Herramienta 1 con aplicación en el aula: *Linguakit*

Pablo Gamallo
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
pablo.gamallo@usc.es

1. Introducción

En el presente capítulo presentaremos la plataforma lingüística *Linguakit*, una caja de herramientas multilingüe diseñada para usuarios de la lengua, basada en un variado conjunto de módulos para el *procesamiento del lenguaje natural* (PLN).[1] El principal objetivo de la plataforma es permitir al usuario un acceso sencillo a módulos lingüísticos tecnológicamente complejos diseñados para explorar, analizar y obtener una mejor información de textos y documentos escritos. Esta plataforma multilingüe, que integra, entre otras herramientas lingüísticas, un resumidor, un analizador de sentimiento o un extractor de las palabras clave que dan sentido a un texto, va dirigida a un amplio abanico de usuarios que hacen de la lengua un uso profesional, educativo o general. *Linguakit* está, por lo tanto, pensado para que toda persona que posea interés lingüístico pueda sacarle el máximo partido a los documentos y textos de su ámbito profesional o educativo. El sistema es multilingüe ya que la mayoría de los módulos se aplican a cuatro lenguas: castellano, portugués, gallego e inglés.

En el ámbito lexicográfico, una plataforma como *Linguakit* puede ser de utilidad tanto para lexicógrafos como para usuarios de diccionarios en la búsqueda de contextos, colocaciones y relaciones entre palabras. En la consulta de una palabra, un usuario exigente, especialmente en el ámbito del aprendizaje de lenguas segundas como en el de la traducción, busca acceder a un historial lingüístico lo más completo posible de la palabra, que incluya una gran variedad de contextos y relaciones. En este sentido, Bowker (1998) observó que los estudiantes con acceso a corpus y herramientas de PLN conseguían mejores traducciones que aquellos con acceso exclusivo a diccionarios tradicionales. En un estudio similar, Montero y Faber (2008) llegaron a la conclusión de que los traductores hacen más búsquedas en corpus con herramientas PLN que

[1] *https://linguakit.com*

búsquedas en diccionarios tradicionales, una tendencia que podría extrapolarse a estudiantes de lenguas segundas. Dados los nuevos hábitos adquiridos por los diferentes tipos de usuarios de diccionarios, comienza a haber cada vez más recursos lexicográficos electrónicos que integran herramientas de PLN y de lingüística de corpus en sus sistemas de búsqueda, como la terminología creada por Castagnoli (2008), donde cada término es un enlace a un corpus y a un sistema de concordancias. Cabe destacar también el auge de los extractores de concordancias, como *WebCorp*[2] o *Monco*[3], que usan la Web como corpus y construyen los contextos de las palabras a partir de todos los documentos indexados por los principales buscadores.

Linguakit es una plataforma que no está, sin embargo, orientada a la búsqueda de información sobre una palabra, sino a la búsqueda y extracción de información a partir de textos o colecciones de documentos. Es el usuario el que, primero, escoge los textos y documentos sobre los que quiere realizar su estudio y, una vez introducidos en el sistema, selecciona la herramienta que necesita para buscar o extraer la información lingüística que más le interese sobre los textos de entrada.

El objetivo del presente capítulo es, por un lado, describir la arquitectura general de *Linguakit* así como sus módulos más relevantes y, por otro, sugerir y presentar tareas didácticas elaboradas a partir de los módulos descritos. Para ello, organizamos el capítulo en las siguientes secciones. En el siguiente apartado (sección 2), describiremos la arquitectura del sistema y su organización en grupos de módulos. En la sección 3 introduciremos con más detalle los módulos que consideramos más útiles para tareas lexicográficas y afines. A continuación (sección 4), describiremos algunos casos de uso que pueden resultar de utilidad en aulas de didáctica de lenguas, y acabaremos el capítulo (sección 5) esbozando las principales conclusiones del trabajo, con especial mención a las dificultades y problemas derivados del uso de este tipo de plataformas.

2. Arquitectura

La página de inicio de *Linguakit* ofrece una muestra de todas las posibilidades que tiene el usuario para trabajar con los textos. La plataforma presenta sus módulos lingüísticos organizados en cuatro apartados orientativos que representan diferentes casos de uso: (1) un primero que atiende a usos genéricos sin un perfil de usuario

[2] *http://www.webcorp.org.uk/*
[3] *http://monitorcorpus.com*

específico, como un conjugador o un resumidor automático. (2) Un segundo grupo de herramientas para un perfil de usuario más ligado al ámbito de la lingüística, con módulos como el etiquetador morfosintáctico, el analizador sintáctico o un sistema de concordancias (palabra de búsqueda en contexto). (3) Un tercer apartado pensado para profesionales de la comunicación y marketing como el analizador de sentimiento o el extractor de palabras clave y de multipalabras. (4) Por último, un apartado experimental donde se presentan nuevas herramientas todavía en proceso de desarrollo pero de más complejidad técnica que las anteriores, por ejemplo, el corrector lingüístico que identifica errores a diferentes niveles de análisis. En la tabla 1 enumeramos los diferentes módulos agrupados en los cuatro apartados mencionados. Todos ellos se basan en técnicas de procesamiento del lenguaje natural y han sido, en su mayoría, diseñados y desarrollados dentro del equipo de investigación ProLNat[4] de la Universidade de Santiago de Compostela.

