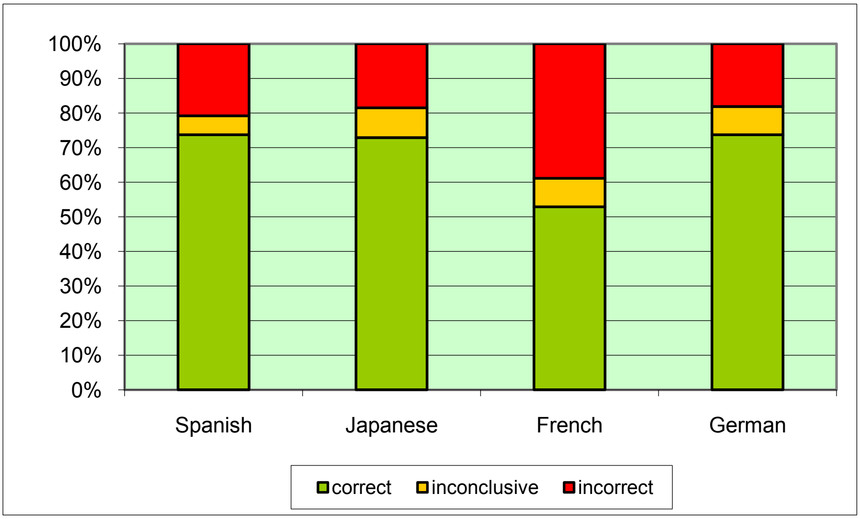

Figure 1: Percentage of correct and incorrect examples across the 4 target languages

We tested the validity of the evaluation by using the (Fleiss) kappa inter-rater measurement to calculate the reliability of the answers provided by the evaluators. The agreement was good for French (K=0.702), German (K=0.630) and Spanish (K=0.641) and moderate for Japanese (K=0.503). The results were satisfactory for two reasons. First, they confirm that the constituent-based approach to evaluation can obtain good inter-rater correlations (Callison-Burch et al., 2007). Second, the constituent-based approach is suitable for evaluating an attribute such as grammaticality and accuracy and does not need to be restricted to a ranking evaluation.

The first group in our classification were titles. The evaluation showed that for French (61%), Japanese (32%) and Spanish (36%) titles were problematic. For German, correct translations were more frequent, but 20% remained incorrect. Two types of problem arose with titles. First, a number of gerund-participles were analysed as participial adjectives and translated as modifiers into all target languages. Second, generation problems were observed, in which apparently correctly analysed gerund-participles were incorrectly translated as gerunds into Spanish, as infinitives and present participles into French and as nouns and gerunds functioning as subjects of the misanalysed plural nouns following the -ing form into Japanese. See Figure 2 for the percentage of correct examples, per category and target language.

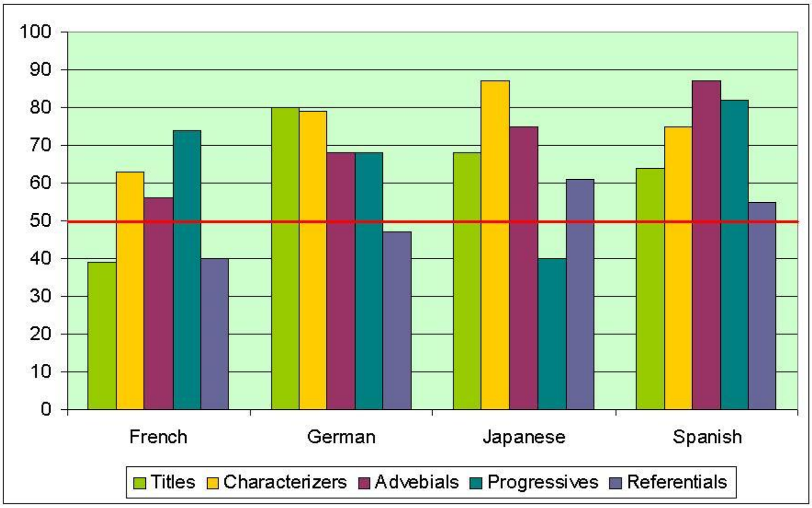

Figure 2: Percentage of correct examples for each -ing constituent category per target language

Characterizers were our second group. Whereas pre-modifiers were generally correctly analysed and translated when the terms were included in the MT system's dictionaries, we observed that post-modifiers presented more problems. The MT system generated post-modifying structures for the target languages; these structures, however, were not completely grammatical. For instance, for French, passive voice reduced relative clauses were translated into a combined structure of passives and participles. For Japanese, post-modifiers tended to show dependency errors. Equally, the MT system often failed (across all four languages) to generate correct prepositions and word classes following adjuncts.

