BUILDING AN ENVIRONMENT FOR UNSUPERVISED AUTOMATIC EMAIL TRANSLATION

Salvador Climent **Joaquim Moré** **Antoni Oliver**
Interdisciplinary Internet Institute (IN3) / Universitat Oberta de Catalunya (UOC)
Av. Tibidabo 39
08035 Barcelona
scliment@uoc.edu jmore@uoc.edu aoliverg@uoc.edu

Abstract

By this paper, we present the INTERLINGUA Project[1]: its design and current work. The goal of the project is achieving fully-automatic (no pre-edition, no post-edition) translation of emails in the virtual campus of the Open University of Catalonia (UOC). The problem of unsupervised machine translation of emails is discussed. Then we describe the strategy designed to build the system, including a multiple-level evaluation process and the building of several automatic pre-edition, post-edition and unknown-word extraction modules. Last, the work carried on so far on building such decision-taking modules is presented.

1 Introduction and Rationale

The UOC (www.uoc.edu) is a virtual University currently offering 17 official university degrees, 1 Ph.D. program, and several dozens of other courses. Communication between students, professors, and supervisors is completely performed via email or email-like means –e.g. kinds of newsgroups– within a virtual campus and a system of virtual classrooms. Courses are taught in Catalan and/or Spanish.

Many of the students are Catalan speakers, which means that, due to the linguistic situation in Catalonia, they are fluent in both Catalan and Spanish. Another part of them are Spanish speakers living in Catalonia, which most of the times can read Catalan but can't write it properly. Last, due to the recent expansion of the UOC to the rest of Spain and South-America, there is a new sector of students who are strict monolingual speakers of Spanish.

Such a situation might lead to gradual substitution of Catalan by Spanish in the classrooms. A statistical study on the language used to write and to reply messages at the UOC, covering 1 year of 4 newsgroups, shows that, although 68.9% of the users can be considered spontaneous users of Catalan, 42.9% of these Catalan-speakers code-switch to Spanish when replying to messages in that language. The INTERLINGUA Project aims to overcome such effect by allowing effective cross-linguistic communication using machine translation (MT).

We foresee that the goal (MT for communication, not for gisting or budget cutting in document translation) is reachable since Catalan and Spanish are two Romance languages structurally quite similar at all levels. This makes that MT systems perform at high levels of quality between the pair, as preliminary tests of translation of pre-edited texts have shown. It seems clear that MT between Catalan and Spanish (and vice versa), when using a knowledge-rich system, just needs of good lexicons and a tuning effort on solving some reluctant ambiguities to produce fully comprehensible and faithful texts.

[1] Funded by the Interdisciplinary Internet Institute; IN3- IR 226

Notwithstanding, it is clear that the special task of machine-translating emails in the environment we have described above shows a number of additional important problems, which can be classified in three main categories:

– Impossibility of human intervention in the MT process
– Specificity of the email register
– Problems posed by the special case of bilingualism and languages in contact

Factors 1 and 2 involve an almost absolute impossibility of any kind of edition or control on the text. On the one side, emails should flow instantly trough the net admitting no delay for human formatting, pre-edition or post-edition. Besides, any try to charge users with some kind of language self-control is led to fail –they just want to write their emails without needing to undergo any kind of a boring process. On the other side, communication by email is strongly characterized by their intensive use of non-standard language –plus some visual information resources, and a wide range of unforeseeable errors (see section 2.2, and also [Fais01] and [Yates93]).

As it is well known, standardization and correction of the input text is a key factor for success in MT – e.g. well-established vocabulary, terminology and abbreviations, well-formed sentences, cohesion and absence of errors or *bizarre* new forms of expressivity. Therefore, in our case, we need of a highly structured effort to customize the system by designing, building and integrating in it a number of decision-taking modules that automatically overcome deviations of the standards in the input text –a sort of automatic language control.

The third factor, languages in contact, adds extra challenges since messages might mix Catalan and Spanish when quoting or linking to previous articles, and there is a range of (mainly lexical but also structural) language interference even in monolingual emails. Besides, for historical and educative reasons, users show different levels of competence in either of the languages –competence in writing Catalan is usually quite lower. This makes us to expect added difficulties on generalizing solutions since either translation direction should be handled differently.

Furthermore, obviously, the environment should manage the usual need for terminological tailoring of the system according to the domain.

