A Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation

Violeta Seretan, Pierrette Bouillon, Johanna Gerlach
Department of Translation Technology
Faculty of Translation and Interpreting
University of Geneva
{Violeta.Seretan, Pierrette.Bouillon, Johanna.Gerlach}@unige.ch

Abstract

The user-generated content represents an increasing share of the information available today. To make this type of content instantly accessible in another language, the ACCEPT project focuses on developing pre-editing technologies for correcting the source text in order to increase its translatability. Linguistically-informed pre-editing rules have been developed for English and French for the two domains considered by the project, namely, the technical domain and the healthcare domain. In this paper, we present the evaluation experiments carried out to assess the impact of the proposed pre-editing rules on translation quality. Results from a large-scale evaluation campaign show that pre-editing helps indeed attain a better translation quality for a high proportion of the data, the difference with the number of cases where the adverse effect is observed being statistically significant. The ACCEPT pre-editing technology is freely available online and can be used in any Web-based environment to enhance the translatability of user-generated content so that it reaches a broader audience.

Keywords: Machine Translation, Authoring Tools, Evaluation Methodologies

1. Introduction

With the advent of the digital era and the Web 2.0 paradigm, an enormous amount of content is nowadays produced by users, and shared in virtual communities grouped around specific areas of interest (e.g., in forums, blogs, social network services). Social media makes it possible to anyone to access and publish information; yet, due to the diversification of languages used on the Internet, the user-generated content – henceforth, UGC – is only accessible to those understanding the language in which messages are posted.

There is a tremendous need to automatically translate user-generated content, in order to help communities share knowledge more effectively across the language barrier. In the ACCEPT project[1], we are pursuing the goal of improving statistical machine translation (SMT) for community content in two main scenarios:

1. technical content produced in the Norton Community forum[2];
2. healthcare content produced by non-governmental organizations such as Doctors Without Borders and translated by the Translators without Borders community of volunteers[3].

There is a lot to be gained from the automatic translation of this type of community content in terms of cross-lingual access to knowledge and faster (human) translation delivery time. But the least to say is that this content is very challenging for machine translation. As shown in Nagarajan and Gamon (2011), there are several characteristics of the community content that pose new processing challenges with respect to traditional content: informal style, slang, abbreviations, specific terminology, irregular grammar and spelling.

In order to improve the translatability of community content in the two given scenarios, the ACCEPT consortium proposes a new integrated approach, which consists of the following main axes of research and development:

1. Development of content pre-editing technology, targeting the most important types of corrections which must be applied to the source content in order to attain a higher translation quality;
2. Development of post-editing technology, allowing to leverage the work of monolingual and bilingual volunteer subject matter experts in order to learn output correction rules and integrate them into the SMT engine;
3. Development of strategies for improving SMT proper, to make it more robust and more efficient for the UGC domain (for which training resources are sparse and heterogeneous). The project focuses on domain adaptation, linguistic backoff and text analytics as means to customise Moses translation systems to our application domains and optimise them for the language pairs considered in the project (English to French/German; French to English).

Past halfway into its research program, the project has accomplished significant progress in all areas mentioned above. The ACCEPT Portal[4], which has recently been released to the broad public, gives access to the pre-editing and post-editing environments created for the purposes of the project, as well as to the software APIs and documentation (see Seretan et al. (2014) for an overview). The pre-editing technology has been installed on the Norton

[1] www.accept-project.eu
[2] community.norton.com/norton
[3] translatorswithoutborders.org
[4] accept-portal.eu

Figure 1: The ACCEPT pre-editing technology embedded in the Norton Community forum (screen capture). The content checking window is displayed when the user clicks on the "ABC" button.

Community forum, so that users can check their posts before submitting them (see Figure 1).

Work is in progress on the second and third of the main axes listed above, whereas the development of the pre-editing technology – i.e., the first axis – is already achieved. In this paper, we outline this technology and the evaluation experiments performed in order to assess its impact on machine translation quality.

The paper is organised as follows. In Section 2 we briefly introduce the pre-editing technology created in the ACCEPT project for improving the translatability of UGC. In Section 3 we describe the experimental setup and the methodology used for evaluating the impact of pre-editing on translation quality. We present the evaluation results for the technical and healthcare domains in Section 4 and Section 5, respectively. Section 6 presents related work, and Section 7 contains final remarks.