The third group covered adverbial clauses with an -ing head. Spanish and Japanese performed better, with respective averages of 87% and 75% correct examples. German obtained a lower number of correct examples, at 68%. The target language most affected by these constituents was French, for which only 56% of the examples were translated correctly. All target languages showed problems with the choice between preposition or subordinate conjunction. Japanese and German, in particular, displayed ambiguity issues with gerund-participles translated as modifiers. Japanese encountered dependency errors, and the German output used incorrect pronouns to refer to the implicit subjects. French performed poorly in the translation of the constituent "*when + -ing*" as, when trying to generate an impersonal subordinate clause, the MT system created gerunds, which is incorrect for this context.

The -ing forms which combine with verbal tenses to introduce progressive aspect were our fourth group. For French and Spanish, this group performed well with respectively 74% and 82% of examples evaluated as correct. The issues found for these target languages were mainly due to the combination of continuous tenses and the passive voice and, in particular for French, the loss of progressive aspect. For German, the number of examples translated correctly was 68%, mainly due to these -ing forms being translated as nouns. For Japanese, the translation of the -ing forms in this group was predominantly incorrect, with only 40% of output correct. Despite the poor performance for Japanese, on average this group performed well across languages.

Finally, let us review the group of referential -ing forms. This was by far the worst-performing group, with 61% correct examples for Japanese, 55% for Spanish, 47% for German and 40% for French. We noticed that most issues were due to lack of translation resources. For instance, gerundial nouns were incorrectly translated in the cases where the MT system did not have the appropriate terminology available. Another example is catenative constituents, in which -ing forms were translated into incorrect word classes, leading to a literal translation that was often incorrect in the target languages. Similar issues were noted for phrasal verbs as for gerundial nouns, whereas prepositional verbs behaved more like catenatives. We observed that the particular constituents within each subgroup performed differently for each target language.

3.1. Correlation between human evaluation and automatic metrics

Our aim was to examine whether the -ing constituent evaluation could be performed using some of the existing automatic metrics. We isolated the constituents and their translations and we calculated the NIST, TER, GTM, METEOR and character edit-distance scores. Because we had four evaluators, the examples were therefore divided into 5 categories in which, in the worst case, none of the evaluators considered the example correct (0), and in the best case, all four evaluators considered the example correct (4). When one evaluator considered a translation to be correct, this corresponds to 1 on our x axis (See Figures 3 and 4); where two said it was correct, this is equal to 2; and so on. We then calculated the average automatic metrics score for each category.[6]

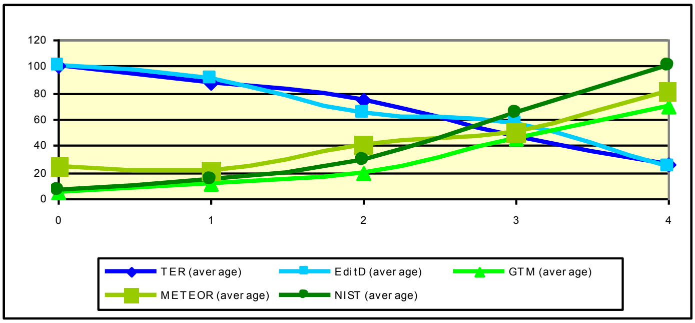

Figure 3: Automatic scores for the -ing constituents classified according to the number of evaluators who considered them correct for French

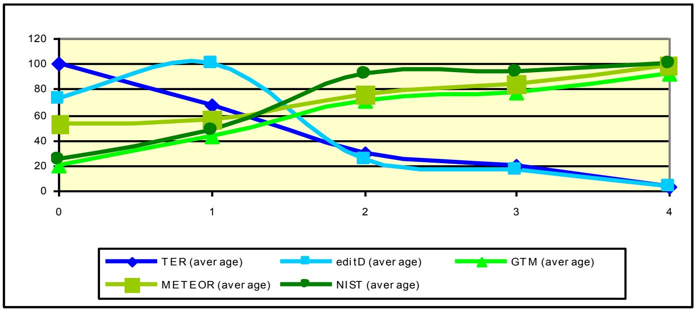

Figure 4: Automatic scores for the -ing constituents classified according to the number of evaluators who considered them correct for Spanish