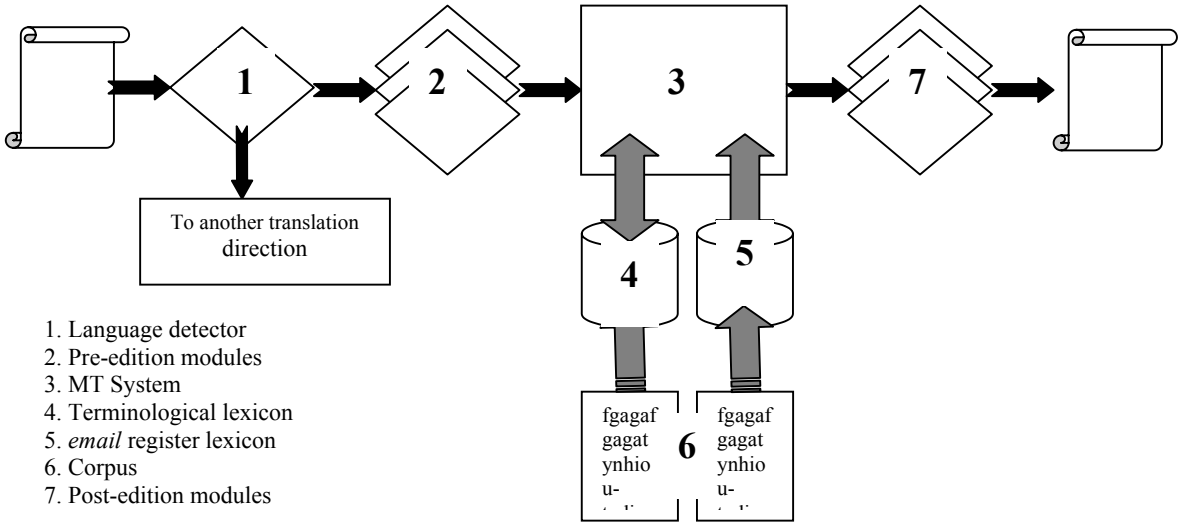
In section 2, we describe our design of the process to get a fully functional prototype of unsupervised email MT in a selected area of the virtual campus of the UOC. Emanating from that design, in section 2.1, we present the evaluation process we have set; in section 2.2 a preliminary typology of errors and problems; and in section 2.3 the modules we foresee will be necessary to be built to face the task. Then, section 3 presents the work done so far on building such modules. Last, section 4 sets some concluding remarks and future work.

2 Outline of the Project

INTERLINGUA is going to address the problem by adapting an MT system in two ways: (a) building before and after it general external modules of both automatic pre-edition and post-edition of the text; and (b) building terminological lexicons for every communicative space according to its domain and a lexicon for the email register vocabulary (eMRV) –the Figure below shows a very simple sketch of the environment. Besides, users will be informed that their emails are going to be machine-translated, so we will set the appropriate informative actions.

The system we have adopted is Sail-Labs Incyta ES/CA, a development of METAL, which, according to preliminary evaluations not to be discussed here, has proved to be the best program to translate from Spanish to Catalan and the other way round.

When we started the project, the first thing we realized is that we needed a sound process of evaluation in order to acknowledge both (a) the actual linguistic effects of the communicative situation related to machine-translation, and (b) the performance of the MT system in that framework. That is, micro-evaluation and macro-evaluation. The task, which is described in section 2.1, has been designed as a complex process at different



Sketch of the environment

levels. As a general approach, we follow the ISLE guidelines [ISLE00] –since they are MT-evaluation specific– adapted to the needs of our project. Incidentally, the process involves the constitution, marking, and alignment of appropriate corpora for each level of evaluation.

Macro-evaluation will provide reliable scores to know and show where we are, what can we expect from here and what do we will eventually reach. Micro-evaluation is carried on to obtain qualitative and quantitative information to decide in which direction should we go first –which kind of modules shall we priorize to achieve a greater impact on the quality of the translation.

To serve as a prototype environment, the so-called *Fòrum d'Informàtica* (Computer-Science Newsgroup) has been chosen. In this newsgroup, students and other members of the community exchange information and opinions about computers, software, and related educational subjects. Messages and replies are posted in Catalan or Spanish indistinctly as it is assumed that all users comprehend both languages. This environment was chosen because it provides a corpus of emails in either languages large enough to carry an evaluation, and because it belongs to a clear terminological domain. Unfortunately, although there are several corpora available for Spanish and Catalan (see [Badia98]) none of them include email texts, so we had to constitute our own corpus.

