Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 45-50 recibido 29-04-2005; aceptado 01-06-2005

Análisis de los fenómenos lingüísticos de los mensajes de correo electrónico en catalán desde la perspectiva de la traducción automática

Joaquim Moré, Salvador Climent, Antoni Oliver
Universitat Oberta de Catalunya
Avda Tibidabo 39, 08035 Barcelona
{jmore, scliment, aoliverg}@uoc.edu

Mariona Taulé
Universitat de Barcelona
Gran Via de les Corts Catalanes
585, 08007 Barcelona
mtaule@ub.edu

Resumen: Los sistemas de traducción automática están preparados para traducir textos que son normativamente correctos. Sin embargo, en la traducción automática de mensajes de correo electrónico hay elementos ajenos a la norma que provocan errores de traducción y es necesario conocerlos si se quiere optimizar un entorno de traducción automática de mensajes de correo electrónico, como el que se diseñó para el proyecto Interlingua desarrollado por la UOC. Una tarea importante de dicho proyecto fue el análisis de los fenómenos lingüísticos no normativos de un corpus de mensajes electrónicos escritos en catalán y su impacto en la calidad de traducción al español. En este artículo presentamos este análisis. Curiosamente los errores de competencia de los emisores causan más errores de traducción que los fenómenos característicos de la comunicación por correo electrónico, como los errores de teclado, vocabulario sms, emoticonos, etc.
Palabras clave: traducción automática, mensajes de correo electrónico, causas de errores de traducción

Abstract: Emails contain linguistic phenomena that deviate from standard language norms and may cause machine translation errors. In order to design an email translation environment for the Interlingua project developed at UOC, we classified deviations from the standard in a corpus of emails in Catalan and analysed their impact on the machine translation quality in Spanish. Here we present this analysis. Curiously, most translation errors are caused by the lack of linguistic competence of the sender. The impact of characteristic email phenomena (smileys, performance errors, sms vocabulary, etc.) is not so strong.
Keywords: machine translation, emails, translation errors

1 Introducción

El funcionamiento correcto de los sistemas de traducción automática (TA) actuales depende en gran parte de la calidad del original. Si éste no tiene un vocabulario bien establecido, sus oraciones no son simples, su estilo no es el estándar y existen errores o expresiones que se apartan de la gramática normativa, la probabilidad de que obtengamos una traducción incomprensible es muy alta. Por tanto, antes de traducir automáticamente un texto éste tendría que ser corregido manualmente con lo cual podría parecer que la intervención de la TA en la comunicación multilingüe online sin supervisión previa de los originales no puede dar resultados satisfactorios. Sin embargo, la Universitat Oberta de Catalunya (UOC) desarrolló el proyecto Interlingua[1] cuyo objetivo era desarrollar un entorno de traducción automática de mensajes de correo electrónico entre estudiantes catalanohablantes y castellanohablantes dentro y fuera de Cataluña (resto del Estado Español y Latinoamérica), lo que les permitiría expresarse con naturalidad en su lengua materna independientemente del interlocutor. En este entorno de aplicación, el objetivo era traducir

[1] www.uoc.edu/in3/interlingua

los mensajes con la máxima calidad posible a pesar de que no pudiera existir ningún proceso de supervisión manual humana de los mensajes originales. Por eso elaboramos una estrategia de supervisión automática con módulos de pre-edición que tenían que corregir los mensajes antes de ser traducidos por el sistema. Para desarrollar estos módulos analizamos los fenómenos lingüísticos no normativos presentes en los mensajes de correo electrónico en las direcciones catalán-español y español-catalán y distinguimos los que provocan más errores de traducción. En este artículo presentamos el análisis cuantificado de los fenómenos relativos a la dirección catalán-español.

2 Detalles del análisis

El sistema de TA utilizado por la UOC es el sistema Comprendium[2] en las direcciones español-catalán, catalán-español. El sistema segmenta el texto en unidades que corresponden más o menos a una oración y traduce segmento a segmento. Preparamos un corpus de 1239 segmentos tomados de 129 mensajes de correo electrónico (12.400 palabras aproximadamente) del Foro de Informática de la UOC. El Foro de Informática es un espacio público del Campus Virtual de la UOC donde los estudiantes piden ayuda a sus compañeros para resolver problemas técnicos de sus equipos, exponen y resuelven dudas relacionadas con sus estudios, anuncian eventos de interés, etc.

