El uso del lenguaje en los diferentes canales de Internet

RESUMEN: *El estilo discursivo es un reflejo de la personalidad del sujeto que lo elabora. La elección de las palabras y el modo en que se combinan, aporta información de dimensiones como el género, la edad e incluso el estado emocional de quién las emite. Pero en Comunicación 2.0 nos enfrentamos a gran variedad de canales y surge la pregunta, ¿define el canal el modo en que se usa el lenguaje? Hemos analizado más de 30 millones de documentos de la Wikipedia, Prensa, Blogs, Foros, Twitter y Facebook, y presentamos cómo se utiliza el lenguaje en cada uno de ellos.*

PALABRAS CLAVE: Palabras de contenido, palabras de estilo, canales de internet, categoría gramatical, frecuencia de palabras

Francisco Manuel Rangel Pardo
Director de Investigación y Desarrollo Tecnológico / Doctorando
Autoritas Consulting SA / Universitat Politècnica de València
C/ Lorenzo Solano Tendero, 7 Madrid
francisco.rangel@autoritas.es

Paolo Rosso
Responsable del Laboratorio de Ingeniería del Lenguaje Natural del grupo de Investigación EliRF
Universitat Politècnica de València
Camino de Vera S/N, Valencia
prosso@dsic.upv.es

Francisco Manuel Rangel Pardo es Director Tecnológico de Autoritas, Ingeniero Superior en Informática e Ingeniero Técnico en Telecomunicaciones especialidad Telemática, Máster en Lingüística Computacional y actual doctorando en Author Profiling, especializado en recuperación y análisis de información de la Web. Premio MAVIR 2007 al mejor trabajo de investigación en procesamiento de lenguaje. Especializa su carrera como Director de I+D en tecnologías lingüísticas y aprendizaje automático, con el objetivo de añadir valor obteniendo conocimiento a partir del tratamiento automatizado de datos, estructurados o desestructurados como la Web. Organizador principal de la tarea Author Profiling del PAN 2013 (Evaluation Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse) en el marco de la conferencia CLEF (Conferences and Labs of the Evaluation Forum) que tiene como objetivo principal la inferencia de la edad y el género de los autores a partir de sus textos escritos http://pan.webis.de

Paolo Rosso es doctor en Computer Science (1999) por el Trinity College de Dublin, Universidad de Irlanda. Actualmente es Profesor Titular en la Universitat de València, España, dónde dirige el Laboratorio de Ingeniería de Lenguaje Natural del grupo de investigación ELiRF. Ha publicado aproximadamente 250 papers en conferencias, talleres y revistas; ha estado involucrado en numerosos proyectos de investigación nacionales e internacionales. Sus intereses de investigación se centran principalmente en detección de plagio, detección de ironía en medios sociales y análisis de textos cortos. Es uno de los organizadores de las actividades del PAN (Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse) en los foros internacionales de evaluación del CLEF (Conference and Labs of the Evaluation Forum): http://pan.webis.de y del FIRE (Forum for Information Retrieval Evaluation): http://www.dsic.upv.es/grupos/nle/clinss.html

El uso del lenguaje en los diferentes canales de Internet

Introducción

Año 2004, campaña presidencial Americana, el presidente George W. Bush se enfrenta a John Kerry, quien representa una seria amenaza especialmente por la cada vez más decadente imagen de Bush en las encuestas de popularidad. Sin embargo, cuando Kerry habla se le percibe distante e incluso arrogante. Su lenguaje corporal es rígido y sus discursos parecen forzados. Los asesores de Kerry trabajan con él para hacerle parecer más cercano, para lo que le hacen emplear cuando habla más la primera persona del plural, (nosotros, nos, nuestro) y menos la primera persona del singular (yo, mi, mío). Craso error.

Cuando los políticos usan la primera persona del plural suenan fríos y distantes, en cambio la primera persona del singular se asocia con la honestidad y lo personal[17][7]. Kerry estaba en un grave problema, estaba usando el doble de veces la primera persona del plural que su rival Bush, y en cambio solo estaba usando la mitad de veces la primera persona del singular.

Lo anterior no es más que una anécdota de la importancia que pueden llegar a tener algunas palabras, palabras vacías de contenido, que por lo general son invisibles y nuestro cerebro procesa sin apenas percibirlo, pero que en cambio producen estados mentales y estereotipos que pueden resultar en percepciones diferentes a la esperada. Estas son las palabras de función, palabras que por contra a las palabras de contenido, no aportan información al discurso pero en cambio permiten organizarlo, conectarlo y dotarlo de estilo.

Este estilo discursivo es un reflejo de la propia personalidad del sujeto que lo elabora, quien decide, de manera inconsciente por lo general, construirlo de una manera y no de otra. Es decir, la elección de las palabras y el modo en que se combinan viene a aportar información relevante sobre la personalidad de quién las emite.

