Revista de Lingüística Informática,
Modelización e Ingeniería Lingüística

Nro 7 - Febrero 2015

ISSN 1851-1996

Publicación Anual del

Grupo INFOSUR

Universidad Nacional de Rosario
Argentina

Editorial Responsable
Grupo INFOSUR, Universidad Nacional de Rosario
Pueyrredón 1175, 4º F - 2000 Rosario, Santa Fe, Argentina
Tel. + 54 341 4211284
mail: zsolana@arnet.com.ar
Web: www.infosurrevista.com.ar

Indice

Inducción temprana de categorías morfosintácticas en español

Fernando Balbachan
Facultad de Filosofía y Letras, Universidad de Buenos Aires (UBA)
Buenos Aires, Argentina
fernando_balbachan@yahoo.com.ar

Abstract

A shortcut in order to defy the validation of the Argument from the Poverty of Stimulus (APS) as garantee for the Universal Grammar (UG) would be to demonstrate that the early task of word categorization, starting point for the comprhensive algorithms of syntax induction, might be induced from the Primary Linguistic Data (PLD) through unsupervised mechanisms of general learning from unspecif domain. Our hypothesis is that the mentioned task can be induced from cues (function words and distributional information). Our experiment reports the feasibilty of inducing morphosyntactic categories from the distributional information of the PLD through a mechanism of general learning based on clustering techniques. In order to do so, our experiment leans on two granted assumptions: the early ability for word and phonological phrases segmentation and the identification of cues (mostly, function words) with no associated typology -it does not matter whether they are prepositions, pronouns, or even conten words-. Although our experiment does not demonstrate that the actual mechanism by which the learner may acquire a natural language grammar involves clustering techniques, we do demonstrate the invalidation of APS as PLD can actually be rich enough to induce a formal grammar (at least, its morphosyntactic categories) only from the distributional information.

Keywords: clustering, categorization, general learning mechanisms, syntax induction, function words.

Resumen

Un atajo argumentativo para desafiar la validez del Argumento de la Pobreza de los Estímulos (APS) como garante de la Gramática Universal (GU) sería demostrar que la etapa temprana de categorización de palabras, punto de partida de los algoritmos integrales de inducción de sintaxis, puede ser inducida a partir de los Datos Lingüísticos Primarios (PLD) mediante mecanismos no supervisados de aprendizaje general no específicos de dominio. La hipótesis de esta investigación es que la tarea de categorización temprana puede ser inducida a partir de indicios facilitadores (palabras funcionales e información distribucional). Nuestro experimento reporta la viabilidad de inducir categorías morfosintácticas a partir de la información distribucional de los PLD mediante un mecanismo general de aprendizaje basado en técnicas de clustering, bajo las siguientes dos premisas: habilidad temprana para reconocer palabras y segmentar oraciones y frases fonológicas e identificación de facilitadores (mayormente palabras funcionales) sin necesidad de una tipología diferenciada (no importa si son preposiciones, pronombres o incluso, palabras de contenido). Aunque no demostramos necesariamente que el mecanismo por el cual se adquiere una gramática de un lenguaje natural involucre técnicas de clustering, sí demostramos la invalidez del APS en cuanto a que los PLD podrían ser suficientemente ricos para inducir una gramática formal -al menos, las categorías morfosintácticas- únicamente a partir de la información distribucional.

Palabras claves: clustering, categorización, mecanismos de aprendizaje general, inducción de sintaxis, palabras funcionales.

1. La modelización de sintaxis como procesos en cascada

1.1 Inducción de gramáticas y categorización de palabras como punto de partida

En la última década aparecieron algunos trabajos dentro del paradigma estadístico que se propusieron atacar el Argumento de la Pobreza de los Estímulos (*Argument from the Poverty of Stimulus* APS) -y consecuentemente, la hipótesis innatista- a partir de la postulación de algún algoritmo general no supervisado de adquisición integral del lenguaje. Parafraseando a Klein y Manning (2004), los estímulos (*Primary Linguistic Data* PLD) no parecen ser tan pobres como se creería:

> "We make no claims as to the cognitive plausibility of the induction mechanisms we present here; however, the ability of these systems to recover substantial linguistic patterns from surface yields alone does speak to the strength of support for these patterns in the data, and hence undermines arguments based on 'the poverty of the stimulus'." [Klein y Manning 2004:478]

	Innatismo	**Empirismo**
Estado inicial	Ricamente estructurado	No estructurado
Algoritmos de aprendizaje	Débiles, de dominio específico	Poderosos, de propósitos generales
Estado final	Proundamente estructurado	Superficial

Tabla 1: Teorías de adquisición del lenguaje enmarcadas en el innatismo y en el empirismo, adaptado de Clark (2002)

Pese a que se proponen confrontar con el APS -refutación argumentativa que se conoce como desafío (*challenging*) en la bibliografía especializada (Johnson 2004)-, estos trabajos enmarcados en el paradigma estadístico de la lingüística compuatcional abordan el problema desde la misma perspectiva inicial que el paradigma simbólico de dicha transdisciplina: la sintaxis como punto de partida para la adquisición del lenguaje y el isomorfismo entre lenguajes formales y lenguajes naturales (Chomsky 1957, Clark 2002). Así pues, la modelización de la adquisición ontogenética de sintaxis se presenta como un proceso en cascada que toma como punto de partida un corpus de lenguaje escrito cuantitativa y cualitativamente homologable a los PLD (Pullum 1996; Clark 2002).

Algunos trabajos que se focalizan sobre el proceso de categorización de palabras toman en cuenta los indicios fonológicos en su modelización (Popova 1973; Levy 1985). En tales casos, será imprescindible que los datos lingüísticos del corpus de entrada al proceso contemplen la especificidad de la oralidad. Si bien dichos trabajos aportan una considerable relevancia al problema de la categorización de palabras, adolecen de un problema insalvable: sus respectivas hipótesis no fueron testeadas en un proceso en cascada para la adquisición integral de sintaxis. En cambio, debido a la naturaleza de la información distribucional que actúa como fuente de información primaria para estos modelos, los trabajos más abarcativos, como los de Clark (2002) y Klein y Manning (2004), optan por experimentar con corpora escritos, asumiendo la habilidad tremprana de procesamiento fonológico y segmentación de palabras y frases que se dan en los niños **en forma previa a la categorización de palabras**, según la abrumadora evidencia proveniente de la psicolingüística (Mehler *et al*. 1998; Jusczyk *et al.* 1999):

> "Taken together, these results (and many others) suggest that when they reach the end of their first year of life, babies have acquired most of the phonology of their mother tongue. In addition, it seems that phonology is acquired before the lexicon contains many items, and in fact helps lexical acquisition (for instance, both phonotactics and typical word pattern may help segmenting sentences into words), rather

> than the converse, whereby phonology would be acquired by considering a number of lexical items." [Mehler *et al*. 1998:63]

Por lo tanto, la categorización de palabras (*Part-Of-Speech tagging*, POS-*tagging* o POS-*etiquetado*) resulta el punto de partida para estos algoritmos de inducción integral de sintaxis.

> "Syntactic categories -lexical and functional categories- are the building blocks of syntax. Some knowledge of these categories would be a prerequisite for acquiring syntax. Therefore, the time when a child possesses the knowledge of syntactic categories would be the earliest possible point in development for his/her knowledge of syntax." [Wang 2012:5]

1.2 Hipótesis: palabras funcionales como facilitadoras de la categorización y de la adquisición de sintaxis

Chomsky (1975) postula una Gramática Universal (GU) ricamente estructurada como estado inicial de la adquisición del lenguaje, un sistema innato de principios que son parametrizados a partir de los PLD bajo la forma de una gramática particular, la cual no puede surgir por inducción a partir de principios simples:

> "Una gramática no es una estructura de conceptos y principios de orden superior elaborados por «abstracción», «generalización» o «inducción» a partir de otros más simples sino una estructura rica, dotada de una forma predeterminada compatible con la experiencia, y de un valor más alto (por una medida de valoración que en sí misma es parte de la GU) que otras estructuras cognitivas que llenan el requisito doble de compatibilidad con los principios estructurales de la GU y con la experiencia relevante. Dentro de tal sistema no existen necesariamente componentes aislables «simples» o «elementales»." [Chomsky 1975:59]

En definitiva, tal vez sea mucho pedir probar la invalidez completa del APS en función de inducir toda una gramática completa de un lenguaje natural a partir de los PLD por medio de métodos no supervisados de aprendizaje de dominio general. El propio Clark (2002), cuya tesis de doctorado es un buen intento de esto mismo, reconoce que las gramáticas (*Probabilistic Context-Free Grammars* PCFG) así generadas no necesariamente se condicen con la totalidad de un lenguaje natural (Clark y Lappin 2011). Un "atajo argumentativo" para desafiar la validez del APS como garante de la GU sería demostrar que la etapa temprana de categorización de palabras, punto de partida de los algoritmos integrales de inducción de sintaxis que mencionamos arriba, sí puede ser inducida a partir de los PLD mediante mecanismos no supervisados de aprendizaje general no específicos de dominio:

> "Syntactic category information is part of the basic knowledge about language that children must learn before they can acquire more complicated structures. It has been claimed that «the properties that the child can detect in the input - such as the serial positions and adjacency and co-occurrence relations among words - are in general linguistically irrelevant.» (Pinker 1984) It will be shown here that relative position of words with respect to each other is sufficient for learning the major syntactic categories." [Schüze 1993:251]

> "A current debate is whether young children possess an abstract representation of functional categories (*e.g.*, determiner, auxiliary and preposition) or whether the representation of functional categories is built gradually in an item-by-item fashion. Strong nativist views held that children are innately endowed with a set of grammatical categories including functional categories. They possess abstract knowledge of grammatical categories since the beginning and use that knowledge to learn their first language.

> Therefore, according to constructivist views, young children do not have abstract knowledge of grammatical categories initially. It is the burden of constructivists to explain how children transform the item-based representation to adult-like grammar." [Wang 2012:3-4]

La hipótesis de esta investigación es demostrar que la tarea de categorización temprana puede ser inducida a través de los PLD a partir de indicios facilitadores (palabras funcionales e información distribucional), con el único pre-requisito del procesamiento fonológico de la segmentación de palabras y frases. De este modo, el APS como garante último de la GU estaría cayendo parcialmente en cuanto a que los PLD no son tan pobres como se creía. En última instancia, la psicolingüística tendrá la última palabra en cuanto a elaborar una teoría ontogenética suficientemente explicativa, pero al menos una modelización formal exitosa resultará una irrefutable prueba empírica de la riqueza estructural de los Datos Lingüísticos Primarios para esta etapa temprana como punto de partida de la sintaxis para la adquisición del lenguaje. Como objetivo secundario, esta investigación se propone demostrar la viabilidad de utilizar la categorización de palabras como punto de partida para un algoritmo integral de sintaxis del español, al estilo de los algoritmos integrales de Clark (2002) y de Klein y Manning (2004).

