Bilingual Newsgroups in Catalonia: A Challenge for Machine Translation

Salvador Climent, Joaquim Moré, Antoni Oliver, Míriam Salvatierra, Imma Sànchez, Mariona Taulé* and Lluïsa Vallmanya

Internet Interdisciplinary Institute
Universitat Oberta de Catalunya

* Department of Linguistics
Universitat de Barcelona

Abstract

This paper presents a linguistic analysis of a corpus of messages written in Catalan and Spanish, which come from several informal newsgroups on the Universitat Oberta de Catalunya (Open University of Catalonia; henceforth, UOC) Virtual Campus. The surrounding environment is one of extensive bilingualism and contact between Spanish and Catalan. The study was carried out as part of the INTERLINGUA project conducted by the UOC's Internet Interdisciplinary Institute (IN3). Its main goal is to ascertain the linguistic characteristics of the e-mail register in the newsgroups in order to assess their implications for the creation of an online machine translation environment. The results shed empirical light on the relevance of characteristics of the e-mail register, the impact of language contact and interference, and their implications for the use of machine translation for CMC data in order to facilitate cross-linguistic communication on the Internet.

Introduction

Catalonia (Spain) is a bilingual country[1] where the native language, Catalan, is being displaced to a great extent by Spanish when Catalan speakers communicate with members of different linguistic groups. Catalan speakers tend to use Spanish if they know the addressee is not a native Catalan speaker, even if the addressee is able to understand Catalan. In addition, they tend to use Spanish as the default language when they are unsure of the addressee's level of understanding of Catalan. In computer-mediated communication (CMC), the tendency of Catalan speakers to use Spanish by default is even more noticeable. Thus we reasoned that the gradual replacement of Catalan by Spanish could be prevented if those writing in Catalan felt that, thanks to human language technologies, they did not need to shift to Spanish. The existence of a machine translation system that could automatically translate e-mails into the language of the addressee would make it unnecessary for users to write in Spanish. Users' confidence in the translation quality would lead to their using their own language regardless of the other's language.

With this idea in mind, we and other colleagues founded the INTERLINGUA project. INTERLINGUA aims to promote machine translation (MT) to enable users to employ their own native languages on the Internet. The project is not limited to Catalan and Spanish, but could be used for any pair of languages.[2] Hence, INTERLINGUA is a project to facilitate communication between people who speak different languages.

However, MT systems are currently not capable of translating spontaneous e-mails accurately. Such systems work reasonably well if the input is in a standard form, but, as we will see in this study, e-mails often contain spelling mistakes, typing errors and the like. Moreover, the systems are not prepared to cope with the creative and spontaneous use of language typical of e-mails, especially in a bilingual community. Thus, e-mail presents certain new challenges for MT. On the one hand, non-standard forms need to be converted into standard forms that can be recognized by the MT system, a challenge that the MT community has not yet addressed (Climent, Moré & Oliver, 2003). On the other hand, the project must also take into consideration the use of language in a bilingual society.

In this paper, we present the results of a study of e-mails written in the UOC Virtual Campus, carried out as part of the INTERLINGUA project. These e-mails were translated from Catalan to Spanish or viceversa using an MT system. According to the translation errors, we classified and quantified the linguistic features in the original messages that caused translation problems. Specifically, we describe the non-standard features that have a negative impact on translation quality, stressing those aspects that involve bilingualism.

The Sociolinguistic Situation in Catalonia

Catalan is still recovering from the extreme measures imposed under the rule of Spain's former dictator Francisco Franco, who declared Spanish the only official language and prohibited Catalan in the press, broadcasting, theatre, schools, street signs, advertising, shop signs, etc. Since the return of democracy after Franco's death, Catalan and Spanish have been co-official languages in Catalonia, and linguistic and educational policies have attempted to redress the situation of Spanish dominance over Catalan. However, according to official statistics, although these policies have improved the levels of people's understanding and ability to speak Catalan, they have not succeeded in improving levels of spontaneous usage. These levels are shown in Table 1. The data on people's understanding and ability to speak are taken from a survey by the Centro de Investigaciones Sociológicas (Sociological Research Centre) (CIS, 1998). The usage data are taken from an analysis by Cerdà (2001), in the section entitled "Lengua predominante y competencia lingüística" (Predominant language and linguistic competence) from the same CIS survey.

Understand Catalan	Speak Catalan	Spontaneously use Catalan	Spontaneously use Spanish	Spontaneously use both
97%	79%	41%	43%	16%

Table 1: Percentage of the population in Catalonia able to understand and speak Catalan and Spanish (Cerdà, 2001; CIS, 1998)

Another survey, carried out by the Institut d'Estadística de Catalunya (Catalonian Institute of Statistics) (IDESCAT, 2001), shows that the percentages for reading and writing Catalan are lower than those for oral comprehension and use. These are summarized in Table 2.

Understand Catalan	Speak Catalan	Read Catalan	Write Catalan
95%	73%	72%	46%

Table 2: Percentage of the population in Catalonia able to speak, read and write Catalan (IDESCAT, 2001)

The ability to read and write is obviously essential in CMC; thus, the growth of Spanish at the expense of Catalan is expected to be greater in this form of communication than in spoken contexts.

These statistics support the impression that Catalans themselves have, namely that the Catalan language is not essential for living in Catalonia. Moreover, the situation has worsened in the last decade due to massive immigration from Morocco, Sub-Saharan Africa and Latin America. According to the CIS survey, 18% of the immigrant population in Catalonia did not understand any Catalan at all in 1998.

Immigrants learn Spanish because it is the official language of Spain and a widespread means of communication. They do not feel compelled to learn Catalan due to the fact that, on the one hand, Spanish is omnipresent in mass media, books, shops, law, etc., and on the other hand, Catalans communicate with them in Spanish. The tendency of Catalan speakers to shift to Spanish when the addressee uses Spanish is an example of *code-switching*. The use of Spanish when addressing someone whose origin is unknown or even someone identified as a Spanish speaker, despite their being able to speak Catalan, is considered a sign of politeness. Many Catalan people also consider Spanish more prestigious, and view it as more important than Catalan because it is a language of wider communication. Some people also code-switch to Spanish because they feel more competent in this language (Pujolar, 2000).[3]

The presence of code-switching on the Internet is particularly apparent, as interactions between Catalans and those from outside Catalonia through chats, e-mail, etc. become increasingly common. Aside from the fact that Spanish has a far

greater presence on the Internet than Catalan, the possibility of addressing a wider group of people from all over Spain and the world encourages the use of Spanish when writing to someone for the first time, or when replying to a message in Spanish, even in cases where the addressee understands Catalan.

The use of new technologies, where Catalan is conspicuous by its absence, has the greatest effect on young, urban dwellers (Castells & Díaz de la Isla, 2001). The young who, with the return of democracy, were expected to lead the way in making Catalan the normal, everyday language of communication, have instead tended to use Spanish. Thus, communication over the Internet in Catalan is increasingly unlikely, in that the main group using this technology is urban youth. This trend goes against the established thought of many scholars who praise the Internet's ability to promote minority languages and multilingualism (Warschauer, 2000; Mehsching, 2000).

Language Choice in a Preliminary E-mail Sample

In order to assess the current status of Catalan on the Internet, we analyzed a sample of e-mails from the UOC-Catalonia[4] Virtual Campus. Catalan is the institutionalized language in the Virtual Campus. That is, Catalan is the language used in educational material as well as by teachers when addressing students in the virtual environments. Although there are no official restrictions on the spontaneous use of other languages, it is assumed that the people who register for instruction through UOC-Catalonia will be fully competent in Catalan.