Conocer la personalidad de los sujetos y cómo esa personalidad se manifiesta a través de sus palabras es clave, como se ilustraba con la anécdota del principio. Conocer la personalidad de los individuos es una de

las tareas indispensables en comunicación, y está tomando gran relevancia dentro de la emergente disciplina conocida como Escucha Activa.

Internet y la Web 2.0 con su gran variedad de canales y la gran cantidad de *prosumers* que la componen, proporciona una fuente inmensa de conocimiento y oportunidades para las empresas y organizaciones, pero en muchas ocasiones la información que los autores proporcionan, principalmente cuando de su perfil se refiere, suele ser engañosa o falsa; su edad, su género, su profesión, incluso su nombre, puede estar ausente, o aún estando presente, ser diferente a la realidad, siendo fiable únicamente aquello que podamos inferir a partir de lo que escriben y de cómo lo escriben sus autores.

Varios son los estudios que infieren dimensiones de la personalidad a partir de textos escritos, dónde encontramos apreciaciones significativas como:

- Género [17][3][10][16]: Los hombres usan los artículos en tasas superiores a las mujeres. Las mujeres tienden a ser más conscientes de sí mismas y más auto-centradas que los hombres. Las mujeres tienden a usar más el nosotros cálido y los hombres el nosotros distante. Los hombres hablan más que las mujeres sobre objetos y cosas. Por contra, las mujeres piensan y hablan más sobre otras personas. Los hombres categorizan su mundo contando, etiquetando y organizando los objetos con los que se encuentran, lo que suele implicar un mayor uso de las preposiciones, que categorizan el mundo de una manera jerárquica y espacial. Las mujeres personalizan sus cosas, hablan de una manera más dinámica, enfocándose en cómo sus cosas cambian, más que en sus cosas en sí.

- Edad [17][3][13]: Los grupos de edad usan las emociones y las palabras de función de manera muy diferente. Los adolescentes usan más los pronombres personales como yo y tú, las palabras cortas y los verbos auxiliares. Los mayores usan palabras más largas, así como más preposiciones y artículos. Según nos hacemos mayores, cambiamos el modo de orientarnos a nuestros amigos y familiares, al sexo, al dinero, la salud, la muerte, entre otras muchas cosas. Así pues, respecto a las emociones, las personas más mayores suelen utilizar palabras que representan emociones positivas, frente a los jóvenes que suelen expresar sentimientos más negativos.

- Clase Social [17]: Quien usa más artículos suelen ser más concienzudos, políticamente conservadores y mayores.

- Estado Emocional [17][11]: Quien usa artículos en altas tasas tienden a ser más organizados y emocionalmente estables.

- Idioma Nativo [17][12]: Es muy difícil cambiar el uso y perfeccionar el modo en que utilizamos las palabras de estilo a partir de los 12 años de edad. Por ello, las personas que de adultos aprenden y utilizan un segundo idioma, aunque lo perfeccionen a tal nivel de no cometer errores en lo que cuentan, suelen hacerlo en el uso de palabras de estilo como las preposiciones, conjunciones y otros marcadores de estilo discursivo.

Otras dimensiones que pueden inferirse de la escritura de los individuos son su nivel de honestidad, tipo de personalidad, grado de formalidad, habilidad de liderazgo, calidad en las relaciones, etcétera[17]. Pero la mayoría de estos estudios han sido realizados para el idioma inglés y limitados a un solo canal, generalmente textos literarios, y más recientemente blogs, surgiendo entonces la pregunta, ¿se utiliza el mismo estilo discursivo en los diferentes canales de Internet?, es decir, ¿podemos aplicar estudios de lingüística computacional como los anteriores de igual manera en Facebook que en Twitter, en la Wikipedia que en un Foro, en un Blog que en un artículo de Prensa?, ¿y podemos aplicar tales estudios al español? Porque no se usan lo mismo los pronombres personales en inglés que en español, y son una de las palabras de función que más interés suscitan [17][8][6]

En el artículo se presenta un estudio sobre el uso de las palabras en diferentes canales de Internet en idioma español, pretendiendo servir de base para posteriores estudios en la extracción de estilos discursivos y la obtención de perfiles de usuario a partir de textos escritos en Internet en español.

En el apartado segundo se proporciona el marco teórico de referencia sobre el que se ha basado el estudio, en el apartado tercero se describe la metodología de investigación que se complementa con el apartado cuarto dónde se presenta el conjunto de datos sobre el que se ha experimentado. En el quinto apartado se muestran y discuten los resultados experimentales obtenidos y se finaliza con las conclusiones y el trabajo futuro en el sexto apartado.

Marco teórico de referencia

Dos son las cuestiones básicas a responder cuando nos enfrentamos a un discurso, QUÉ decir y CÓMO decirlo. La primera cuestión responde al objeto que queremos comunicar y viene a definir el contenido del discurso, de qué trata el mismo. No es una cuestión personal de quién comunica sino una cuestión referida a lo que va a comunicar. La segunda cuestión por su parte responde a la manera en que se va a comunicar dicho contenido, a cómo se va a articular el discurso y es por lo tanto una cuestión inherente al comunicador, a su forma de comunicar, y esta forma de comunicar viene determinada en esencia por su personalidad. Desde el punto de vista del receptor, las preguntas se convierten en QUÉ se dice, y QUIÉN lo dice.