2.1 The evaluation process

As told above, we realized that the evaluation should be performed at different levels –each one mixing with the rest. This implies constituting, marking, translating, aligning, and evaluating different versions of the corpora. Each corpus consists of 130 emails and 12,000 words. Such levels are the following:

– Translation direction. Both directions, SPA-CAT and CAT-SPA must be evaluated.

– Granularity. Email-to-Email and Sentence-to-Sentence. On the one hand, complete emails must be translated as-they-are, without additional segmentation or spacing and punctuation correction. This will shed light on the performance of the MT system related to email structural problems. On the other hand, translation of the manually segmented emails will show how the system performs as it is naturally prepared to work –that is, Sentence-to-Sentence.

– Limits. We want to evaluate the system's performance (a) without any kind of correction or edition of the input text; and (b) once the input has been manually pre-edited according to a sheet of

style basically oriented to correct punctuation, *typos* and lexical errors. The former will set the baseline of the project –where we are when we start: what could we get without any intervention, just plugging in the MT system to the email server. The later will settle the uppermost level of expected performance. That is, as by now we can't make internal changes on the MT system (we can only act on its surroundings), the environment can hardly aspirate to surpass that top-level performance –except by improvements reachable by vocabulary enrichment.

The first step in the evaluation process consists in selecting the most suitable items from the ISLE guidelines for both macro-evaluation and micro-evaluation. For macro-evaluation, we choose *intelligibility* and *fidelity*, and also *terminological precision* because the e-mails of the newsgroup are terminologically rich. *Style* is also selected because we want translations to keep the informal flavour of the original e-mails. We reject ISLE items such as *clarity*, *coherence*, *consistency*, *informativeness*, and *readability* because, in our opinion, these items would be suitable if inputs were well written and coherent but informal e-mails are often characterized by the lack of these qualities. We prefer to outstand *intelligibility* and, if the e-mail is clear, consistent, or readable, we consider these qualities as factors that improve intelligibility.

The items of the micro-evaluation are the errors to be solved in the future. These items are grouped in what the ISLE calls *characteristics of the input* and *characteristics of the output.* The characteristics of the input cover errors made by e-mail writers and are classified in *performance errors* and *language-competence errors.* The performance errors are typing errors. The language-competence errors are *syntactic error, spelling error, intentional lexical error, non-intentional lexical error, expression error and language interference.* By *intentional lexical error,* we mean lexical items that deviate from the standard and are used intentionally by the writer (e.g. 'holassss' instead of 'hola' ('hello')). However, in *non-intentional lexical errors* the writer is unaware of his/her wrong use of a lexical item. *Expression errors* are wrong uses of an expression in a certain context and also preposition errors within an expression. *Language interference* is the influence of the writer's knowledge of the target language on the use of words and expressions that are not correct in the source language. Another case of *language interference* is when the writer prefers to use phrases, terms, etc. in the target language or any other language because in the computer-science domain these words, terms, and phrases are more commonly used.

The characteristics of the output are those errors that are imputable to how the MT system translates. The items related to the characteristics of the output are syntactic error, morphological error, word not translated, word badly translated, expression not translated, expression badly translated.

After having selected the items of the evaluation, we have developed a tool for the judges to evaluate the translation of e-mail segments. By this tool, the evaluation of each segment is carried out in five steps. Firstly, the evaluators must judge whether the translation is intelligible or not without seeing the source segment. Then they see the source segment as well and must decide whether the translation is faithful to the original in content and style. If the translation is not fully intelligible or faithful, the judge must grade the errors responsible for it.

