Normalización del microtexto. Nuevos desafíos en PLN para el gallego

Microtext normalization. New NLP challenges for Galician language

Sandra Álvarez García
Estefanía Mosquera Castro
Universidade da Coruña
salvarezg@udc.es / emosquera@udc.es

Resumen: *Este trabajo analiza los nuevos retos a los que se enfrenta el gallego en el ámbito del procesamiento del lenguaje natural (PLN). En particular, centra su interés en el microtexto: una nueva modalidad gráfica utilizada en la interacción electrónica (SMS, Whatsapp, Twitter, etc.), caracterizada por diversas licencias lingüísticas que la distancian del código estándar, lo que supone también un desafío para las herramientas de PLN empleadas en la recuperación de información de texto normativo. Para su procesamiento automático es necesario normalizar los microtextos y este constituye un proceso altamente dependiente del idioma. Y aunque ya existen algunas propuestas para otras lenguas, en el ámbito gallego no ha transcendido demasiado este tipo de investigaciones. Por ello, en este artículo prestaremos atención al microtexto gallego: indicaremos sus problemas metodológicos, determinaremos sus similitudes y particularidades con respecto a otras lenguas y abordaremos el diseño conceptual de un normalizador específico para el gallego.*

Palabras clave: *microtexto, PLN, gallego, normalizador, dinamización lingüística*

Abstract: *This paper analyzes the new challenges facing the Galician language in the field of Natural Language Processing (NLP). In particular, it focuses its interest in the Microtext: a new graphic modality used in electronic interaction (SMS, Whatsapp, Twitter, etc.) and characterized by several linguistic licenses that distance it from the standard code, which also implies a challenge for classical NLP tools used in well-written text information retrieval. Therefore, for its automatic processing, normalizing such microtext becomes necessary and this process is highly dependent on the language. And even though for other languages there are some proposals, such research has not substantiated in the Galician context. For this reason, in this paper we will pay attention to Galician microtext: we will indicate methodological problems, determine their similarities and particularities regarding to other languages and address, from a conceptual perspective, the design of a specific normalizer for Galician.*

Keywords: *microtext, NLP, galician, model normalization, linguistic revitalization*

1. El gallego y la lingüística computacional. Conquista y nuevos retos

En los últimos años, con el establecimiento masivo de nuevas formas de comunicación, una gran cantidad de contenidos son generados diariamente en Internet; estos, procesados adecuadamente, suponen un recurso de información muy rico que puede ser utilizado en multitud de ámbitos. Al tratarse de un soporte de base fundamentalmente textual también supuso numerosas transformaciones en la esfera lingüística, donde se desarrolló un nuevo campo de investigación científica enmarcado

en la lingüística computacional, que aborda el tratamiento automatizado de información en lenguaje natural. Como aplicaciones de este nuevo campo destacan los programas de traducción automática, los sintetizadores de voz, los lematizadores, los corpus de textos anotados, los diccionarios, etc., que facilitan el estudio, la comprensión y la adquisición de las lenguas y que, además, son fundamentales para la universalización y para la mejora de accesibilidad de los contenidos en la web.

En las últimas décadas los avances tecnológicos y la universalización del acceso a Internet han multiplicado la cantidad de información y, consecuentemente, de aplicaciones de este tipo para todas las lenguas; sin embargo, esta expansión no es homogénea. En los extremos encontramos lenguas como el inglés o el español que cuentan con numerosos recursos —principalmente económicos— que facilitan esta proyección, u otras como el mbara o el ocaina, que se localizan en partes del mundo menos desarrolladas (Chad y Perú, respectivamente) y que en algunos casos ni siquiera tienen representación escrita (Moseley, 2010). A medio camino, se sitúan las lenguas minorizadas como el gallego que, a pesar de la escasez de medios, consiguen avanzar en este campo y favorecen el proceso de normalización lingüística (Crystal, 2004: 87-92). Para esta lengua destacan diversas iniciativas como las versiones de los sistemas operativos *Windows*, *Linux*, *Trisquel GNU* o *Galinux* y los programas y aplicaciones como *OpenOffice*.org, *LibreOffice.org* o *Gnome*. También disponen de versión gallega los navegadores *Firefox*, *Internet Explorer 9* y *Google Chrome*, los buscadores *Google* y *Bing* o clientes de correo como *Gmail*, entre otros.

Son notables las iniciativas de *software* libre para esta lengua, todas ellas concentradas en el portal Mancomún de la Consellería de Innovación e Industria de la Xunta de Galicia, entre las que resalta la traducción de Pidgin, un cliente de mensajería instantánea, o Gimp, un programa de edición de imágenes con funciones similares a las de Photoshop. Asimismo, es significativo el aumento de corpus de referencia como el *CORGA* (*Corpus de Referencia do Galego Actual*), el *TILG* (*Tesouro Informatizado da Lingua Galega*) o el *CLUVI* (*Corpus Lingüístico da Universidade de Vigo*); de diccionarios como el *Estraviz*, el de la RAG (Real Academia Galega), el diccionario jurídico gallego, el diccionario de pronunciación o el *DigaTIC* (*Dicionario galego de Telecomunicación*); de correctores ortográficos en línea como *Ortogal* y *Galgo* u otros propios de algunos paquetes ofimáticos como *Golfiño* para *OpenOffice*; de traductores como *Tradgal*, *Esgl*, *Traducíndote* o *Apertium*; de conjugadores verbales como *CILENIS* o de lematizadores como *XIADA*. De igual modo, también es reseñable la enciclopedia colaborativa *Wikipedia*, que en su versión gallega (*Galipedia*) incluye ya más de 100.000 artículos en gallego.