Paralelizamos cada segmento original con su correspondiente traducción para formar el corpus de análisis. Cuatro lingüistas del Servei Lingüístic de la UOC- departamento encargado de la corrección lingüística y estilística de los documentos generados por la institución- identificaron los fenómenos lingüísticos no normativos y detectaron los causantes de errores de traducción. Además de la detección, evaluaron los segmentos traducidos con una escala de inteligibilidad y una escala de fidelidad al original cuyos valores extremos eran, para la inteligibilidad, "completamente inteligible" y "completamente ininteligible" y para la fidelidad "totalmente fiel al original" y "totalmente infiel al original". La escala de inteligibilidad también recogía valores intermedios entre la inteligibilidad y la ininteligibilidad completas. Al menos dos evaluadores examinaron cada segmento original con su traducción. Los resultados se recogieron, se compararon para corregir discrepancias entre evaluadores, y se clasificaron según la tipología que presentamos a continuación.

3 Descripción de la clasificación

Los datos obtenidos del análisis están organizados en tres grandes grupos (ver Tabla 1). El primer grupo es el de los errores presentes en los originales que no han sido cometidos voluntariamente por el emisor (errores no-intencionales). El segundo grupo lo constituyen los fenómenos lingüísticos que manifiestan la voluntad del usuario por transgredir o alejarse de la norma, y también un uso creativo del lenguaje. Finalmente, el tercer grupo lo componen los errores terminológicos. La distinción especial de los fenómenos voluntarios causantes de errores se inspiró en la taxonomía para la evaluación de sistemas de TA del International Standards for Language Engineering (ISLE)[3] y también en la bibliografía sobre comunicación mediada por ordenador que suele considerar el uso intencionado de nuevas formas de expresión como una de las características principales de los mensajes de correo electrónico (Herring, 2001), (Yates y Orlikowski, 2003), (Folguerà y Tebé, 2002), (Fais y Ogura, 2001) y (Murria, 2000).

Errores no intencionales

Los errores no intencionales se subdividen en ***errores de actuación*** y ***errores de competencia***. Los errores de actuación son los errores cometidos por pulsar teclas vecinas (**Cstalunya*), pulsar teclas extra (**Caatalunya*) o pulsar teclas en orden inverso (**Catlaunya*). También consideramos como errores de actuación las teclas no pulsadas en una palabra (**Ctalunya*) o en un conjunto de palabras (**aCatalunya*) y el uso incorrecto de un símbolo por otro, como el uso de un acento en vez de un apóstrofe. Los errores de competencia son los errores cometidos involuntariamente por el usuario porque éste no es consciente de que está infligiendo una norma de la lengua que está usando. Entre los errores de competencia detectados distinguimos los errores ortográficos, los errores debidos al uso no normativo de unidades léxicas, los errores sintácticos y los errores de cohesión textual.