Más allá del diseño específico de las etapas de un algoritmo integral de inducción de sintaxis que modelice la adquisición del lenguaje, resulta evidente que una de las primeras tareas lingüísticas que debe llevar a cabo exitosamente el adquiriente es la categorización de palabras; es decir, la habilidad de agrupar ítems léxicos por sus características morfosintácticas diferenciales como piezas fundamentales para las reglas sintácticas combinatorias de todo lenguaje natural. La necesidad de algún mecanismo de mapeo de ítems léxicos a "protocategorías" morfosintácticas de palabras hace que resulte imprescindible postular esta habilidad tempranamente en los niños, aun en el caso de los innatistas, con el único pre-requisito estricto de una exitosa habilidad para segmentar palabras, lo cual ocurre -por lo menos para el inglés- desde los 10 meses de edad (Mehler *et al*. 1998; Jusczyk *et al.* 1999):

> "Even if we hypothesise that these closed class categories are innate, a difficult assumption given the high cross-linguistic variability in the set of lexical categories, the infant learner is still faced with the difficulty of working out which words correspond to which classes – the so-called linkage problem." [Clark 2002:57-58]

> "Even if young children are predisposed with notions of abstract functional categories, they still have to assign the word forms in the target language to those categories because word forms and members of a category differ between languages and have to be learned from the input. In other words, a child has to map words in the target language to the right categories." [Wang 2012:32]

2. Consideraciones acerca de la pertinencia de las técnicas de clustering para la categorización de palabras

En la mayoría de los trabajos de inducción de categorías morfosintácticas a partir de información distribucional mediante técnicas de clustering se recurre a una misma premisa: para analizar la distribución del contexto de ocurrencia de cada palabra (*target*) usaremos una unidad denominada bigrama: co-ocurrencia de pares de ítems léxicos en una relación fija contigua. Dicha relación puede ser, por ejemplo, la contigüidad que existe entre una palabra *target* (es decir, la palabra que se pretende estudiar) y su contexto inmediato (la palabra inmediatamente siguiente o anterior), relación denominada comúnmente *ventana de análisis* y en particular, bigrama hacia la derecha o bigrama hacia la izquierda, respectivamente. Por ejemplo, si todo el *corpus* consistiera en una única frase "*la vaca salta sobre la cerca*", la siguiente tabla representaría el vector de ocho dimensiones

del contexto correspondiente a la palabra *salta* (Manning y Schütze 1999; Zhitomirsky-Geffet y Dagan 2009):

Target	Contexto (bigramas a la derecha)			
	-la	*-vaca*	*-sobre*	*-cerca*
salta	0	0	1	0
Target	Contexto (bigramas a la izquierda)			
	la-	*vaca-*	*sobre-*	*cerca-*
salta	0	1	0	0

Tabla 2: Ejemplo de vector de bigramas hacia la derecha y hacia la izquierda para la palabra "*salta*" en la oración "*la vaca salta sobre la cerca*"

Este vector de '*salta*' (0,0,1,0,0,1,0,0) representaría, en este corpus de una única oración, una suerte de ADN de la palabra target respecto de su combinatoria con las 4 únicas palabras de este vocabulario, en términos de bigramas hacia la derecha y bigramas hacia la izquierda, respectivamente. Eventualmente, la relación de determinación del tipo de palabra entre una palabra target y sus vecinos del contexto (*context*) puede extenderse hasta abarcar a los vecinos más alejados (trigramas, tetragramas, etc.). No obstante, se ha demostrado que la influencia ejercida sobre el tipo de palabra target por parte de la ventana de análisis disminuye notablemente con las unidades mayores a bigramas (Redington *et al*. 1998).

En corpora masivos es de esperar que los ítems lexicales que pertenecen a una misma categoría morfosintáctica tengan una distribución similar, lo cual se traduce en una cercanía en el espacio vectorial (Manning y Schütze 1999) susceptible de ser descubierta a partir de técnicas de clustering. El mapeo de categorías sintácticas sobre un espacio vectorial multidimensional asume que hay una manera de dividir esas mismas categorías bajo un criterio geométrico: tradicionalmente se han propuesto modelos donde la frontera es discreta, y otros donde es prototípica o basada en similitudes entre ítems lexicales individuales.

Por supuesto, resulta inadecuada la idea de que el perfil de ocurrencias distribucionales de una palabra target en un corpus masivo involucra combinaciones a izquierda y a derecha con cada una de las palabras del vocabulario de una lengua. Esto se verifica con la concepción misma de la sintaxis subyacente a dichas combinaciones, independientemente de la extensión del corpus a relevar. Sólo por mencionar un ejemplo, en una misma frase fonológica la combinación de dos sustantivos en español -sin palabra funcional de por medio que los articule- está prohibida. Esto nos lleva a considerar la intuición de que resultaría inadecuada una caracterización vectorial de una palabra *target* respecto de todas las combinaciones posibles, lo cual redundaría en vectores de 40.000 dimensiones en un vocabulario de 20.000 palabras a derecha y a izquierda, y de 800.040.000 dimensiones en el caso de considerar bigramas y trigramas. Desde un punto de vista matemático resulta inviable modelizar un espacio vectorial de decenas de miles e incluso millones de dimensiones. Incluso así, la inmensa mayoría de dichas dimensiones aportaría cero ocurrencias al vector, en virtud de las prohibiciones sintácticas combinatorias -dispersión de eventos en el espacio vectorial (*sparsity*). Estas consideraciones matemáticas han derivado necesariamente en la idea de la reducción de la dimensionalidad de los vectores, ya sea a partir de la identificación de cierta palabras "definitorias" de la palabra target -lo que la bibliografía especializada da en llamar *cue* (Redington *et al*. 1998; Clark 2002) o *feature words* (Nath *et al*. 2008)-, o bien a partir de la simplificación de la matriz resultante de los vectores, desestimando las submatrices *anuladas* en

cero – a partir de técnicas como *Single Value Decomposition* (SVD) (Deerwester *et al*. 1990; Schütze 1993) o *Principal Component Analysis* (PCA) (Böhm *et al*. 2006).

Este procedimiento algebraico de reducción de la dimensionalidad del espacio vectorial a partir de la identificación de palabras marcas (*cues*) tiene su perfecto correlato en la evidencia psicolingüística ontogenética de la adquisición de la habilidad temprana de categorización de palabras: aprendemos a categorizar palabras en función de cierta información facilitadora (*cues*), la cual bien puede estar representada por ciertos *descriptores* preferenciales (Redington *et al*. 1998; Clark 2002) para todos los tipos de palabras. Como mencionamos anteriormente, la hipótesis central de este trabajo sostiene que dicho papel sería desempeñado mayormente por las palabras funcionales de un idioma, en virtud de su ocurrencia masiva y de sus propiedades distribucionales y articulatorias (actúan como bisagras) respecto de las restantes palabras. Dos grandes desafíos se derivan de esta hipótesis central: demostrar que estas *cues* están disponibles para el adquiriente de un lenguaje en forma previa a los tipos de palabras morfosintácticas a inducir -si no como palabras plenamente adquiridas, al menos como marcas formales en los PLD- y demostrar que esta inducción puede ser llevada a cabo mediante mecanismos generales (no de dominio específico) de aprendizaje no supervisado.

Justamente, todas estas consideraciones nos llevan a contemplar algunos aspectos de modelización que deben ser cuidadosamente analizados para este tipo de enfoques en experimentos de clustering. Algunas consideraciones son inherentes a la naturaleza del problema de la categorización de palabras y otras, en cambio, atañen a las técnicas de clustering empleadas como metodología para la presente investigación.

3. Inducción no supervisada de categorías morfosintácticas mediante clustering a partir de palabras funcionales sin tipología diferenciada

3.1 Motivación de las decisiones de diseño

En función de las fortalezas y las críticas relevadas para los trabajos que durante las últimas dos décadas atacaron el problema de cómo los adquirientes de una lengua conforman clases morfosintácticas de palabras, nuestro experimento se propone como un enfoque computacional compatible con la evidencia empírica de la psicolingüística, con mayor una adecuación explicativa. Así pues, nuestra propuesta de modelo de adquisición de categorías morfosintácticas del español responde a los siguientes lineamientos:

1) Para el marco epistemológico general, optamos por el paradigma estadístico de la lingüística computacional, en detrimento del paradigma simbólico. A pesar de que algunos modelos enmarcados en el paradigma simbólico son compatibles con nuestra hipótesis de un sesgo débil (Lappin y Shieber 2007; Clark y Lappin 2013) para inducir sintaxis a partir de un mecanismo de aprendizaje general, consideramos que los modelos disponibles de marcos frecuentes (Mintz 2003; Chemla *et al*. 2009) y de protoconstituyentes (Christophe *et al*. 2008) presentan insalvables cuestionamientos a la adecuación descriptiva y a la adecuación explicativa, respectivamente.

2) Desde el paradigma estadístico de la lingüística computacional, nos inclinamos hacia las técnicas de clustering con un enfoque tradicional, sin el agregado de técnicas avanzadas de *machine learning*. Esto nos garantiza una aceptable cobertura del fenómeno a elucidar, sin contradecir la hipótesis de un mecanismo general de aprendizaje, ya que algunos modelos actuales logran una mayor efectividad en inducir categorías sintácticas a partir de considerar

features como la distinción mayúscula/minúscula (Berg-Kirkpatrick *et al*. 2010) , algo que obviamente nos está vedado en función de mantener las condiciones de aprendibilidad de una teoría formal de inducción de sintaxis (Pinker 1979).

3) Para el algoritmo de clustering en particular, elegimos el clustering no jerárquico K-means con distancia euclideana sobre los centroides. Nos proponemos "historizar" el proceso iterativo de inducción de categorías hasta hallar una distribución óptima en función del conjunto de datos iniciales y una parametrización creciente de los números de clusters desde *K=2* hasta *K=nº máximo de cues*. Esta historización sería inviable con un algoritmo de clustering jerárquico. Además, K-means ofrece otra ventaja: la menor complejidad de poder de cómputo. La distancia euclideana como criterio de similitud de objetos en el espacio vectorial se nos presenta más intuitivamente correcta que la distancia Manhattan para garantizar la plausibilidad de un mecanismo de aprendizaje general, a pesar de que se considera que esta última resulta menos sensible que la primera a la influencia de los objetos apartados (*outliers*) en el espacio vectorial (Manning y Schütze 1999).