2. The ACCEPT Pre-editing Technology

The pre-editing technology developed in the framework of the ACCEPT project in order to enhance the translation quality for the community content scenario is founded on the Acrolinx lingware, a suite of tools and resources for supporting authoring through spelling, grammar, terminology and style checking based on shallow language processing (Bredenkamp et al., 2000).

Correction rules are manually defined using the Acrolinx formalism, by using regular expressions over partial parsing output for specifying error triggers and text reformulations. The rules can be either automatically applied by the system if the reformulation is deterministic, or will require the user intervention if the reformulation is non-deterministic (as in the case of spelling suggestions showing multiple reformulation candidates) or no automatic reformulation can be proposed.

The definition of Acrolinx rules for the ACCEPT scenarios involved the corpus-driven manual identification of potential correction rules and their individual evaluation, in order to come up with a selection of the most efficient rules for each of the source languages and domains considered (English/French; technical/healthcare domain).

The English rules have been created by adapting the existing Acrolinx rule set for the general domain to our target domains. The phenomena targeted are casing and spelling issues, punctuation usage, missing spaces, duplicate words, homophone confusion (e.g., *their*/*there*, *to*/*too*) grammatical issues (such as lack of agreement, incorrect verb form, use of incorrect prepositions), as well as style issues (e.g., long sentences). Additional reformulation rules target specifically the SMT engine, without necessarily improving the source text (e.g., *have to*→*must* leads to a better translation to German).

For French, the pre-editing rules have been created from scratch. The main phenomena targeted are homophones, word confusion, wrong verb forms, elision and punctuation (in particular, hyphenation, e.g., *avez vous* → *Avez-vous* 'Have you'); grammar and style issues (in particular, informal language); and specific reformulation rules for the SMT engine (e.g., changing the second person from informal to formal as in *tu as* → *vous avez* 'you have').

The final selection of rules is presented in detail in the ACCEPT Deliverable 2.2 (2013). The rule evaluation on a case-by-case basis is described in Roturier et al. (2012) and Gerlach et al. (2013).

Table 1 shows sample pre-editing rules, illustrating the impact of their application on machine translation output.[5]

"its/it is" confusion	
Original version	
Source:	How much longer until *its* fixed?
MT output:	Wie viel länger, bis seine behoben?
Pre-edited version	
Source:	How much longer until *it's* fixed?
MT output:	Wie viel länger, bis es behoben?
ça vs sa	
Original version	
Source:	oups j'ai oublié, j'ai *sa* aussi.
MT output:	Oops I forgot, I have its also.
Pre-edited version	
Source:	oups j'ai oublié, j'ai *ça* aussi.
MT output:	I have forgotten, I have this too.

Table 1: Sample pre-editing rules for English and for French and their effect on machine translation output

3. Evaluation Experiments

Once the development of the pre-editing rules has been achieved, we proceeded to the large-scale systematic evaluation of the combined effect of the rules on the quality of machine translation output. Evaluation experiments were conducted for both scenarios considered in the project, i.e., on data from the technical domain and from the healthcare domain. As the evaluation methodology was the same in both cases, for the sake of clarity in this section we will focus on a single experiment, with technical data. We will report on the second experiment with healthcare data in Section 5.

3.1. Data

For the technical domain scenario, the evaluation experiment was performed on data provided by one of the project partners, Symantec. The data consists of posts from the English and French Norton Community forum. The evaluation testset was built by randomly selecting 1,000 forum posts, half in English and half in French, from a withheld portion of data, which was not used for development purposes. We converted the posts from the original HTML format into text format, more suitable for machine translation. Table 2 shows a sample post in HTML and in MT-ready format.

Re: restoring a bootable operating drive from an independent recovery point<P>Check these instructions by Brian</P><P>There is a quirk that it fails the first time.</P><P><A href="http://community.norton.com/t5/Other-Norton-Products/Network-restore-with-Ghost-15/m-p/579844/highlight/true#M41167" target="_blan">http://community.norton.com/t5/Other-Norton-Products/Network-restore-with-Ghost-15/m-p/579844/highlight/true#M41167</A></P>
Re: restoring a bootable operating drive from an independent recovery point Check these instructions by Brian There is a quirk that it fails the first time.<URL>

Table 2: Sample forum post (original and cleaned version).