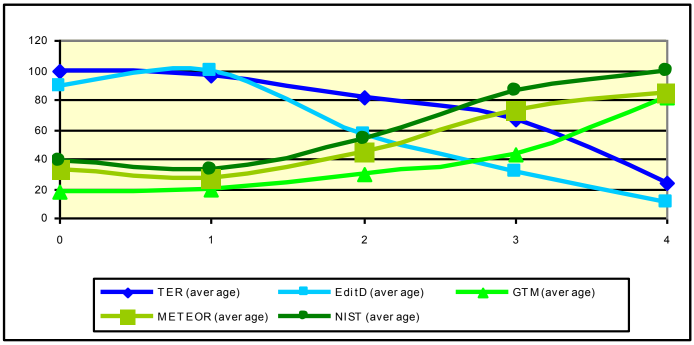

Figure 5: Automatic scores for the -ing constituents classified according to the number of evaluators who considered them correct for German

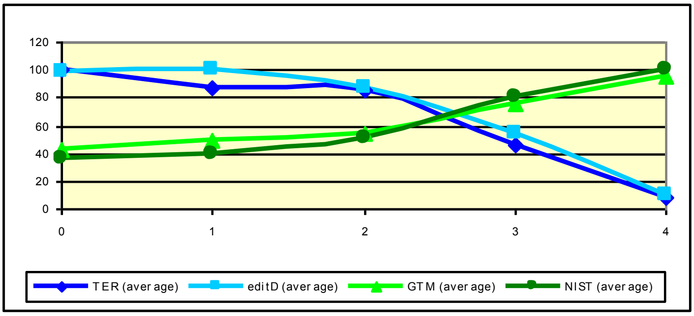

Figure 6: Automatic scores for the -ing constituents classfied according to the number of evaluators who considered them correct for Japanese

From the results we can clearly see that the tendency for each category correlates with the response of the human evaluators. According to the automatic metrics, while the examples classified as incorrect by all the evaluators (0) need higher numbers of changes to convert to the reference translation, the examples classified as correct by all four evaluators (4) need hardly any changes. We calculated the Pearson r-correlation between the mean human scores and the mean automatic metric scores to see if we could verify the trend shown above (see Table 2).

Table 2: Pearson r-correlation for the human evaluation score means (H) and automatic metrics score means (NIST, TER, GTM, METEOR, EditD) where all results are significant at the 0.01 level.

Pearson r-correlation	French	Spanish	German	Japanese
H / NIST	0.97	0.93	0.97	0.96
H / TER	-0.99	-0.97	-0.93	-0.94
H / GTM	0.96	0.97	0.92	0.96
H / METEOR	0.93	0.98	0.92	N/A
H / EditD	-0.98	-0.86	-0.94	-0.92

It is agreed that correlation is weak if the coefficient is less than 0.5 and strong if the coefficient is greater than 0.8. Our results are in the region of 0.86 to 0.99. Therefore, we observe that, even if the difference between them is statistically significant at 0.01, the agreement between human scores and automatic metrics is strong regardless of the automatic metric and the target language used.[7]

4. Conclusions

-ing forms are functionally very flexible, yet we conclude that they do not deserve their reputation for being classified as highly problematic for RBMT. The MT system has proven to be able to translate -ing forms grammatically and accurately 72-73% of the time for German, Japanese and Spanish. French performed worse, achieving correct translation for half the samples. However, closer examination allowed us to pinpoint the reason for the 20-point difference. Two highly frequent constituents were found to be systematically incorrect for French but not for the other target languages. Had these two constructs obtained similar results to those of the other target languages, the overall results would have been similar for all four.

A comparison between the human evaluations and NIST, GTM, TER and Edit-distance, showed good correlations. This may be an interesting avenue for further investigation.

A fined-grained analysis of the translation of the -ing constituents helped us detect the most problematic categories. The issues we found varied in type and we have considered solutions that could be implemented at different stages in the machine translation process. Firstly, we considered the use of controlled language at the content authoring stage. CL is most beneficial for the issues shared across all languages. Such was the case for titles, reduced relative clauses and prepositional phrases, and we have fine-tuned existing rules in the CL rule set for some of these categories.

Not all our -ing categories were problematic across all languages and, therefore, CL rules are not an appropriate solution. Hence, alternative ways should be explored and additional pre-processing stages are suggested. For example, the RBMT system we tested detects participial adjectives correctly but occasionally translates gerund-participles as modifiers. Our current research examines whether it would be possible to tag gerund-participles in such a way that the MT system could understand the tags and disambiguate appropriately.