Partiendo de lo anterior se pueden dividir las palabras en dos tipos: aquellas orientadas a comunicar contenidos, a responder al QUÉ, y aquellas orientadas a conectar el discurso, a darle forma y estilo, a responder al CÓMO-QUIÉN.

Las palabras de contenido son aquellas que tienen un significado, generalmente un significado compartido entre diferentes culturas porque nombran a un objeto o a una acción. Dentro de las palabras de contenido se encuentran las categorías gramaticales correspondientes a los sustantivos (perro, mesa, niño, lápiz), a los verbos regulares y de acción (comer, decir, dar), a los adjetivos calificativos (rojo, alto, grande, bonito) y a muchos adverbios (aquí, pronto, bien, poco).

Por su parte, las palabras de estilo son palabras que conectan, dan forma y estilo y organizan el discurso. No aportan información, no dicen nada sobre ningún objeto del mundo real ni sobre ninguna acción. Se corresponden con las palabras de las categorías gramaticales de los pronombres (yo, tú, él), los artículos (el, la, los, las), las preposiciones (a, ante, bajo, cabe), los verbos auxiliares (haber, ser), las negaciones (no, ni), las conjunciones (y, o), los cuantificadores (un, uno, dos, muchos) y otros muchos adverbios. Sus características más distintivas son [17][6]:

- Son muy frecuentes
- Son cortas y difíciles de detectar
- Son muy, muy sociales
- Se procesan por el cerebro de manera diferente a las palabras de contenido

A finales del siglo XIX y como resultado de diversos estudios sobre la afasia, el neurólogo y psiquiatra alemán Karl Wernicke y el médico, anatomista y antropólogo francés Paul Pierre Broca definieron dos áreas del cerebro

involucradas en la comprensión y producción del habla. Estas son el Área de Broca [14] y el Área de Wernicke [19], llamadas así en su honor.

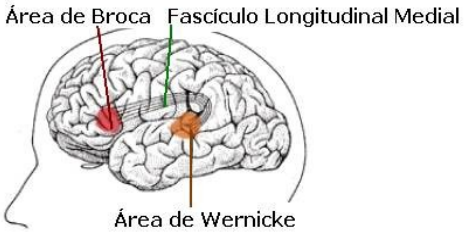

Fig. 1. Áreas involucradas en el procesamiento de la lengua

El área de Broca se encuentra en la tercera circunvolución frontal inferior, en el lóbulo frontal del hemisferio izquierdo del cerebro, que para la mayoría de personas es el hemisferio dominante para el lenguaje. Controla gran parte de las capacidades sociales de la persona, procesa la gramática, está involucrada en la producción del habla, en el procesamiento del lenguaje y en la comprensión. Controla la habilidad para expresarse y conciliar emociones, la habilidad de leer las expresiones faciales de otras personas, las emociones y la habilidad para establecer relaciones sociales. Es la que procesa las palabras de estilo [17].

El área de Wenicke se encuentra situada en la corteza cerebral en la mitad posterior de la circunvolución temporal superior y en la parte adyacente del circunvolución temporal media, y su papel fundamental radica en la decodificación auditiva de la función lingüística, que se la relaciona con la comprensión del lenguaje y con el control de las palabras de contenido [17].

El Fascículo Longitudinal Medial es una ruta neuronal que clásicamente se pensaba que conecta ambas áreas, pero que en la actualidad se piensa que conecta con diversas funciones motoras y no sólo con las funciones lingüísticas.

Las palabras de función no son fácilmente controlables, no podemos controlar cuándo y cómo las usamos. Es difícil percibirlas en los demás y controlarlas en nosotros mismos. Se procesan de una manera extremadamente rápida y eficiente en nuestros cerebros.

Según el estudio realizado por [1] las palabras de función aparecen un 3,08% frente a las palabras de contenido que aparecen un 96,92%. Dentro de las palabras de contenido, los sustantivos se utilizan un 54%, los verbos un 22% y los adjetivos un 18%. Según el estudio realizado por [17] las palabras de función aparecen un 0,04% frente al 99,96% de las palabras de contenido, tratándose en este caso de un hablante inglés.

Metodología de Investigación

Se ha determinado un conjunto de fuentes o canales de información de Internet para comparar el uso del lenguaje en cada uno de ellos. La selección corresponde a los siguientes canales representativos:

- Wikipedia: Enciclopedia abierta y colaborativa dónde todo el mundo, salvo excepciones, puede editar su contenido. Es por lo tanto un medio de información formal, aunque escrito por gran cantidad de personas diferentes con estilos discursivos propios.