Todas estas iniciativas contribuyen a aumentar la presencia de la lengua gallega en Internet, mitigando en cierto modo los temores apuntados por Romero y Vaquero (2001: 59) sobre la potenciación del proceso de minorización lingüística y cultural con la aparición de las TIC (tecnologías de la información y la comunicación). El tratamiento de las características propias de las variedades estándares de las lenguas naturales, asociado de modo sutil al contexto sociocultural de sus hablantes, ha sido

el reto al que se han enfrentado los algoritmos clásicos de procesamiento de lenguaje natural (Xue *et al.*, 2011: 74-75). No obstante, la web no solo está formada por la lengua estándar, sino que gran parte de la información del soporte electrónico está escrita utilizando una variedad gráfica informal, que viola la ortografía y la gramática normativa y que usa de forma deliberada numerosas técnicas de reducción y de expresión en donde los símbolos son utilizados de modo no convencional; estas son especialmente utilizadas en el contexto de los SMS y en microblogs. Conforme afirman Kobus *et al.* (2008: 441), "adressed to relatives or peers, written in the spur of the moment, using interfaces, each with its specific constraints (computer keyboards, PDAs, mobile phones keypads)", estos textos se caracterizan por desviaciones masivas y sistemáticas con respecto a la lengua estándar y, en consecuencia, las herramientas de PLN clásicas no sirven para procesarlas. Aw *et al.* (2006: 34) inciden en la complejidad de este proceso:

> General text normalization deals with Non-Standard Words (NSWs) and has been well-studied in text-to-speech [...] while SMS normalization deals with Non-Words (NWs) or lingoes and has seldom been studied before-NSWs, such as digit sequences, acronyms, mixed case words (WiNT, sunOS), abbreviations and so on, are gramatically correct in linguistics. However lingoes, such as "b4" (before) and bf (boyfriend), which are usually self-created and only accepted by young SMS users, are not yet formalized in linguistics.

Sin embargo, con el uso masificado de microblogs como Twitter, se vuelve de vital importancia procesar esta gran cantidad de información generada diariamente para numerosas aplicaciones como pueden ser, entre las más significativas, el análisis de sentimientos o la detección de eventos. Por esta razón, surge la necesidad de normalizar este tipo de textos (o microtextos, como comúnmente son denominados en este ámbito[1]) para poder procesarlo con herramientas PLN clásicas. Este proceso de normalización tiene multitud de aplicaciones: en primer lugar, permitiría al usuario no adaptar su grafía en función del destinatario —principiante o experto— al facilitar a los no iniciados la posibilidad de "traducir" su contenido. En segundo lugar, podría aplicarse a la corrección automática en textos —no siempre interactivos— que presenten este tipo de para-ortografía, como los blogs o los correos electrónicos. Además, permitiría a los motores de búsqueda ampliar su comportamiento para soportar búsquedas de forma independiente a la variación orto-tipográfica utilizada (Choudbury *et al.*, 2007: 64). Por último, también contribuiría a mejorar los sistemas de conversión del texto escrito en habla, lo que permitiría a las personas con problemas visuales —o a aquellas no iniciadas en la escritura— acceder a contenidos en mi-

1 Véase, a modo de ejemplo, Gouws *et al.* (2011).

crotexto, como ocurre en ocasiones con las redes sociales; e inclusive posibilitaría la conversión del microtexto en mensajes de voz.

En la actualidad podemos encontrar algunos normalizadores, principalmente en lengua inglesa y francesa, que se apoyan en la similitud de este problema con otros más estudiados en procesamiento de lenguaje natural, como analizaremos en la sección siguiente. Sin embargo, desconocemos la existencia de investigaciones similares fundamentadas en la lengua gallega o en la portuguesa, que por sus similitudes estructurales podría constituir también un referente que habría que tener en cuenta. En consecuencia, consideramos necesario integrar el gallego a este nuevo campo de estudio para que esta lengua pueda situarse al mismo nivel que otras no minorizadas, favoreciendo, de este modo, su normalización.

2. Normalizadores de microtexto. Estado de la cuestión

Como señalamos en §1, la reciente expansión del microtexto lo convierte en una importante fuente de información, pero a la vez supone un reto para las herramientas clásicas de procesamiento de lenguaje natural. Por este motivo, surge una nueva área de estudio: la normalización de microtexto, cuyo objetivo consiste en reescribir este tipo de textos a una ortografía más convencional para facilitar las tareas de Recuperación de Información (RI). En este contexto surgen diversas iniciativas normalizadoras; no se trata de aplicaciones definitivas, sino de aproximaciones que todavía pueden ser mejoradas. De hecho, no todas se centran en la misma tipología de microtexto: los trabajos del ámbito anglófono y francófono se basan principalmente en los SMS, mientras que los realizados para el español están más enfocados a resolver la problemática específica observada en los tuits. Sin embargo, hasta donde sabemos, no existe ninguna propuesta de normalizador para microtextos gallegos; por esta razón, en esta sección describiremos brevemente algunos de los normalizadores más relevantes en otras lenguas para comprobar en qué medida las técnicas aplicadas pueden ser de utilidad para las características propias del gallego. Asimismo, no todas las investigaciones se guían por los mismos procedimientos; en este sentido, los normalizadores que encontramos al revisar el estado de la cuestión se pueden clasificar en tres grandes grupos en función de las técnicas de PLN que utilizan: corrección ortográfica, traducción o técnicas de reconocimiento de voz.

Entre los normalizadores basados en técnicas de corrección ortográfica podemos destacar el trabajo realizado por Choudbury *et al.*(2007) que modela el problema de la normalización de microtextos en lengua inglesa como un proceso de eliminación de ruido. Utilizando un conjunto de SMS ya alineados palabra por palabra con su versión normalizada, genera un modelo de Markov para cada palabra observada, teniendo en cuenta sus posibles desviaciones (desviaciones fonéticas, elisiones, etc.) y calibrándolo con los alineamientos observados en el corpus de entrenamiento. Así, este sistema obtiene una eficacia del 80% a la hora de normalizar una palabra. No

obstante, cabe destacar que el problema de cómo separar las palabras en el mensaje no es tratado en este normalizador.