4) El espacio vectorial multidimensional quedará definido por un procedimiento de identificación no arbitraria y no apriorística de las marcas sintácticas (*cues*) (Elghamry 2004) que habrán de sentar las bases del posterior modelado vectorial de las palabras targets en función de su contexto distribucional inmediato. Así pues, la única premisa lingüística que damos por sentada en esta modelización es la habilidad exitosa de segmentación de palabras, frases fonológicas y oraciones o enunciados (Mehler *et al*. 1998; Jusczyk *et al.* 1999), dejando de lado el acceso a indicios morfológicos de las palabras target y a indicios prosódicos para la identificación de palabras funcionales (Wang 2012), indicios sobre cuya disponibilidad no hay un consenso absoluto (Clark 2000, 2002, 2003)-. Al igual que Clark (2002), no renegamos, en principio, de la plausibilidad de dichas fuentes de información en el proceso de facilitación (*bootstrapping*) de la habilidad de categorización temprana de palabras. Simplemente, demostraremos que las propiedades distribucionales del corpus que modeliza los PLD son suficientes para inducir la categorización de palabras sólo a partir de postular la habilidad de segmentación de palabras y frases fonológicas. La convergencia de indicios provenientes de otras fuentes de información no hará sino robustecer nuestro argumento *a fortiori.*

5) La información distribucional con la que trabajaremos son los bigramas a derecha y a izquierda de las palabras target respecto de cada una de las dimensiones (*cues*) que conformarán el perfil distribucional de dicha palabra target. En todos los trabajos de clustering relevados, la mayor informatividad de la ventana de análisis sobre el contexto distribucional de la palabra target se focaliza en la relación de bigramas por sobre contextos más mediatos (trigramas, tetragramas). Esta decisión de diseño nos encolumna detrás de los clásicos trabajos del campo (Brown *et al*. 1992; Schütze 1993; Redington *et al*. 1998; Clark 2002), pero nos obliga a considerar mecanismos no arbitrarios de identificación de cues (Elghamry 2004) y de reducción de la dimensionalidad del espacio vectorial (Schütze 1993).

6) En cuanto a la escalabilidad del algoritmo, seguiremos a Redington *et al*. (1998) y plantearemos un escenario con un vocabulario reducido de aproximadamente 1000 palabras target. De hecho, esa cantidad de palabras resulta esperable para la finalización de la etapa ontogenética que nos interesa modelizar: la explosión léxica (*vocabulary spurt*) (Dromi 1987) que se da en los niños entre los 2 y 3 años de edad. Por supuesto, este corte en las palabras target nos aleja de enfoques exhaustivos como los de Clark (2002). Sin embargo, consideramos que el aprnedizaje no supervisado basado en técnicas de clustering es

especialmente eficaz en agrupar eventos con una cierta ocurrencia frecuente en el espacio vectorial (Martin *et al*. 1998). A su vez, esta decisión de diseño se condice con la plausibilidad de la evidencia empírica psicolingüística y con la robustez de los modelos matemáticos postulados en dichas técnicas de clustering, reduciendo los costos implementativos:

> "Although the child might not have access to 1,000 vocabulary items, if the child applies distributional analysis over its small productive vocabulary, this will work successfully, because this vocabulary consists almost entirely of content words. Moreover, prior to the vocabulary spurt, the child's syntax, and thus, presumably, knowledge of syntactic categories is extremely limited, and hence even modest amounts of distributional information may be sufficient to account for the child's knowledge. By the third year, the child's productive vocabulary will be approaching 1,000 items (*e.g*., Bates *et al*. 1994, found that the median productive vocabulary for 28 month olds was just under 600 words) and hence could in principle exploit the full power of the method.
> **It is also possible that, even when children's productive vocabularies are small, they may have a more extensive knowledge of the word forms in the language. It is possible that the child may be able to segment the speech signal into a large number of identifiable units**, before understanding the meaning of the units (Jusczyk 1997)." [Redington *et al*. 1998:454] (*las negritas y el subrayado son nuestros*)

> "In practical systems, it is usual to not actually calculate n-grams for all words. Rather, the *n*-grams are calculated as usual only for the most common *k* words […] Because of the Zipfian distribution of words, cutting out low frequency items will greatly reduce the parameter space (and the memory requirements of the system being built), while not appreciably affecting the model quality (*hapax legomena* often constitute half of the types, but only a fraction of the tokens)." [Manning y Schütze 1999:199]

7) El inglés es un idioma con orden fijo de constituyentes sintácticos, los cuales mayormente siguen el orden canónico SVO. Este mecanismo actúa para desambiguar morfosintácticamente formas léxicas idénticas, a falta de marcación morfológica enriquecida. Gran parte del vocabulario inglés puede funcionar indistintamente como verbo o sustantivo. Esto justificaba el tratamiento de la ambigüedad del tipo de palabra morfosintáctica que se observa en Schütze (1993) y en Clark (2002) como un problema de *soft clustering* (posibilidad de asignar un miembro a más de una clase) (Manning y Schütze 1999). Sin embargo, éste no es el caso del español, un idioma morfológicamente rico. Si bien existen en español numerosas formas POS-ambiguas, incluso entre las palabras más frecuentes de cualquier corpus (por ejemplo '*como*', '*para*', '*era*', etc.), consideramos que esta problemática no está tan extendida como en inglés (Graça *et al*. 2011). Por eso, al igual que Redington *et al*. (1998), implementaremos un mecanismo de desambigüación morfosintáctica para tales casos, basado en un corpus de referencia. Es decir, nuestro algoritmo trabajará con un *hard clustering* que asignará cada miembro de las palabras *target* a una única clase o cluster.

8) El corpus con el que se trabajará contará con una extensión compatible con los experimentos de Redington *et al*. (1998) del orden de 2 millones de tokens, respetando criterios de balance y plausibilidad de modelización de los PLD (Chomsky 1959; Pullum 1996). Si bien Clark (2002) sostiene que un corpus que modelice los PLD debe ir desde 10 millones de tokens a 100 millones de tokens para los cuatro años de estímulos linigüísticos que abarcan el período de surgimiento de una gramática de un lenguaje natural, preferimos reducir la complejidad combinatoria de nuestro experimento y demostrar que dichos corpus reducidos ya ofrecen las condiciones suficientes para la categorización de palabras mediante la información distribucional. Si nuestro objetivo se verifica, la hipótesis será validada *a fortiori* para un corpus más masivo.

9) Para la evaluación de nuestro experimento exploraremos diversas alternativas, pero podemos adelantar que nos basaremos principalmente en la métrica *many-to-1* (Christodoulopoulos *et al*. 2010). También seguiremos a Redington *et al*. (1998) en una evaluación discriminada para cada tipo de categoría inducida y postularemos nuestra propia justificación algebraica del agrupamiento de clusters (*cluster merging*) (Böhm *et al*. 2006) en *hiperclusters* a partir del mapeo *many-to-1*.

Básicamente el algoritmo propuesto se muestra en el siguiente esquema:

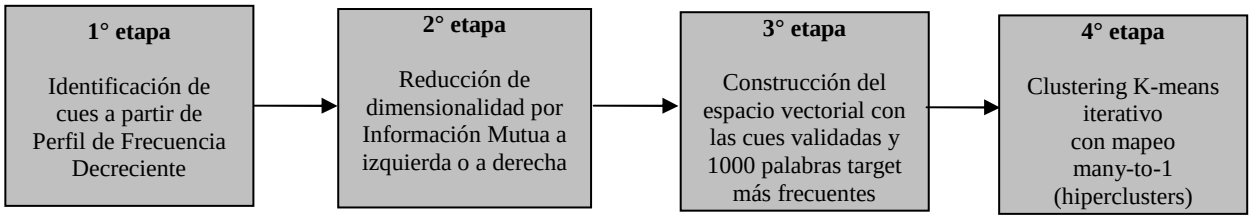

Figura 1: Esquema del algoritmo de categorización de palabras propuesto

3.2 La medida justa: mapeo many-to-1 e hiperclusters

Ante la posibilidad de que algunas categorías del gold standard aparezcan repartidas en varios clusters en función de la granularidad morfosintáctica del tag, la mayor parte de los trabajos de clustering recurren a un mapeo de varios clusters en una única categoría, criterio denominado mapeo *many-to-1*:

> "Many-to-one mapping accuracy (also known as *cluster purity*) maps each cluster to the gold standard tag that is most common for the words in that cluster (henceforth, the *preferred tag*), and then computes the proportion of words tagged correctly. More than one cluster may be mapped to the same gold standard tag. This is the most commonly used metric across the literature as it is intuitive and creates a meaningful POS sequence out of the cluster identifiers. However, it tends to yield higher scores as $|C|$ [number of clusters] increases, making comparisons difficult when $|C|$ can vary." [Christodoulopoulos et al. 2010:577]

En nuestro experimento adoptamos esta decisión de diseño. Más allá de la justificación metodológica, existe una inituición gramatical en adoptar este criterio de evaluación general de la distribución de un ciclo de clustering. Es de esperar que la ubicación de los clusters en el espacio vectorial refleje en alguna medida el criterio de agrupamiento de clusters en función de la similitud de los miembros preeminentes en cada uno de ellos. Así, pues a dos o más clusters del mismo tipo (indicado por el valor del *cluster_tag*) corresponde un mismo *hipercluster*.

Si un centroide representa prototípicamente la ubicación espacial de un cluster, al menos en cuanto a la concentración mayoritaria de sus miembros, entonces al computar la distancia euclideana de los centroides entre sí podemos darnos una idea de qué clusters están más cercanos o más alejados entres sí. Nuestra intucición metodológica de los hiperclusters podría verse justificada empíricamente si, por ejemplo, los clusters *sustantivos singulares* NN1 que conforman el hipercluster NN1 aparecen de algún modo más cercanos entre sí, en comparación con, por ejemplo los clusters que conforman el hipercluster *verbos en infinitivo* VVI.

El concepto de hipercluster, tal como denominamos en este trabajo al agrupamiento de clusters, resulta muy significativo. Desde un punto de vista metodológico permite una evaluación que resuelve el problema del mapeo de un número creciente de clusters inducidos en las categorías del gold standard. Desde un punto de vista algebraico el hipercluster se ve justificado en gran medida

por la ubicación en el espacio vectorial de los centroides de los clusters que lo conforman, lo cual, a su vez, refleja particularidades morfosintácticas propias del dominio lingüístico al que pertenecen los datos.