- Prensa: Publicaciones formales emitidas por los medios de comunicación con el objetivo de informar. Son escritas por los reporteros y siguen un proceso de revisión y filtrado antes de su publicación final, siguiendo las pautas del periódico que las publica.

- Blogs: Sitios web que se actualizan de manera periodica con contenidos publicados por uno o más autores según su propio criterio. Abarcan un amplio abanico de posibilidades, desde blogs corporativos hasta personales, pasando por medios de comunicación similares a la prensa.

- Foros: Sitios web que permiten a los usuarios discutir y compartir información relevante a un tema. Es un sitio de discusión libre e informal, a veces moderada en cuanto a la posibilidad de publicación o no, pero dejando libertad a los autores para expresarse en su propio estilo.

- Twitter: Es un servicio de microblogging dónde los usuarios pueden compartir lo que hacen o están pensando con la limitación de un máximo de 140 caracteres por actualización (o tweet). Se utilizan desde perfiles personales hasta perfiles corporativos, incluyendo bots automáticos de generación de contenido. La restricción en el número de caracteres y la velocidad de generación de contenidos hace que sea un canal especialmente dinámico dónde prima la síntesis y el ahorro ortográfico.

- Facebook: Es una red social compuesta por usuarios, páginas, grupos y eventos y la relación entre todos ellos. Un usuario de esta red social puede tener un perfil personal y conectarlo mediante amistad a otros perfiles personales. También puede crear una página o hacer “Me gusta” en otras páginas. Puede crear un grupo o pertenecer a grupos creados por otros usuarios. Y puede generar un evento o apuntarse a eventos generados por otros usuarios. El

contenido es dinámico, libre y totalmente personal. El usuario decide qué comparte y qué no comparte con los demás. El uso de Facebook es tanto personal como corporativo.

Para la recopilación de los datos experimentales se ha procedido de la siguiente manera:

- Wikipedia: se ha descargado el fichero oficial de la Wikipedia en español de fecha 27/12/2012[1] de dónde se ha extraído aquellos ítems referentes a páginas. No se han tenido en cuenta ni revisiones, ni el histórico de las páginas, solo el contenido de las páginas actualmente en uso.

- Prensa: se ha recopilado durante un periodo de seis meses todo lo emitido en la prensa de 6.581 diarios online de España, Argentina, México, Chile y Panamá en idioma español.

- Blogs: se ha recopilado durante un periodo de seis meses todo lo emitido en un total de 78.289 blogs de España, Argentina, México, Chile y Panamá en idioma español.

- Foros: se ha recopilado durante un periodo de seis meses todo lo emitido en un total de 900 foros de España, Argentina, México, Chile y Panama en idioma español.

- Twitter: se ha recopilado durante un periodo de dos años un stream de Twitter de diversas ciudades de España, todo en español.

- Facebook: se ha recopilado de un total de más de 250 proyectos de escucha activa de organizaciones e instituciones de España, Argentina, México, Chile y Panamá todo lo mencionado en páginas, grupos, muros y comentarios relativo a dichos proyectos, en español.

Se ha realizado un análisis computerizado mediante técnicas de procesamiento de lenguaje natural. Concretamente, se ha realizado una extracción de la terminología de los documentos y se ha etiquetado de manera automática la categoría gramatical a la que pertenece cada palabra mediante un etiquetador de la parte del discurso (postagger: part of spech tagger).

1. http://dumps.wikimedia.org/eswiki/20121227/eswiki-20121227-pages-meta-current.xml.bz2

Las categorías gramaticales que se han tenido en consideración son las presnetadas en la Fig. 2.

Adjetivo	Adverbio	Conjunción
Cuantitativo	Determinante	Interjección
Marcador del discurso	Preposición	Pronombre
Sustantivo	Verbo	

Fig. 2. Categorías gramaticales

La selección de las categorías se ha realizado por su función sintáctica, siguiendo los estudios de [15].

Para los pronombres personales se ha obtenido los rasgos morfológicos de persona (primera, segunda y tercera) y número (singular y plural). Debido a que en español se elude por norma el uso de pronombres personales antes de los verbos (sujeto elíptico), se ha procedido a obtener para estos últimos también la persona y el número.

Es preciso tener en cuenta que no se ha efectuado ningún tipo de corrección ortográfica ni reducción de términos deformados (hoooola, ke, t, kiero), analizándose únicamente los términos que tienen una entrada correcta en una de las derivaciones válidas del español.

Se ha efectuado la contabilización de la frecuencia de aparición de cada categoría gramatical para cada canal de Internet y se muestran y discuten los resultados en el apartado 5.

La recopilación, indexación y análisis de datos se ha realizado con la herramienta Cosmos para la Escucha Activa de Autoritas Consulting.

Datos experimentales

Para cada canal se ha obtenido el número de documentos indicado en la Fig. 3.