[2] www.comprendium.com

[3] http://www.issco.unige.ch/projects/isle/femti/

	CATALÁN		IT
	FA	**FR**	
1. errores no-intencionales	**512**	**46.66**	
1.1 errores de actuación	***92***	***8.38***	A
1.2 errores de competencia	***420***	***38.27***	
1.2.1 ortográficos	***296***	***26.97***	
1.2.1.1 acentos	233	21.23	A
1.2.1.2 confusión fonema-grafema	49	4.46	A
1.2.1.3 composición y separación de símbolos	3	0.27	A
1.2.1.4 capitalización	9	0.82	B
1.2.1.5 errores en abreviaturas y acrónimos	2	0.18	B
1.2.2 léxicos	***54***	***4.92***	
1.2.2.1 barbarismos	17	1.54	A
1.2.2.2 confusiones recurrentes	5	0.45	A
1.2.2.3 reproducción oral	29	2.64	A
1.2.2.4 adaptación de préstamos a la normativa	3	0.27	M
1.2.3 sintácticos	***36***	***3.28***	A
1.2.4 cohesión	***34***	***3.09***	
1.2.4.1 errores de tiempo verbal	8	0.72	M
1.2.4.2 errores anafóricos	1	0.09	A
1.2.4.3 errores de puntuación	25	2.27	A
2. desviaciones intencionales	**155**	**14.12**	
2.1 cambio de lengua	***24***	***2.18***	
2.1.1 léxico	***24***	***2.18***	
2.1.1.1 expresivo	5	0.45	M
2.1.1.2 terminológico	19	1.73	B
2.1.2 frasal	0	0.00	M
2.2 nuevas formas de expresividad textual	***131***	***11.93***	
2.2.1 ortográficas	***71***	***6.47***	
2.2.1.1 innovaciones ortográficas	53	4.83	M
2.2.1.2 falta de acentuación sistemática	18	1.64	A
2.2.2 léxicas	***36***	***3.28***	
2.2.2.1 vocabulario propio de usuarios de internet	8	0.72	B
2.2.2.2 registro informal (parecido al lenguaje oral)	9	0.82	M
2.2.2.3 reprodución prosódica	6	0.54	A
2.2.2.4 vocabulario sms	13	1.18	M
2.2.3 visuales	***9***	***0.82***	B
2.2.4 pragmáticas	***2***	***0.18***	B
2.2.5 puntuación simplificada	***2***	***0.18***	A
2.2.6 sintaxis simplificada	***11***	***1.00***	A
3. terminología	**396**	**36.08**	
3.1 terminología del dominio	***268***	***24.42***	B
3.2 terminología de la comunidad	***128***	***11.66***	M

Tabla 1: Cuantificación de errores no intencionales y desviaciones intencionales y su impacto en la calidad de la traducción

Los errores ortográficos más destacables son los errores de acentuación como la falta de acentuación, las equivocaciones en el uso del acento abierto y cerrado, y la acentuación por influencia del español (****cóm*** en oraciones interrogativas directas e indirectas). Tras la acentuación le sigue la confusión grafema-morfema como el uso incorrecto de *a* o *e* para representar el fonema /ə/, los usos incorrectos de *o* o *u* para el fonema /u/, de *b* o *v* para el fonema /b/, o los usos *s, c* o *ç* para el fonema /s/ (**andavant*, **adreçes*, **trovar*, **dunar* en vez de *endavant*, *adreces, trobar i donar* que son las formas correctas). También hay que destacar la combinación incorrecta de clíticos (**tant se m'en dóna* en vez de *tant se me'n dóna*), los errores en el uso de mayúsculas y minúsculas (**ametlla del vallès* por *Ametlla del Vallès*) y en la ortotipografía de los acrónimos (**sos* por *S.O.S.*) que provocan que el sistema traduzca los nombres propios y los acrónimos como nombres comunes (**almendra del vallès, *sus*).

Por lo que respecta a los errores léxicos destacamos el uso de barbarismos (**insertar* por *inserir*, **recent* por *acabat de fer*). También existen casos de confusiones recurrentes en el uso de palabras similares en la forma pero distintas en el significado (**sinó /si no, *perquè/ per què, *al que fos/el que fos, *s'hem queda curta/se'm queda curta*). Otra clase importante de error léxico es el causado por la reproducción escrita de la pronunciación de la palabra. Casos típicos son **vos* que reproduce una variante dialectal de la pronunciación de *vols* (quieres), *avere* que reproduce la pronunciación de *a veure* (*veamos*), **dongués* que contiene la consonante velar epéntica y que es una variante dialectal de la pronunciación de *donés*. Finalmente encontramos algunos casos de errores producidos por la adaptación a la normativa catalana de préstamos de otras lenguas (**cookis* por *cookies*, **Acces* por *Access*).

Los errores sintácticos son generalmente usos incorrectos u omisiones de preposiciones y pronombres (**abonat les lleis de Murphy/ abonat a les lleis de Murphy*). Algunas de las omisiones se deben a la influencia del español (**jo vull* en vez de *jo en vull* (quiero (de eso)). Otros errores sintácticos son los errores de concordancia sujeto-verbo, determinante-nombre y la elección de un modo verbal inapropiado.