We establish 4 levels of error, based on Green's Rating Scale [Green77]: 1- minor error (error that affects style), 2- error which does not impair comprehension of the segment, 3- error which leads to ambiguity, 4- serious errors (error that makes the translation unintelligible). From this rating scale we can infer most of Van Slype´s grading scales of intelligibility and fidelity [VanSlype79], so if the judge detects serious errors we can infer that the translation is unintelligible, if the errors are minor or do not affect the meaning of the sentence the segment is fairly intelligible and if the error leads to ambiguity, the segment is unfaithful. From our point of view, in e-mail translation the important thing is the global feeling of 'intelligibility' and 'fidelity' not the grades of it. Because of this and for simplicity's sake, we decided not to make judges evaluate grades of intelligibility and fidelity so they grade errors straightforwardly. The fourth step is to analyze the original and the translation and to typify the error as either an input error (of the user) or an output error (of the system). If an input error, they must state whether this error is a *syntactic error***,** a *spelling error*, a *lexical error*

(*intentional* or *non-intentional*)**,** an *expression error,* or a piece of language *interference*. If an output error, they must state whether the error is morphological or syntactical or whether there are words, terms or expressions not translated or badly translated. After having performed these steps, the judge can write comments that will be an important source of information about future improvements and data for investigations on e-mail writing and MT-translation.

At this moment all of the corpora have been constituted, treated and translated, and sent to the judges.

2.2 Preliminary typology of errors and problems

At the moment of the public presentation of this paper, the evaluation process will be completed. Therefore, we will be able to show results that classify, quantify and rank in order of actual impact all errors and pieces of linguistic deviation of formal texts which cause malfunctioning of the MT system.

By now, preliminary examination of the corpora allows to present the following typology of problems to be found in our domain of input email texts. We have detected three main categories: (1) non-intentional errors; (2) intentional deviations of the standards; and (3) lexical gaps in the system.

1. non-intentional errors
1.1 performance errors (*typos*, involuntary word repetition)
1.2 competence errors[2]
1.2.1 orthographic (spelling mistakes)
1.2.2 lexical (specially wide-spread Spanish-Catalan interference –*barbarisms*)
1.2.3 syntactic (specially typical incorrect use of some functional words in Catalan by influence of Spanish)
1.2.4 cohesion errors (such as incorrect anaphoric agreement)
2. intentional deviations
2.1 language shift
2.1.1 lexical (usually Catalan words in Spanish texts or vice versa, but also English words in both)
2.1.2 phrasal (longer texts chunks of other languages in the email)
2.2 new forms of expressivity typical of the email register
2.2.1 lexical (e.g. SMS-like shortenings, orthographic innovations such as *tod@s* or *todos/as,* phonetic reproduction as in *wow*, capitalization to show emphasis as in *"It was NOT me"*[3], eMRV: words usual in speech but not normative –therefore they are not in dictionaries)
2.2.2 visual (e.g. smileys, multiple marking as in *Is it true???!!!!*)
2.2.3 pragmatic (use of *linking* –fragments of the mail one is answering to simulate a dialogue)
2.2.4 simplified punctuation (intentional lack of punctuation marks, accents...)
2.2.5 simplified syntax (e.g. sentence-shortening by preposition drop, composition by symbols instead of words as in *Apache+Tomcat*)
3. lexical gaps (vocabulary missing in the system: domain terminology, speech community's terminology, acronyms, dialectal vocabulary, standard words still missing in the system's database)

In addition to such problems coming from the input text, there are some systematic cases that largely cause malfunction of the MT system and that can be straightforwardly detected just looking at the output:

– (Inappropriate) translation of proper nouns – especially in the "From" and "To" fields.

– Systematic lack of disambiguation (therefore, usual bad translation) of a number of typical homographs –specially grammatical words, as ho/el, per/per a, en/a...

– Non-translated terminology

Such cases are retrievable from the output text since they come out tagged as problematic. Therefore, we shouldn't wait to complete the evaluation process to realize we ought to build appropriate post-edition modules to solve them –see next section, specially for the case of terminology which

[2] Some of them are caused by linguistic interference: influence of the Spanish norms on writers in Catalan or vice versa. The extent of the problem is to be analyzed further but in any case they are classified in principle as competence errors.

[3] Although these probably won't cause errors in translation, they should be faced in order to try to preserve their pragmatic function in the translated email.

still remains untranslated even after having built new terminological dictionaries.

2.3 Foreseen decision-taking modules

As discussed above, the evaluation will be the key factor to eventually decide what modules will be developed, and, remarkably, which are the priorities –work will concentrate on those that are expected to have greater impact on the quality of the results. Nevertheless, at this stage we foresee that the following modules and tasks will be needed:

(a) Language detector

This first module is very important because it will decide the direction of the MT system (SPA-CAT or CAT-SPA). If we fail to detect the language of the e-mail, obviously, the result of the MT process will be completely useless.