	WIKI	PRENSA	BLOGS	FOROS	TWITTER	FB
DOCUMENTOS	3.987.179	5.191.694	1.083.709	673.664	23.873.371	576.723
TÉRMINOS	267.465.810	499.477.658	122.509.753	21.026.388	163.188.448	28.974.716
TÉRMINOS ÚNICOS ANALIZADOS	162.357	157.457	162.412	93.145	128.147	110.040

Fig. 3. Documentos, términos y términos únicos por canal

La primera fila (documentos) indica el número de documentos analizados para cada uno de los canales estudiados. La segunda fila (términos) indica el número de palabras totales extraídos del conjunto global de documentos. La tercera fila (términos únicos analizados) identifica el conjunto de palabras identificadas y analizadas como una de las categorías gramaticales propuestas.

Resultados experimentales

Una primera aproximación a lo que sucede en cada uno de los canales es el número de términos únicos analizados por cada canal mostrado en la Fig. 4.

WIKI	PRENSA	BLOGS	FOROS	TWITTER	FB
162.357	157.457	162.412	93.145	128.147	110.040

Fig. 4. Número de términos únicos por canal

Si podemos valorar la riqueza léxica de una comunicación a partir del número de palabras diferentes empleadas, podemos decir que en este sentido los canales Wikipedia (162.357), Prensa (157.457) y Blogs (162.412) son los que más riqueza léxica tienen, algo inerente a la función principalmente informativa de estos canales. De los tres anteriores, es el canal Prensa el que menos variedad léxica presenta, quizás por el esfuerzo extra de los autores de blogs y los colaboradores en Wikipedia por comunicar de una manera más elaborada frente a un estilo de comunicación más objetivo y directo de la Prensa.

Es preciso notar el número de términos diferentes utilizados en Twitter (128.147), teniendo en cuenta su limitación a 140 caracteres (7 palabras de media como se ve en la Fig. 6).

Por último comentar el resultado en Foros (93.145), canal extremamente informal dónde el objetivo se persigue mediante un uso del lenguaje directo, en muchas ocasiones basado en preguntas / respuestas sobre un tema, y que en cierta medida viene reflejado en la variedad léxica utilizada.

Teniendo en cuenta que el español parte de un lemario de aproximadamente 85.918 lemas según la 22º edición del Diccionario de la Real Academia Española, podemos obtener un ratio a partir de las cifras anteriores y el número de lemas del español, permitiendo una comparación relativa a la riqueza de la lengua, como se presenta en la Fig. 5.

WIKI	PRENSA	BLOGS	FOROS	TWITTER	FB
1,89	1,83	1,89	1,08	1,49	1,28

Fig. 5. Ratio entre número palabras únicas y lemas del español

La Fig. 6 presenta el ratio entre el número de términos totales y el número de documentos por canal, lo que permite obtener la longitud media en palabras de los textos en cada uno de los canales.

WIKI	PRENSA	BLOGS	FOROS	TWITTER	FB
67	96	113	31	7	50

Fig. 6. Longitud media en palabras por canal

De igual manera que los canales de la Wikipedia (67 palabras), la Prensa (96 palabras) y los Blogs (113 palabras) disponían de una mayor variedad léxica, también disponen de una longitud media por documento mayor, siendo el caso de los Blogs el que supera a los demás. Estos tres canales se erigen como los canales que más información pretenden aportar.

Twitter es el canal que tiene una longitud significativamente menor, de 7 palabras de media por documento, debido a la limitación de los 140 caracteres. Esta limitación del número de caracteres implica una necesidad de síntesis mayor para la transmisión de información.

De nuevo los Foros (31 palabras) se ajustan a su característica como canal de información concisa y directa alrededor de un tema. Facebook (50 palabras) por su parte muestra un canal intermedio entre los Foros y los canales de carácter más informativos como la Wikipedia, la Prensa y los Blogs.

En la Fig. 7. se presentan los porcentajes de palabras etiquetadas en cada una de las categorías gramaticales (filas) para cada uno de los canales (columnas). Se marcan en negrita los porcentajes que presentan una variación significativa y que merecen mención especial:

CAT	WIKI	PRENSA	BLOG	FORO	TW	FB
ADJ	13,57%	12,50%	13,67%	**9,27%**	**6,62%**	12,06%
ADV	2,78%	3,46%	3,87%	4,74%	**6,30%**	3,49%
CONJ	1,52%	2,10%	1,80%	4,18%	**7,00%**	2,65%
Q	3,34%	4,47%	4,15%	5,34%	5,53%	4,29%
DET	2,88%	3,48%	2,78%	4,18%	**6,40%**	4,02%
INTJ	0,35%	0,04%	0,06%	0,42%	0,38%	0,07%
MD	0,01%	0,03%	0,02%	0,00%	0,00%	0,00%
PREP	4,00%	5,49%	5,07%	**8,94%**	**13,81%**	6,15%
PRON	0,65%	0,92%	1,12%	2,22%	**3,32%**	1,39%
NOUN	**50,33%**	47,05%	46,59%	42,63%	**34,08%**	47,04%
VERB	20,55%	20,47%	20,88%	18,08%	**16,56%**	18,83%

Fig. 7. Análisis de frecuencia de uso de categoría gramatical por canal

Analizando por categoría, tenemos lo siguiente. Para los adjetivos podemos observar que el canal Twitter (6,62%) es el que los utiliza en menor proporción, reduciendo su uso a prácticamente la mitad del resto de canales, excepto Foros (9,27%), con un valor intermedio. El adjetivo es la

categoría gramatical que acompaña al sustantivo determinándolo, calificándolo o indicando quién lo posee. Ayudan por lo tanto a describir. Según la tabla, su uso es inferior en Twitter y Foros, canales dónde menos descripción detallada se realiza de las cosas.