3.3 Evaluación iterativa de todos los ciclos de clustering con la métrica many-to-1

Ahora que explicamos en detalle en qué consistió nuestro experimento de clustering para inducción de categorías sintácticas en español, su plausibilidad de modelización, sus lineamientos de diseño y sus métricas de evaluación, llegó el momento de analizar la salida completa de los 106 ciclos. Recordemos que el experimento corre iterativamente en ciclos que van desde K=2 clusters hasta K = 106 clusters. Si bien el corte inicial era de 1000 palabras target, 89 de esas palabras correspondía a categorías morfosintácticas marginales: categorías funcionales de poquísimos miembros y de prevalencia intermitente (en muy asialdas ocasiones) en los clusters (REL, AJC, CJC, CJS, etc.). Las restantes 911 palabras target, entonces, se distribuyeron entre 16 categorías de inducción casi permanente a lo largo de todo el experimento, con elevados valores de pureza consolidados a partir de los ciclos medios.

TOTALES	*n*	**Baseline = n/1000**	Probabilidad de acertar el POS-tag por azar
AJ1	106	0,106	
AJ2	38	0,038	
AV0	55	0,055	
CRD	14	0,014	
DPS	7	0,007	Si no se pondera el promedio, la probabilidad de acertar el POS-tag es 1/16, lo cual sigue siendo muy bajo (0,0625 = 6,25%)
DT1	7	0,007	
DT2	7	0,007	
NN1	342	0,342	
NN2	92	0,092	
NNP	43	0,043	
PND	5	0,005	
PRP	8	0,008	
VMZ	14	0,014	
VVI	42	0,042	
VVN	14	0,014	
VVZ	117	0,117	
	Total = 911	**0,0569 = 5,7%**	**Baseline ponderado**

Tabla 3: Palabras target a ser clusterizadas según POS-tag de corpus de referencia y baseline de cada POS-tag

En cada ciclo calculamos Precisión, Cobertura y medida F para cada uno de los 16 POS-tags, prevalezcan o no como el *cluster_tag*, en cada uno de los hiperclusters inducidos. Sobre estas 16 medidas F calculamos el promedio común y el promedio ponderado (según el peso de cada POS-tag en la distribución de 911 palabras target).

Es de destacar que a partir de los ciclos medios (ciclo 52 en adelante), las medidas F de la mitad de los POS-tag se presentan consolidadas en valores relativamente estables, especialmente para las categorías mayores de sustantivos y verbos (NN1, NN2, VVZ, VMZ, VVI, VVN), lo cual significa que a partir de cierto momento de la "historización" de la inducción, las clases están mayormente consolidadas en cuanto a la pertenencia de sus miembros (con mínimas fluctuaciones).

Esta convergencia en las distribuciones de los hiperclusters otorgaría una mayor robustez a nuestro enfoque, ya que no sería necesario postular un parámetro inicial de K clusters, para incializar el modelo, en virtud de la iteración convergente a partir de los ciclos medios. Este punto de consolidación de los ciclos de agrupamiento dependería exclusivamente de la cantidad de cues identifacadas en el corpus. Esto reforzaría la plausibilidad algorítmica del modelo, en tanto no demandaría de un mecanismo de evaluación basado en mínimos o máximos locales sino que la mera iteración convergería a distribuciones consolidadas.

	CICLO 87											
Hipercluster	n		TP	FP	TN	FN	**Precision**	**Recall**	**Fscore**			
AJ1	106		41	37	xxxxxx	65	0,525641026	0,386792453	**0,44565217**	AJ1	0,05185415	
AJ2	38		18	34	xxxxxx	20	0,346153846	0,473684211	**0,4**	AJ2	0,016684962	
AV0	55		32	70	xxxxxx	23	0,31372549	0,581818182	**0,40764331**	AV0	0,024610738	
CRD	14		10	1	xxxxxx	4	0,909090909	0,714285714	**0,8**	CRD	0,012294182	
DPS	7				xxxxxx	7	0	0	**0**	DPS	0	
DT1	7		3	2	xxxxxx	4	0,6	0,428571429	**0,5**	DT1	0,003841932	
DT2	7		4	9	xxxxxx	3	0,307692308	0,571428571	**0,4**	DT2	0,003073546	
NN1	342		304	49	xxxxxx	38	0,861189802	0,888888889	**0,87482014**	NN1	0,328417661	
NN2	92		64	9	xxxxxx	28	0,876712329	0,695652174	**0,77575758**	NN2	0,078342148	
NNP	43		19	33	xxxxxx	24	0,365384615	0,441860465	**0,4**	NNP	0,018880351	
PND	5				xxxxxx	5	0	0	**0**	PND	0	
PRP	8		4	1	xxxxxx	4	0,8	0,5	**0,61538462**	PRP	0,005404036	
VMZ	14		3	2	xxxxxx	11	0,6	0,214285714	**0,31578947**	VMZ	0,004852967	
VVI	42		32	32	xxxxxx	10	0,5	0,761904762	**0,60377358**	VVI	0,027835884	
VVN	14		9	3	xxxxxx	5	0,75	0,642857143	**0,69230769**	VVN	0,010639196	
VVZ	117		103	38	xxxxxx	14	0,730496454	0,88034188	**0,79844961**	VVZ	0,10254512	
INDECIDIBLES	16 clusters con 29 miembros								**0,50184864**	PROMEDIO	**0,68927687**	PONDERADO

Tabla 4: Detalle de evaluación de ciclo 87

4. Discusión de los resultados y conclusiones

4.1 Consideraciones cuantitativas y cualitativas

1) Todas las categorías sintácticas mayores fueron inducidas con un alto grado de pureza. Se observan refinamientos granulares en rasgos de género y número (para sustantivos) y en otras caracterizaciones morfosintácticas (verbos modales VMZ vs. verbos léxicos VVZ).
2) Al igual que en Redington *et al*. (1998), las categorías sintácticas mayores, coincidentes con palabras de contenido (verbos y sustantivos), reportan medidas F altísimas, del orden del 80% y hasta 90%.
3) En el otro extremo, uno de los hiperclusters con menor medida F (40,7%) son los adverbios (AV0). Este grupo quedó confinado a un cluster único y masivo de 95 miembros muy heterogéneos, con objetos claramente marginales (caracteres únicos como ‘*d*’, ‘*p*’, ‘*v*’, etc.). Como reporta Nath *et al*. (2008), es normal que en el clustering partitivo quede en cada ciclo uno o dos clusters masivos que actúan como receptáculo indiferenciado de objetos del espacio vectorial. Posiblemente éste sea el caso.
4) Si bien los adjetivos presentan medidas F bajas, en muchos casos el refinamiento por cluster es sumamente interesante. En uno de los cluster aparecen adjetivos que en general son usados con una proposición (“*es preciso que…*”, “*es necesario que…*”, etc.).
5) En todos los casos, es notable la consolidación de los agrupamientos a partir de los ciclos medios (ciclo 52 en adelante).

4.2 Plausibilidad psicolingüística de la modelización

Recapitulando todo lo expuesto hasta ahora, podemos consignar que nuestro experimento reporta exitosamente la viabilidad de inducir categorías morfosintácticas a partir de la información distribucional de los PLD mediante un mecanismo general de aprendizaje, bajo las siguientes dos premisas:

1) Habilidad temprana para reconocer palabras y segmentar oraciones y frases fonológicas (Mehler *et al*. 1998; Jusczyk *et al.* 1999). Evidencia de disponibilidad a partir de los 10 meses.

2) Identificación de las cues (mayormente palabras funcionales) sin necesidad de una tipología diferenciada (no importa si son preposiciones, pronombres o incluso palabras de contenido). Aunque Wang (2012) sostiene que las palabras funcionales pueden estar representadas en forma temprana en el léxico de un modo abstracto, identificadas a partir de indicios prosódicos pero sin acceso a su significado o tipología, en nuestro experimento basta con su reconocimiento como marcas muy frecuentes en los PLD y sus propiedades articulatorias (*pivot*) respecto de las palabras target. (Elghamry 2004). Evidencia de disponibilidad a partir de los 14 meses.

Estas condiciones están plausiblemente dadas incluso bastante antes de la explosión léxica (*vovabulary spurt*) (Dromi 1987) que se da alrededor de los dos años y ciertamente para los 15 meses en donde se verifican los primeros juicios de categorización (Shi *et al.* 1999), por lo que nuestro algoritmo resulta compatible con la evidencia empírica psicolingüística. Lo que demuestra nuestro algoritmo, entonces, es la suficiencia de los PLD mismos para aportar la información necesaria en el proceso de categorización de palabras, sin necesidad de postular conocimiento innato específico de dominio.

En resumen, tomando el trabajo de Redington *et al.* (1998) como punto de partida, nos propusimos encarar un experimento que incorpore sustanciales mejoras en el diseño del algoritmo. A su vez, también éramos concientes de los casi inexistentes intentos previos de llevar a cabo procedimientos sistemáticos de clustering sobre corpora en español. El objetivo del experimento fue demostrar que la información distribucional es una poderosa herramienta suficiente para la inducción de juicios de pertenencia de ítems lexicales a categorías sintácticas. Como se remarcó a lo largo de todo este artículo, el diseño general del experimento respondió a una necesidad de compatibilizar la modelización algorítmica con la plausibilidad psicolingüística del proceso ontogenético de la categorización temprana de palabras.

5. Trabajo a futuro para el experimento de categorización

Los experimentos aludidos en este artículo son una versión resumida de nuestra tesis de doctorado y nos revelan una importante veta de indagación científica que obliga a replantearse cuestiones tan sensibles para la lingüística como la naturaleza del lenguaje y los mecanismos de adquisición del mismo, a la luz de las promisorias técnicas de aprendizaje de máquina y de los procesos de inducción de gramáticas.

> "This problem does not entail that formal learning theory has nothing to offer the study of language acquisition. On the contrary, it is highly relevant. However, we argue that the crucial problems are not information theoretic, as suggested in the Gold results. Instead, they are complexity theoretic. By modeling the computational complexity of the learning process, we can, under standard assumptions, derive interesting result concerning the types of representations (or grammars) that are efficiently learnable. It is uncontroversial that the human capacity

> to learn is bounded by the same computational limitations that restrict human abilities in other cognitive domains. The interaction of this condition with the complexity of inducing certain types of representations from available data constitutes a fruitful object of study." [Clark y Lappin 2013:90-91]

El progreso de las técnicas estadísticas y el avance de las investigaciones sobre corpora abarcativos revelan que incluso los más simples mecanismos estadísticos pueden contribuir al esclarecimiento del proceso de adquisición del lenguaje. En particular, el conjunto de marcas e indicios provistos por la información distribucional constituye una herramienta válida para la inducción de juicios acerca de la pertenencia de palabras a categorías morfosintácticas. Hemos demostrado empíricamente la estrecha correlación entre palabras cue vs. palabras target, distinción operativamente homologable a las nociones lingüísticas de palabras funcionales vs. palabras de contenido, y hemos señalado el importante papel que podrían desempeñar dichas palabras funcionales en la adquisición del lenguaje, aunando las respectivas agendas de investigación de la lingüística computacional y de la psicolingüística. Justamente, una deuda pendiente en el campo de la psicolingüística es la necesidad de compatibilizar evidencia contradictoria acerca del momento ontogenético de la adquisición de las palabras funcionales en producción y en comprensión, lo cual contribuirá a la mayor adecuación explicativa de los enfoques computacionales, en función de los diversos pre-requisitos de modelización (el pre-requisito son las cues, no la categorización de las cue).