As a test bed for the research, several so-called *Fòrums d'Informàtica* (computer science newsgroups) were chosen. In these informal newsgroups, students exchange information and opinions related to computers, software, bugs, cheats, academic subjects, etc. These newsgroups are not, by themselves, representative of text-based, asynchronous computer-mediated communication in Catalonia as a whole. Both linguistically and sociolinguistically, the communicative situation in Catalonia is too diverse and complex to be represented by a single group. Nonetheless, we believe that these newsgroups are good examples of contemporary practice, and that by analyzing them we can learn many things about how languages are currently used in Catalonia in online communication.

Although the official language of UOC-Catalonia is Catalan, messages and replies are posted in the forums in both Catalan and Spanish, sometimes mixing the two languages. In an attempt to quantify the degree of code-switching between messages (not within messages), we analyzed all messages mailed to the forums between July and December 2002: 533 messages sent by 254 users (an average of 2.1 messages per user). In this sample[5], 76% of the messages were in Catalan and 24% were in Spanish. To infer the degree of code-switching in the group, we took into account only 189 e-mails: the ones that were replies made by those users defined as *spontaneous* in one or the other language. The criteria for defining

spontaneous users are given in Figure 1. We considered those users who wrote initiating e-mails (i.e., not replies to other e-mails) in both languages indistinctly as *indifferent*. Finally, we considered those users who replied to e-mails originally written in A in the same language, but who did not write any initiating e-mails, as *undetermined*.

Users (U) are regarded as spontaneous in language A, if:

- They wrote only in A and not all of their e-mails were replies to initiating e-mails written in A (no code-switching)
- They replied in A to e-mails written in B; and
- They generally wrote in A, although they replied in B to certain e-mails written in B.

Figure 1: Definition of a "spontaneous user" of a particular language.

Table 3 shows the results of applying this classification scheme to the messages in the sample.

<table>
<tr><td colspan="2">Spontaneous Catalan Users

68.9%</td><td colspan="2">Spontaneous Spanish Users

18.1%</td><td>Indifferent

1.2%</td><td>Undetermined

11.8%</td></tr>
<tr><td colspan="2">Reply to e-mails written in Spanish</td><td colspan="2">Reply to e-mails written in Catalan</td><td rowspan="3"></td><td rowspan="3"></td></tr>
<tr><td>In Catalan</td><td>In Spanish</td><td>In Catalan</td><td>In Spanish</td></tr>
<tr><td>57.1%</td><td>42.9%</td><td>15.4%</td><td>84.6%</td></tr>
</table>

Table 3: Classification of messages in terms of spontaneous language use

Although these results may not be statistically significant because the sample size is small, they suggest that code-switching between messages is an important phenomenon among spontaneous users of Catalan (43% of messages), but less so among spontaneous users of Spanish (15% of messages). This, and the fact that only 69% of messages were sent by spontaneous Catalan users, appears paradoxical when we recall that, in the environment studied, Catalan should be the only language used.

In other UOC-Catalonia environments, the dominance of Spanish over Catalan is even more pronounced. Many users of the Ph.D. virtual classrooms are students who do not live in Catalonia (but rather come from other parts of Spain or South America), so they need not be competent in Catalan. As a result, activities in 14 of

the 15 Ph.D. classrooms are carried out exclusively in Spanish, despite the fact that structural and institutional information is given in Catalan.

All of this evidence suggests a trend that is endangering Catalan as a language for communication on the Internet.

The Role of Machine Translation

MT systems allow users to employ their own language, regardless of whether the addressee can read it. If the quality of the translation is sufficiently high, the sender can trust the MT system and avoid having to code-switch in order to guarantee communication. In this way, the better the quality of the translation, the more likely the sender is to use his or her native language.

However, to guarantee quality in the translation of e-mails, certain factors need to be taken into account. The MT systems currently in operation depend on the input text being correct and standardized; i.e., the systems' rules and lexical databases are only able to recognize standard words and correctly written texts. Even when working with standardized texts, machine translation systems make mistakes; the greater the structural differences in the languages involved, the more mistakes the systems make. Thus it is generally assumed that MT systems are not able to produce perfectly correct translations, merely approximate translations that allow the addressee to understand the gist of the text. In the case of non-standardized texts, such as e-mails, the quality is expected to be even lower. Moreover, messages written by bilingual users may contain further deviations: for example, these messages might mix languages when quoting or linking to previous articles, the user may employ words of the other language, make spelling and grammatical mistakes due to language interference, and so on.

Currently, MT specialists take for granted that any text to be submitted for automatic translation should be manually pre-edited to overcome errors and deviations from the standards, as well as post-edited to correct the remaining mistakes. However, in e-mail communication, human pre- and post-editing are not feasible, as the system has to work in real time and completely automatically.

For this reason, an in-depth analysis of the e-mail register is needed to shed light on the specific problems an MT system might be required to overcome in order to produce a good quality translation in an unsupervised environment (no pre- or post-editing).

The E-mail Register

Researchers in the field of CMC have studied various aspects of the e-mail register. We will focus here on the relationship between the linguistic structure of e-mails and non-standard features.[6]

According to Herring (2001), most non-standard features in English e-mails are deliberate choices made by users to:

- Economize on typing effort
- Mimic spoken language features
- Express themselves creatively

Murray (2000) claims that CMC in general uses what she calls “simplified registers," characterized by (among other features) short sentences, special lexicon and feedback devices that facilitate the reader's comprehension, as well as simplifications that may include the use of abbreviations and the omission of articles, pronouns, and copula. According to Murray, the technology constrains time and space. CMC relies on typing, computer, and network speed, and CMC gives no visual paralinguistic or nonverbal cues. Consequently, CMC users employ strategies that reduce the time needed to write the message or substitute for the lack of paralinguistic and nonverbal cues.

According to Yates and Orlikowsky (1993), the mimicking of spoken language features in e-mail results in unconventional orthography, such as textual indication of emphasis (e.g., *If an implementation DOES support vectors...*), informal words typically used in speech (e.g., *groove, stuff*) or syntactic informality often taking the form of incomplete sentences and conversational cadences, usually combined with word choice and punctuation in order to simulate oral communication, as in *Hmm, I see...*

As for the creative use of language, Alonso, Folguerà and Tebé (2000), focusing on the Catalan lexicon used in the Internet (“the Internet slang”), identify a category of informal, expressive lexical elements, such as *correu tortuga* (“snail mail”), *emili* (“e-mail" referred to humorously, due to its resemblance to the proper noun *Emile*). These lexical items are common in the Internet lexicon because they embody a recreational, creative, ironic and informal dimension.

For Fais and Ogura (2001), there are features that are exclusive to e-mail[7], causing it to differ significantly from both formal textual and spoken language. These are visual and discourse-level phenomena such as the following:

1. A highly idiosyncratic use of indentation and spacing to mark paragraph shifts, in such a way that a difference in paragraphing is typically interpreted as a cue to a different topic.

2. Openings and closings: “Closings are typically formalized and devoid of meaning. Openings, on the other hand, contain information about the

addressee(s). (...) The variability in format for openings and closings also makes their recognition a difficult problem."

3. Use of visual strategies to capture aspects of spoken utterances, such as:

   - Non-standard punctuation: "(In Japanese) Center dots are the most frequent type of non-standard visual device. (...) They represent a "hanging intonation" which invites the listener/reader to draw inferences, supplementing the explicit meaning in the text."