Xue *et al.* (2011) proponen un normalizador con componentes procedentes de técnicas de corrección ortográfica y de reconocimiento de voz; en este sentido, modelan el microtexto como un canal con ruido que distorsiona el mensaje original a través de cuatro canales: distorsión ortográfica (responsable de las elisiones vocálicas, entre otras), distorsión de pronunciación, contexto (entendido como los unigramas, bigramas o trigramas que preceden a cada palabra del mensaje) y acrónimos (responsable de las abreviaturas). Estos modelos son combinados para calcular, para cada palabra del mensaje, sus posibles normalizaciones y seleccionar el mejor candidato. La importancia de cada canal en el cálculo del término normalizado está ponderada, de forma que diferentes configuraciones varían la precisión final del normalizador, que puede adaptarse a las características del microtexto. La precisión obtenida es de 93% para mensajes SMS, y del 50% para tuits. La baja precisión observada en el caso de los tuits se debe principalmente al gran número de nombres propios y de entidades que contienen y a que son normalizados incorrectamente en esta propuesta, reduciendo así la precisión global. Cabe destacar que el problema de la detección de entidades es un reto que continúa vigente en la investigación actual y que, además, es altamente dependiente del idioma del mensaje y del contexto social en el que este se produce.

En uno de los normalizadores de mayor relevancia en lengua francesa, Kobus *et al.* (2008) proponen otra aproximación híbrida. En primer lugar, se emplean estrategias clásicas de traducción para obtener una primera normalización. Posteriormente, y teniendo en cuenta que en muchas ocasiones el microtexto es un lenguaje más próximo a la fonética del mensaje, se usan técnicas procedentes de los estudios en reconocimiento de voz para obtener posibles traducciones para aquellas palabras que no fueron localizadas en el vocabulario de traducción. Obtienen un valor en la escala BLEU de 0.8[2].

En español, los trabajos son mucho menos numerosos, si bien existen iniciativas propias para tratar la normalización como, por ejemplo, el Tweet-Norm[3]. Destaca el realizado por Oliva *et al.* (2013), que supone una nueva aproximación híbrida, ya que utiliza técnicas de reconocimiento de voz con información semántica típica de las herramientas de traducción. El proceso de normalización se realiza a través de tres módulos. El primer módulo de separación de palabras está específicamente diseñado para SMS en español e identifica las diferentes palabras que aparecen en un SMS. El siguiente módulo genera las posibles normalizaciones de cada palabra, a través del uso de un diccionario SMS y un diccionario de fonética español que tiene en cuenta

2 BLEU es una medida de evaluación utilizada habitualmente para medir la calidad de una traducción, comparando la traducción obtenida por el sistema con una traducción que se considera *correcta*. El resultado de la medida BLEU será un valor entre 0 y 1 y se considera que la traducción tiene una mayor calidad cuanto más cercano esté a 1(Papineni et al., 2002).

3 Véase http://komunitatea.elhuyar.org/tweet-norm.

las posibles abreviaturas por similitud fonética, y que aplica una distancia de Levenshtein modificada[4]. Finalmente, el módulo de desambiguación descarta candidatos según su rol léxico y su semántica en el contexto de la frase. Con este normalizador, se obtiene una puntuación en la escala BLEU de 0.81.

También en lengua española, Vilares *et al*. (2013) proponen un sistema normalizador que aborda el problema de la normalización desde una perspectiva léxica. Este genera a partir de la palabra observada en el mensaje las traducciones candidatas teniendo en cuenta algunos de los mecanismos de desviación habituales en microtextos: repetición de caracteres, escritura fonética, errores en los diacríticos y otros de carácter ortográfico. Su principal fortaleza es que a pesar de su simplicidad y de los pocos recursos utilizados, sus resultados pueden considerarse positivos.

En cuanto a normalizadores propuestos para lenguas minoritarias, sobresale el trabajo realizado por Costa-Jussa y Banchs (2012) para el idioma haitiano. La normalización del microtexto parte, en este caso, de una combinación de dos sistemas. En primer lugar, un sistema de traducción estadístico orientado a frases con probabilidades extraídas del corpus de entrenamiento, que contempla aspectos adicionales como el reordenamiento de palabras dentro de la frase o la frecuencia de una palabra globalmente en el idioma. Este sistema de traducción se combina con unas reglas, que a pesar de haber sido extraídas automáticamente del conjunto de entrenamiento, como ya indicamos, han sido igualmente enriquecidas con un diccionario global haitiano.

3. El microtexto gallego. Particularidades (socio)lingüísticas

Las circunstancias (socio)lingüísticas en las que se enmarca la lengua gallega difieren, en cierta medida, de aquellas en las que se encuadran las aproximaciones anteriores y, en consecuencia, también el microtexto. La situación de minorización lingüística que sufre el gallego plantea diversos obstáculos que dificultan este tipo de iniciativas normalizadoras: el primero tiene que ver con la caracterización del microtexto y de sus propiedades lingüísticas que, desde una óptica descriptiva, tiene que partir, necesariamente, de un corpus de textos lo suficientemente representativo. A nivel general existen muy pocos bancos de datos que recojan muestras de este tipo, como *sms4science* (http://www.sms4science.org/) o *treasuremytext* (http://www.treasuremytext.com/) y, en particular, ninguno de los existentes incorpora a día de hoy la lengua gallega. Y aunque constatamos alguna iniciativa en este sentido en el ámbito gallego, la cantidad de textos manejados es aún limitada tanto desde una perspectiva cualitativa –está restringida a una temática concreta– como cuantitativa –al contener un número reducido de muestras–. Consecuentemente, considera-