En este sentido, y sin menoscabo de otros mecanismos de aprendizaje que podrían actuar simultáneamente, se puede concluir que la información distribucional se perfila como un enfoque enriquecedor. El paradigma estadístico se propone como un promisorio marco epistemológico de investigación que requerirá una amplia gama de herramientas y experimentos para explorar cabalmente todo su potencial. Valga, pues, la aclaración de que el experimento delineado en este trabajo representa una mera prueba de concepto que debe ser exhaustivamente mejorada a futuro.

Finalmente, resulta imperioso situar este tipo de investigaciones en el marco más general de un proyecto de inducción integral de sintaxis (Clark 2002; Klein y Manning 2004). El aprendizaje no supervisado de sintaxis o, en otras palabras, el problema de la inducción de una gramática a partir de un corpus sin anotaciones, todavía presenta interesantes desafíos desde el punto de vista de la lingüística teórica y de sus aplicaciones prácticas.

Por otro lado, los investigadores del campo reconocen que es necesaria una mayor evidencia translingüística que apoye la plausibilidad psicolingüística de un aprendizaje general no supervisado de una gramática formal a partir de técnicas estadísticas. En la actualidad no existen trabajos que se hayan propuesto probar tales enfoques para la inducción integral de sintaxis en lenguas flexivas y con orden libre de constituyentes como el español. Así pues, en última instancia el objetivo final de nuestro trabajo a futuro será aportar dicha evidencia translingüística, estudiando la factibilidad de inducir fenómenos sintácticos del español mediante técnicas estadísticas a partir de corpus no estructurado y modelos formales de aprendizaje no supervisado.

Aunque no demostramos necesariamente que el mecanismo por el cual se adquiere una gramática de un lenguaje natural involucre técnicas de clustering, sí demostramos la invalidez del APS en cuanto a que los PLD son suficientemente ricos para inducir una gramática formal (al menos, las categorías POS-tags) únicamente a partir de la información distribucional. Asimismo, dirigimos nuestra atención al debate epistemológico en torno del APS, tratando de arrojar cierta luz sobre confusiones generalizadas en cuanto a los mecanismos lógicos inductivos que podrían actuar como el sustrato cognitivo de los mecanismos generales de aprendizaje que modelizamos en nuestra investigación.

Consideramos entonces que el mérito de la presente investigación es abarcar modelos de inducción de fenómenos sintácticos que puedan aportar renovada evidencia al debate acerca de la adquisición

del lenguaje; en especial, si consideramos que este tipo de enfoques para el español –un idioma particularmente desafiante por el orden libre de sus constituyentes sintácticos– ha venido escaseando durante la última década en el panorama global del estado del arte dentro del paradigma estadístico de la lingüística computacional. En última instancia, la evidencia psicolingüística debería ser refrendada por la neurología, las ciencias cognitivas o incluso la biolingüística, pero la plausibilidad de dicha evidencia mediante una modelización efectiva es claramente un asunto para la agenda actual de la lingüística computacional.

Referencias bibliográficas

1. Berg-Kirkpatrick, Taylor, Alexandre Côté, John Denero y Dan Klein. 2010. Painless unsupervised learning with features. En *Proceedings of NAACL 2010*, pp.582-590. Los Angeles.

2. Böhm, Christian, Christos Faloutsos, JiaYu Pan y Claudia Plant. 2006. Robust information-theoretic clustering. En *Proceedings of the 12th ACM SIGKDD International Conference knowledge discovery and data mining*, pp.65-75. Philadelphia.

3. Brown, Peter, Vincent Della Pietra, Peter Desouza, Jennifer Lai y Robert Mercer. 1992. Class-based n-gram models of natural language. En *Computational Linguistics* 18(4):467-479.

4. Chomsky, Noam. 1957. *Estructuras sintácticas*. México. Siglo XXI.

5. ----------.1959. A review of B.F. Skinner's verbal behavior. En *Language* (35):26-58.

6. ----------. 1975. *Reflexiones sobre el lenguaje*. Buenos Aires. Sudamericana.

7. Christophe, Anne, Séverine Milotte, Savita Bernal y Jeffrey Lidz. 2008. Bootstrapping lexical and syntactic acquisition. En *Language and Speech* (51):61-75.

8. Christodoulopoulos, Christos, Sharon Goldwater y Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? En *Proceedings of the 2010 conference on empirical methods in Natural Language Processing*: 575–584. Cambridge, Massachusetts.

9. Clark, Alexander. 2000. *Inducing syntactic categories by context distribution clustering*. En Proceeding of the CoNLL-2000 and LLL-2000, pp.91-94. Lisboa

10. ----------. 2002. *Unsupervised language acquisition: theory and practice*. Tesis de doctorado. University of Sussex.

11. ----------. 2003. Combining distributional and morphological information for part of speech induction. En *Proceedings of EACL 2003*, pp.59-66. Morristown.

12. Clark, Alexander y Shalom Lappin. 2011. Computational learning theory and language acquisition. En Ruth Kempson, Tim Fernando, y Nicholas Asher (eds.). *Handbook of the philosophy of science*. Volumen 14: Philosophy of Linguistics, pp.1-34. Oxford. Elsevier.

13. Clark, Alexander y Shalom Lappin. 2013. Complexity in language acquisition. En *Topics in Cognitive Science* (5):89-110.

14. Deerwester, Scott, Susan Dumais, George Furnas, Thomas Landauer y Richard Harshman. 1990. Indexing by Latent Semantic Analysis. En *Journal of American Society of Information Sciences* 1(6):391-407.

15. Dromi, Esther. 1987. *Early lexical development*. Nueva York. Cambridge University Press.

16. Elghamry, Khaled. 2004. *A generalized cue-based approach to the automatic acquisition of subcategorization frames*. Tesis de doctorado. Indiana University.

17. Graça, João, Kuzman Ganchev, Luísa Coheur, Fernando Pereira y Ben Taskar. 2011. Controlling Complexity in Part-of-Speech Induction. En *Journal of Artificial Intelligence Research* (41):527-551.

18. Johnson, Kent. 2004. Gold's theorem and cognitive scoences. En *Philosophy of Science* (71):571-592.

19. Jusczyk, Peter, Derek Houston y Mary Newsome. 1999. The beginnings of word segmentation in English-learning infants. En *Cognitive Psychology* (39):159-207.

20. Klein, Dan y Christopher Manning. 2004. Corpus based induction of syntactic structure: models of dependency and constituency. En *Proceedings of ACL 2004*, pp.478-485. Barcelona.

21. Lappin, Shalom y Stuart Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. En *Linguistics* (43):393-427.

22. Levy, Yonata. 1985. It's frogs all the way down. En *Cognition* (15):75-93.

23. Manning, Christopher y Hinrich Schütze. 1999. *Foundations of statistical Natural Language Processing*. Cambridge, Massachusetts. MIT Press.

24. Martin, Sven, Jörg Liermann y Hermann Ney. 1998. Algorithms for bigram and trigram word clustering. En *Speech Communication* (24):19-37.

25. Mehler, Jacques, Anne Christophe y Franck Ramus. 1998. What we know about the initial state of language. En *Proceedings of the 1st mind-brain articulation project symposium*, pp.51-75. Tokio.

26. Mintz, Toben. 2003. Frequent frames as a cue for grammatical categories in child directed speech. En *Cognition* 90(1):91-117.

27. Nath, Joydeep, Monojit Choudhury, Animesh Mukherjee, Chris Biemann y Niloy Ganguly. 2008. Unsupervised Parts-of-Speech induction for Bengali. En *Proceedings of LREC'08, European Language Resources Association (ELRA)*, pp.1220-1227. Marrakesh.

28. Pinker, Steven. 1979. Formal models of language learning. En *Cognition* (7):217-282.

29. Popova, Maria. 1973. Grammatical elements of language in the speech of pre-school children. En Ferguson, Charles y Dan Slobin (eds.). *Studies of child language developments*. Nueva York. Holt, Rinehart & Winston.

30. Pullum, Geoffrey. 1996. Learnability, hyperlearning and the argument from the poverty of the stimulus. En *Parasession on learnability, 22nd annual meeting of the Berkeley Linguistics Society*, pp.498-513. Berkeley, California.

31. Redington, Martin, Nick Charter y Steven Finch. 1998. Distributional information: a powerful cue for acquiring syntactic categories. En *Cognitive Science* 22(4):425-469.

32. Schütze, Hinrich. 1993. Part-of-speech induction from scratch. En *Proceedings of the 31st annual conference of the Association for Computational Linguistics*, pp.251-258. Columbus.

33. Shi, Rushen, Janet Werker y James Morgan. 1999. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. En *Cognition* 72(2):11-21.

34. Wang, Hao. 2012. *Acquisition of functional categories*. Tesis de doctorado. University of Southern California.

35. Zhitomirsky-Geffet, Maayan e Ido Dagan. 2009. Bootstrapping distributional feature vector quality. En *Computational Linguistics* (35):435-461.