   - Non-standard spelling (for example, elongating one sound by repeating the letter several times) is used to emphasize a word or to mimic an emphatic pronunciation.
   - Discourse characteristics: "Authors also attempt to capture the flavor of speech, and employ typically spoken discourse markers to do so," e.g., using *um* and *ah*, sometimes called fillers or filled pauses.

The aforementioned studies tend to focus on new, intentionally expressive devices. This may be because, as Herring (2001) concludes,

> actually, although computer-mediated language often contains non-standard features, only a relatively small percentage of such features appears to be errors caused by inattention or lack of knowledge of the standard language forms. (Herring, 2001, p.616)

However, in the e-mail register, if the user writes quickly and carelessly, texts may contain many unintentional language mistakes. Likewise, the assertion that lack of knowledge of standard language forms is relatively minor may be true in monolingual English-speaking environments; however, in a bilingual community, lack of knowledge of standard forms of one language or the other can be significant. For instance, in Catalonia, some users writing in Catalan have less of a command of Catalan than Spanish, in that they may not have studied Catalan at school or may not be used to reading Catalan. Their lack of knowledge may be displayed, unintentionally, in e-mails. Another factor that has been inadequately discussed in the CMC literature is interference between two languages in the messages of bilinguals.

Research Questions and Methodology

The study was motivated by the formulation of the following questions:

1. Which non-standard linguistic features are responsible for bad translations when using an MT system?
2. To what extent are they intentional?
3. To what extent are they related to lack of knowledge of language norms?

4. Do users of our universe of study write in Catalan as well as they do in Spanish?
5. What is the influence of language interference in the features that cause translation problems?

According to this, we planned the study with the following steps:

1. MT System's performance evaluation
2. Linguistic classification of non-standard features that cause translation problems
3. Quantification
4. Interpretation of the results

The evaluation was performed in order to answer question 1. The linguistic classification was crucial to find out and delimitate problem spaces before quantifying them. The quantification of results provided answers to questions 2 to 5; that is, we would know if senders have a good knowledge of the language they are using and whether, for this reason, the translation quality would depend mainly on intentional non-standard features, typical of the e-mail register. In the case that the level of knowledge were not equal for Catalan and Spanish, we would know which language deserves more effort on linguistic correction (orthography, etc.) and to what extent the gaps in the knowledge of the language used are due to interference by the knowledge of the second language.

Evaluating Translation Quality: The Impact of Non-Standard Features

The MT system used in the INTERLINGUA project is Sail-Labs Incyta ES/CA, an application to translate between Catalan and Spanish and vice versa which is based on the prestigious METAL system, developed by Siemens. This system was evaluated in the first stage of the study in order to find out how it worked and what its shortcomings were. By analyzing the results of the evaluation we would know to what extent non-standard features were responsible for poor quality translations, and to what extent these features were attributable to specific aspects of the e-mail-register or, on the contrary, to the user's lack of awareness of standard language forms.

The evaluation of the system followed the ISLE international standards for MT evaluation (ISLE, 2000) and consisted of two processes: macro- and micro-evaluation (Van Slype, 1979).The macro-evaluation provided information about the acceptance of the translation system in a global perspective (intelligibility, fidelity, readibility of the e-mails translated). The goal of the macro-evaluation was just the validation of the translations performed in order to assess the usability of the system, without detailing its limitations. On the other hand, the micro-evaluation showed the system's limitations and was required to establish a strategy for

improvements. The micro-evaluation provided information about the origin of translation errors, whether they came from the system's shortcomings or from the user's writing. In this paper, we focus on the micro-evaluation.

In as much as the system translated segment by segment (roughly sentence by sentence), the micro-evaluation was carried out on text segments. The corpus prepared for the micro-evaluation amounted to 1239 segments in Catalan and 1128 segments in Spanish, taken from 129 randomly-selected e-mails for each language. The number of words amounted to a total of around 25,000. All segments were sent to the MT system so that the corpus was constituted by parallellizing each segment with its automatic translation. The source of the corpus was the postings between July and December 2002 in four computer science newsgroups, where students ask for assistance, offer advice, announce events and so on. The reason for choosing these newsgroups is that they are the most active and provide the largest corpus size

We developed a tool to perform the micro-evaluation. Using this tool, the evaluation of each segment was carried out in five steps. First, the evaluator had to judge whether the translation was intelligible or not without reference to the source segment. Second, given both the source segment and the translation, the evaluator had to decide whether the translation was faithful to the original in content, intelligibility and style. Third, if the translation was not fully intelligible or faithful, the evaluator had to grade the translation errors that led to the problem. We established four levels of translation error, based on Green's Rating Scale (Green, 1977): 1) a minor error (an error that affects the style), 2) an error that does not impair comprehension of the segment, 3) an error that leads to ambiguity, 4) serious errors (errors that make translation unintelligible). The fourth step was to analyze the original and the translation and to classify the translation error as either caused by the writer of the input, or by inadequate functioning of the system. If caused by the writer, the evaluator had to state whether the writer had expressed him- or herself badly, written a syntactically incorrect sentence, or used a non-standard language form (a typing error, a spelling error, or an intentional or unintentional lexical deviation). The evaluator also had to consider "language interference," a category that affects expression, syntax and lexicon. If the translation error was caused by the system, the evaluator had to state whether the translation error was morphological or syntactical or whether there were words, terms or expressions that were not translated or translated badly. Fifth, after having performed these steps, the evaluator could write comments that would be an important source of information for future improvements and research into e-mail writing and MT.

The evaluation was carried out by six Linguistic Service's technicians such that at least two evaluators examined each segment of text. The results were then collected and analyzed.

Data Analysis

In this paper, we only show the results concerning what the evaluators regarded as translation errors caused by the input. The errors caused by malfunctioning of the MT system are not presented here.

The analysis of the data was carried out by systematizing the classification and comments by evaluators regarding the translation errors, and quantifying the results. Our aim was to determine, on the one hand, which characteristics of the text were attributable to the writer's desire to use language that differed from formal norms and, on the other hand, which characteristics were attributable to other factors, focusing on language contact, a significant aspect of the area under study. Another important aim was to quantify each type and subtype of phenomenon. We believe this approach to be innovative, as the CMC literature to date has often pointed out certain phenomena without quantifying their actual relevance. We think that certain phenomena that may have been overvalued because of their novelty actually have little effect, while, on the contrary, other important phenomena have been neglected.

Due to the size of the corpus studied, the background of the users, and the specific nature of their communication needs, the conclusions we have drawn from the analysis of the results cannot describe e-mail communication in general. However, as will be seen below, we can infer certain interesting points.

Classification Scheme

We have empirically classified the linguistic characteristics that cause translation errors into three broad areas: (1) **unintentional non-standard features,** (2) **intentional non-standard features** and (3) **terminology**.

The full classification scheme used for the empirical analysis of the corpus is summarized below. Following that, we describe each category and subcategory in turn.