Another obvious avenue of exploration for language-specific issues is to simply post-edit the MT output. We are investigating how to semi-automate the post-editing process so that recurring problems can be quickly fixed using find-and-replace rules crafted for each target language, based on our knowledge from this research. Another possibility we are considering is to "post-edit the source text" (Somers, 1997). This would involve editing the source text to eliminate known problems for specific target languages. This is different from implementing CL rules, which are normally implemented by technical writers at the time of writing, and the resulting documentation is published in English and is machine translated. A major advantage of post-editing the source text is that the modified source, because it would not be published, could contain any sort of "ungrammatical" changes in the source text, which would, hopefully, produce grammatical MT output.

Our future work will involve implementing and testing the effectiveness of these proposed solutions for the different categories of -ing forms across all four target languages.

Bibliography

Balkan, L., Netterz, K., Arnold, D. & Meijer, S. (1994). Test suites for natural language processing. In Proceedings of Language Engineering Convention (pp. 17-22);, Paris, July 6-7, 1994.

Banerjee, S. & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005 (pp. 65-72); Ann Arbor, Michigan, June 29, 2005.

Bernth, A. & Gdaniec, C. (2001). Mtranslatability. *Machine Translation, 16*, 175-218.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bowker, L. & Pearson, J. (2002). *Working with specialized language. A practical guide to using corpora*. London/New York, NY: Routledge.

Callison-Burch, C., Osborne, M. & Koehn, P. (2006). *Re-evaluating the role of BLEU in machine translation research*. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (pp. 249-256); Trento, Italy, April 3-7, 2006.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. (2007). *(Meta-)Evaluation of machine translation*. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 136-158); Prague, Czech Republic, June 23, 2007.

Cieri, C., Strassel, S., Glenn, M. L., & Friedman, L. (2007, September). *Linguistic resources in support of various evaluation metrics*. Presentation at MT Summit XI Workshop: Automatic Procedures in MT Evaluation, Copenhagen, Denmark.

Coch, J. (1996). Evaluating and comparing three text-production strategies. In Proceedings of the 16th International Conference on Computational Linguistics (COLING 96) (pp. 249-254); Copenhagen, Denmark, August 5-9, 1996.

Dervišević, D. & Steensland, H. (2005). *Controlled languages in software user documentation*. M.A. thesis, Department of Computer and Information Science, Linköpings University. Linköping, Sweden.

Elliott, D., Hartley, A. & Atwell, E. (2004). A fluency error categorization scheme to guide automated machine translation evaluation. In R.E. Frederking & K.B. Taylor (Eds.), *AMTA 2004* (pp. 64-73). Berlin/Heidelberg: Springer-Verlag.

Estrella, P., Popescu-Belis, A. & King, M. (2007). A new method for the study of correlations between MT evaluation metrics. In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (pp. 55-64); Skövde, Sweden, September 7-9, 2007.

Huddleston, R. & Pullum, G. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Huijsen, W.O. (1998). Controlled language – An introduction. In Proceedings of the Second Controlled Language Applications Workshop (CLAW 1998) (pp. 1-15); Pittsburgh, Pennsylvania, May 21-22, 1998.

Izquierdo, M. (2006). *Análisis contrastive y traducción al español de la forma -ing verbal inglesa*. M.A. thesis, Department of Modern Philology, University of León, León, Spain.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London/New York, NY: Longman.

LDC (2003). *Linguistic data annotation specification: Assessment of fluency and adequacy in translation*. Project LDC2003T17.

McEnnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies*. London/New York, NY: Routledge.

Microsoft Corporation (1998). *Microsoft manual of style for technical publications* (2nd ed.). Redmond, WA: Microsoft Press.

Mutton, A., Dras, M., Wan, S. & Dale, R. (2007). GLEU: Automatic evaluation of sentence-level fluency. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 344-351); Prague, Czech Republic, June 23-30, 2007.

National Institute of Standards and Technology (2002). *Automatic evaluation of machine translation quality using N-gram co-occurrence statistics*. Retreived January 26, 2009 from http://www.nist.gov/speech/tests/ mt/2008/doc/ngram-study.pdf

NLTK (Natural Language Toolkit). Retrieved January 26, 2009 from http://www.nltk.org/

O'Brien, S. (2003). Controlling controlled English: An analysis of several controlled language rule sets. In Proceedings of the 4th Controlled Language Applications Workshop (CLAW 2003) (pp. 105-114); Dublin, Irelend, May 15-17 ,2003.

Olohan, M. (2004). *Introducing corpora in translation studies*. Abingdon/New York, NY: Routledge.