4 La distancia de Levenshtein es utilizada como medida de similitud entre dos palabras. Esta distancia viene determinada por el número de operaciones (de sustitución, inserción y borrado) que es necesario aplicar sobre una de las palabras para obtener la otra palabra sobre la que se está calculando la distancia. En su versión más básica, cada una de estas operaciones tiene valor 1, y la distancia coincide por tanto con el número de operaciones aplicadas.

mos que un paso previo al diseño de un normalizador específico para el gallego era crear un corpus *ad hoc* para esta lengua, que recogiese el mayor número de muestras posibles. En este momento, nuestro corpus consta de 812 mensajes[5] que fueron recopilados a través de varios sistemas: integración de los disponibles en la red, selección de diversos informantes repartidos por toda la geografía gallega y envío de un formulario electrónico de participación abierta.

Aunque se trata de un corpus todavía en construcción consideramos que el número de muestras puede ser representativo para un análisis inicial del microtexto, si bien concordamos con Pérez Guerra (1998: 28) en que cuanto mayor sea el volumen de datos, mayores serán también las oportunidades de estudio. A este respecto, debemos tener en cuenta la dificultad de compilar un corpus de este tipo, sobre todo si se fundamenta en mensajes telefónicos (SMS, Whatsapp, etc.), ya que no resulta fácil acceder a textos que se inscriben en la esfera privada de las personas (Fairon y Paumier, 2006: 351; Crystal, 2008: 103-104). No obstante, se trata del microtexto más complejo desde el punto de vista del procesamiento lingüístico y, por tanto, el más atractivo para los investigadores porque al concentrar una mayor cantidad de desviaciones, se convierte en el que más desafíos formula a la hora de normalizar (Choudhury *et al*., 2007:65). En este sentido, diseñar un normalizador de mensajes de texto para el gallego ofrecería mayores posibilidades de aplicación a otros microtextos con menores desviaciones lingüísticas, como Twitter, si este alcanzase un porcentaje alto de éxito.

En segundo lugar, es necesario señalar que a pesar de los avances de la Lingüística Computacional en el ámbito gallego (ver §1), la mayoría de estas iniciativas tienen un carácter privado. Es decir, pueden ser utilizadas de forma puntual, como ocurre con los diccionarios o los traductores automáticos –en algunos casos previo pago o bajo suscripción como el diccionario jurídico gallego, en otros de forma gratuita como el de la RAG–, pero no es posible acceder a su estructura, al número total de entradas y a sus clasificaciones gramaticales. Estas aplicaciones, al proceder de organismos lingüísticos oficiales como el Instuto da Lingua Galega (ILG), la Real Academia Galega (RAG) o el Centro Ramón Piñeiro, cuentan además con una información muy detallada y completa, de modo que serían enormemente útiles para el desarrollo de herramientas lingüísticas de esta clase si estuviesen accesibles para toda la comunidad investigadora. A su vez, los recursos en formato abierto –ciertamente abundantes para las lenguas analizadas en el apartado 2–, son todavía escasos para el gallego, en donde tan solo disponemos, además de los señalados en §1, de los diccionarios *GNU ASPELL*, el corrector ortográfico *Hunspell* (utilizado en herramientas como *OpenOffice.org*, *Firefox Mozilla* o *Google Chrome*) y el analizador lingüístico *Freeling*, que comparte datos sobre aspectos morfológicos del gallego. E, igualmente, existen también lagunas y restricciones de naturaleza terminológica en lo referente a los len-

5 812 mensajes con una cantidad media de palabras de 17,11 y un número medio de caracteres de 83,10 que se corresponden con textos normalizados de 96 caracteres de media.

guajes de especialidad en contraste con otras lenguas (Benedito, 2003; Belda Medina, 2003), para las que incluso se han elaborado diccionarios y glosarios SMS (Oliva *et al.*, 2013: 126). Esta situación dificulta en cierto modo el tratamiento mecánico de los textos que, en ocasiones, tienen que ser revisados de forma manual en búsqueda de patrones específicos que puedan ser automatizados.

Un reto habitual al que se tienen que enfrentar las herramientas tradicionales en recuperación de información es el reconocimiento de entidades, como nombres propios, ubicaciones geográficas…, pero también nombres de negocios o de películas, que son altamente dependientes del idioma utilizado. Como ya señalamos, cuando tratamos el problema de la normalización de SMS, esta cuestión se vuelve más compleja, puesto que tales entidades deben también ser normalizadas. En este contexto, se hacen necesarios conjuntos de datos de entidades, externos a los diccionarios, que puedan servir como base para su normalización. En la actualidad, y gracias a las iniciativas en el ámbito de Linked Data (Bizer *et al*., 2009), que tienen como objetivo procesar e integrar diferentes fuentes de datos disponibles en Internet, existen numerosos conjuntos de datos de libre acceso que podrán utilizarse para este reconocimiento de entidades usadas en el corpus y que ayudarán a su normalización. Por ejemplo, la base de datos Geonames (http://www.geonames.org/) puede ser empleada en el reconocimiento de elementos geográficos e incluye gran parte de la geografía gallega. La DBpedia (http://dbpedia.org) supone una versión estructurada de la Wikipedia –y por extensión de la Galipedia–, y nos permite extraer entidades de diferente tipo.