Por último destacamos los errores de cohesión textual, que son: uso incorrecto de signos de puntuación, uso incoherente de tiempos verbales para expresar relaciones temporales y falta de concordancia entre los pronombres y sus antecedentes.

Desviaciones intencionales

De los datos que revelan una intención comunicativa voluntaria del usuario y que se desvían del uso normativo del catalán, destacamos el uso de palabras o expresiones en otras lenguas. Esto es usual en la comunicación informal: al emisor le viene a la mente una expresión en inglés, pongamos por caso, y la usa porque le parece más expresiva y, evidentemente, cree que el receptor será capaz de apreciarlo. Por ejemplo, hemos visto cómo los estudiantes maldicen con un *joder*, agradecen con un *merci*, se despiden con un *ciao* y piden ayuda escribiendo *help*. Sin embargo, el uso de palabras o expresiones en otras lenguas no siempre es expresivo. A menudo se utilizan términos que los estudiantes han aprendido o que son de uso común en inglés u otra lengua a pesar de que exista una forma normativa en catalán. Un ejemplo típico es el uso de *software* en vez del término normativo catalán *programari*. Uno podría creer que los comunicantes simplemente ignoran el término normativo pero creemos que en realidad se trata de un uso intencionado. En el caso de que los estudiantes no sepan el término catalán son conscientes de que éste debe existir pero no quieren interrumpir la escritura del mensaje para consultar el diccionario o una base de datos terminológica en línea. En cambio, los términos que no tienen equivalente en catalán los hemos clasificado como terminología del dominio.

El segundo fenómeno destacado en las desviaciones intencionales son las nuevas formas de expresividad textual que, como hemos dicho, caracterizan el registro de los mensajes de correo electrónico según la bibliografía. Hemos encontrado recursos visuales (típicamente emoticonos), aunque también hay nuevas formas de expresividad que se manifiestan en la innovación ortográfica y en el léxico. Entre los casos de innovación ortográfica están el uso de mayúsculas para enfatizar (*necessito ajuda URGENT*), las formas con arroba como tod@s, el uso de *'s* para pluralizar los acrónimos (*CD's*) y acrónimos como A10 que se lee com *adéu* en catalán (adiós).

La ausencia total de acentos ortográficos en los mensajes no la hemos clasificado como un error sino como una nueva forma de expresividad. Creemos que así el emisor expresa su voluntad de mantener un tono informal y desenfadado, sin convencionalismos, con su receptor. Sin embargo, si en el mensaje aparecen acentos, la falta de acentuación de una palabra ha sido considerada como un error.

Por lo que respecta a las nuevas formas de expresividad léxica encontramos vocabulario de uso general entre internautas como *nick*, *àlies* o *xat*. No clasificamos este vocabulario como un cambio de lengua ni tampoco como unidades terminológicas ya que pertenecen a un registro de habla emergente más que a un dominio específico. Otro subtipo de formas expresivas léxicas lo forman las palabras que son de uso general en la comunicación informal pero que no aparecen nunca en textos formales (*mates*, *profe* o *yuyu*). Un tercer subtipo es el de las palabras que reproducen un efecto prosódico como *modesssno*, *humm, psé*. Las formas SMS como *tb* (también), o *k* (que) son la última categoría de expresividad léxica.

Otras formas expresivas del registro de los mensajes electrónicos son la sintaxis y la puntuación simplificada. Consideramos como sintaxis simplificada la ausencia de palabras funcionales para crear intencionadamente un estilo telegráfico. En cuanto a la puntuación simplificada, la distinguimos de la puntuación incorrecta cuando la falta de puntuación afecta a todo el mensaje entero.

Terminología

En nuestro corpus de análisis hemos encontrado un buen número de vocabulario terminológico. Su detección en el ínput es crucial porque suelen ser palabras que no están en las bases de datos léxicas del sistema y, por esta razón, pueden provocar errores de traducción. La mayoría de la terminología está relacionada con la informática debido a que pertenecen al foro público de esta asignatura. (*XML*, *disc dur* ("disco duro"), *script*, etc.) pero hemos encontrado unos términos que deben ser diferenciados: aquellos que pertenecen a la comunidad de estudiantes de la UOC. Algunos de estos términos son *pla docent* (plan docente) y los acrónimos *PAC* (Prueba de Evaluación Continua) o *MIC* que es el acrónimo de la asignatura *Multimèdia i Comunicació* (Multimedia y Comunicación).