Clasificación por grupos	Módulos
Módulos de uso genérico	• Procesamiento completo • Resumidor automático • Conjugador
Módulos de análisis	• Análisis de frecuencias • Etiquetador morfosintáctico • Analizador sintáctico • Concordancias
Módulos de extracción	• Extractor de opinión / sentimiento • Extractor de términos básicos (palabras clave) • Extractor de términos multipalabra • Reconocedor y clasificador de entidades
Módulos experimentales	• Corrector y evaluador de lengua • Extractor de tripletas semánticas • Extractor de conceptos y anotador semántico

Tabla 1: Organización de los módulos de *Linguakit*

[4] *http://gramatica.usc.es/pln/*

Cabe destacar el primer módulo citado en la tabla, *procesamiento completo*, que utiliza el resto de módulos y elabora un informe detallado a partir de los resultados obtenidos. Permite conocer el número de palabras y frases del texto, su riqueza léxica, las clases de palabras más frecuentas, ofrece un resumen de su contenido, así como el sentimiento (positivo, negativo o neutro) de ese resumen. Además, el informe también facilita las cinco palabras y multipalabras más relevantes del texto, las entidades más importantes que allí se mencionan, las palabras más frecuentes y el contexto en el que aparece la palabra clave escogida.

3. Módulos lingüísticos

Entre los diferentes módulos que ofrece la plataforma, a continuación describimos con más detalle aquellos que presentan más utilidad en el ámbito de la lexicografía, en sentido amplio, y en el aprendizaje de una lengua segunda.

3.1. Extracción de términos relevantes (básicos y multipalabra)

Se trata de dos módulos de extracción que realizan dos tareas similares que tienen en común la identificación y selección de los términos clave o relevantes de un texto. Los términos relevantes de un texto son las expresiones más importantes que se usan como claves o índices para la detección inmediata del tema o tópico, para el etiquetado automático o la clasificación documental. Con cada vez más frecuencia encontramos términos relevantes en los diarios digitales para destacar, con palabras de distinto tamaño, los temas tratados, por ejemplo, en un debate político. Por otro lado, los dos módulos de extracción se diferencian en el tipo de términos relevantes que extraen: básicos y multipalabra.

3.1.1. Términos básicos

Llamamos términos básicos a unidades léxicas relevantes para un texto que se codifican como nombres comunes, nombres propios, adjetivos y verbos. Excepto los nombres propios, que pueden ser expresiones compuestas pluriléxicas (p. ej. Nueva York, Universidad de Santiago de Compostela, etc.), los términos básicos son palabras simples monoléxicas. El método de extracción se lleva a cabo en dos fases. La primera

identifica todos los candidatos a ser términos básicos utilizando un etiquetador morfosintáctico (Garcia y Gamallo 2015). De esta manera, se seleccionan como candidatos todos las unidades léxicas que fueron etiquetadas como nombres (comunes y propios), adjetivos y verbos. En la segunda fase, los términos se ordenan por relevancia y se escogen los *N* primeros, siendo *N* un valor numérico parametrizable por el usuario. Para calcular la relevancia de los términos básicos recurrimos a la noción de *termhood*, es decir el grado en el que una unidad lingüística está relacionada con conceptos específicos del dominio (Kageura y Umino, 1996). Esta noción se puede ver también como la probabilidad de que un término forme parte del dominio.

El *termhood* no es, por lo tanto, una medida discreta sino continua. En consecuencia, medimos la relevancia de un término básico (*termhood*) por medio de un peso estadístico que se calcula contrastando las frecuencias de los candidatos en el texto dado (datos observados) y en un corpus de referencia (datos esperados). Más concretamente, el peso de un término es el valor chi-cuadrado que mide la divergencia entre los datos observados y los esperados. Estos últimos son los datos obtenidos a partir de un corpus de referencia de 100M de tamaño, construido en el seno del grupo, que abarca varios géneros y dominios: periodístico, técnico, literario, redes sociales, etc. Finalmente, los términos se ordenan por peso, de mayor a menor, y el usuario escoge los *N* más relevantes en función del tamaño del texto y de sus necesidades de estudio.