The evaluation unit is the whole post, as opposed to the sentence. The reason behind this choice is that posts are relatively short[6], easier to evaluate than isolated sentences, and there is no need for sentence segmentation (a challenging task in itself). More importantly, we are interested in studying the global impact of pre-editing rule application and not only the local impact, since the positive impact on one sentence may be counterbalanced by a negative impact on another sentence.

3.2. Pre-editing

The cleaned posts were pre-edited both automatically and manually, and then translated using the baseline SMT systems built for the project (ACCEPT Deliverable 4.1, 2012). The manual pre-editing was performed in each source language by a native speaker, Master student in translation, paid for the task. Guidelines were distributed to pre-editors and a post-task questionnaire survey of about 20 questions was carried out to elicit editor's feedback about the task. The answers showed that the editors perceived the task as quick and easy; they did not encounter technical issues; the instructions received were clear and they knew how to edit.

[5] The MT systems used are the ACCEPT baseline systems, referenced later in the paper (Section 3).

[6] On average, 93.7 words for English and 78.6 for French.

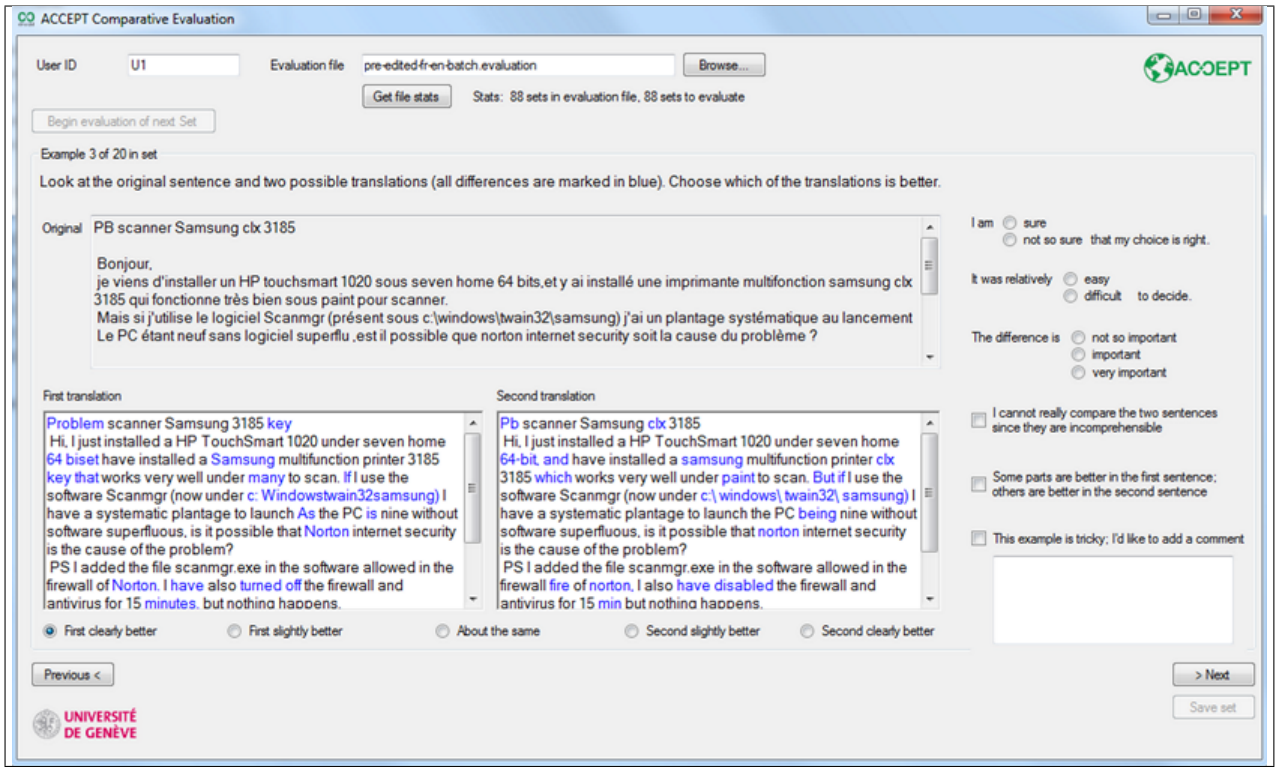

Figure 2: Interface of the evaluation tool used in the experiments (screen capture).

They understood the system suggestions and the rule description. They perceived their edits as useful, had a positive overall experience and would perform similar tasks in the future. However, their detailed comments revealed that they found the task was repetitive and they wished they could correct more mistakes than those highlighted by the system ("It's not easy not to correct everything").