Por razones similares, también nos parece relevante traducir el corpus, ya que la enorme variabilidad en la forma de las palabras hace complejo trabajar con textos de estas características. Así, contar con una versión normativa del corpus facilita el análisis de los mensajes y de sus características. De acuerdo con Fairon y Paumier (2006: 52), el único límite en la variación parece restringirse a la imaginación de los usuarios y a la necesidad –a veces limitada– de que tenga sentido: por tanto, las variaciones gráficas de una palabra son totalmente impredecibles. Como se indica en Mosquera Castro (2012), las características principales de esta modalidad escrita son similares en todos los idiomas y aunque pueda parecer un sistema caótico y carente de normas reguladoras, lo cierto es que existe cierto grado de estandarización: elisiones vocálicas y consonánticas, uso de abreviaturas, siglas y acrónimos, omisión de espacios, simplificación de la sintaxis, uso de símbolos y números con valor fonológico, estiramientos gráficos, onomatopeyas, emoticonos, etc. Serán los usuarios los que conociendo estas estrategias lingüísticas decidan utilizarlas en unas palabras y no en otras o en unas ocasiones sí y en otras no.

En este sentido, en el diseño de cualquier normalizador de microtexto será necesario tener en cuenta diversos patrones comunes que se detectan en la mayor parte de las lenguas. Por un lado, habrá que contemplar de modo genérico los reordenamientos de letras producidos por errores tipográficos, frecuentes en una modalidad de escritura en la que prima la velocidad de los intercambios y en la que se desatien-

de la corrección ortográfica (**arblo* por *árbol*, por ejemplo). Del mismo modo, será necesario incluir los signos diacríticos como los acentos gráficos, los apóstrofos o la diéresis, habitualmente ausentes en los microtextos, incluso en aquellos escritos en la variedad estándar. Por otro lado, habrá que considerar determinadas frecuencias en la configuración del normalizador: así, suele ser habitual que la elisión y la repetición de caracteres se produzca especialmente en final de palabra, que cuanto mayor sea la palabra más elisiones se detecten e, igualmente, que existan más probabilidades de que la elisión o la reduplicación sea vocálica antes que consonántica (Choudhury *et al.*, 2007: 162-164). Esta serie de pautas serán importantes a la hora de automatizar la inclusión o la reducción de caracteres para que los procesos resulten verdaderamente efectivos.

No obstante, a pesar de la universalidad de este nuevo modelo de escritura, las propiedades estructurales y sociolingüísticas del gallego implicarán la preeminencia de unas estrategias sobre otras y la aparición de fenómenos no siempre contemplados por los normalizadores analizados en el apartado 2. Así, por ejemplo, la elisión de vocales será mucho más frecuente en lenguas como el español o el gallego (en contraposición al inglés o al francés), donde pueden ser recuperadas fácilmente por el contexto o por las consonantes que forman la palabra. En esta línea, la vocal que se elide en mayor medida es la "e", favorecida porque su pronunciación ya viene dada por gran parte de los nombres de las consonantes ("be", "ce", "de", etc. Ej. *bsar* por *besar*). Por tanto, el normalizador deberá concederle mayor probabilidad en la traducción. Y en el sentido opuesto, hay que destacar que respecto a otros idiomas como el inglés o el francés, la escritura fonética no tiene demasiadas repercusiones en el microtexto gallego, dado que por regla general la grafía reproduce con gran fidelidad la oralidad.

Las implicaciones fonéticas en el gallego se reducen a la simplificación de la correspondencia fonema/letra en casos particulares para los que a un fonema le corresponden dos o más grafías. De este modo, habrá que tener en cuenta que se utilizará indistintamente la "b" y la "v" para representar el fonema /b/ y la "c" la "k" y la "q" para /k/. Con todo, la distancia que separa *ksa* de *casa*, por ejemplo, dista mucho de aquella que debe ser considerada para que palabras como *rite*, *gud* o *nite* –abreviaturas habituales en inglés– se normalicen en *right*, *good* o *night*. De igual forma, la eliminación de grafías mudas en gallego afecta fundamentalmente a la "h" y con más frecuencia en posición inicial –98,68%– y, en menor medida, medial –1,31%– (*oxe* [hoxe], *proibido* [prohibido]), al contrario que en francés, por ejemplo, donde existe una frecuencia más alta de elisiones en posición final (*pa* [pas], *cour* [cours]), como documenta Anis (2006). En gallego la reducción de grafías mudas en posición final tan solo afecta en nuestro corpus a la palabra *Madrid*, que aparece escrita como *Madri*. Sin embargo, su incidencia es mínima y no se puede afirmar que represente un fenómeno sistemático de este idioma.

En cuanto al uso de números y símbolos con valor fonológico, debemos tener en consideración que, aunque se trata de un recurso muy frecuente en los microtex-

tos –aparece en el 57,2% de los textos del corpus–, en gallego muestra diversos patrones específicos. Así, no siempre estas sustituciones se realizan por equivalencia acústica, como sucede en *m8* ('moito'), en *7mbro* ('setembro') o en *xq* ('porque'), sino también por proximidad: por ejemplo, el símbolo *x* ('por') también goza de rendimiento en secuencias con 'per' como en *xo* ('pero') o *xdn* ('perdón') o los numerales *100* ('cen') o *5* ('cinco'), que son utilizados en ocasiones para sustituir secuencias con *sen* y *sin co*, como ocurre en *100pre* e *5mpromiso*. Los usos de estos numerales derivan de la reprodución del seseo, variante dialectal que, en la actualidad, no se recoge en la normativa gráfica del gallego (RAG y ILG, 2005), aunque sí se promueve para su estándar oral (Regueira, 2013).