3.1.2. Términos multipalabra

Los términos multipalabra son expresiones relevantes codificadas como compuestos pluriléxicos que instancian patrones específicos de etiquetas morfosintácticas. Por ejemplo, *lenguaje natural*, *procesamiento del lenguaje*, *tecnologías de la lengua* o *analizador sintáctico* pueden ser multipalabras relevantes dentro un texto de dominio científico que verse sobre temas de PLN. Como en el caso de los términos básicos, el proceso de extracción de multipalabras se divide en dos fases: selección de candidatos y ordenación de los mismos por relevancia. Sin embargo, tanto la selección de candidatos como la ordenación se efectúan en base a diferentes estrategias. En la primera fase, utilizamos un conjunto de patrones de etiquetas (ver tabla 2) para identificar todas aquellas expresiones pluriléxicas que los instancien. Los artículos y determinantes de las expresiones no se toman en cuenta en la instanciación.

El conjunto fue diseñado para la identificación de multipalabras en cuatro lenguas: castellano, portugués, gallego e inglés. Este método es similar al de otros sistemas de extracción terminológica (Vivaldi y Rodríguez, 2001; Sánchez y Moreno, 2006).

Conjunto de patrones de etiquetas
nombre-adjetivo
adjetivo-nombre
nombre-nombre
nombre-preposición-nombre
nombre-preposición-adjetivo-nombre
nombre-preposición-nombre-adjetivo
adjetivo-nombre-preposición-nombre
nombre-adjetivo-preposición-nombre
adjetivo-nombre-preposición-nombre-adjetivo
nombre-adjetivo-preposición-nombre-adjetivo
adjetivo-nombre-preposición-adjetivo-nombre
nombre-adjetivo-preposición-adjetivo-nombre

Tabla 2: Conjunto de patrones (sintagmas nominales) utilizado en la identificación de candidatos a multipalabra.

En la segunda fase, la ordenación por relevancia, utilizamos una estrategia diferente a la empleada en la ordenación de términos básicos. Mientras que estos se ordenan en función de la noción de *termhood*, la relevancia de las expresiones multipalabra se define por medio del concepto de *unithood*. Esta noción hace referencia a cómo asociamos o no secuencias de palabras con unidades léxicas estables. Más formalmente, *unithood* se refiere al grado de fuerza y cohesión entre las unidades léxicas que constituyen los sintagmas y colocaciones (Kageura y Umino, 1996). La

unithood solo se aplica, por tanto, a unidades pluriléxicas con cierta cohesión interna y no a unidades monoléxicas.

Peso	Multipalabra	Patrón de etiquetas
9.95	dación en pago	*nombre - preposición - nombre*
7.94	viviendas vacías	*nombre - adjetivo*
7.27	renta básica	*nombre - adjetivo*
5.27	iniciativas legislativas	*nombre - adjetivo*
2.99	reuniones de representantes	*nombre - preposición - nombre*

Tabla 3: Las cinco multipalabras más relevantes (*unithood*) extraídas del programa electoral de Podemos para las elecciones del 20D/2015.

Peso	Multipalabra	Patrón de etiquetas
20.37	inversores extranjeros	*nombre - adjetivo*
11.44	creación de empleo	*nombre - preposición - nombre*
9.75	competitividad de economía	*nombre - preposición - nombre*
7.73	reducción de impuestos	*nombre - preposición - nombre*
2.93	ciudadanos españoles	*nombre - adjetivo*

Tabla 4: Las cinco multipalabras más relevantes (*unithood*) extraídas del programa electoral del Partido Popular para las elecciones del 20D/2015.

El grado de cohesión o *unithood* se puede calcular con diferentes medidas estadísticas. En nuestro modelo, hemos vuelto a usar el chi-cuadrado, el cual se interpreta en este contexto como un test de asociación entre los constituyentes de la multipalabra. Concretamente, este coeficiente se aplica para verificar si los constituyentes coocurren en un mismo sintagma aleatoriamente o por atracción. Así, los valores observados se corresponden con la frecuencia de la multipalabra en el texto de entrada, mientras que los esperados se calculan a partir de las frecuencias de los

constituyentes por separado. Las tablas 3 y 4 muestran dos ejemplos reales de extracción de multipalabras de dos programas de partidos políticos españoles.