Para los adverbios observamos el efecto contrario, es en Twitter (6,30%), y en Foros (4,74%) en menor medida, dónde se observa un uso mayor en comparación con el resto de canales. La función principal del adverbio es la de acompañar y modificar el significado del verbo, aunque también en ocasiones a otros adjetivos y adverbios. La modificación del verbo se produce en cuanto a tiempo, lugar, modo, intensidad… lo que indica, por su mayor uso en estos canales, que se procura aportar mayor información sobre el contexto de la acción, dónde sucede, cómo sucede, cuándo sucede algo.

Ligado a lo anterior tiene que ver el uso de las preposiciones, significativamente superior de nuevo en los canales Twitter (13,81%) y Foros (8,94%). Las preposiciones tienen como principal función permitir una categorización jerárquica y espacial, complemento natural de la modificación adverbial realizada en estos canales.

En cuanto a los sustantivos es interesante notar el mayor uso de los mismos realizado en la Wikipedia (50,33%), algo que se espera por tratarse de una enciclopedia dónde se describen objetos, lugares y personas, aunque a un ratio similar al resto de canales, excepto Twitter (34,08%) que es significativamente inferior, lo que muestra un interés menor en hablar de cosas en este último canal.

De igual modo y referido a los verbos se detecta un uso significativamente menor en el canal Twitter (16,56%), y en los canales Foro (18,08%) y Facebook (18,83%) frente al resto de canales. El uso de los verbos implica acción, presión o estado, pero también es preciso para la formación de oraciones. Lo primero puede explicar el mayor uso de los verbos en los canales que efectúan descripciones de cosas (Wikipedia, Blogs) y acciones o sucesos (Prensa). Así mismo, el menor uso de los verbos en el canal Twitter se puede deber a la necesidad de reducción y síntesis de lo que se comenta, debido a su limitación en espacio, primando frases cortas a oraciones elaboradas. Comentar el mínimo uso de los pronombres en todos los canales, pero siendo significativo el mayor uso en el canal Twitter (3,32%) frente al resto.

Para el caso de los pronombres (Fig. 8) y los verbos (Fig. 9) se ha incorporado los ragos morfológicos de persona y número. En el caso de los pronombres se ha obtenido la persona y el número para los personales, y se

ha añadido una fila más (OT) con los porcentajes de los otros pronombres sin persona y/o número (pe. Relativos):

CAT	PER	NUM	WIKI	PRENSA	BLOG	FORO	TW	FB
PRON	1	SIN	13,61%	14,58%	18,85%	**54,47%**	**65,81%**	22,30%
		PLU	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	2	SIN	4,58%	1,18%	2,23%	1,54%	3,53%	3,95%
		PLU	1,92%	1,75%	5,31%	4,61%	5,62%	3,49%
	3	SIN	**55,06%**	**50,75%**	39,26%	24,08%	12,70%	34,68%
		PLU	13,42%	18,22%	16,93%	8,91%	**3,35%**	17,14%
	OT		11,41%	13,52%	17,42%	6,39%	8,99%	18,44%

Fig. 8. Análisis de frecuencia de uso de persona y número de los pronombres por canal

Se aprecia un incremento significativo en el uso de la primera persona del singular en el canal Foro (54,47%) y Twitter (65,81%), describiendo a estos canales como los más egocéntricos en el sentido que emiten más información desde el YO.

Es interesante cómo en todos los canales el uso de la primera persona del plural queda completamente en desuso. Así mismo, es importante notar el uso de la tercera persona del singular (EL) cuando se trata de los canales Wikipedia (55,06%) y Prensa (50,75%) reafirmando de nuevo a estos canales como descriptores de lo que pasa a las cosas o a terceros.

Nótese el decremento significativo del uso de la tercera persona del plural en el caso de Twitter (3,35%), así como el menor uso de otros pronombres en Foros (6,39%) y Twitter (8,99%), reafirmando de nuevo el uso directo y personal de estos canales.