3.3. Annotators

The translations corresponding to the raw and pre-edited post versions have been comparatively evaluated by human judges. Groups of three judges annotated the data for each language pair. A total of 9 judges participated to the experiment. The judges were all Master students in translation, native speakers of the target language and fluent in the source language. They have been paid for the task. An in-house tool was used to carry out the evaluation and to record the time spent *(1)* judging the translation pairs and *(2)* providing feedback on the evaluation of each pair. The interface of the evaluation tool is shown in Figure 2. Figure 3 displays the total time that annotators took to perform the evaluation task (outliers where removed, i.e., times higher than 1,000 seconds were removed).

3.4. Evaluation categories

The evaluation used a 5-point Likert scale with the following categories: *first clearly better*, *first slightly better*, *about the same*, *second slightly better*, and *second clearly better*, where *first* and *second* refer to the two compared translations, one corresponding to the raw source version and one

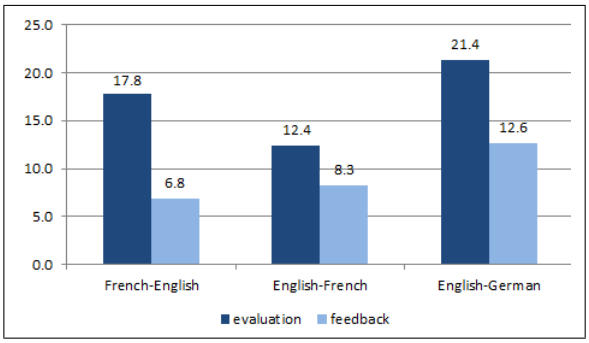

Figure 3: Total time needed by annotators for the task (in hours).

to the pre-edited version of a post.[7] The two translations were shown to the judges in randomised order, to avoid bias. The source post was also displayed for reference. To help evaluators, the differences in the two translations were automatically highlighted using a different color.
Additional variables measured in the experiment were: *confidence* (how sure the evaluators are that their choice is right), *difficulty* (how difficult it was for them to decide), *importance* (how important the difference between the two translations is), *low quality* (the two translations cannot really be compared because they are incomprehensible), *conflicts* (some parts are better in the first translation, others are better in the second), and *flag* (used for marking

[7]The evaluation in terms of relative ranking of translations is seen in the literature as more reliable that traditional evaluation metrics, like fluency and adequacy (Koehn, 2010).

Impact of Pre-editing	French-English	English-French	English-German
better	68.9%	51.5%	56.4%
same	16.3%	21.7%	14.4%
worse	14.8%	26.9%	29.2%
N	472	443	459

Table 3: Impact of pre-editing on machine translation quality (percentages of the total number of cases where a majority judgement exists, *N*).

tricky examples and to add comments). Guidelines were distributed to evaluators, which explained the task and the context of the work. Evaluators were instructed to base their decision on the usefulness criterion, more precisely, to select the translation which they would prefer to post-edit.

3.5. Post-task questionnaire

An anonymous post-task survey containing about 20 questions has been conducted two weeks after the evaluation campaign, in order to elicit feedback from evaluators about the task they performed. All 9 evaluators took part to the survey. Their answers generally show a positive attitude towards the task performed (they enjoyed the task, the instructions were clear, the tool was easy to use, they did not encounter technical issues, they felt comfortable with the evaluation scale as well as with the feedback questions, and they perceived their work as useful).

However, they disagreed that the comparative evaluation task was quick and easy and the amount of data to evaluate was convenient for them. Detailed comments highlighted, in particular, that the poor quality of the text was a main cause for frustration ("What made the task very tiring for me was the fact that the source sentences were often already written really badly"). Still, most evaluators (8 out of 9) stated that they are willing to perform similar tasks in the future.

4. Evaluation Results

The impact of pre-editing strategies on SMT output quality is reported by taking into account the mode[8] of the three judgements collected for each post. To even out subjective differences between judgements, we make no distinction between *slightly* and *clearly* in the evaluation categories. Therefore, we report the impact of pre-editing on a 3-point scale, as either *better*, *same* or *worse*, according to the majority vote obtained for the pre-edited source version after grouping the *slightly* and *clearly* categories. We discard the cases where there is complete disagreement between the three judgements of a post. Table 3 displays the results.