3.1.3. Extractor de conceptos

Además de los dos módulos de extracción terminológica, *Linguakit* contiene un módulo experimental que utiliza la extracción terminológica para enlazar los térnimos mencionados en el texto con conceptos indexados en una base externa enciclopédica, concretamente la DBpedia[5]. Este método se conoce como "enlace de entidades" (*entity linking*), y cuenta con numeros trabajos a caballo entre PLN y la Web Semántica, dos líneas de investigación complementarias (Mendes et al., 2011; Hachey et al., 2013). El enlace de entidades es una operación con mayor complejidad que la extracción terminológica, pues para enlazar un término del texto de entrada con un concepto del repositorio externo es necesario desambiguar el sentido del término previamente. Por ejemplo, supongamos que el extractor selecciona el término *PP* dentro del texto de entrada. Este término es ambiguo ya que se puede enlazar a dos conceptos: el *Partido Popular español* y el *Partido Progresista de Brasil*. Para poder realizar correctamente el enlace, es necesario tomar en cuenta el contexto en el que se encuentra el término. El contexto determina la interpretación del término y, por consiguiente, su correcto enlace con el concepto que denota en ese contexto. El concepto enlazado no es más que la entrada del recurso conceptual/enciclopédico que el sistema usa para relacionar términos mencionados con conceptos preestablecidos.[6]

3.2. Corrector y evaluador

El módulo de corrección y evaluación automática ha sido desarrollado en su fase experimental en lengua gallega. Es un sistema que analiza el texto buscando errores ortográficos, léxicos, gramaticales, sintácticos o de estilo, y ofrece información sobre la clase de error localizado, sus posibles soluciones y una explicación detallada del fenómeno lingüístico subyacente. La investigación en corrección automática ha crecido en popularidad e interés en los últimos años como lo muestra la organización de numerosos congresos, workshops y *Shared Tasks*, centrados en el diseño y evaluación

[5] *http://wiki.dbpedia.org/*

[6] El extractor de conceptos es un módulo experimental que todavía no se ha integrado en *Linguakit*, pero que está disponible en la siguiente dirección web: *http://fegalaz.usc.es/~gamallo/demos/semantic-demo/*

de sistemas corrección automática de textos en inglés escritos por aprendices no nativos. Entre ellos, destacamos los siguientes: Helping Our Own (HOO-2011-2012) (Dale y Kilgarriff, 2010; Dale y Kilgarriff, 2011; Dale et al., 2012), *Conference on Computational Natural Language Learning* (CoNLL-2013, CoNLL-2014) (Ng et al., 2013).

En cuanto a la arquitectura, nuestro sistema utiliza herramientas de PLN que facilitan el trabajo de identificación de errores lingüísticos (identificador de entidades, etiquetador morfosintáctico, analizador sintáctico...). Y posee también los recursos lingüísticos necesarios para una corrección y evaluación automáticas: léxicos de formas y lemas, lista de errores léxicos y neologismos, reglas de errores gramaticales, corpus de entrenamiento, etc. La lista de errores léxicos contiene, entre otros, falsos amigos y extranjerismos.

El sistema de corrección y evaluación está pensado para aplicarse en dos contextos propios del aprendizaje de lenguas: el individual o (autoaprendizaje) y el colectivo. En el plano individual, permite que el estudiante visualice, interiorice y tome consciencia de los errores de una manera autónoma, sin esperar la corrección del profesor, pudiendo así buscar las estrategias de superación del error más eficaces y progresando a un ritmo que se adapte a sus necesidades. Se trata de un método de autoaprendizaje basado en el *feedback* correctivo. En el plano colectivo, el sistema minimiza el tiempo de corrección y evaluación puramente lingüísticas. El profesor puede conocer rápidamente los tipos de error más frecuentes de sus alumnos, pudiendo así dedicar más tiempo a la preparación de material y actividades adecuadas para trabajar los casos más problemáticos. Por último, una institución u organización estatal puede seleccionar una muestra de centros escolares representativos, y conocer sin grandes inversiones materiales y humanas el nivel de lengua de la población de estudiantes objeto de la evaluación.

En cuanto al *feedback* correctivo, que es la base del modelo de aprendizaje subyacente a este tipo de sistemas, se debe indicar que existen críticas de autores que no ven claro su papel en el aprendizaje de una segunda lengua (Truscott, 1996, Truscott & Hsu 2008, Liu 2008, Hartshorn et al. 2010). Sin embargo, numerosos investigadores como Ferris (1999), Hyland y Hyland (2006) o Russell y Spada (2006) argumentan claramente a favor del uso de la toma de consciencia del error en el aprendizaje de una segunda lengua, aunque dejando claro que no es el único factor responsable en la

mejora de la competencia lingüística en el proceso de aprendizaje. Se trata de una ayuda importante pero no decisiva. Lo mismo debemos decir desde el punto de vista del docente. La corrección del error no es más que un elemento en el proceso de evaluación así como un factor más a tomar en cuenta a la hora de planificar los objetivos de aprendizaje.