1. Unintentional non-standard features
    - 1.1 Mistyping
    - 1.2 Deviations from prescriptive language norms
        - 1.2.1 Orthographic
            - 1.2.1.1 Accents
            - 1.2.1.2 Phoneme-grapheme confusion
            - 1.2.1.3 Composition and separation symbols
            - 1.2.1.4 Capitalization
            - 1.2.1.5 Errors in abbreviations and acronyms
        - 1.2.2 Lexical
            - 1.2.2.1 Barbarisms
            - 1.2.2.2 Recurrent mix-ups

- 1.2.2.3 Oral reproduction
- 1.2.2.4 Loan-word errors
- 1.2.3 Syntactic
- 1.2.4 Cohesion
  - 1.2.4.1 Verb tense errors
  - 1.2.4.2 Anaphoric errors
  - 1.2.4.3 Punctuation errors

2. Intentional non-standard features
   - 2.1 Language shift
     - 2.1.1 Lexical
       - 2.1.1.1 Expressive
       - 2.1.1.2 Terminological
     - 2.1.2 Phrasal
   - 2.2 New forms of textual expressivity (characteristic of the e-mail register)
     - 2.2.1 Orthographic
       - 2.2.1.1 Orthographic innovations
       - 2.2.1.2 Systematic lack of accentuation
     - 2.2.2 Lexical
       - 2.2.2.1 Internet user vocabulary
       - 2.2.2.2 Informal (oral-like) language
       - 2.2.2.3 Prosodic reproduction
       - 2.2.2.4 Shortenings
     - 2.2.3 Visual
     - 2.2.4 Pragmatic
     - 2.2.5 Simplified punctuation
     - 2.2.6 Simplified syntax
3. Terminology
   - 3.1 Domain terminology
   - 3.2 User community terminology

Main categories: Unintentional non-standard features (1), intentional non-standard features (2) and terminology (3)

Our first main category is **unintentional non-standard features** (1). These differ from **intentional non-standard features** (2) in that they are not deliberately chosen by the writer. In some cases, doubt arises as to whether a non-standard feature is intentional or not. These cases are considered in relation to the context of the e-mail as a whole. If the case appears to be embedded in a system of coherent *odd* features, it is classified not as unintentional but as intentional (2). For instance, with regard to accents, if only one or a few words in the e-mail lack the necessary accentuation, we classify it as unintentional. However, when the user does not use accents at all in the message, or where the lack of accentuation is consistent with a rationale, then we regard this as a kind of voluntary deviation, i.e., *systematic lack of accentuation* (2.2.1.2).

Accordingly, our second main category, **intentional non-standard features** (2), is characterized by deliberate choice. One main group of intentional deviations is that which, according to the literature, defines the e-mail register itself: **new forms of expressivity** (2.2)—oral patterns, shortenings, simplified punctuation or syntax, specific pragmatic resources, visual information, etc. The other main group is **language shift** (2.1), i.e., the use of words or constructions from other languages, even though well-known equivalents exist in the language in which the text is written. Both the categories and their subcategories are explained below.

Last, we classified as **terminology** (3) that vocabulary which is specific to either the domain of knowledge and communication (in this case, computer science) or the particular user community under consideration (in this case, UOC students). This differs from the vocabulary that can be considered part of the general register of Internet users, classified as (2.2.2.1) under *new forms of expressivity*.

Mistyping (1.1)

These are mainly deviations caused by neighboring key strikes (**Cstalonia* instead of *Catalonia*), extra strikes (**Caatalonia*), inverted strikes (**Catlaonia*), missing strikes in a word (**Ctalonia*), or connecting two words (**toCatalonia*). We also include here the mistyping of a symbol that is similar to one that the user intends to strike; a typical example is the use of accents instead of apostrophes.

Deviations from prescriptive language norms (1.2)

In this case, the deviation is caused by users not being aware of a rule or a norm of the language they are using. This occurs at different linguistic levels, as described below. In Catalonia, a number of these deviations can be seen to be caused by language interference, although it is difficult to say exactly how many, since it depends greatly on the social and educational backgrounds of each individual. We return to this point later.

Orthographic deviations (1.2.1)

We have found different types of orthographic deviations, from erroneous capitalization (i.e., asystematic non-capitalization) to errors in writing acronyms and abbreviations or in the use of certain characters (e.g., apostrophes and hyphens in Catalan to affix clitics, as in **dona**'me-l*** for *dona**-me'l***, "give + to me + it"). However, the most common orthographic deviation comes from accentuation and phoneme-grapheme confusions (****a****ndavant; *adre****ç****es; *tro****v****ar* instead of ***e****ndavant, adre****c****es, tro****b****ar*), typically when one phoneme can be spelled by many graphemes. For instance, both 'a' and 'e' can represent the schwa sound in Catalan, and writers sometimes choose the wrong letter, as in the case of ****a****ndavant* (forward), which should be written ***e****ndavant.* Similarly, 's,' 'c' and 'ç' can all represent the /s/ sound, and writers sometimes choose the wrong option, e.g., **adre****ç****es* should be *adre****c****es* (addresses).

This happens in our corpus in four cases: (i) a/e and o/u alternation to represent the schwa sound and the unstressed /u/ sound; (ii) c/s/ç to spell /s/; (iii) b/v for /b/; and (iv) confusion in the use of digraphs: s/ss to spell /s/, and l/l·l and n/nn. The digraphs 'l·l' and 'nn' represent a combination of two /l/ and two /n/ respectively, which are prescribed by spelling norms but are hardly ever pronounced in the oral language. For instance, speakers pronounce a single /l/ when saying 'pel·licula' (film).

Lexical deviations (1.2.2)

We have found four types of lexical units that differ from standard language use. The first are what in Catalonia are called *barbarisms* (1.2.2.1), words or lexical constructions that the speaker believes are genuinely Catalan but which in fact are Spanish. Examples of these are **insertar* (instead of *inserir*, "to insert") and **recent* (instead of *acabat de fer*, "fresh"). These are archetypal cases of interference between languages in contact. In Catalonia, they also occur in Spanish due to the influence of Catalan, as in **antes de nada* instead of *en primer lugar* ("first of all"). This particular example is caused by the speaker translating the lexical construction word for word from the Catalan equivalent *abans de res* (*abans* = *antes* = "before," *de* = *de* = "of," *res* = *nada* = "nothing").

The second type is lexical mix-ups (1.2.2.2) which are caused by similarity of form but difference of meaning, e.g., *si no/sinó* ("but"/"otherwise"), *per què/perquè* ("why"/"because"), *per/per a* ("for" or "by"/"in order to") in Catalan, and *a parte/aparte* ("in part"/"apart") in Spanish. These are very common in some people's writing. Some of these mix-ups may also be caused by language interference due to false analogies between similar forms in Spanish and Catalan.

The third type of lexical deviation is caused by attempts to reproduce oral forms in writing (1.2.2.3). There are different subtypes, but all of them are distinct from phoneme-grapheme confusion (1.2.1.2), in which one phoneme can be spelled by two or three alternative letters, thus constraining the deviation in terms of available options. The scope is wider in the case of oral spelling, in as much as it might affect several phonemes/graphemes, or the whole word, thus changing the overall form of the lexical unit; for this reason it has been classified as 'lexical.' Typical cases are *vols* or *a veure*, Catalan words which some speakers pronounce /bos/ and /abere/, so that those speakers sometimes mistakenly write them as **vos* and **avere*. Another case is the pronunciation of *donés* /dunes/ (a subjunctive form of "to give") with an epenthetic velar consonant, /dunges/, thus leading the word to be written as **dongués*. An example in Spanish is **osea* instead of *o sea*; in this case the oral reproduction consists of converting two words into one, thus reproducing the seeming lack of spacing between words in continuous speech. Deviations caused by oral reproduction are dialect dependant in as much as pronunciation in different dialects resembles to a greater or lesser extent the standard in Catalan or Spanish.