The inter-annotator agreement is reported in terms of Fleiss κ (Fleiss, 1981) both for the 5-point and the 3-point evaluation scale (Table 4). Fleiss' κ is the equivalent of Cohen's κ agreement statistics (Cohen, 1960) for more than two raters. As Cohen's κ, it basically subtracts from the observed agreement the agreement which is due to chance. Values range in the interval [-1, 1], with positive values indicating agreement and negative values indicating disagreement. Values close to 0 correspond to agreement due to chance, whereas a value of 1 represents perfect agreement and a value of -1 perfect disagreement. The following scale is used to interpret the absolute values of κ: *slight* (0–0.2), *fair* (0.2–0.4), *moderate* (0.4–0.6), *substantial* (0.6–0.8), and *almost perfect* (0.8–1) (Landis and Koch, 1977).

The values obtained in our experiment for Fleiss κ are between 0.19 and 0.43, corresponding to *slight* to *moderate* agreement. These relatively low values are indicative of the difficulty and subjectivity of the task. From the judges' comments, it became apparent that both the low quality of the text and the length of the posts made the task very tedious (see also the post-task survey results in Section 3).

	3-point scale	5-point scale
French-English	0.43	0.30
English-French	0.20	0.19
English-German	0.38	0.20

Table 4: Inter-annotator agreement statistics (Fleiss κ).

A McNemar test (McNemar, 1947) was conducted to compare the number of cases in which pre-editing had a positive *vs.* a negative impact on translation quality. For all language pairs, the difference is statistically significant ($p < 0.001$). The number of posts that benefit from pre-editing significantly outweighs that of posts whose translation is degraded through pre-editing. While the systematic error analysis is currently in progress, the preliminary results indicate that automatic spelling correction of proper names is one of the main issues leading to worse translations. As the recognition of usernames is a real challenge in UGC (Bontcheva et al., 2013), in future experiments we will perform evaluations of pre-editing excluding spelling corrections.

Furthermore, Spearman's rho correlation coefficients are computed between all the dependent variables of the experiment. A significant correlation was observed between most of the variables measured ($p < 0.01$). A strong negative correlation was found, for instance, between *difficulty* and *confidence* (rho = -0.694). Also, a weak positive correlation was observed between *difficulty* and *conflicts* (rho = 0.237) and between *conflicts* and *evaluation_time* (rho = -0.242). These findings are in line with the comments from evaluators, who emphasised the difficulty of evaluating long, poor quality texts with conflicting changes. In future work, we plan to conduct an additional evaluation at the sentence level (as opposed to the post level), in order to facilitate evaluators' work and attain a higher inter-annotator agreement.

[8]The mode is the most frequent value in a dataset.

5. Results for the Healthcare Domain

The same evaluation methodology has been used to study the impact of pre-editing on user-generated content from a different domain, namely, the healthcare domain, which is also targeted in our project in addition to the technical forum domain.

Indeed, the second application scenario of the project is the NGO scenario. The ACCEPT project support non-governmental organisations such as Doctors without Borders which need to deliver critical information in the right language at the right time.

Our project partner Lexcelera, the founder of the Translators without Borders Organisation, provided data which allowed for the development of machine translation systems and pre-editing technology adapted to the healthcare domain.

The impact of pre-editing on translation quality has been evaluated on a testset of 200 sentences randomly extracted from medical field reports in French. The average sentence size is 29.1 words.

Two annotators compared the English translations obtained for the original and the pre-edited sentence versions. The inter-annotator agreement was slightly higher for this domain (Cohen's κ = 0.54 for the 3-point scale; κ = 0.39 for the 5-point scale).

As can be seen from Table 5, the impact of pre-editing is comparable with the one observed for the technical domain. According to the McNemar test, the difference between the number of *better* and *worse* cases is statistically significant ($p < 0.05$).

Impact of Pre-editing	French-English
better	50.0%
same	24.3%
worse	25.7%
N	70

Table 5: Results for the healthcare domain (N = number of cases on which the two judges agreed).

6. Related Work

Pre-editing texts to improve human readability or MT performance is an old topic (Ruffino, 1981). Pre-editing can take different forms: spelling and grammar checking; lexical normalisation (Han and Baldwin, 2011; Banerjee et al., 2012); controlled natural language (O'Brien, 2003; Kuhn, 2013); or reordering (Wang et al., 2007; Genzel, 2010).