La figura 1 muestra un ejemplo de corrección de una frase en gallego que contiene 8 errores, pero solo uno de ellos detectable por un corrector ortográfico (*spell checker*) tradicional.

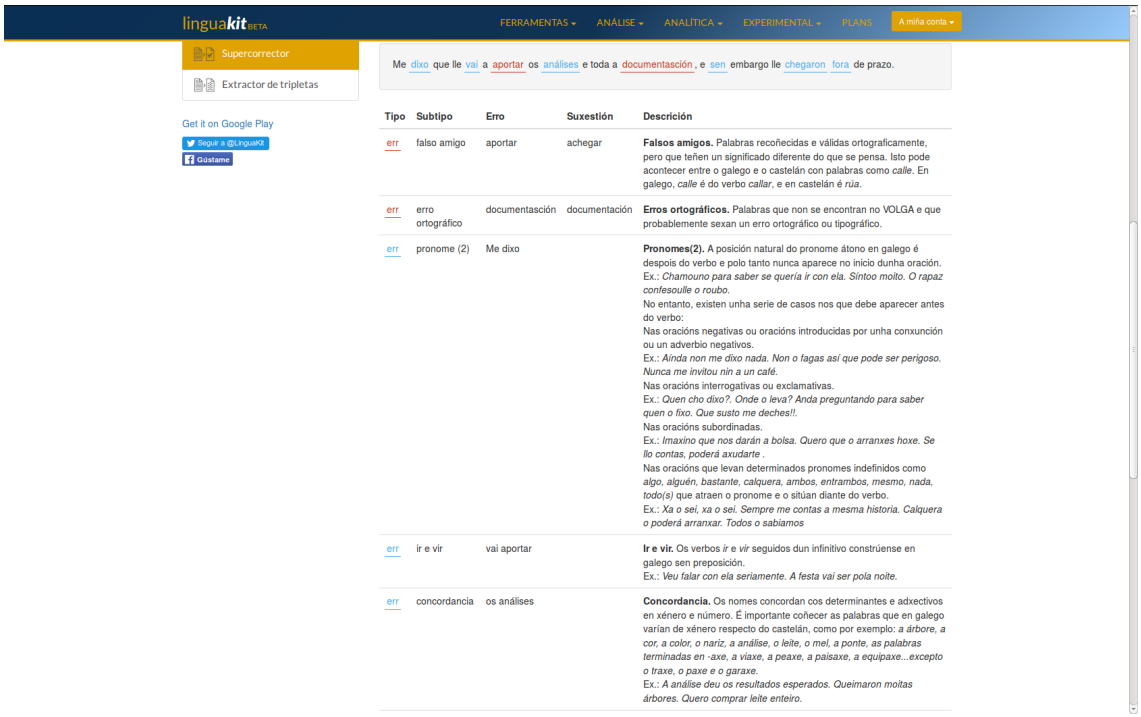

Figura 1: Ejemplo de corrección de una secuencia en gallego con ocho errores ortográficos, léxicos y gramaticales.

La secuencia analizada que muestra la figura es la siguiente:

> *Me dixo que lle vai a aportar os análises e toda a documentasción, e sen embargo lle chegaron fora de prazo.*[7]

El único error detectable por un corrector tradicional es *documentasción*, ya que se trata de una palabra mal escrita que no se encuentra en el diccionario. El resto de errores son

[7] En castellano: *Me dijo que le va a aportar los análisis y toda la documentación, y sin embargo le llegaron fuera de plazo*.

gramaticales (fallos de concordancia: *os análises*), sintácticos (posición del clítico: *Me dixo*, *lle chegaron*), o léxicos. Entre estos últimos, consideramos diferentes subtipos: falsos amigos (*aportar* en vez de *achegar*), castellanismos (*sen embargo*) o acentos diacríticos (*fora* como adverbio se escribe *fóra*). En todos estos casos, las palabras o expresiones mal formuladas se encuentran en cualquier diccionario gallego. Para detectar la mayoría de estos errores, invisibles para los correctores automáticos tradicionales, es necesario apoyarse en un analizador sintáctico y en recursos léxicos especiales que tengan en cuenta fenómenos como los falsos amigos. Por último, las explicaciones que acompañan a cada tipo de error ayudan al estudiante a procesar con más eficacia el *feedback* correctivo.