Anexo - Listado completo de etiquetas morfosintácticas

Nro.	*Tag*	Ejemplos
1	AJ0	adjetivo neutro en número (**bello** en "*lo bello*")
2	AJ1	adjetivo singular (**amable**)
3	AJ2	adjetivo plural (**amables**)
4	AJC	adjetivo comparativo (**peor**)
5	AJS	adjetivo superlativo (**pésimo**)
6	AT0	artículo neutro (**lo**)
7	AT1	artículo singular (**la**)
8	AT2	artículo plural (**los**)
9	AV0	adverbio (**seguidamente**)
10	AVQ	adverbio interrogativo (**cuándo**)
11	CJC	conjunción coordinante (**y**)
12	CJS	conjunción subordinante (excepto *que*) (**cuando**)
13	CJT	conjunción subordinante (**que** en "*dijo que...*")
14	CRD	adjetivo numeral cardinal (**tres**)
15	DPS	determinante posesivo (**su**, **mi**)
16	DT1	determinante definido singular (**aquel** en "*aquel hombre*")
17	DT2	determinante definido plural ("***aquellos** hombres*", "***todos** los hombres*")
18	EX0	existencial (**hay**)
19	ITJ	interjección (**ah**, **ehmm**)
20	NN0	sustantivo neutro en número (**virus**)
22	NN1	sustantivo singular (**lápiz**)
22	NN2	sustantivo plural (**lápices**)
23	NNP	sustantivo propio (**Rafael**)
24	ORD	adjetivo numeral ordinal (**sexto**)
25	PND	pronombre demostrativo (**éste**, **esto**)
26	PNI	pronombre indefinido (**ninguno**, **todo**)
27	PNP	pronombre personal (**tú**)
28	PNQ	pronombre interrogativo (**quién**)
29	POS	pronombre posesivo (**mío**)
30	PPE	pronombre personal enclítico (*dar*-**lo**, se cuasi-reflejo ("*morir**se***", "*él **se** cayó*")
31	PRP	preposición (excepto *de*) (**sin**)
32	REL	pronombre relativo (**quien** en "*el presidente, quien avisó...*")
33	SEP	se pasivo ("***se** venden casas*") e impersonal ("***se** reprimió a los manifestantes*")
34	VBG	gerundio de verbo cópula (**siendo**)
35	VBI	infinitivo de verbo cópula (**ser**)
36	VBN	participio de verbo cópula (**sido**)
37	VBZ	verbo cópula conjugado (**es**)
38	VM0	infinitivo de verbo modal (**soler**)
39	VMZ	verbo modal conjugado (**debía**)
40	VMG	gerundio de verbo modal (**pudiendo**)
41	VMN	participio de verbo modal (**podido**)
42	VVG	gerundio de verbo léxico (**obrando**)
43	VVI	infinitivo de verbo léxico (**vivir**)
44	VVN	participio de verbo léxico (**cifrado**)
45	VVZ	verbo léxico conjugado (**vive**)
46	XX0	adverbio de negación (**no**)
47	$$$	fin de oración

Técnicas estadísticas de clasificación. Una aplicación en la clasificación de textos según el género: Textos Científicos y Textos No Científicos

Statistical Techniques in Classification. An Application to Text Classification According to Gender: Scientific – Non scientific

Celina Beltrán
Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina
beltranc36@yahoo.com.ar

Abstract

The problem of unit classification in groups or known population samples offers great interest to Statistics; as a consequence of this several techniques have been developed to fulfill this purpose. This work is aimed to classify scientific and non scientific texts comparing the analysis of classification (CT) and the logistic regression (LR) trees. The scientific texts are abstracts of papers published on journals and conference proceedings coming from different disciplines. The non scientific texts are news report of general interest published on the web page of Argentine newspapers. The information obtained from the morphological analysis of these texts is employed as explanatory variable of the multivariable technique applied in this work. The performance of the techniques was measured by means of the false classification rate (FCR), the precision rate (PR) and the coverage rate (CO) estimated by a sample text not included in the prediction model neither in the tree construction. The classification tree showed a FCR lower than the logistic model while the scientific text samples showed a major precision.

For the CT, the FCR, the PR and the CO resulted in 4%, 84% and 96% for the scientific texts and 28%, 92% and 72% for the non scientific texts, respectively.

For the LR model, the FCR, the PR and the CO resulted in 14%, 83% and 86% for the scientific texts and 26%, 77% and 74% for the non scientific texts, respectively.

Key words: Multivariable logistic regression – Classification Trees – Automatic text analysis– Text classification.

Resumen

El problema de la clasificación de unidades en grupos o poblaciones conocidas es de gran interés en estadística, por esta razón se han desarrollado varias técnicas para cumplir este propósito. Este trabajo se propone la clasificación de textos científicos y no científicos comparando las técnicas de Árboles de Clasificación (AC) y Regresión logística (RL). Los textos científicos corresponden a resúmenes de publicaciones en revistas científicas y actas de congresos de distintas disciplinas y los textos no científicos corresponden a noticias periodísticas de interés general publicadas en páginas web de periódicos argentinos. La información resultante del análisis morfológico de dichos textos es utilizada como variables explicativas en las técnicas multivariadas aplicadas en este trabajo. El

desempeño de las técnicas fue medido con la tasa de mala clasificación (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos no incluidos en la estimación del modelo y construcción del árbol. El árbol de clasificación presentó una TMC inferior a la del modelo logístico logrando clasificar con mayor precisión los textos científicos.

Para el AC la TMC, PR y CO resultaron 4%, 84% y 96% para los textos científicos y 28%, 92% y 72% para los textos no científicos, respectivamente.

Para el modelo de RL la TMC, PR y CO resultaron 14%, 83% y 86% para los textos científicos y 26%, 77% y 74% para los textos no científicos, respectivamente.

Palabras claves: Regresión logística multivariada, árboles de clasificación, análisis automático de textos, clasificación de textos.

1. INTRODUCCION

Este trabajo se propone la realización del análisis automático de distintos tipos de textos: científicos y no científicos. Los textos científicos corresponden a resúmenes de publicaciones en revistas científicas y actas de congresos de distintas disciplinas y los textos no científicos corresponden a noticias periodísticas de interés general publicadas en páginas web de periódicos argentinos. Se recurre al analizador morfológico Smorph, implementado como etiquetador, para asignar categoría a todas las ocurrencias lingüísticas.

La información resultante del análisis morfológico de dichos textos es utilizada para conformar una base de datos sobre la cual se aplican las técnicas multivariadas de clasificación: Regresión logística y Árboles de clasificación.

El desempeño de las técnicas es evaluado con tres medidas: la tasa de mala clasificación (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos, muestra de prueba, no incluidos en la estimación del modelo y construcción del árbol.

2. MATERIAL Y METODOS

2.1. Diseño de la muestra

El marco muestral para la selección de la muestra de los textos científicos está compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a distintas disciplinas. Los textos periodísticos fueron seleccionados de un corpus mayor utilizado por el equipo de investigación INFOSUR. Este corpus se construyó con noticias extraídas de las páginas web de periódicos argentinos (noticias de tipo general, no especializadas en español). La unidad de muestreo fue el texto y la selección de la muestra se llevó a cabo empleando un diseño muestral estratificado.

Luego de obtener las muestras de los estratos, fueron evaluadas y comparadas respecto al número medio de palabras por texto. Se requiere esta evaluación para evitar que la comparación entre los géneros se vea afectada por el tamaño de los textos.

La muestra final para este trabajo quedó conformada de la siguiente manera:

Tabla 1. Conformación de la muestra final

Muestra	Nro. de textos	Cantidad de palabras
Científico	90	14.554
No científico	60	8.080

2.2. Etiquetado de los textos

El software Smorph, analizador y generador morfosintáctico desarrollado en el Groupe de Recherche dans les Industries de la Langue (Universidad Blaise-.Pascal, Clermont II) por Salah Aït-Mokhtar (1998) realiza en una sola etapa la tokenización y el análisis morfológico. A partir de un texto de entrada se obtiene un texto lematizado con las formas correspondientes a cada lema (o a un subconjunto de lemas) con los valores correspondientes. Se trata de una herramienta declarativa, la información que utiliza está separada de la maquinaria algorítmica, en consecuencia, puede adaptarse a distintos usos. Con el mismo software se puede tratar cualquier lengua si se modifica la información lingüística declarada en sus archivos.

Smorph compila, minimiza y compacta la información lingüística que queda disponible en un archivo binario. Los códigos fuente se dividen en cinco archivos: Códigos ASCII, Rasgos, Terminaciones, Modelos y Entradas.

En el archivo **entradas**, se declaran los ítems léxicos acompañados por el modelo correspondiente. Este indicador de modelo oficia de enlace con el archivo modelos, en el que se especifica la información morfológica y las terminaciones que se requieren en cada ítem.

En el archivo **modelos**, se introduce la información correspondiente a los modelos de flexiones morfológicas. A un conjunto de terminaciones se le asocia el correspondiente conjunto de definiciones morfológicas. En el archivo **terminaciones** es necesario declarar todas las terminaciones que son necesarias para definir los modelos de flexión, se declaran una a continuación de otra, separadas por un punto. Para construir los modelos se recurre a rasgos morfológico- sintácticos. En el archivo **rasgos**, se organizan jerárquicamente las etiquetas. En el archivo de códigos **ASCII** se especifican, entre otros, los caracteres separadores, las equivalencias entre mayúsculas y minúsculas. El archivo “data”, contiene los nombres de cada uno de los cinco archivos descriptos anteriormente.

El módulo post-smorph **MPS** es un analizador que recibe en entrada una salida Smorph (en formato Prolog) y puede modificar las estructuras de datos recibidos. Ejecuta dos funciones principales: la Recomposición y la Correspondencia, que serán útiles para resolver las ambigüedades que resulten del análisis de Smorph.

La información contenida en estos archivos es la presentada en Beltrán (2009).

2.3. Diseño y desarrollo de la base de datos

La información que contiene la base de datos es el resultado del análisis de Smorph-Mps almacenada en un archivo de texto. La información resultante del análisis morfológico se dispuso en una matriz de dimensión: tantas filas como cantidad de objetos lingüísticos tenga el texto y tantas columnas como ocurrencia+lema+valores. De esta manera se obtuvo una base de datos que posee la información del texto, ocurrencia, lema y etiqueta asignada. Luego, a partir de esta base de datos por palabra (cada unidad o fila es una palabra analizada del texto), se confeccionó la base de datos

por documento que es analizada estadísticamente. La información registrada en esta base corresponde a las siguientes variables:

- CORPUS: Corpus al que pertenece el texto
- TEXTO: Identificador del texto dentro del corpus
- Adj: cantidad de adjetivos del texto
- Adv: cantidad de adverbios del texto
- Cl: cantidad de clíticos del texto
- Cop: cantidad de copulativos del texto
- Det: cantidad de determinantes del texto
- Nom: cantidad de nombres (sustantivos) del texto
- Prep: cantidad de preposiciones del texto
- V: cantidad de verbos del texto
- Otro: cantidad de otras etiquetas del texto
- Total_pal: cantidad total de palabras del texto

2.4. Metodología Estadística

Uno de los problemas de los que se ocupan las técnicas estadísticas multivariadas es la clasificación de objetos o unidades en grupos o poblaciones. Dos enfoquen agrupan a las técnicas de clasificación. Uno de ellos es cuando se conocen los grupos o categorías y se pretende ubicar los individuos dentro de estas categorías a partir de los valores de ciertas variables. El segundo enfoque, que no es el utilizado en este trabajo, ocurre cuando no se conocen los grupos de antemano y se pretende establecerlos a partir de los datos observados.