While controlled natural language has mostly been associated with rule based machine translation (Pym, 1988; Bernth and Gdaniec, 2001; O'Brien and Roturier, 2007; Temnikova, 2011) spell-checking, normalisation and reordering are frequently used as pre-processing steps for SMT.

There are few pre-editing scenarios that actually combine these different approaches. In ACCEPT, we have chosen an approach in which we consider of all pre-editing forms mentioned above to deal with the particularities of community content.

7. Conclusion

The experiments described in this paper show that computationally-light pre-editing strategies, such as the ones designed for correcting user-generated content in the framework of the ACCEPT project, may lead to a significant increase in the quality of statistical machine translation output.

The ACCEPT pre-editing technology can be downloaded from the ACCEPT Portal (`www.accept-portal.eu`) and installed on social media platforms or any Web-based environment. It may be used to enhance the translatability of community content, helping post authors reach a broader audience and allowing more people to access knowledge in a language which is not their own.

The pre-editing technology is limited to the languages for which shallow parsers are available. Many of the rules which are defined for the specific languages of the project are portable to new languages, as they often encode general principles for improving text readability. Further efforts are, however, needed to customise the rules for specific target languages and specific domains. The interaction between pre-editing, domain adaptation for SMT and post-editing are in the focus of ongoing work in the ACCEPT project and will help developed technologies for an increased accessibility to user-generated content via automatic translation.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 288769.

8. References

Banerjee, P., Naskar, S. K., Roturier, J., Way, A., and van Genabith, J. (2012). Domain adaptation in SMT of user-generated forum content guided by OOV word reduction: Normalization and/or supplementary data? In *Proceedings of EAMT*.

Bernth, A. and Gdaniec, C. (2001). MTranslatability. *Machine Translation*, 16(3):175–218.

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M., Maynard, D., and Aswani, N. (2013). Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 83–90, Hissar, Bulgaria, September.

Bredenkamp, A., Crysmann, B., and Petrea, M. (2000). Looking for errors: A declarative formalism for resource-adaptive language checking. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

(2013). ACCEPT deliverable D 2.2 Definition of pre-editing rules for English and French (final version). `http://www.accept.unige.ch/Products/D2_2_Definition_of_Pre-Editing_`

Rules_for_English_and_French_with_appendixes.pdf.

(2012). ACCEPT deliverable D 4.1 Baseline MT systems. http://www.accept.unige.ch/Products/D_4_1_Baseline_MT_systems.pdf.

Fleiss, J. L. (1981). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Genzel, D. (2010). Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 376–384, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gerlach, J., Porro, V., Bouillon, P., and Lehmann, S. (2013). La préédition avec des règles peu coûteuses, utile pour la TA statistique des forums ? In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 539–546, Les Sables d'Olonne, France.

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a #Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Kuhn, T. (2013). A survey and classification of controlled natural languages. *Computational Linguistics*.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

Nagarajan, M. and Gamon, M., editors. (2011). *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. Association for Computational Linguistics, Portland, Oregon, June.

O'Brien, S. and Roturier, J. (2007). How portable are controlled languages rules? a comparison of two empirical MT studies. In *Proceedings of MT Summit XI*, pages 105–114.

O'Brien, S. (2003). Controlling controlled English: An analysis of several controlled language rule sets. In *Proceedings of EAMT-CLAW-03*, pages 105–114.

Pym, P. J. (1988). Pre-editing and the use of simplified writing for MT: an engineer's experience of operating an MT system. In *Translating and the Computer 10:The translation environment 10 years on*, pages 80–96.

Roturier, J., Mitchell, L., Grabowski, R., and Siegel, M. (2012). Using automatic machine translation metrics to analyze the impact of source reformulations. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, USA, October.

Ruffino, J. R. (1981). Coping with machine translation. In *Practical experience of machine translation. Proceedings of a conference*, pages 57–60, November.

Seretan, V., Roturier, J., Silva, D., and Bouillon, P. (2014). The ACCEPT Portal: An online framework for the pre-editing and post-editing of user-generated content. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation (HaCat 2014)*, Gothenburg, Sweden, April. Association for Computational Linguistics.

Temnikova, I. (2011). Establishing implementation priorities in aiding writers of controlled crisis management texts. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 654–659, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Wang, C., Collins, M., and Koehn, P. (2007). Chinese syntactic reordering for statistical machine translation. In *In Proceedings of EMNLP*, pages 737–745.