En español el uso del pronombre personal queda reducido en muchas ocasiones por la elipsis del mismo, así que se hace imprescindible el análisis de la persona desde el análisis de las formas verbales. A continuación (Fig. 9) se muestra el porcentaje, para cada canal, de uso de las personas y los números en los verbos:

CAT	PER	NUM	WIKI	PRENSA	BLOG	FORO	TW	FB
VERB	1	SIN	19,95%	17,41%	17,50%	**28,94%**	**24,00%**	16,61%
		PLU	2,10%	2,42%	4,19%	2,68%	4,68%	4,89%
	2	SIN	6,02%	1,55%	3,58%	3,55%	6,77%	2,95%
		PLU	0,46%	0,42%	0,69%	0,98%	1,65%	0,76%
	3	SIN	31,40%	34,00%	29,92%	28,80%	31,21%	31,21%
		PLU	40,07%	44,20%	45,11%	35,05%	31,69%	43,59%

Fig. 9. Análisis de frecuencia de uso de persona y número de los verbos por canal

En este caso es el canal Foro (28,94%) el canal que mayor uso realiza de la primera persona del singural, seguido de Twitter (24,00%), en ambos casos significativamente superior al resto de canales. También notar el uso de la primera persona del plural que frente al no uso del pronombre en este tiempo, muestra que en todos los canales esta es la persona y el número en el que en mayor medida se produce la elipsis.

En la Fig. 10. se muestra el detalle de las 15 palabras más frecuentemente utilizadas por cada uno de los canales:

WIKI	PRENSA	BLOG	FORO	TW	FB
de	de	a	**de**	**de**	de
en	la	de	y	que	la
la	el	la	que	a	el
y	en	en	a	la	en
el	a	el	la	el	y
por	que	y	el	y	a
un	y	que	en	en	que
una	del	del	un	**no**	los
que	los	los	**no**	**me**	del
a	por	un	**pregunta**	un	por
los	un	por	**es**	**es**	para
del	**se**	**se**	por	**se**	un
es	con	con	**abierta**	**lo**	con
las	las	para	**se**	con	**se**
con	para	las	para	por	**no**

Fig. 10. Top 15 palabras más frecuentemente utilizadas por canal

La Fig. 10. corrobora la afirmación de [17] de la alta frecuencia de uso de las palabras de función, pues prácticamente ocupan todas las posiciones. Se han marcado en negrita aquellas palabras que llaman la atención por ser de categoría gramatical diferente a las más frecuentes, por lo general preposiciones.

Es de notar que el pronombre reflexivo se aparece en prácticamente todos los canales, en casos como la prensa para sustituir el uso de una persona en la configuración del discurso. De igual modo el verbo ser en su tercera persona es aparece con cierta frecuencia en todos los canales.

Un primero punto de máximo interés es la aparición de dos sustantivos en el canal foro, canal que por su parte tenía el menor uso de sustantivos (42,63%) junto con Twitter, y que el primero de ellos (pregunta) demuestra el uso principal de este canal como mediador entre usuarios que necesitan información y usuarios que la proporcionan.

Un segundo punto también de máximo interés es la aparición de dos pronombres personales en el canal Twitter, uno de primera persona (me) y

otro de tercera (lo), lo que posiciona este canal como el más personal y egocentrado.

En un análisis por canal se aprecia la similitud entre los canales de la Wikipedia, la Prensa y los Blogs en cuanto al uso de categorías gramaticales como los sustantivos, los verbos y los adjetivos para la descripción de objetos, personas, lugares y situaciones.

Es interesante notar la similitud entre el canal Facebook y el canal Prensa, lo que viene determinado por el uso que se realiza del primero de ellos, en una frecuencia muy elevada, de compartir las noticias y comentarlas. Además, el canal Facebook no destaca significativamente en ningún otro aspecto, lo que de nuevo posiciona a este canal como el gran desconocido.

El canal Twitter destaca por el uso de los pronombres, especialmente en primera persona al igual que los verbos, los adverbios y las preposiciones, lo que conforma y hace honor a lo que viene a representar, “en qué estás pensando”, “qué estás haciendo” o “qué está sucediendo”.

El canal Foro destaca por su estilo directo (baja variedad léxica) involucrado en un proceso de obtención de información guiado por preguntas respuestas.

Conclusiones y trabajo futuro

La presente investigación ha puesto de manifiesto las variaciones en el uso del lenguaje según el canal de Internet dónde se está efectuando la comunicación.

Así mismo debemos tener en cuenta que el propio canal acentúa el uso de determinadas categorías gramaticales, por ejemplo los pronombres en primera persona en Twitter, que a su vez son identificativas de rasgos de personalidad.

Las tablas de frecuencia presentadas tienen un carácter informativo y además deben servir de base para introducir agentes correctivos en los estudios de personalidad basados en texto cuando se realicen sobre los diferentes canales de Internet.

Para ello, debemos realizar un análisis de significación estadística que nos permita determinar la igualdad o diferencia de los diferentes canales, con cierto rigor científico. Una primera aproximación realizada pero no expuesta en el presente trabajo ha sido la comparativa de parejas de canales para

cada una de las categorías gramaticales, estudio no exento de problemas por las elevadas cifras de la muestra y los problemas de sensibilidad de este tipo de test cuando la población es elevada. El siguiente paso puede ser el estudio de posibles correlaciones en el uso de diferentes categorías gramaticales, como la intuición nos sugiere del nombre y el adjetivo, o el verbo y el adverbio.