Referido al primer enfoque, una de las técnicas más utilizadas es la Regresión Logística. La regresión logística estima la probabilidad de un suceso en función de un conjunto de variables explicativas. Este modelo expresa matemáticamente la probabilidad de pertenencia a uno de los grupos, de manera que es posible calcularlas y asignar cada unidad al grupo cuya probabilidad de pertenencia estimada sea mayor. Otra técnica muy utilizada son los Árboles de Clasificación que crean una serie de reglas basadas en las variables predictoras que permiten asignar una nueva observación a una de las categorías o grupo.

2.4.1. Árboles de Clasificación

Los árboles de clasificación (AC) se emplean para asignar unidades experimentales a las clases de una variable dependiente a partir de sus mediciones en uno o más predictores. En esta aplicación, los AC se emplean para asignar textos al género al que corresponde: CIENTIFICO – NO CIENTIFICO a partir de la información relevada en el análisis morfológico automático de los textos.

Es un algoritmo que genera un árbol de decisión en el cual las ramas representan las decisiones y cada una de ellas genera reglas sucesivas que permiten continuar la clasificación. Estas particiones

recursivas logran formar grupos homogéneos respecto a la variable respuesta (en este caso el género a la que pertenece el texto). El árbol determinado puede ser utilizado para clasificar nuevos textos.

Es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. Comienza el algoritmo con un nodo inicial o raíz a partir del cual se divide en dos sub-grupos o sub-nodos, esto es, se elige una variable y se determina el punto de corte de modo que las unidades pertenecientes a cada nuevo grupo definido sean lo más homogéneas posible. Luego se aplica el mismo procedimiento de partición a cada sub-nodo por separado. En cada uno de estos nodos se vuelve a repetir el proceso de seleccionar una variable y un punto de corte para dividir la muestra en dos partes homogéneas.

Las divisiones sucesivas se determinan de modo que la heterogeneidad o impureza de los sub-nodos sea menor que la del nodo de la etapa anterior a partir del cual se forman. El objetivo es particionar la respuesta en grupos homogéneos manteniendo el árbol lo más pequeño posible.

La función de impureza es una medida que permite determinar la calidad de un nodo, esta se denota por i(t). Si bien existen varias medidas de impureza utilizadas como criterios de partición, una de las más utilizadas está relacionada con el concepto de entropía

$$i(t) = \sum_{j=1}^{K} p(j/t).\ln p(j/t)$$

donde j = 1, ..., k es el número de clases de la variable respuesta categórica y p(j|t) la probabilidad de clasificación correcta para la clase j en el nodo t. El objetivo es buscar la partición que maximiza

$$\Delta i(t) = -\sum_{j=1}^{K} p(j/t).\ln p(j/t).$$

De todos los árboles que se pueden definir es necesario elegir el óptimo. El árbol óptimo será aquel árbol de menor complejidad cuya tasa de mala clasificación (TMC) es mínima. Generalmente esta búsqueda se realiza comparando árboles anidados mediante validación cruzada. La validación cruzada consiste, en líneas generales, en sacar de la muestra de aprendizaje o entrenamiento una muestra de prueba, con los datos de la muestra de aprendizaje se calculan los estimadores y el subconjunto excluido es usado para verificar el desempeño de los estimadores obtenidos utilizándolos como “datos nuevos”.

2.4.2. Regresión logística

La regresión logística es utilizada en situaciones en las cuales el objetivo es describir la relación entre una variable respuesta categórica, en este caso dicotómica, y un conjunto de variables explicativas que pueden ser tanto categóricas como cuantitativas.

Sea **x** un vector de p variables independientes, esto es, $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. La probabilidad condicional de que la variable *y* tome el valor 1 (presencia de la característica estudiada), dado valores de las covariables **x** es:

$$P(y = 1/X) = \pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

donde

$$g(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

β_0 es la constante del modelo o término independiente

p el número de covariables

β_i los coeficientes de las covariables

x_i las covariables que forman parte del modelo.

Si alguna de las variables independientes es una variable discreta con k niveles se debe incluir en el modelo como un conjunto de k-1 "variables de diseño" o "variables dummy". El cociente de las probabilidades correspondientes a los dos niveles de la variable respuesta se denomina odds y su expresión es:

$$\frac{P(y=1|X)}{1-P(y=1|X)} = e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}$$

Si se aplica el logaritmo natural, se obtiene el modelo de regresión logística:

$$\log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right) = \log(e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

En la expresión anterior el primer término se denomina logit, esto es, el logaritmo de la razón de proporciones de los niveles de la variable respuesta.

Los coeficientes del modelo se estiman por el método de Máxima Verosimilitud y la significación estadística de cada uno de los coeficientes de regresión en el modelo se puede verificar utilizando, entre otros, el test de Wald y el test de razón de verosimilitudes.

Una cuestión importante en este tipo de análisis es determinar si todas las variables consideradas en la función de discriminante contienen información útil y si solamente algunas de ellas son suficientes para diferenciar los grupos. Dado que las variables utilizadas para explicar la respuesta es probable que estén correlacionadas, es posible también que compartan información. Por lo tanto, se puede buscar un subgrupo de variables mediante algún criterio de modo tal que las variables excluidas no contengan ninguna información adicional. Existen varios algoritmos de selección de variables, entre ellos podemos citar:

Método forward: comienza por seleccionar la variable más importante y continúa seleccionando las variables más importantes una por vez, usando un criterio bien definido. Uno de estos criterios involucra el logro de un nivel de significación deseado pre-establecido. El proceso termina cuando ninguna de las variables restantes encuentra el criterio pre-especificado.

Método backward: comienza con el modelo más grande posible. En cada paso descarta la variable menos importante, una por vez, usando un criterio similar a la selección forward. Continúa hasta que ninguna de las variables pueda ser descartada.

Selección paso a paso: combina los dos procedimientos anteriores. En un paso, una variable puede entrar o salir desde la lista de variables importantes, de acuerdo a algún criterio pre-establecido.

En este trabajo se utilizó como evaluación del ajuste del modelo el test de Hosmer-Lemeshow y, dado que el modelo es utilizado para clasificar unidades, se utilizó también la tasa de mala clasificación calculada sobre la muestra independiente excluida en la etapa de estimación.

3. RESULTADOS

En Beltrán 2013 se realizó un análisis exploratorio en el cual se evidencian las características que discriminan los corpus de textos en estudio. En dicho estudio se evidenció que existen diferencias significativas entre los corpus respecto al tamaño de los textos (número de palabras por texto). Esta situación llevó a realizar las sucesivas comparaciones sobre los porcentajes o proporciones de las categorías gramaticales, hallando diferencias significativas ($p<0.05$) para todas las categorías gramáticas excepto la proporción de clíticos y de verbos en los documentos analizados. Asimismo, en un análisis de componentes principales, se dispusieron los textos en el plano de proyección demostrando que los textos procedentes del corpus No Científico presentan un mayor número de adverbios, respecto a las restantes categorías, que los textos Científicos.

3.1. Árboles de Clasificación

Se aplicó la técnica de Árboles de Clasificación para obtener reglas de clasificación que permitan asignar los textos en dos poblaciones, definidas por el género al que pertenecen: CIENTÍFICO y NO CIENTÍFICO. De la misma manera que en el apartado previo, los predictores utilizados corresponden a la distribución de las distintas categorías morfológicas halladas en el análisis automático de los textos (proporción de cada categoría morfológica).

Las variables que mostraron una buena discriminación de los grupos son la proporción de adverbios, nombres, adjetivos, preposición y verbos. El árbol final presenta 10 nodos terminales.

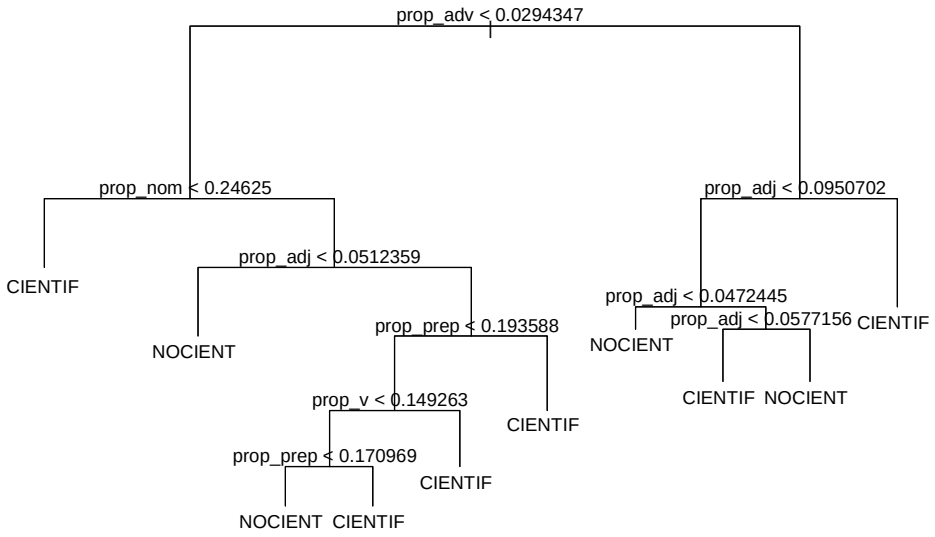

Gráfico 1: Árbol de clasificación

El gráfico 1 muestra el árbol resultante de esta aplicación. La variable regresora más fuertemente asociada con el género es la proporción de adverbios en el texto, categorizada como superiores e inferiores a 0.029 (2.9%). El corpus general queda así dividido en dos grupos, los que presentan más del 2.9% de adverbios y los que no. Luego intervienen en las sucesivas subdivisiones el número de nombres, adjetivos, verbos y preposiciones. Interpretando el árbol resultante, se

encuentran 10 perfiles de textos (que corresponden a los 10 nodos terminales) asociados con una de los dos géneros. Estos son:

- Textos con un porcentaje de adverbios inferior al 2.9% y un porcentaje de nombres menor a 25% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25% y de adjetivos inferior al 5% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición menor al 17% y de verbos menor al 15% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición entre 17% y 19%, y de verbos menor al 15% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición inferior al 19%, y de verbos mayor al 15% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición mayor al 19% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos inferior al 4.7% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos entre 4.7% y 5.7% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos entre 5.7% y 9.5% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos superior al 9.5% son clasificados como textos CIENTÍFICOS.

El árbol final fue evaluado utilizando la muestra de prueba, no fue utilizada en la construcción del mismo, hallando una tasa de mala clasificación del 14%, siendo 4% para los textos científicos y 28% para los no científicos. Respecto a la precisión y cobertura fueron de 84% y 96% para el género CIENTÍFICO y de 92% y 72% para los textos NO CIENTÍFICOS, respectivamente.