El contexto de conflicto lingüístico que se vive en Galicia entre el gallego y el español es también la causa de algunos fenómenos particulares. Continuando con los numerales, es frecuente el uso de 2 para reemplazar secuencias fónicas que equivalen a su realización en español ('dos') y no en gallego ('dous'), como se documenta en *to2* ('todos'). Igualmente, debemos tener en cuenta que será también habitual la alternancia de códigos; y si bien es cierto que no es inusual el uso de la lengua inglesa –y en menor medida otras como el francés, el portugués o el catalán– en textos escritos en gallego, lo más frecuente es que la alternancia se produzca entre el gallego y el español –de hecho, este representa el 47,3% de los casos–. Así los usuarios escogerán en determinados casos un término en español si este es más funcional por razones expresivas o de brevedad. En una línea similar se sitúan los españolismos, esto es, formas híbridas y espurias que surgen por la interferencia del español sobre el gallego. Esta situación deberá tenerse en cuenta a la hora de diseñar un normalizador, que necesitará implementar reglas específicas para intentar solucionar esta mezcla idiomática, ya que no se contempla en ninguno de los normalizadores analizados en la sección 2.

Grosso modo, las citadas aquí son algunas de las particularidades (socio) lingüísticas que detectamos en el análisis preliminar del corpus de microtextos que manejamos y que deben ser consideradas en el diseño de un normalizador para la lengua gallega. De la misma forma, será necesario tener en cuenta los diversos modelos de normalizadores descritos en el apartado 2 y escoger como referentes aquellos que se adapten mejor a las características estructurales de este idioma o incluso integrar las diversas propuestas para crear un prototipo de normalizador que atenúe en la medida de lo posible las limitaciones anteriormente señaladas y que permita alcanzar los resultados deseados.

---

6 Probablemente esta es la posición donde se insertan *frente*, *cima*, o *bajo*, en los elementos de las series adverbiales *debajo* y *abajo*. *Vid.* Romeu (2014).

4. Propuesta de normalizador para el microtexto gallego

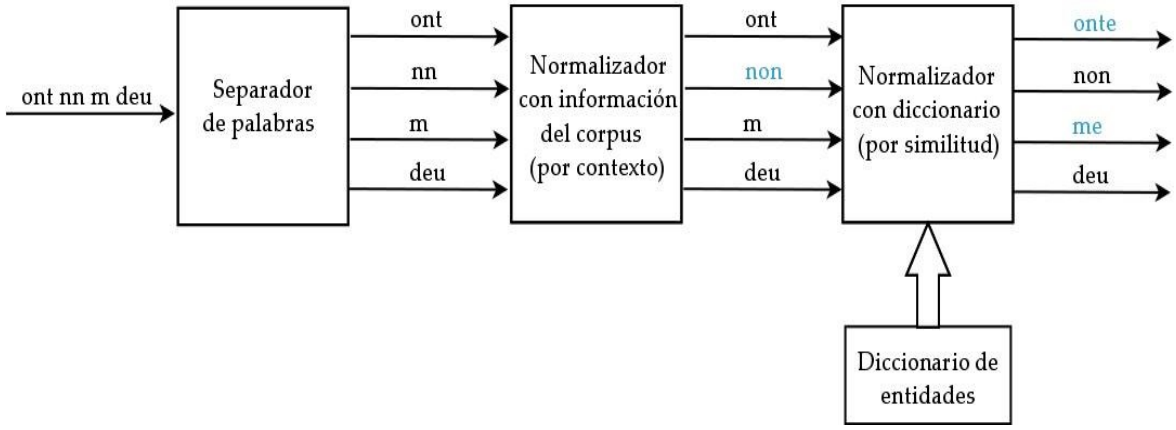

Fig. 1: Modelo de normalizador para el microtexto gallego

A la vista de las peculiaridades lingüísticas observadas en el gallego, proponemos la implementación de un normalizador inspirado en las técnicas habituales apuntadas en el estado de la cuestión de normalizadores de microtexto (ver §2), pero adaptándolas a las especificidades del microtexto gallego, esto es, atendiendo a sus desviaciones ortográficas características; esto, en nuestra opinión, mejorará la precisión obtenida en el proceso de normalización. En la Figura 1 se puede observar el esquema conceptual del normalizador que proponemos.

Como paso previo a la normalización del microtexto se realizará un proceso de separación, en el que el microtexto se divide en elementos individuales (idealmente palabras) que serán normalizados en varios pasos. Aunque existen herramientas disponibles que abordan el problema de la separación en palabras específicamente en microtexto (https://github.com/brendano/tweetmotif), consideramos que los patrones específicos del gallego hacen necesario adaptar este proceso con conocimiento específico de nuestro idioma de estudio. Como ejemplo de esta especificidad podemos observar el tratamiento de números. Una estrategia habitual en microtexto consiste en omitir espacios entre palabras, especialmente si una de ellas es un numeral como en el caso de *1kfe* ('un café'), por lo que el proceso de separación tenderá a aislar este elemento en dos palabras: *1 kfe*. Sin embargo, en el caso del gallego encontramos excepciones a esta regla, como el número *100*, que en ocasiones se utilizará por su transcripción fonética dentro de una palabra. Así, en *100pre* ('sempre'), el segmentador debería producir un único elemento *100pre*. A modo de ejemplo, en la Figura 1 podemos observar cómo el separador de palabras recibe como entrada la frase *ont nn m deu* y produce como salida cuatro elementos diferenciados.