4. Casos de uso

Linguakit es una caja de herramientas lingüísticas especialmente útil para preparar material didáctico en clases de lengua. Sin embargo, sus aplicaciones son muy diversas, pudiendo servir en ámbitos tan diversos como la traducción o las ciencias de la información. Un ejemplo interesante de uso de nuestra herramienta en el ámbito periodístico es el artículo firmado por el periodista Alberto Quian en Galicia Confidencial, donde utiliza varios módulos de *Linguakit*, concretamente extractores de términos, para comparar el uso del lenguaje en los programas de dos partidos políticos de Galicia durante las elecciones del 20 de diciembre de 2015.[8] No obstante, no es fácil separar los diferentes ámbitos de aplicación, ya que trabajos como el citado pueden ser perfectamente actividades lingüísticas para mejorar las competencias de aprendices de lengua segunda de niveles avanzados, donde los conocimientos de cultura, sociedad y lengua se entrecruzan e interrelacionan muy profundamente. A continuación, vamos a describir dos actividades diferentes, una de ellas basada en el extractor de expresiones multipalabra, y la otra en el corrector lingüístico.

4.1. Uso del extractor multipalabra

Una actividad de lengua y cultura avanzada podría ser la siguiente: obtener los principales contextos en los que aparece la palabra Galicia en diferentes periódicos gallegos y españoles y analizar las diferencias, si las hubiese. A modo de ejemplo,

[8] *http://www.galiciaconfidencial.com/noticia/27170-son-galiza-galicia-marea*

juntamos numerosas noticias publicadas *on-line* en tres periódicos: La Voz de Galicia[9], Portal Galego da Língua[10] y ABC[11]. Las noticias de cada periódico se agregaron en un único fichero con aproximadamente 10M de tamaño. A cada fichero se le aplicó el extractor de términos multipalabra, se identificaron los términos que contenían la secuencia *Galicia* (o *Galiza* en el caso del Portal Galego da Língua) y se seleccionaron los 15 términos con más peso. El resultado se muestra en la tabla 5.

Portal Galego da Língua[12]	La Voz de Galicia	ABC
semanário Sermos Galiza	presidente de Augas de Galicia	litoral de Galicia
português em Galiza	Plaza de Galicia	mejillón de Galicia
Galiza contemporânea	avenida de Galicia	procede de Galicia
língua em Galiza	capital de Galicia	cultivado en Galicia
português de Galiza	zona Galicia-Costa	puntos de Galicia
língua de Galiza	presidente de Xunta de Galicia	presidentes de Galicia
trabalho em Galiza	Volta a Galicia	interior de Galicia
dia de Galiza Mártir	responsable de Augas de Galicia	chefs de Galicia
Sempre em Galiza	parte de Galicia	oeste de Galicia
Galiza sem petróleo	sur de Galicia	precipitaciones en Galicia
invasom de Galiza	predicción de Meteo Galicia	lluvias en Galicia
dissolução de Junta de Galiza	norte de Galicia	costa de Galicia
Made in Galiza	Delegación del Gobierno en Galicia	carretera en Galicia
libertaçom de Galiza	ciudades de Galicia	Policlínico Lucense en Galicia
Galiza barroca	bueyes de Galicia	metros en Galicia

Tabla 5: Extracción de las 15 expresiones multipalabra más representativas que contienen la secuencia Galicia/Galiza en tres periódicos diferentes.

Un análisis superficial de los resultados permite entrever que el Portal Galego da Língua relaciona el término *Galiza* (*Galicia*) con asuntos culturales, lingüísticos, económicos e históricos. En contraste, La Voz de Galicia y el ABC resaltan otros aspectos y facetas del término: lugares, geografía, autoridades y clima. De hecho, ambos periódicos se asemejan mucho en su aproximación al término *Galicia* al presentar como más representativos contextos muy similares. La principal diferencia entre estos dos periódicos es que La Voz de Galicia destaca un contexto de carácter deportivo (*Volta a*

[9] *http://www.lavozdegalicia.es/*
[10] *http://pgl.gal/*
[11] *http://www.abc.es/*
[12] La lengua escrita de esta publicación es el gallego en norma AGAL (cercana al portugués).

Galicia) mientras que el ABC destaca la faceta gastronómica en varios contextos: *mejillones de Galicia* y *chefs de Galicia*.

4.2. Uso del corrector lingüístico para el gallego

El módulo de corrección solo se aplica a la lengua gallega dadas las dificultades y el elevado coste humano que supone diseñar y crear los recursos necesarios (p. ej. gramática y léxico de errores) para las otras lenguas. Además de usarse para monitorizar el avance de los aprendices en gallego, el sistema de corrección puede también servir para diseñar actividades didácticas de diferentes tipos. Por ejemplo, se puede analizar el tipo de errores que cometen las administraciones locales y autonómicas o el sector de la comunicación en su uso del gallego. La lengua gallega se encuentra todavía en proceso de normalización y su uso, incluso en contextos formales, da lugar a innumerables interferencias con el castellano, entre otros problemas. Un ejercicio concreto que vamos a describir a continuación consiste en analizar los problemas lingüísticos más frecuentes que se encuentran en artículos periodísticos escritos en gallego en la prensa gallega que escribe habitualmente en castellano. Concretamente, hemos aplicado el corrector sobre los primeros 47 artículos escritos en gallego del Correo Gallego[13], publicados en enero de 2016. El sistema encontró 113 errores, de los cuales 93 eran claramente lingüísticos (faltas de ortografía, problemas léxicos o gramaticales) y 20 simples confusiones tipográficas (letra duplicada, intercambio de letras, etc.). Aunque de media, hemos detectado 1 error por cada 11 líneas, en algunos casos hemos encontrado hasta 4 problemas lingüísticos en una única frase, como la que mostramos a continuación:

> *Non faltaba **sen embargo** quen coidaba, con toda razón, que este ano o belén do Obradoiro **cambiábase** por un **tiovivo** cuberto pola noite con plásticos e **mais** por unhas árbores encantadas, do bosque máxico.*
>
> | | |
> |---|---|
> | Error léxico (castellanismo): ***sen embargo*** | Corrección: ***no entanto*.** |
> | Error sintáctico: ***cambiábase*** | Corrección: ***se cambiaba*.** |
> | Error léxico (castellanismo)*:* **tiovivo** | Corrección: **carrusel.** |
> | Error ortográfico: **mais** | Corrección: **máis.** |

Solo uno de estos cuatro errores se puede detectar con un corrector automático tradicional, concretamente *tiovivo*, que no se encuentra en ningún diccionario gallego.

[13] http://www.elcorreogallego.es/

Sin embargo, nuestro sistema ofrece más información sobre este caso, ya que clasifica el error como un castellanismo y ofrece su equivalente en gallego (*carrusel*), soluciones que no puede aportar un corrector tradicional.

El principal interés de esta actividad es, por un lado, mostrar la situación real del gallego escrito, el cual se encuentra en un estado de debilidad crónica y fosilización incluso a niveles de uso formales y, por otro, resaltar que los tipos de errores que emergen fuera del contexto del aula son parecidos a los que se comenten en el aula. El objetivo es motivar a las estudiantes para que tomen consciencia de que una mejora de su nivel de lengua, no solo es una mejora de sus competencias individuales, sino también, en el caso del gallego, de la competencia colectiva de la lengua y de su prestigio a nivel social.

5. Discusión

Linguakit es un portal con un paquete de módulos y herramientas de análisis lingüístico y de extracción textual para poder explorar, analizar y obtener una mejor información de los textos y documentos escritos. En este capítulo, hemos descrito los módulos que pueden resultar más útiles a estudiantes y profesores de lengua, proporcionando algunos ejemplos de actividades didácticas que fácilmente pueden llevarse a cabo en un aula de enseñanza de una lengua segunda.

A pesar del potencial de *Linguakit* y de los aspectos positivos que hemos destacado a lo largo del capítulo, este tipo de herramientas también presenta ciertos problemas que es preciso tomar en cuenta y no obviar. Uno de ellos es la baja calidad lingüística de algunos módulos. Al tratarse de procesos automáticos, la precisión y cobertura de los sistemas de PLN en los que se basan la mayoría de estos módulos está lejos de ser del cien por cien. Los usuarios deben tener en cuenta estas limitaciones a la hora de valorar los resultados del análisis y de la extracción. Otro problema de gran calado es la baja eficiencia computacional del sistema. El tamaño de los textos de entrada está limitado a un número de caracteres que puede ser demasiado restrictivo para determinadas tareas. La solución a este problema vendrá de la mano del uso de tecnologías *Big Data* que permitan la paralelización de los procesos y, por consiguiente, la posibilidad de procesar más texto en menos tiempo. No hay que desdeñar tampoco las cuestiones relativas a la ergonomía y al diseño de las interfaces gráficas, que se podrán

ir mejorando con el *feedback* de los usuarios a través de sus experiencias y comentarios. Finalmente, una problemática que no podemos ignorar es la búsqueda de nuevas funcionalidades y de nuevos módulos que cubran esas nuevas necesidades. El conjunto de herramientas disponible, si bien es amplio y numeroso, está muy lejos de cubrir todos los requerimientos de todos los profesionales y usuarios de la lengua. Una posible ampliación de *Linguakit* podría realizarse en dos sentidos complementarios: (1) diseño y creación de nuevos módulos, por ejemplo, análisis del discurso o generación del lenguaje; (2) combinación de diferentes módulos ya existentes para construir una nueva funcionalidad, por ejemplo, utilizar el analizador y el sistema de concordancias para diseñar un buscador de contextos léxico-sintácticos de una palabra: ¿con qué sujetos aparece el verbo *dimitir*?, ¿qué adjetivos modifican al sustantivo *español*?, etc.