(b) Automatic pre-edition

(b.1) Punctuation recovery

Many people write e-mails without any kind of punctuation marks. Without such information the MT system has no way to track sentence limits –a problem related to segmentation–, leading to important errors in translation.

(b.2) Typing mistakes recovery

Mails usually contain several orthographic errors due to *typos* –users know how to spell the word but fail to write it due to rapid writing. We foresee it will be important to detect this kind of errors, although it is dangerous for our system to perform fully automatic spelling correction, since the input text is full other kinds of unknown words.

(b.3) Accent recovery

Users tend to lack accentuation in emails. This is a big source of ambiguity in SPA and CAT since the lack of accents dramatically enlarges the number of homographs –one of the main causes of lexical transfer errors.

(c) Lexical modules

(c.1) Techniques of rapid terminology extraction

We will develop subject-specific (computer-science) glossaries by combining different NLP techniques (see Section 3).

(c.2) eMail Register Vocabulary (eMRV)

The other main class of unknown words in our environment is eMRV. Different to terminology, it is not domain-specific but register-specific (email register – close to speech). We shall build a lexicon module for eMRV using similar techniques that those used for Terminology extraction. The main problem would be getting an email corpus large enough for the task and the need for morphological inflection and derivation.

(d) Automatic post-edition

(d.1) Homograph disambiguation

The MT system in some cases can't disambiguate translation of high-frequency homographs, therefore it tags the output for the option: e.g. SPA (original): "llevar el temario al día" → CAT (MT-translated): "portar el temari al/en dia". This kind of ambiguities are a well-known problem in CAT<->SPA translation [Canals02]. We plan to develop an algorithm based on Machine Learning [Knight97] [Màrquez00] to disambiguate the most productive cases.

(d.2) Terminology on demand

We want to extend the algorithms developed for rapid terminology resolution to work "on line" with the MT-system as a post-edition module. This module (*TonD*) tries to detect an untranslated string as an unknown terminological entry and find it's translation on a multilingual corpus. There are many problems behind this simple idea: the terminological unit not always correspond to the untranslated string and may extend some words before of after it, the untranslated string may correspond to an misspelled word not detected in the pre-edition modules, etc.

(e) Proper Noun Resolution.

Translation (or non-translation) of Proper Nouns is a problem that mixes with that of confusion between proper nouns and other kinds of capitalized words (at the beginning of a sentence, for emphasis or for other reasons). We still have to perform tests to decide about dealing with it as a kind of post-edition error-recovering module (since possible PNs come output-tagged by the MT system) or as a pre-edition one –as a more standard PN-detection module–.

3 Current work on decision-taking modules

We have adapted van Noord's TextCat language identificator[4], which is an implementation of [Cavnar94]. The straight application of this identificator on our corpus of emails gives a precision score of 93.8%[5]. Applying it to the pre-edited corpus, precision improves slightly (94.6%). The relative low precision of the detector is mainly due to the short length of emails and to the fact that some of them mix languages.

As for automatic pre-edition, we are testing Machine Learning approaches on the tasks of accent and punctuation recovery [Beeferman98]. The task of punctuation recovery has connections with that of capitalization recovery and proper noun detection. In order to train the Machine Learning algorithms we need a larger corpus than the one used for evaluation, so we are using the same corpus we have developed for terminology extraction.

We are developing a module to detect typing errors based on minimal edit distance and supported by subject lexicons and subject specific corpora. The module will try to correct an unknown word only if it's not present in the subject lexicon of any of the implied languages —Spanish, Catalan, and English. This query will be extended to subject specific corpora for the same languages. The module will take into account the relative position of characters in a standard Spanish-Catalan keyboard [Schulz01].

[4] http://odur.let.rug.nl/~vannoord/TextCat/index.html

[5] Using the language models of Spanish, Catalan, French and English and performing the detection on the body of the mail

At the moment, we are approaching all pre-edition problems separately. Nevertheless, our goal is now to find the method to deal with all of them in an integrated way.