Cada uno de los elementos resultantes del proceso de separación se intentará normalizar utilizando como base la información del corpus de mensajes de texto en gallego disponible. Este corpus de 812 mensajes ha sido normalizado y alineado manualmente, y de él podemos extraer patrones habituales de normalización. El proceso

de normalización será secuencial, de forma que al normalizar una palabra, conoceremos el resultado de la normalización de las palabras anteriores, que compondrán el contexto de la palabra actual. De este modo, normalizaremos la palabra con aquella solución más probable conociendo el contexto previo (las dos palabras previas ya normalizadas, es decir, un contexto de orden 2 en base al corpus alineado del que disponemos. Supongamos, por ejemplo, el siguiente microtexto *fn d smana*, en el que ya se han normalizado las dos primeras palabras dando como resultado *fin de*. El elemento *smana* será normalizado como *semana* si en nuestro corpus alineado la normalización más probable para el elemento *smana* precedido del contexto normalizado *fin de* es *semana*. En el caso de que no se encuentre esa ocurrencia en nuestro corpus, se intentará el mismo proceso con un contexto de orden 1 (sólo la palabra previa *de*) y finalmente, con un contexto de orden 0, que supone comprobar en el corpus, para un término, cuál es su normalización más frecuente. En caso de que para un contexto de orden 0 no se disponga de información para realizar la normalización, el elemento permanecerá intacto y se intentará su normalización en el siguiente paso. En la Figura 1, solo el elemento *nn* se ha podido normalizar en la fase de "Normalización con el corpus", obteniendo la palabra *non.* Los restantes elementos, por tanto, permanecerán igual por el momento (*ont, m, deu*).

Como hemos podido comprobar, es muy probable que aquellas palabras o desviaciones menos frecuentes no aparezcan en el corpus de entrenamiento. Estas palabras se normalizarán, de forma independiente al contexto en el que aparecen, utilizando un diccionario completo de gallego. En primer lugar, dado un término sin normalizar, se generarán los candidatos normalizados a través de elisiones e inserciones en el término. Seleccionaremos el mejor candidato utilizando como medida una distancia de Levenshtein modificada para adecuarse a las características del microtexto gallego. El candidato que esté a una menor distancia de un término existente en el vocabulario será la normalización escogida. No todas las inserciones y borrados tendrán peso similar al calcular esta distancia de Levenshtein, sino que dependerán de lo probable de esa desviación en gallego y tendrán en cuenta la posición en la que tiene lugar en el término. Una aproximación similiar fue seguida también en Oliva (2013) para el español. En nuestro caso como medida de similitud ponderamos las posibles desviaciones estableciendo 4 pesos diferentes en función de su frecuencia en el microtexto gallego. Así, asignamos el menor peso (1) a la inserción de acentos en las vocales (por ejemplo *que* → *qué*), y otras conversiones habituales (como *b* ↔ *v*, *q* → *que*, *x* → *por*, *x* → *che*, *y* → *ll* o *w* → *g*). Como siguiente rango de pesos consideramos las inserciones de vocales, ya que suelen ser los caracteres eliminados más habitualmente, documentadas en el 78,8% de los textos. Este peso variará entre 1 y 2 en función de la posición del carácter y de su contexto; así, por ejemplo, en el término *prdoa*, *perdoa* es más probable que *prdoea*. Un rango de peso mayor (2-3) se le asignará a la inserción de consonantes teniendo en cuenta la probabilidad de elisión de las diferentes consonantes extraída de lo observado en nuestro corpus de entrenamiento. Finalmente, con peso 4 tendríamos el borrado de caracteres, que trataría el problema

de los estiramientos gráficos (como ocurre en *bicosss*). De esta forma, dado un término generaremos candidatos aplicando estas reglas de inserción, elisión y sustitución iterativamente, y seleccionaremos el candidato con la menor distancia en base a la métrica que hemos definido. Volviendo nuevamente a la Figura 1, se puede observar que los términos *nn* y *m* han sido normalizados utilizando esta comparación con el diccionario. Así, el término *m* ha sido normalizado mediante la inserción de la vocal *e* y el mismo proceso ha operado en o*nt* para su normalización en *onte*. Nótese que dado que *deu* es una palabra válida en el diccionario, ya no se generarán candidatos ni se sustituirá por otra palabra.

El diccionario gallego, a pesar de ser completo y flexivo (formas verbales, número y género, etc.), no será suficiente a la hora de normalizar entidades como nombres propios o localizaciones, de uso muy frecuente en el microtexto. Con el objetivo de mejorar la precisión obtenida en mensajes que incluyan estas entidades, proponemos enriquecer este último paso en el normalizador con la incorporación de información sobre entidades (ciudades, nombres propios, e incluso libros o películas). Nótese que el vocabulario de estas entidades también es específico del idioma, ya que en el caso de localizaciones, lo habitual es que se utilicen en su forma gallega, y los nombres propios mencionados también son específicos de cada región. El diccionario de entidades que elaboraremos estará compuesto por conjuntos de datos RDF (Resource Description Framework, Brickley *et al.* 2004) como la DBpedia en su versión gallega (para localizaciones) y otros recursos disponibles, como un conjunto de nombres propios gallegos (http://www.amesanl.org/gl/nomeseapelidos/nomes). Este diccionario de entidades nos permitirá realizar un proceso de generación de candidatos y evaluación según la distancia de Levenshtein similar al descrito previamente, pero cuyos pesos podrán adecuarse para el caso específico de las entidades, ya que en general estos términos sufren un menor proceso de desvío ortográfico en el microtexto. De esta forma, para cada palabra seleccionaremos el mejor candidato normalizado de entre los propuestos por los dos sistemas (diccionario gallego y diccionario de entidades).

El normalizador expuesto en esta sección constituye un modelo conceptual en el que ya estamos trabajando y que, una vez implementado y a la vista de las debilidades que puedan detectarse, podrá modificarse en una u otra dirección para mejorar su precisión y alcanzar un mayor porcentaje de éxito.

5. Conclusiones y trabajo futuro

El éxito del microtexto en el soporte electrónico ha instaurado nuevos desafíos en el área del procesamiento del lenguaje natural y desde algunas lenguas ya se han realizado diversas aplicaciones que tratan de normalizar esta modalidad gráfica. Las condiciones sociolingüísticas en las que se genera el microtexto gallego, así como sus particularidades estructurales hacen necesaria una aproximación específica que, aun

aprovechando las características de los normalizadores ya existentes, las mejore y las adapte a sus propiedades intrínsecas.