Last, we also found cases of errors in the spelling of loan words (1.2.2.4), e.g., mistakenly writing the English word *cookies* as **cookis*, or *Access* (the database software) as **Acces*.

Syntactic deviations (1.2.3)

This category covers the non-prescriptive use of grammatical categories (e.g. the wrong choice of verbal mood, as in the use of infinitive instead of imperative in **decirme* instead of *decidme* ["tell me"] in Spanish) and other cases of syntactic ill-formedness. Relevant cases for the latter are the omission or addition of pronouns, prepositions or other function words, as in Catalan's **jo vull* ("I + want") instead of *jo* ***en*** *vull* ("I" + direct object pronoun + "want") to mean "I want (that thing)" or, in Spanish, **pienso* ***de*** *ir* for *pienso ir* ("I think I'll go"); and also typical cases of lack of agreement (subject-verb, determiner-noun).

Although it is difficult to systematize due to the sparseness of this type of data in the corpus, it is clear that at least some syntactic deviations are caused by language interference, as in the first two examples above, where (i) the incorrect omission of the pronoun *en* in Catalan reflects Spanish norms; and (ii) the incorrect addition of the preposition *de* in Spanish reflects Catalan norms.

Cohesion deviations (1.2.4)

Textual cohesion is affected in our corpus by inappropriate use of punctuation marks (colons, semicolons, hyphens, etc.), incorrect choices of verbal tenses to express temporal relations, and lack of concordance between pronouns and their antecedents.

Language shift (2.1)

As mentioned above, intentional deviations from language standards have been classified into two main categories, the first being language shift, i.e., the voluntary use of words or phrases from other languages. In Catalonia, where there is close contact between Catalan and Spanish, language shift is very common in informal speech since all speakers have a degree of knowledge of both languages. A consequence is that sometimes, when speaking in language A, a lexical choice corresponding to language B comes naturally to the speaker's mind. Since the language shift does not usually affect communication, in as much as the interlocutor is also bilingual, the speaker uses the other language's word, or even sometimes a phrase, not by mistake, but for the sake of fluency or other expressive reasons. For instance, it is typical to swear in Catalan using Spanish *jo* or *joder* ("fuck") or to say goodbye in Spanish using Catalan *adéu*.

This also happens to Catalan and Spanish speakers with third languages, especially English. Sometimes people say goodbye by using Italian *ciao*, express gratitude by using French *merci*, or ask for aid with English *help*.

Not every intentional use of language shift is expressive: Many shifts involve terminology which has either been learned or is better established in another language. A typical example is the use of English *software* instead of Catalan *programari*. This case is debatable, in as much as some speakers might simply be unaware of the existence of the Catalan term. However, we have classified these cases as intentional since we assume that our users either know the terminology of their field in their language (but that they still prefer using English terms), or else are aware that there must exist a word in their language for the term (but they do not wish to stop to think about it or look it up in a dictionary when they write e-mails). Finally, those foreign terms that lack a well-known equivalent in Spanish or Catalan have been classified as domain terminology (3.1).

Two characteristics should be highlighted about such language shifts in e-mails: (i) they are related to the written reproduction of informal speech, and (ii) they are related to language interference.

New forms of textual expressivity (2.2)

These are the features that, according to the literature, best define the e-mail register. We find here simple categories, such as visual resources (2.2.3)—typically, smileys; the pragmatic resource of dialogue simulation in *quoting* part of a previous message (2.2.4); and simplified punctuation (2.2.5) or syntax (2.2.6). We classified as simplified syntax cases involving the lack of a function word in intentionally telegraphic constructions, e.g., the lack of an article in *M'adreço a aquest fòrum amb l'esperança de trobar tècnic disposat a...* (instead of *...l'esperança de trobar **un** tècnic...*) ("...I'm hoping to find [a] technician..."). To distinguish between punctuation errors (1.2.4.3) and simplified punctuation, we counted as the latter any lack of (expected) punctuation marks in e-mails lacking any punctuation at all.

Similarly, for accentuation, we counted those e-mails which did not have any accents at all separately, so that any lack of an accent within them was classified as a case of systematic lack of accentuation (2.2.1.2). Otherwise, when occurring in e-mails that did have accents, lack of an accent was counted as an error (1.2.1.1).

Systematic lack of accentuation is one subtype of "new orthography." The other main class (2.2.1.1) includes a wide range of innovations such as capitalization or the use of a range symbols to show emphasis (*necessito ajuda URGENT...* "I need help URGENTLY;" *no funciona!!??!!* "it doesn't work!!??!!"), use of symbols as meaning components in words (*tod@s* covering both masculine and feminine genders instead of "*todos y todas*"), or the use of [*'s*] to pluralize acronyms, as in *CD's*.

The other main class under "new forms of expressivity" includes a variety of lexical units which are not found in formal texts (2.2.2). First, we have colloquial Internet user vocabulary (2.2.2.1) such as *online, hoax, nick, àlies* ("nickname") or *xat*

("chat"). These are usually English terms or adaptations from English. We have not classified English terms as language shifts or terminology as they clearly belong to an emerging Internet register more than to the specific domain of computer science.

The second subtype includes general-purpose informal vocabulary (2.2.2.2), typically used in speech but never in formal texts, e.g., *mates* ("maths") for *matemáticas* ("mathematics"), *profe* for *profesor* ("teacher, lecturer") or *yuyu* (a colloquial term in Catalan and Spanish for either feeling under the weather or unusual behavior).

Another class is that of intentional reproduction of spoken prosody (2.2.2.3) used as an expressive resource. For instance, *modessssno* contains a graphical reproduction of a very long [s] sound; this "word," which represents *moderno* ("modern/fashionable"), means something or someone pretending to be fashionable but who in fact seems ridiculous. We also include here reproduction of oral sounds such as *hmmm* (expressing doubt) or *psé* (indifference).

The last category is that of SMS message-like shortenings (2.2.2.4), such as *tb* instead of *també* ("as well") or *k* for *que* ("who, what, which...").

Terminology (3)

As expected, we found many examples in our corpus of terminology. This finding is crucial for machine translation, since these are words that are usually missing in the lexical databases of MT systems, and could therefore cause errors in translation. Most of the time, such terminology is associated with a specialized knowledge domain: in the case of our newsgroups, computer science. Thus we find words such as *XML*, *disc dur* ("hard disk"), and *script*. However, we also found terms particular to the community of users, UOC students. These include *PACs* (a kind of academic assignment) and *MIC* (an acronym for an academic subject, *Multimèdia i comunicació*, "Multimedia and communication").

Such terms cannot be considered characteristic of the CMC register: CMC features are expected to be found in any kind of newsgroup; however, in a newsgroup devoted to medicine, for example, we will find medical terms instead of terms for computer science. Furthermore, in a newsgroup devoted to computer science in another kind of community, e.g., professionals instead of students, we would not expect to find student vocabulary such as *MIC* or *PAC*.

Quantification of Features that Cause Translation Problems

Having classified the errors and deviations from standard language use found in the corpus, we present the quantitative results of the classification in Table 4. The columns labeled AF (absolute frequency) show the total number of occurrences of each category in the corpus. RF (relative frequency) shows the number of occurrences of each category per thousand words in the corpus. IT (impact on

translation) indicates the high (H), medium (M) or low (L) expected impact of the category on the quality of translation, independent of the number of occurrences. The IT ratings were produced according to the level of translation error caused by each category. This level was established by the evaluators during the evaluation process, when they had to grade the translation error according to Green's Rating Scale.