4 Cuantificación

Una vez clasificados los errores y las desviaciones intencionales que hemos encontrado en el corpus de mensajes de correo electrónico, cuantificamos los datos. El resultado de la cuantificación se muestra en la Tabla 1. FA (frecuencia absoluta) muestra el número total de ocurrencias de cada categoría del corpus. FR (frecuencia relativa) muestra el número de ocurrencias de cada categoría por cada mil palabras del corpus. IT (impacto en la traducción) indica el impacto alto (A), medio (M) o bajo (B) de cada categoría en la calidad de traducción, independientemente de la cantidad de ocurrencias. El cálculo del impacto se realiza a partir de los valores de inteligibilidad y fidelidad que los evaluadores asignaron a los segmentos traducidos. Los errores de impacto alto son los responsables de la ininteligibilidad de la traducción o bien son los que la convierten en totalmente infiel al original. Los de impacto medio no motivaron al evaluador a poner un valor negativo extremo de inteligibilidad y fidelidad, y los de impacto bajo apenas tienen una incidencia negativa en la inteligibilidad y fidelidad de la traducción.

5 Comentario de los resultados

A partir de los datos cuantificados arriba interpretamos lo siguiente:

Los mensajes analizados no se caracterizan principalmente por la presencia de desviaciones intencionales sino por los errores de competencia lingüística. Además, tanto los errores de actuación motivados por la rapidez de la comunicación como los errores terminológicos no superan los errores de competencia. Esto es sorprendente teniendo en cuenta que hemos analizado mensajes de estudiantes universitarios a los que se les supone un conocimiento correcto del catalán.

Los errores de competencia del usuario son los que provocan un impacto mayor en la calidad de la traducción. La mayoría de errores son errores ortográficos, sobretodo de acentuación, lo cual provoca que el motor de traducción no identifique la palabra mal escrita en su diccionario y la deje sin traducir o bien, si ésta coincide con otra palabra recogida en el diccionario pero con un significado distinto, se genere una traducción absurda. Los errores de actuación como el orden incorrecto de dos letras, y los errores léxicos (reproducciones orales, barbarismos, confusiones recurrentes), que siguen a los errores ortográficos en los errores de competencia, provocan malas traducciones por la misma razón. Los errores sintácticos y de cohesión y algunos errores de actuación como el uso de un acento por un apóstrofe entorpecen la tarea del analizador del motor de traducción, lo cual provoca una representación mala o incompleta de la frase original que desemboca lógicamente en una traducción incorrecta.

El número de desviaciones intencionales que provocan errores importantes no es muy grande. Esto también es sorprendente porque teníamos la expectativa de que encontraríamos la causa de los problemas más graves en las desviaciones intencionales. Los fenómenos que provocan los errores más serios son los que manifiestan por parte del emisor un uso intencionadamente no-normativo (falta de acentuación sistemática, reproducción prosódica, puntuación y sintaxis simplificada). Sus efectos son los mismos que los provocados por los errores de competencia como las faltas de ortografía, la reproducción oral o los errores sintácticos. En cambio, los fenómenos restantes tienen un impacto medio o bajo. Creemos que ello se debe a que estos fenómenos aparecen en la traducción sin cambios. Hemos de tener en cuenta que las desviaciones intencionales del emisor cuentan con la complicidad del receptor. Si el emisor recurre a una expresión en otra lengua, utiliza vocabulario sms y del registro de Internet, incluso recurre a la ortografía creativa o a los emoticones es porque está convencido que no va más allá del lenguaje que comparte con el receptor y no se producirán graves problemas de comprensión.