As for terminology, we have developed an extraction module and a parallel corpus (a compendium of manuals and technical documents) on computer technology. We are applying some different techniques of terminology extraction: purely statistical, statistical with entropy-based scores and a linguistically-based approach. The statistical approach [Church90] is based on frequency and results are filtered out with a list of stop words. Entropy-based methods [Merkel00] provide useful information to discriminate those multi-word units than can be terminological. The linguistic approach [Kupiec93] works with a POS tagged corpus. In order to POS-tag the corpora we are using tools and techniques developed by [Padró96] [Padró97] and [Màrquez97]. Such techniques are used to extract monolingual glossaries from subject-specific corpora. Furthermore, we plan to extract terminology translation from aligned, equivalent and comparable corpora.

We have also developed a module that automatically detects untranslated terminology units in the output. The next step is to link these modules to configure *TonD*. Related tot this, at the moment, we are applying EBMT methods on aligned corpora [Nagao84] [Niremburg95] giving good results for high frequency terms. In a next step, these methods will be compared to those of [Allen98].

Last, with respect to eMRV inflection, we have developed techniques that have proved to be highly effective for other morphologically rich languages [Oliver02].

4 Concluding remarks and future work

In this paper, we have presented the INTERLINGUA Project and its design. Although the development of problem solutions is on a preliminary stage, we think that the proposal of modules that monitor automatically all the translation process for a real application that demands an unsupervised process is important enough. This proc-

ess involves decision-taking actions such as choosing translation direction, recovering accents and punctuation, stating proper noun interpretations, disambiguating homographs, and finding the right term in the target language even when it is not in the system's dictionary.

Besides, our approach to e-mail MT is based on a sound investigation of the peculiarities of this register and we take into account new aspects such as bilingualism.

The streamlines of the future work will be based on the results of the evaluation, after realizing (a) whether the approaches we are taking describe and solve the most relevant problems or we must face problems not expected so far and (b) what lines must be prioritized in order to optimize results. In the case our evaluation approach proves to be insufficient we will test other translation metrics also applicable to MT such as those described in [IJLP00].

As for automatic pre-edition and post-edition, we will also explore the works by Hogan and others (e.g. [Lenzo98]) on accent mark reinsertion and [Allen00,02], [Krings01] and [Knight94] on error recovering and text repairing.

References

[Allen98] Allen J. and C. Hogan (1998) Expanding lexical coverage of parallel corpora for the EBMT approach. *Proceedings of the 1st. International Language Resources and Evaluation Conference (LREC98) vol. 2, pp. 747-754.* Granada http://www-2.cs.cmu.edu/~chogan/Publications.html

[Allen 00] Allen J. and C. Hogan (2000) Towards the development of a post-editing module for MT raw output: a new productivity tool for processing controlled language. *Proceedings of CLAW2000.* http://www.controled-language.org

[Allen02] Allen J. (2002) Review of Repairing Texts: Empirical Investigations of MT Post-Editing Processes. *Multilingual Computing and Technology 13.2, 27-29.* www.multilingual.com/allen46.htm

[Atserias98] Atserias J. and Rodriguez H. (1998) TACAT: Tagged Corpus Text Analyzer. *Technical Report.* Software Department (LSI), Technical University of Catalonia (UPC). Barcelona.

[Badia98] Badia, T., T. Cabré, M. Pujol, A. Tuells, J. Vivaldi, Ll. de Yzaguirre (1998) IULA's LSP Multilingual Corpus: compilation and processing, Proceedings of *the 1st. International Language Resources and Evaluation Conference (LREC98)*, pp. 29-31. Granada

[Beeferman98] Beeferman D, A. Berger and J. Lafferty. (1998) Cyberpunk: A lightweight punctuation annotation system for speech. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Seattle, WA.

[Canals02] Canals R., A. Esteve, A. Garrido, M.I. Guardiola, A. Iturraspe, S. Montserrat, S. Ortiz, H. Pastor, P.M. Pérez & M.L. Forcada (2002) The Spanish<->Catalan machine translation system interNOSTRUM. *Proceedings of MT Summit VIII*. Santiago de Compostela, Spain.

[Carmona98] Carmona J., Cervell S., Màrquez L., Martí MA., Padró L., Placer R., Rodríguez H., Taulé M. and Turmo J. (1998) An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the 1st. Conference on Language Resources and Evaluation. LREC'98*, pp. 915-922. Granada.