	CATALAN		SPANISH		IT
	AF	RF	AF	RF	
1. Unintentional non-standard features	**512**	**46.7**	**322**	**30.7**	
1.1 Mistyping	92	8.4	55	5.2	H
1.2 Deviations from prescriptive language norms	420	38.3	267	25.4	
1.2.1 Orthographic	296	27.0	169	16.1	
1.2.1.1 Accents	233	21.2	149	14.2	H
1.2.1.2 Phoneme-grapheme confusion	49	4.5	2	0.2	H
1.2.1.3 Composition and separation symbols	3	0.3	0	0.0	H
1.2.1.4 Capitalization	9	0.8	7	0.7	L
1.2.1.5 Errors in abbreviations and acronyms	2	0.2	11	1.0	L
1.2.2 Lexical	54	4.9	19	1.8	
1.2.2.1 Barbarisms	17	1.5	8	0.7	H
1.2.2.2 Recurrent mix-ups	5	0.4	4	0.4	H
1.2.2.3 Oral reproduction	29	2.6	7	0.7	H
1.2.2.4 Loan-word errors	3	0.3	0	0.0	M
1.2.3 Syntactic	36	3.3	48	4.6	H
1.2.4 Cohesion	34	3.1	31	2.9	
1.2.4.1 Verb tense errors	8	0.7	3	0.3	M
1.2.4.2 Anaphoric errors	1	0.1	9	0.8	H
1.2.4.3 Punctuation errors	25	2.3	19	1.8	H
2. Intentional non-standard features	**155**	**14.1**	**346**	**32.9**	
2.1 Language shift	24	2.2	46	4.4	
2.1.1 Lexical	24	2.2	45	4.3	

2.1.1.1 Expressive	5	0.4	4	0.4	M
2.1.1.2 Terminological	19	1.7	41	3.9	L
2.1.2 Phrasal	0	0.0	1	0.1	M
2.2 New forms of textual expressivity	131	11.9	300	28.6	
2.2.1 Orthographic	71	6.5	250	23.8	
2.2.1.1 Orthographic innovations	53	4.8	86	8.2	M
2.2.1.2 Systematic lack of accentuation	18	1.6	164	15.6	H
2.2.2 Lexical	36	3.3	39	3.7	
2.2.2.1 Internet user vocabulary	8	0.7	18	1.7	L
2.2.2.2 Informal (oral-like) language	9	0.8	8	0.8	M
2.2.2.3 Prosodic reproduction	6	0.5	5	0.5	H
2.2.2.4 Shortenings	13	1.2	8	0.7	M
2.2.3 Visual	9	0.8	3	0.3	L
2.2.4 Pragmatic	2	0.2	3	0.3	L
2.2.5 Simplified punctuation	2	0.2	0	0.0	H
2.2.6 Simplified syntax	11	1.0	5	0.5	H
3. Terminology	**396**	**36.1**	**437**	**41.6**	
3.1 Domain terminology	268	24.4	293	27.0	L
3.2 User community terminology	128	11.7	144	13.7	M
TOTAL	**1063**	**96.8**	**1105**	**105.2**	

Table 4: Quantitative results of features that caused translation errors

Discussion

It appears that the e-mails in the sample are not only characterized by new forms of expressivity, as is often claimed in English CMC research, but also by at least as many unintentional infelicities, mistypings and deviations from prescriptive norms. Most of these deviations seem to be due to a weak awareness of the language, especially Catalan, as the data show that the number of deviations from prescriptive norms is noticeably higher in this language than in Spanish. We have analyzed only a specific group of highly-educated users, university students. It is likely that among less well educated users, the ratio of unintentional felicities would be higher.

Analogy with Spanish sheds light on a number of deviations from the norms in Catalan, such as accentuation of the common ending *–ia* (e.g., **enginyería* instead

of *enginyeria*, “engineering”) or phoneme-grapheme confusions such as the failure to use ‘ss’ to represent the phoneme /s/, as in **a**s**ociació* instead of *a**ss**ociació*, “association.” The interference is clear in such cases, since norms in Spanish demand both accentuation of ‘–ía’ and use of ‘s’ instead of ‘ss.' Language interference is also evident in the lexicon (barbarisms) and explains certain recurrent confusions. In Spanish, language interferences are mainly cases of analogy in accentuation as well (Catalan: *ex**à**mens* (“exams”)–Spanish: **ex**á**men* instead of *ex**a**men*). However, the impact of these spelling confusions is not as great as it is in Catalan. The number of barbarisms in Spanish e-mails is also lower, and in recurrent mix-ups the difference disappears.

Many types of deviations cannot be explained in terms of interference. However, language interference is the single most important influence on the deviations found in the corpus. We would hypothesize that each of the following types of deviation was caused mostly or entirely by linguistic interference: incorrect accentuation, grapheme-phoneme confusions, barbarisms, recurrent mix-ups, syntactic errors and language-shift deviations. This hypothesis is supported by the comments of the language experts who evaluated the data, all of whom registered statements to this effect. Adding up these categories results in an estimate that 49.1% of the errors in Catalan and 31.3% of the errors in Spanish were caused by language interference. If we simply focus on deviations from norms (not counting language shifts), we find that as many as 59.4% of deviations from prescriptive norms in Catalan and 50.6% in Spanish were plausibly caused by interference. The high incidence of interference-induced errors and deviations is no doubt due to the fact that the users in our sample must deal with language contact on a daily basis. This situation represents a unique challenge for the application of MT to e-mail communication in Catalonia.

On the other hand, the ratio of intentional language-shifts is not very significant (only 2.2 per thousand words in the corpus). Therefore, it appears that the main influence of linguistic interference in spontaneous texts is that it causes deviations from prescriptive norms.

Adaptation to intentional deviations through customization of the MT system in both directions would be worthwhile as well. Among new forms of expressivity, the feature which best characterizes the register is *new orthography.* New orthography is much more noticeable in Spanish (83.3% of intentional features) than in Catalan (54.1%). Another interesting point is that there seems to be greater impact in terms of intentionality and the creative use of language in Spanish than in Catalan. In Spanish, fully intentional non-accentuation is more evident. Likewise, there are more orthographic innovations in Spanish than in Catalan.

Lexical forms of new expressivity, considered together, have some effect in characterizing the register, although their relative frequency in terms of the corpus as a whole is low (3.3 for Catalan and 3.7 for Spanish). In contrast, visual and pragmatic resources, simplified syntax and simplified punctuation, despite having

been paid a great deal of attention in CMC literature (Herring, 1999; Murray, 2000), appear to be scarcely significant in either direction.

Another aspect often characterized as significant is oral patterns. However, the impact of the features related to this in our study did not reach expected levels. The features selected as oral patterns were as follows: barbarisms, oral reproduction, language shift, informal oral-like language and prosodic reproduction. It is not completely clear whether all of the barbarisms or language shift reproduce oral behavior, but they are included here as they refer to vocabulary that is used when speaking, but not usually when writing a formal text. Counting all of these features as oral patterns, they represent 12.7%, in Catalan, and 11.0%, in Spanish, of all of the non-standard features. If unintentional features are set aside and we concentrate on intentional aspects, oral patterns are, in Catalan, 25.1%, and, in Spanish, 17.0% of all of the intentional deviations. Compared to the total number of words in the corpus, oral patterns are 7.7 per thousand words in Catalan, and 7.0 per thousand words in Spanish. Therefore, it seems that in our sample e-mails are to a large extent textual or written in nature, with little evidence of oral patterns. However, oral features deserve special attention due to their negative impact on translation quality.