Sin embargo, hay que reconocer que hay condiciones favorables para que estos fenómenos tengan un impacto menor. Una de estas condiciones es que las expresiones en otras lenguas y el vocabulario del registro Internet que hemos analizado son bastante universales. Además, los acrónimos como A10 llegan a ser populares muy rápidamente entre usuarios acostumbrados a comunicarse por Internet, incluso si no son hablantes de la lengua de donde proceden dichos acrónimos. Prueba de ello es que muchos conocen el significado de los acrónimos ingleses U2, P2P, etc. Otra condición favorable es la similitud entre las dos lenguas implicadas. Por ejemplo, muchas formas abreviadas en catalán coinciden o son muy similares en castellano (pq, k). Por ello sería interesante analizar el uso de las

desviaciones intencionales de un emisor que escribe un mensaje en catalán a un receptor del Reino Unido que no sabe esta lengua.

La terminología presente en los mensajes que no está recogida en el léxico del sistema no produce graves errores de comprensión. El hecho de que dicha terminología aparezca sin traducción no supone muchos problemas para unos usuarios acostumbrados a verlos y a referirse a ellos en su forma original. Curiosamente, la traducción de la terminología es la causa de algunos problemas que hemos detectado. Por ejemplo, el término de la comunidad de estudiantes de la UOC pla docent se traduce como *llano docente y no como plan docente. El motivo es que pla se puede traducir como llano o plan y se ha producido una mala selección léxica. Un problema similar ocurre con términos que el segmentador del sistema interpreta como palabras compuestas y traduce los segmentos que coinciden con una palabra común. Por ejemplo, excel se traduce como *excielo porque el sistema ha traducido cel.

6 Conclusiones y trabajo futuro

Hemos visto que los errores de competencia del usuario en su uso del catalán provocan la mayoría de los errores de traducción de los mensajes analizados. También hemos visto que los elementos no normativos característicos de los mensajes de correo electrónico según la bibliografía no tienen la presencia que habíamos esperado y su impacto en la traducción automática no es tan fuerte como el de los errores de competencia lingüística.

Por ello es necesario priorizar el desarrollo de un módulo automático de pre-edición que corrija los errores de competencia- especialmente errores ortográficos- sin descuidar la corrección de errores de actuación y otros fenómenos típicos de los mensajes de correo electrónico como la ortografía creativa o la reproducción prosódica.

Un trabajo similar al que presentamos en este artículo se debería realizar con hablantes monolingües de una lengua sin dificultades de normalización. Quizá los errores debidos a la escritura rápida y a las nuevas formas de expresión superarían los errores de competencia. También sería interesante ver en qué medida la comunicación con un receptor hablante de una lengua muy distinta de la que usa el emisor condiciona el uso de estas nuevas formas de expresión.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el MCYT de España mediante el proyecto AMEDIDA (Programa PROFIT, FIT 350201-2004-6).

Bibliografía

Alonso, A., R. Folguerà y C. Tebé. 2000. Del tecnolecte al sociolecte: consideracions sobre l'argot tècnic en català. En M. Torres, Ll. Jardí, N. Alturo, Ll. Payrató & F.X. Vila (Eds.) *CMO-Cat I Jornada sobre comunicació mediatitzada per ordinador en català*.

Climent S., J. Moré y A. Oliver. 2002. Building an environment for unsupervised automatic email translation. En *Proceedings of the EAMT-CLAW* 2003. Dublin. http://www.uoc.edu/in3/dt/20292/index.html. Fecha de consulta: 3 de junio de 2005

Climent S., P. Gispert-Saüch, J. Moré, A. Oliver, M. Salvatierra, I. Sànchez, M. Taulé y Ll. Vallmanya. 2003. Bilingual Newsgroups in Catalonia: A Challenge for Machine Translation. En *Journal of Computer-Mediated Communication 9* (1). http://jcmc.indiana.edu/vol9/issue1/climent.html. Fecha de consulta: 3 de junio de 2005

Fais, L. y K. Ogura. 2001. Discourse issues in the translation of Japanese e-mail. En *Conference of the Pacific Association for Computational Linguistics, PACLING* 2001, Kitakyushu, Japan: Proceedings.

Herring, S. C. 1999. Interactional coherence in CMC. En T. Erickson (Ed.), *Journal of Computer-Mediated Communication, 4* (4), special issue on Persistent Conversation. http://www.ascusc.org/jcmc/vol4/issue4/herring.html. Fecha de consulta: 3 de junio de 2005

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