[Cavnar94] Cavnar W.B. and J. M. Trenkle (1994). {N}-Gram-Based Text Categorization. *Proceedings of {SDAIR}-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas, US

[Church90] Church, K.W. and P. Hanks (1990). Word association norms, mutual information and lexicography. *Computational Linguistics* 16(1): 22-29

[Fais01] Fais L. and Ogura K. (2001) Discourse Issues in the Translation of Japanese Email. *Proceedings of PACLING 2001.* http://afnlp.org/pacling2001/pdf/fais.pdf

[Green77] Green R. (1977) Analysis of errors. *CEC, memorandum,* October, 5+5 p, Luxembourg

[IJLD00] International Journal for Language and Documentation 3 (2000) Translation Quality Evaluation. http://www.crux.be

[ISLE00] ISLE. International Standards for Language Engineering. (2000) *The Isle Classification of Machine Translation Evaluations*. http://www.isi.edu/natural-language/mteval

[Knight94] Knight K. and I. Chander (1994) Automatic Post-Editing of Documents. *Proceedings of AAAI 1994.* http://www.isi.edu/natural.language/people/knight.html

[Knight97] Knight K. (1997) Automating Knowledge Acquisition for Machine Translation. *AI Magazine* v. 18 n. 4. pp. 81-96. citeseer.nj.nec.com/knight97automating.html

[Krings01] Krings H. (2001) Repairing Texts: Empirical Investigations of MT Post-Editing Processes. *Translation Studies Series*. Kent State University Press. Ohio. http://bookmasters.com/ksu-press/ksu071.htm

[Kupiec93] Kupiec, J. (1993). An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics* (ACL-93):17-22

[Lenzo98] Lenzo K., C. Hogan and J. Allen. Rapid-Deployment Text-to-Speech in the DIPLOMAT System. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP '98)* volume 5, pp. 1999-2002. Sydney. http://www-2.cs.cmu.edu/~chogan/Publications.html

[Màrquez97] Màrquez L. and L. Padró (1997) A Flexible POS Tagger Using an Automatically Acquired Language Model. *Proceedings of EACL/ACL 1997*. Madrid, Spain.

[Màrquez00] Màrquez L. (2000). Machine Learning and Natural Language Processing. *@techreport{ marquez00, Machine Learning and Natural Language Processing {LSI-00-45-R},* "Departament de Llenguatges i Sistemes Informàtics (LSI), Universitat Politècnica de Catalunya (UPC). Barcelona, Spain. citeseer.nj.nec.com/marquez00machine.html

[Merkel00] Merkel M. and M. Andersson. (2000) Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. *In Proceedings of Recherche d'Informations Assistee par Ordinateur 2000* (RIAO'2000).

[Nagao84] Nagao, M. (1984) A Framework of a Mechanical Translation System by Analogy Principle. En A. Elithorn & R. Banerji (eds.) *Artificial and Human Intelligence*. Amsterdam: Elsevier Science Publishers, 173-180.

[Niremburg95] Niremburg, S. (ed.) (1995) The Pangloss Machine Translation System. *Joint Technical Report, Computing Research Laboratory (New Mexico State University), Center for Machine Translation (Carnegie Mellon University), Information Sciences Institute (University of Southern California).*

[Oliver02] Oliver A., L. Màrquez & Castellón I. (2002) Adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar: aplicación al serbocroata y ruso. *Proceedings of SEPLN 2002*. Valladolid, Spain.

[Padró96] Padró L. (1996) POS Tagging Using Relaxation Labelling. *Proceedings of COLING 1996*. Copenhagen, Denmark.

[Padró97] Padró L. (1997) A Hybrid Environment for Syntax-Semantic Tagging. *Ph.D. thesis*. Software Department (LSI), Technical University of Catalonia (UPC). Barcelona.

[Schulz01] Schulz K. and S. Mihov (2001) Fast String Correction with Levenshtein-Automata. citeseer.nj.nec.com/501807.html

[VanSlype79] Van Slype G. (1979) Critical Study of Methods for Evaluating the Quality of Machine Translation. *Commission of the European Communities Directorate General Scientific and Technical Information and Information Management Report BR 19142* http://www.ling.ed.ac.uk/~beatrice/bibliography.htm

[Yates93] Yates JA. and Orlikowski W.J. (1993) Kneejerk Anti-LOOPism and other Email Phenomena: Oral, Written, and Electronic Patterns in Computer-Mediated Communication. *MIT Sloan School Working Paper 3578-93*. Center for Coordination Science Technical Report 150.