In any case, the results indicate that successful application of MT to the online translation of e-mails in our environment would require customizing the MT system, taking into account the following:

- For Catalan, the main efforts would have to focus on automatic correction of unintentional language norm deviations, mainly mistyping, orthography and mistakes caused by language contact. The results show that such unintentional deviations represent more than three times the amount of intentional non-standard features.
- The situation for Spanish is more balanced, which means that, in terms of frequency, efforts would have to be focused on feeding the MT system with terminology as its ratio is slightly higher than those of unintentional and intentional non-standard features.

For both languages, both unintentional deviations and terminology as well as adaptation to intentional deviations are issues worthy enough to require customizing the system. As for terminology, both domain terminology and user community terminology have to be dealt with, but the impact on the translation is much greater in the case of user community terminology, as most of the domain terminology is in English and does not usually need to be translated to be understood. The impact on translation of Internet users' vocabulary, which can be considered a special kind of terminology, is not significant, because the terms are widespread and commonly understood (note that they are usually English terms). Apart from adaptation to the terminology, the main sources of problems for translating e-mails in both languages are orthography and the lexicon. Errors and deviations in syntax and pragmatics are scarcely significant. In terms of

orthography, the most common problem is accentuation. 37.6%, in Catalan, and 46.8%, in Spanish, of all errors and deviations involve accentuation, regardless of the fact that in Spanish this seems, for the most part, to be intentional (systematic lack of accentuation) and in Catalan, to be unintentional mistakes.

Conclusions and Future Directions

In this paper, we have presented an in-depth linguistic evaluation of a corpus of about 260 e-mails and 25,000 words, written in Catalan and Spanish, which came from four informal computer science newsgroups at the Universitat Oberta de Catalunya Virtual Campus. The messages were produced within a situation of bilingualism and language contact, where Spanish is progressively substituting for Catalan as the language of daily use.

The main goal of the study was to identify the linguistic characteristics of the e-mail register for our universe of study in order to assess their impact on machine translation, and based on our findings, to take decisions for improving the quality of translation. So, the e-mails of our corpus were translated by an MT-system and, while we evaluated the translations in order to state an improvement strategy, we classified and analyzed the features of the originals that caused problems. The study is part of the INTERLINGUA project, which aims to adapt a system for online unsupervised translation of e-mails from Catalan to Spanish and vice versa, to avoid the marginalization of Catalan as a language for communication on the Internet. The main conclusions we can draw are the following:

• For our sample, the e-mail register is characterized, on the one hand, by unintentional mistyping and deviations from prescriptive language norms and, on the other, by the intentional use of features usually considered typical of e-mails. In Spanish, the two kinds of features are balanced, while in Catalan, unintentional non-standard features outnumber their intentional counterparts by three to one.

• One of the main reasons for non-standard input that results in translation errors in the sample is the interference of one language with the other. Language interference can account for up to half of the errors; however, some of these may be due to a lack of awareness of prescriptive norms. Interference affects Catalan more than Spanish, thus confirming the marginalization of the former with respect to the latter. The analysis has shown that, despite educational efforts made over recent decades, there still exist gaps in many people's knowledge of Catalan.

• The intentional non-standard feature that best defines the e-mail register is non-standard orthography, in particular, orthographical innovations in Catalan and systematic lack of accentuation in Spanish. The use of visual information, new pragmatic resources, simplified syntax and simplified punctuation is not significant in quantitative terms.

• From an MT perspective, the extremely high ratio of spelling mistakes, barbarisms, etc. in e-mails severely threatens the feasibility of online automatic translation, given that MT systems are not currently prepared to deal with noisy input. The MT community has not addressed such a challenge as yet. Therefore, we believe that this study and the project that it forms a part of represent an important innovation in both the fields of MT and CMC.

An important implication is that the MT system must be fine-tuned in order to build software modules for the automatic correction of input, and of accents in particular. To the extent that the feasibility of incorporating minority languages such as Catalan into the multilingual Internet will depend on natural language processing, which usually deals with standard texts, difficulties might be expected for these languages in the future. But two important questions arise: Should MT systems bear the responsibility for correcting input in order to preserve users' spontaneity when writing, or should users be more careful in their use of language? In the latter case, might the effort to write accurately help increase the status of minority languages in CMC, which, ironically, is a medium that encourages non-standard writing and expressive innovation? In the environment studied, we cannot expect users to always write carefully, taking the effort of looking up words in the dictionary, using spell and grammar checkers, etc. Effort is against the trend in CMC for relaxation. So if users feel constrained to write accurately, the results will be the reverse of what we expect. Hence, we think that, nowadays, the tuning of MT systems is essential for the survival of Catalan in newsgroups, chats and the so.

In the future, our efforts will be extended to purely monolingual environments in particular to Spanish in non-bilingual territories, e.g. Madrid. This will allow us to assess more precisely the effects of language interference when bilingual users write e-mails, i.e. by comparing the Spanish written by Catalans to the Spanish written by Madrilians. In this connection, it should be pointed out that the ratio of errors is expected to be lower in monolingual environments; more precisely, unintentional non-standard features should play a less important role in the e-mail register for communities of users that do not have to deal with a situation of language contact. At the same time, communities that are not made up of university students or professionals (whether bilingual or monolingual) are expected to have a lower level of linguistic knowledge, which should lead to a higher ratio of errors. In addition, it might be hypothesized that e-mails written by other communities of Internet users, e.g., teenagers, would show greater use of intentional new forms of expressivity. Thus, the impact of expressivity on the characterization of the register would increase. Moreover, as mentioned above, teenagers would also be expected to make more errors due to a lack of awareness of the language, in a way that would confirm or even reinforce the tendencies shown in the present study.

The future of Catalan depends on its users but, in the multilingual internet, it also depends on the help of technological tools such as MT systems. We have seen that the gaps in the knowledge of the language are a very serious handicap in the

usability of these systems. We have also realized that these gaps are significantly present in one of the principal groups of Internet users, university students, and it is foreseeable that it will also be significant in the other principal group, teenagers. It is also quite obvious that email is a medium that encourages linguistic innovations and creativity and that this is a trend which is not likely to change in the future. Consequently, we would encourage MT developers to tune their systems to make them error-proof and also flexible enough to deal with the non-standard features typical of the e-mail register. Otherwise, that is, if MT continues to be only focused on controlled, standard, well written texts, with no deviations from the linguistic norms, Catalan is likely to disappear from collective CMC in few years.

Acknowledgements

We are grateful to Susan Herring, Brenda Danet and the anonymous reviewers of this paper for their useful critical comments and observations. Needless to say, all remaining errors are our own.

The research presented in this paper has been partially funded by the project IR-226 (Internet Interdisciplinary Institute, IN3/UOC: project *INTERLINGUA*).

Notes

[1] Catalonia is an "autonomous community" within Spain that holds self-governing powers guaranteed by the Spanish Constitution and the Catalan Statute of Autonomy. Although its actual political status is largely controversial, it can be seen as a kind of a “stateless nation” or “a nation within a nation”. Moreover, some people use the word ‘Catalonia’ to refer to the three Spanish autonomous communities (Catalonia, Balearic Islands, Valencia) and other territories in the South of France whose native language is Catalan. For simplicity’s sake, in this paper, we will use the term ‘Catalonia’ to refer to the strict autonomous community as defined and bounded by the Spanish Constitution (i.e. the region in northeast Spain whose capital is Barcelona); and we will call it “a country”.