La investigación continúa mediante la división de las categorías gramaticales en un nivel de detalle mayor, por ejemplo, verbos transitivos, intransitivos, copulativos, auxiliares, o adverbios de contenido frente adverbios de función, etcétera, con el objetivo de poder determinar características significativas a la hora de representar un texto para la construcción de modelos de aprendizaje automático.

Un punto de mejora necesario es la introducción de correctores ortográficos y de derivaciones y deformaciones de la lengua presentes de una manera elevada en canales como los Foros y Twitter, de modo que se amplíen las posibilidades de procesado de textos.

Igualmente es de interés continuar la investigación analizando la diferencia de estilo en cada canal según la edad y el género de los autores, posíblemente se podrá realizar para canales dónde dicha información esté disponible, o en pequeñas muestras etiquetadas a mano o en las que participen sus propios autores (encuestas), en la línea de la competición sobre "Author Profiling" organizada en el laboratorio PAN 2013[2] de la conferencia CLEF para la detección de la edad y el género de los autores a partir de conjuntos de textos anónimos, que siguiendo los estudios de la bibliografía adjunta, pretende servir de foro de intercambio de aproximaciones por diferentes equipos de investigación sobre cómo a partir de características de los textos, se pueden inferir rasgos de personalidad como la edad y el género de sus autores, en esta edición, o perfil emocional, idioma nativo, rasgos de personalidad, etcétera, en posibles futuras ediciones.

Agradecimientos

El trabajo de investigación para la consecución de Cosmos ha sido parcialmente financiado por los proyectos del ministerio ITC/464/2008, TSI-020100-2011-56 e IPT-2012-1220-430000

2 http://pan.webis.de

Referencias

1. Almela, R., Cantos, P., Sánchez, A., Sarmiento, R., Almela, M. Frecuencias del Español. Diccionario y estudios léxicos y morfológicos. Universitas. 2005
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J. Automatically profilling the author of an anonymous text. Communications of the Association for Computering Machinery (CACM). 2009
3. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J. Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday, 12 (peer-reviewed journal on the Internet). 2007
4. Brown, R. Words and things: An introduction to language. New York: Free Press. 1968
5. Chung, C. K., Pennebaker, J. W. Linguistic Inquiry and Word Count (LIWC): pronounced “Luke”... and other useful facts. McCarthy, P., & Boonthum, C (eds.). Applied natural language processing and content analysis: Identification, investigation and resolution. Hershey, Pennsylvania: IGI Global. 2012
6. Chung, C. K., Pennebaker, J. W. The psycological functions of function words. Fiedler, K. (ed.) Social communications. New York: Psycology Press. 2007
7. Davis, D., Brock, T.C. Use of first person pronouns as a function of increased objective self-awareness and performance feedback. Journal of Experimental Social Psychology. 1975
8. Francis, W.N., Kucera, H. Frequency analyses of Enflish usage: Lexicon and grammar. Boston: Houghton Mifflin. 1982
9. Gardner, W. L., Gabriel, S., Lee, A.Y. “I” value freedom, but “we” value relationships: Self-construal priming mirrors cultural differences in judgment. Psycological Science. 1999
10. Ireland, M.E., Pennebaker, J.W. Men imitate life, women underestimate it: Sex differences in scripwriters' portrayal of naturalistic dialogue. Manuscript under review. 2011
11. Kahn, J.H., Tobin, R.M., Massey, A.E., Anderson, J.A. Measuring emotional expression with the Linguistic Inquiry and Word Count. The American Journal of Psychology. 2007
12. Koppel, M., Schler, J., Zigdon, K. Determining an author's native language by mining a text for errors (short paper). Proceedings of KDD, Chigaco IL. 2005
13. Loehlin, J.C, Martin, N.G. Age changes in personality traits and their heritabilities during the adult years: Evidence from Australian twin registry samples. Personality and Individual Differences. 2001
14. Mesulam M-M. Large-scale neurocognitive networks and distributed processing for attention, language, and memory. Ann. Neural. 1990
15. Moreno, A., Guirao, J.M. Morpho-syntactic Tagging of the Spanish CORAL-ROM Corpus: Methodology, Tools and Evaluation. In Spoken Language Corpus and Linguistic Informatics. John Benjamin. 2006
16. Newman, M.L, Groom, C.J., Handelman, L.D., Pennebaker, J.W. Gender differences in language use: An analysis of 14,000 text samples. Discourse Processes. 2008
17. Pennebaker, J.W. The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury Press. 2011
18. Simmons, R.A., Chambless, D.L, Gordon, P.C. How do hostile and emotionally overinvolved relatives view relationships? What relatives' pronoun use tells us. Family Process. 2008
19. Wernicke C. De aphasische symptomenkomplex. Breslau. Cohen and Weigart. 1874