Pan, S. & Shaw, J. (2004). Segue: A hybrid case-based surface natural language generator. In Proceedings of the International Conference on Natural Language Generation (INLG04) (pp. 130-140); Brighton, UK, July 14-16, 2004.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.J. (2002). *BLEU: A method for automatic evaluation of machine translation*. In Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL-2002) (pp 311-318), Philadelphia, PA, July 7-12, 2002.

Pierce, J.R., et al. (1966). *Language and machines: Computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee*. .Washington, D.C.: National Academy of Sciences – National Research Council.

Przybocki, M., Sanders, G. & Le, A. (2006). *Edit distance: A metric for machine translation*. In Proceedings of LREC 2006: Fifth International Conference on Language Resources and Evaluation (pp. 2038-2043); Genoa, Italy, May 24-26, 2006.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Snover, M. & Dorr, B. (2006). A study of translation edit rate with targeted human annotation. In *Visions for the Future of Machine Translation* (pp. 8-12): Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006) (pp. 223-231);,Cambridge, UK, August 8-12 ,2006.

Somers, H. (1997). A practical approach to using MT software – "Post-editing" the source text. *The Translator, 3(2),* 193-212.

Turian, J., Shen, L. & Melamed, I.D. (2003). Evaluation of machine translation and its evaluation. In Proceedings of the Machine Translation Summit IX (pp. 386-393); New Orleans, LA September 23-27, 2003.

Vilar, D., Leusch, G., Ney, H. & Banchs, R.E. (2007). Human Evaluation of Machine Translation Through Binary System Comparisons. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 96-103); Prague, Czech Republic, June 23, 2007.

Appendix: Numbers of Extracted –ing Forms and Their Classifications

titles **2,603**	**beginning with -ing**	**no quotations**	**1,255**				
		within quotations	**BOS**	**530**			
			embedded in sentence	**620**			
	beginning with about + -ing	**no quotations**	**100**				
		within quotations	**BOS**	**60**			
			embedded in sentence	**38**			
referentials **594**	**nouns**	**252**					
	comparatives	**46**					
	objects of prepositional verbs	**116**					
	objects of phrasal verbs	**13**					
	catenatives	**167**					
characterizers **2,488**	**pre-modifiers**	**participial adjectives**	**1,873**				
	post-modifiers	**reduced relatives**	**377**				
		nominal adjuncts	**226**				
		adjectival adjuncts	**12**				
progressives **661**	**present**	**active**	**501**				
		passive	**117**				
		questions	**2**				
	past	**active**	**9**				
		passive	**3**				
	future	**2**					
	modal	**22**					
	infinitive	**5**					
adverbials **1,970**	**manner** **763**	**by - 516**	**without - 88**	**ϕ - 159**			
	contrast **11**	**instead of - 11**					
	time **728**	**before-179**	**after – 139**	**when – 313**	**while – 65**	**between – 4**	**upon - 1**

		on – 8	**in – 2**	**during – 2**	**prior – 1**	**along with – 2**	**in the middle of - 2**
		from – 2	**through – 5**	**ϕ - 3**			
	concession **1**	**besides - 1**					
	place **1**	**where - 1**					
	purpose **444**	**for – 443**	**in - 1**				
	condition **20**	**if - 20**					
	cause **2**	**because 2**					

[1] For an overview of the different functional categories of –ing words, see Izquierdo 2006.

[2] We draw on our experience here with the editors and technical writers in Symantec.

[3] CL checkers are software programs that allow checking for specific syntactic or lexical occurences which are disallowed, according to the specific CL rule set.

[4] The unclassified 20.17% can be accounted for by -ing forms which do not fall into any of the syntactic patterns proposed by Izquierdo due to long-distance occurrence, that is, the -ing form is not directly followed/preceded by the syntactic anchors used to retrieve them using the CL checker. We also expect that there is a large number of gerundial nouns acting as subjects or objects in the remaining group. No specific search was carried out for this group for two reasons. Firstly, because unless they are preceded by a determiner they are difficult to find automatically. Secondly, because they behave as genuine nouns and should be included in the RBMT system's dictionary, thus not creating problems for translation.

[5] Note that we use "-ing form" to refer to words that end in -ing whereas we use "-ing constituents" when we refer to the -ing form in context.

[6] Given that the automatic scores express the results on different scales, we normalized them in order to compare the trends. Note that there is no upper bound for the TER and Edit-distance scores. For those metrics the highest score, i.e. the worst-performing score, was taken as the upper bound. Note also that these two metrics score best when the result is zero, as opposed to NIST, GTM and METEOR, for which a zero score is the worst possible result. This is the reason why the metrics appear to go in opposite directions on the graphs.

[7] Note that the negative sign refers to the direction of the correlation.