[2] The project is expected to be extended in the future to Catalan/English and Spanish/English.

[3] For further information on the sociolinguistic situation in Catalonia, see Pujolar (2000) and Strubell and Hall (1992).

[4] We term it UOC-Catalonia, as the institution has recently opened a line of studies in Spanish for the rest of Spain and Latin America.

[5] Please notice that this corpus is larger than the one used for the second phase of this study: the linguistic evaluation. The corpus used in the second phase is a subset of that used in the first one.

[6] See Danet (2001, ch.1) for a more comprehensive and in-depth review of recent literature on this topic.

[7] Nevertheless, it may be discussed whether they are exclusive to e-mail or they may also belong to other forms of text-based CMC

References

ÀLATAC. (2003). Website of the *Servei de Llengües i Terminologia de la Universitat Politècnica de Catalunya i l'Associació del Voluntariat Lingüístic*. Retrieved July 18, 2003 from: http://www.upc.es/slt/alatac/cat/dades/catala.html

Alonso, A., Folguerà, R., & Tebé, C. (2000). Del tecnolecte al sociolecte: consideracions sobre l'argot tècnic en català. In M. Torres, Ll. Jardí, N. Alturo, Ll. Payrató & F.X. Vila (Eds.) *CMO-Cat I Jornada sobre Comunicació Mediatitzada per Ordinador en Català.* December 1, 2000 Barcelona, Spain: Proceedings. Retrieved July 18, 2003 from: http://www.ub.es/lincat/cmo-cat/tebe-alonso-folguera.htm

Badia, J., Bertran, C., & Castells, A., (2001). *És possible viure en català?* Barcelona: Angle.

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

Castells, M., & Díaz de la Isla, M. I. (2001). Diffusion and uses of Internet in Catalonia and Spain. *Project Internet Catalonia (PIC) Working Paper Series,* PICWP 1201. Barcelona, Spain: Universitat Oberta de Catalunya / IN3. Retrieved July 18, 2003 from: http://www.uoc.edu/in3/wp/picwp1201/

Cerdà, R. (2001). Castellano y catalán en Cataluña y las Islas Baleares. *II Congreso Internacional de la Lengua Española.* 2001 Valladolid, Spain: Proceedings. Retrieved July 18, 2003 from: http://cvc.cervantes.es/obref/congresos/valladolid

CIS: Centro de Investigaciones Sociológicas. (1998). Uso de lenguas en comunidades bilingües (II): Cataluña. In *Catálogo del banco de datos del Centro de Investigaciones Sociológicas, estudio 2298*, Madrid, Spain. Retrieved July 18, 2003 from: http://www.cis.es/estudio.asp?nest=2298

Climent, S., Moré, J., & Oliver, A. (2003). Building an environment for unsupervised automatic email translation. *EAMT-CLAW 2003, Joint Conference combining the 7th international workshop of the European Association for Machine Translation*

and the 4th Controlled Language Applications Workshop, Dublin: Proceedings. Retrieved July 18, 2003 from http://www.uoc.edu/in3/dt/20292/index.html

Danet, B. (2001). *Cyberpl@y: Communicating online*. Oxford: Berg.

EC: European Commission. (1998). *The Euromap Report.* Luxembourg: Linglink.

Fais, L., & Ogura K. (2001). Discourse issues in the translation of Japanese e-mail. *Conference of the Pacific Association for Computational Linguistics, PACLING 2001*, Kitakyushu, Japan: Proceedings. Retrieved March 27, 2003 from: http://afnlp.org/pacling2001/pdf/fais.pdf

Ferrara, K., Brunner, H., & Whitemore, G. (1990). Interactive written discourse as an emergent register. *Written Communication, 8*, 8-34.

Green, R. (1977). *Analysis of errors.* Commission of the European Communities (*CEC) memorandum*, October, 5+5 p. Luxembourg.

Herring, S. C. (1999). Interactional coherence in CMC. In T. Erickson (Ed.), *Journal of Computer-Mediated Communication,* 4 (4), special issue on *Persistent Conversation*. Retrieved March 27, 2003 from: http://www.ascusc.org/jcmc/vol4/issue4/herring.html

Herring, S. C. (2001). Computer-mediated discourse. In D. Tannen, D. Schiffin & H. Hamilton (Eds.), *Handbook of discourse analysis* (pp. 612-634). Oxford: Blackwell.

IDESCAT, Institut d'Estadística de Catalunya. (2001). Interactive Statistics Request performed at the IDESCAT web site: http://www.idescat.es. Retrieved March 27, 2003 from: http://www.idescat.es/scripts/sqldequavi.dll?TC=444&V0=8&V1=6

ISLE: International Standards for Language Engineering. (2000). *The ISLE classification of machine translation evaluations.* Retrieved March 27, 2003 from:http://www.isi.edu/natural-language/mteval

Mehsching, G. (2000). *The internet as a rescue tool of endangered languages: Sardinian*. Retrieved July 29, 2003 from: http://www.gaia.es/multilinguae/pdf/Guido.PDF

Murray, D. E. (2000). Protean communication: The language of computer-mediated communication. *TESOL Quarterly, 34 (3*), 397-421.

Payà, M. (2000). Com responem els missatges de correu electrònic? Noves formes de diàleg. In M. Torres, Ll. Jardí, N. Alturo, Ll. Payrató & F.X. Vila (Eds.), *CMO-Cat I Jornada sobre Comunicació Mediatitzada per Ordinador en Català.*

December 1, 2000 Barcelona, Spain: Proceedings. Retrieved July 18, 2003 from: http://www.ub.es/lincat/cmo-cat/paya.htm

Pujolar, J. (2000). *Gender, heteroglossia and power. A sociolinguistic study of youth culture.* Berlin: Mouton de Gruyter.

Siguán, M. (1999). *Conocimiento y uso de las lenguas*. Madrid, Spain: *Centro de Investigacions Sociológicas, CIS, 22.*

Siguán, M. (2001). El Español como lengua en contacto con otras lenguas. *II Congreso Internacional de la Lengua Española.* 2001 Valladolid, Spain: Proceedings. Retrieved March 27, 2003 from: http://cvc.cervantes.es/obref/congresos/valladolid

Strubell, M., & Hall, J. (1992). Problems and prospects of small linguistic societies – Catalonia. *EMI Education Media International, (29)* 1, 26-37.

Turell, Mª T. (Ed.) (2001). *Multilingualism in Spain: Sociolinguistic and psycholinguistic aspects of linguistic minority groups.* Clevedon: Multilingual Matters.

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.* Retrieved March 27, 2003 from: http://www.ling.ed.ac.uk/~beatrice/bibliography.htm

Warschauer, M. (2000). Language, identity, and the Internet. In B. Kolko, L. Nakamura, & G. Rodman (Eds.), *Race in Cyberspace* . New York: Routledge. Also retrievable from: http://www.gse.uci.edu/markw/lang.htm (accessed July 29, 2003)

Yates, J. A., & Orlikowski, W. J. (1993). Knee-jerk anti-LOOPism and other e-mail phenomena: Oral, written, and electronic patterns in computer-mediated communication. *MIT Sloan School Working Paper 3578-93, Center for Coordination Science Technical Report 150.* Retrieved March 27. 2003 from http://ccs.mit.edu/papers/CCSWP150.html