Por último, es necesario reflexionar con más profundidad para qué y para quién puede ser práctico usar herramientas de PLN como las que se integran en *Linguakit*. Conviene definir mejor los objetivos y contenidos que se pueden asociar a cada perfil de usuario, ya sea lingüista, lexicógrafo, periodista, editor, gestor, analista, publicista, o bien cualquier usuario en general con interés en explorar una lengua a través de textos, documentos o libros que le interesen.

Referencias bibliográficas

Bowker, L. (1998): «Using Specialized Monolingual Native-Language Corpora as a Translation Resource», Meta*,* 43(4), pág. 631-651.

Castagnoli, S. (2008): «Corpus et bases de données terminologiques: l'interpretation au service des usagers», en F. Maniez, P. Dury, N. Arlin y C. Rougemont (coords.), *Corpus et dictionnaires de langues de spécialité,*Bresson: Presses Universitaires de Grenoble, pág. 213-230.

Dale, R. y A. Kilgarriff (2010): «Helping Our Own: Text Massaging for Computational Linguistics as a New Shared Task», en *Proceedings of the 6th International Natural Language Generation Conference (NLG'10)*, pág. 263-267.

Dale, R., I. Anisimoff y G. Narroway (2012): «HOO 2012: A report on the preposition and determiner error correction shared task», en *Proceedings of the Seventh*

Workshop on Innovative Use of NLP for Building Educational Applications, Montréal, Québec, Canada, pág. 54-62.

Dale, R. y A. Kilgarriff (2011). «Helping Our Own: The HOO 2011 Pilot Shared Task», en Belz, A., Evans, R., Gatt, A. and K. Striegnitz (coords.), *Proceedings of the 13th European Workshop on Natural Language Generation (NLG'11),* Nancy, France, pág. 242-249.

Ferris, D. (1999): «The case for grammar correction in L2 writing classes: a response to Truscott», *Journal of Second Language Writing*, 8, pág. 1-11.

Garcia, M. y Gamallo, P. (2015): «Yet another suite of multilingual NLP tools», en José-Luis Sierra-Rodríguez, José Paulo Leal and Alberto Simões (coords.), *Languages, Applications and Technologies,* Communications in Computer and Information Science, 563. Switzerland: Springer, pág. 65-75.

Hachey, Ben, Will Radford, Joel Nothman, Matthew Honnibal y James R. Curran (2013): «Evaluating entity linking with wikipedia», *Artificial Intelligence*, 194. pág. 130-150.

Hartshorn, K. J., N. Evans, P.F. Merrill, R.R. Sudweeks, D. Strong-Krause y N.J. Anderson, (2010): «Effects of dynamic corrective feedback on ESL wiring accuracy», *TESOL Quarterly,* 44*,* pág. 84-109.

Hyland, K. y F. Hyland (2006): «State of the art article: Feedback on Second Language students' writing», *Language Teaching* 39, (2), pág. 83-101.

Kageura, Kyo y Bin Umino (1996): «Methods of automatic term recognition: A review», *Terminology*, 3(1), pág. 259-289.

Liu, Y. (2008): «The effects of error feedback in second language writing», *Arizona working papers in SLA & Teaching* 15, pág. 65-79.

Mendes, Pablo, Max Jakob, Andrés García-Silva y Christian Bizer (2011) «Dbpedia spotlight: Shedding light on the web of documents», en *7th International Conference on Semantic Systems*, pág. 1-8.

Montero, S. y P. Faber (2008): *Terminología para traductores e intérpretes*, Granada: Tragacanto.

Ng, H., S. Wu, Y. Wu, C. Hadiwinoto y J. Tetreault (2013). The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task (CoNLL-2013 Shared Task)*, Sofia, Bulgaria, pág. 1-14.

Russell, J., y N. Spada (2006): «The effectiveness of corrective feedback for the acquisition of L2 grammar. A metaanalysis of the research», en J. M. Norris y L. Ortega (coords.) *Synthesizing research on language learning and teaching*, Amsterdam/Philadelphia: John Benjamins, pág. 133-164.

Sánchez, David y Antonio Moreno (2006): «A methodology for knowledge acquisition from the web», *Journal of Knowledge-Based and Intelligent Engineering Systems*, 10(6), pág. 453-475.

Truscott, J., y A. Y. Hsu (2008): «Error correction, revision, and learning», *Journal of Second Language Writing* 17, pág. 292-305.

Truscott, J. (1996): «The case against grammar correction in L2 writing classes», *Language Learning* 46, pág. 327-369.

Vivaldi, Jordi y Horacio Rodríguez (2001): «Improving term extraction by combining different techniques», *Terminology*, 7(1), pág. 31-47.