Por esta razón, es nuestra intención desenvolver en el futuro un normalizador que se ajuste a las necesidades del microtexto gallego y de sus usuarios y para ello proponemos un proceso de normalización en etapas que se adecua a las características propias de la lengua gallega. Su aplicabilidad en un campo emergente y de vital importancia como el procesamiento de la información convertiría a esta herramienta y, por extensión, a la lengua gallega, en un referente para otras lenguas minorizadas y, del mismo modo, contribuiría también a su dinamización.

Bibliografía

Anis, Jean (2006): *Communication électronique scripturale et formes langagières* [en línea], disponible en <http://rhrt.edel.univ-poitiers.fr/document.php?id=547> [consultado en marzo de 2014].

Aw, Aiti, Min Zhang, Juan Xiao y Jian Su (2006): "A phrase-based Statistical Model for SMS Text Normalization". En *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Stroudsburg: Association for Computational Linguistics, 33-40.

Belda Medina, José Ramón (2003): *El lenguaje de la informática e Internet y su traducción*. Alicante: Universidad de Alicante.

Benedito, Joviana. (2003): *Dicionário da Internet e do Telemóvel*. Lisboa: Centro Atlântico.

Bizer, Christian, Tom Heath y Tim Berners-Lee (2009): "Linked data-the story so far". *International Journal on Semantic Web and Information Systems*, 5, 1-22.

Brickley, Dan y Ramanathan Guha (eds.) (2004): *Resource description framework (rdf) schema specification 1.0.* [en línea], disponible en <http://www.w3.org/TR/2004/REC-rdf-schema-20040210/> [consultado en marzo de 2014].

Choudhury, Monojit, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar y Anupam Basu (2007): "Investigation and Modeling of the Structure of Texting Language". *International Journal of Document Analysis and Recognition*, 10, 157-174.

Costa-Jussa, Marta y Rafael Banchs (2012): "Automatic normalization of short texts by combining statistical and rule-based techniques". *Language Resources and Evaluation*, 47, 179-193.

Crystal, David (2004): *The language Revolution*. Cambridge: Polity Press.

Crystal, David (2008): *Txtng: The gr8 db8*. Oxford: Oxford University Press.

Fairon, Cédrick y Sébastien Paumier (2006): "A translated corups of 30,000 French SMS". *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 351-354. [en línea], disponible en <http://www.sms4science.org/userfiles/A%20translated%20corpus.pdf> [consultado en marzo de 2014].

Gouws, Stephan, Donald Metzler, Congxin Cai y Eduard Hovy (2011): "Contextual Bearing on Linguistic Variation in Social Media", en *Proceedings of the Workshop on Language in Social Media*. Stroudsburg: Association for Computational Linguistics, 20-29.

Kobus, Catherine, François Yvon y Géraldine Damnati (2008): "Normalizing sms: are two metaphors better than one?", en *Proceedings of the 22nd International Conference on Computational Linguistics*, 1. Stroudsburg: Association for Computational Linguistics, 441-448.

Moseley, Christopher (2010): *Atlas of the World's Languages in Danger* [en línea], disponible en <http://www.unesco.org/culture/languages-atlas/> [consultado en marzo de 2014].

Mosquera Castro, Estefanía (2012): "Novos retos da Lingüística: as textualidades electrónicas. Consideracións sobre a escrita dos chats e das SMS". *LLJournal*, 7 (1), disponible en <http://ojs.gc.cuny.edu/index.php/lljournal/article/view/1147/1250> [consultado en marzo de 2014].

Oliva, Jesús, José Ángel Castillo y Ángel Iglesias (2013): "A SMS normalization system integrating multiple grammatical resources". *Natural Language Engineering*, 19, 121-141.

Papineni, Kishore, Salim Roukos, Todd Ward y Wei-Jing Zhu (2002): "BLEU: a method for automatic evaluation of machine translation", en *Proceedings of the 40th annual meeting on association for computational linguistics*. Stroudsburg: Association for Computational Linguistics, 311-318.

Pérez Guerra, Javier (1998): *Introducción a la lingüística de corpus. Un ejercicio con herramientas informáticas aplicadas al análisis textual*. Santiago de Compostela: Tórculo Edicións.

[RAG y ILG] = Real Academia Galega e Instituto da Lingua Galega (eds.) (2005): *Normas ortográficas e morfolóxicas do idioma galego*. A Coruña: Real Academia Galega.

Regueira, Xosé Luís (2013): "Estándar oral e modelos de lingua". *A letra miúda. Revista de sociolingúística para o ensino*, 2. [en línea], disponible en <http://coordinadoraendl.org/aletramiuda/artigos/art2_n2.pdf> [consultado en marzo de 2014].

Romero, Daniel e Isabel Vaquero (2001): *Da periferia á rede*. Vigo: Xerais.

Vilares, Jesús, Miguel Ángel Alonso y David Vilares (2013): "Prototipado Rápido de un Sistema de Tuits: Una Aproximación Léxica", en Alberto Díaz Esteban, Iñaki Alegría Loinaz y Julio Villena Román (eds.): *XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural (SEPLN 2013). Tweet Normalization Workshop at SEPLN 2013*, 76-80.

Xue, Zhenzhen, Dawei Yin y Brian Davison (2011): "Normalizing Microtext". *Proceedings of* The *AAAI-11 Workshop on Analyzing Microtext"*, 74-79. [en línea], disponible en <http://www.cse.lehigh.edu/~brian/pubs/2011/AAAI/normalizing-microtext.pdf> [consultado en marzo de 2014].

Fecha de recepción: 28/04/2014

Fecha de aceptación: 17/09/2014