Tabla 2: Tasa de error estimada, Precisión y Cobertura

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	4%	28%
Precisión	84%	92%
Cobertura	96%	72%

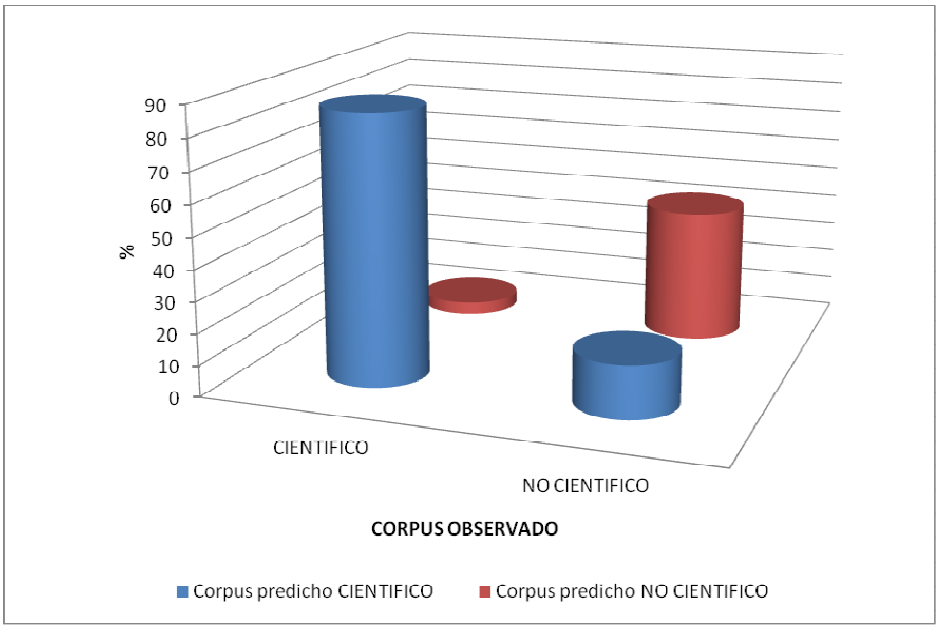

Grafico 2: Clasificación de textos según género mediante Árboles de Clasificación

3.2. Análisis de Regresión Logística

Se realizó un análisis de regresión logística para obtener una regla de clasificación que permita asignar los textos en estas dos poblaciones, definidas por el género al que pertenecen (Científico / No científico), en base a la frecuencia de cada categoría gramatical en el texto.

Para determinar cuáles categorías gramaticales son las responsables de la discriminación se utilizaron los tres algoritmos de selección de variables. En todos los casos se llega al mismo resultado: las variables que logran diferenciar a los dos corpus de textos analizados son el número de adjetivos, adverbios, conjunciones copulativas, determinantes, nombres y preposiciones.

Tabla 3: Coeficientes del modelo de regresión logística

Estimación máximo verosímil					
Coeficiente	**gl**	**Estimador**	**Error estándar**	**Est. Chi-cuadrado**	**Prob. asociada**
Intercepto	1	-0.6868	0.5507	1.5554	0.2123
adjetivos	1	0.1694	0.0562	9.0777	0.0026
adverbios	1	-0.3106	0.0800	15.0862	0.0001
Conj. Cop.	1	0.2769	0.1073	6.6566	0.0099
determinantes	1	0.1216	0.0464	6.8795	0.0087
nombres	1	-0.1995	0.0464	18.5044	<.0001
preposiciones	1	0.1575	0.0544	8.3925	0.0038

Este modelo permite, mediante la utilización de los coeficientes estimados, calcular para cada texto la probabilidad de pertenecer a cada uno de los corpus definidos por el género.

$$P(\in Cien / X) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}}$$

$$P(\in No\ Cien / X) = \pi(x) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}}$$

Con este criterio un texto es asignado al corpus cuya probabilidad es máxima.

La bondad del ajuste se evaluó mediante el test de Hosmer-Lemeshow y la tasa de error de clasificación. Con el modelo de regresión logística obtenido durante la selección de variables se obtuvo una tasa de error global, estimada sobre el corpus de prueba, del 20% y la probabilidad asociada en el test de bondad de ajuste es p=0.9696 evidenciando lo adecuado del modelo. La tabla 4 presenta las medidas de precisión, cobertura y tasa de error para cada género.

Tabla 4: Tasa de error estimada, Precisión y Cobertura

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	14%	26%
Precisión	83%	77%
Cobertura	86%	74%

Tabla 5: Razones de odds estimadas

Razón de odds			
Efecto	**Estimación puntual**	**IC 95%**	
adjetivos	1.185	1.061	1.323
adverbios	0.733	0.627	0.857
Conj. Cop.	1.319	1.069	1.628
determinantes	1.129	1.031	1.237
nombres	0.819	0.748	0.897
preposiciones	1.171	1.052	1.302

Los coeficientes del modelo de regresión logística permiten la interpretación de la misma. La razón de odds para el número de adjetivos es 1.19 lo cual indica que la chance de clasificar a un texto como Científico se incrementa en un 19% al aumentar en número de adjetivos en una unidad. Con respecto al número de adverbios la razón de odds es menor a la unidad por lo tanto si se interpreta el recíproco, 1/0.73=1.36, significa que la chance de clasificar un texto en el corpus No Científico aumenta un 36% al incrementarse en una unidad el número de adverbios. Si analizamos el efecto de las conjunciones copulativas, determinantes y preposiciones, al incrementar en una unidad cada una de estas categorías gramaticales, la chance de clasificar un texto como Científico se incrementa en un 32%, 13% y 17% respectivamente. Al igual que el efecto del número de adverbios, la

probabilidad de clasificar un texto como No Científico se incrementa en un 22% (1/0.82=0.22) al aumentar en una unidad la cantidad de nombres en el texto.

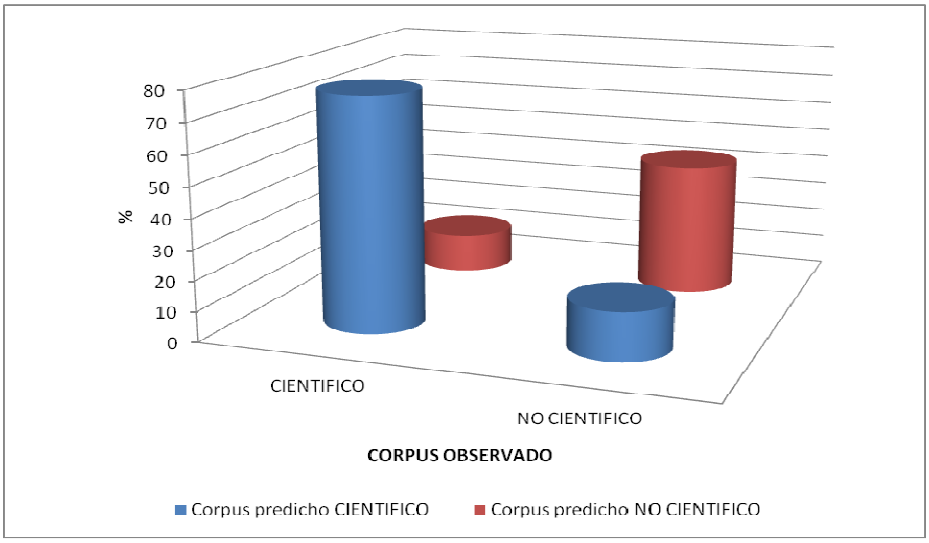

Grafico 3: Clasificación de textos según género mediante Regresión logística

4. CONCLUSIONES

Los resultados del análisis morfológico de los textos se analizaron teniendo en cuenta simultáneamente todas las mediciones realizadas sobre ellos.

El desempeño de las técnicas fue medido con la tasa de mala clasificación (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos no incluidos en la estimación del modelo y construcción del árbol. El árbol de clasificación presentó una TMC inferior a la del modelo logístico logrando clasificar con mayor precisión los textos científicos. Para el AC la TMC, PR y CO resultaron 4%, 84% y 96% para los textos científicos y 28%, 92% y 72% para los textos no científicos, respectivamente. Para el modelo de RL la TMC, PR y CO resultaron 14%, 83% y 86% para los textos científicos y 26%, 77% y 74% para los textos no científicos, respectivamente.

La diferencia en la tasa de mala clasificación sólo se diferenció en el corpus de textos científicos para el cual con el árbol se obtuvo un 4% de mala clasificación versus un 14% para el modelo de regresión logística.

En ambos tipos de análisis, las diferencias entre los dos tipos de textos están centradas principalmente en el porcentaje de adverbios, adjetivos, nombres y preposiciones presentes. Sin embargo, en el modelo de regresión logística han intervenido otras variables en la discriminación como los determinantes y conjunciones copulativas; mientras que el árbol de clasificación utiliza el porcentaje de verbos, categoría morfológica no utilizada en la regresión.

Una ventaja observada en el árbol de clasificación es la adaptación para recoger el comportamiento no aditivo de las variables predictoras, de manera que las interacciones se incluyen de manera automática. Sin embargo, en esta técnica se pierde información al tratar a las variables predictoras continuas como variables dicotómicas.

Referencias

Aitchison J. 1983. The Statistical Analysis of Compositional Data. Chapman & Hall, London.

Beltrán, C., Bender, C., Bonino, R., Deco, C., Koza, W., Méndez, B., Moro, Stella Maris. 2008 Recursos informáticos para el tratamiento lingüístico de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2009 Modelización lingüística y análisis estadístico en el análisis automático de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2010 Estudio y comparación de distintos tipos de textos académicos: Biometría y Filosofía. Revista de Epistemología y Ciencias Humanas. Grupo IANUS. Rosario.

Beltrán, C. 2013 Estudio exploratorio para la comparación de distintos tipos de textos: Textos Científicos y Textos No Científicos. Revista de Epistemología y Ciencias Humanas. Grupo IANUS. Rosario.

Bès,Gabriel, Solana, Z y Beltrán, C. 2005 Conocimiento de la lengua y técnicas estadísticas en el análisis lingüístico en Desarrollo, implementación y uso de modelos para el procesamiento automático de textos (ed. Víctor Castel) Facultad de Filosofía y Letras, UNCUYO

Cuadras, C.M. 2008 Nuevos métodos de análisis multivariante. CMC Editions. Barcelona, España.

Johnson R.A. y Wichern D.W. 1992 Applied Multivariate Statistical Análisis. Prentice-Hall International Inc.

Khattre R. y Naik D. 1999 Applied Multivariate Statistics with SAS Software. SAS. Institute Inc. Cary, NC. USA.

Rodrigo Mateos, José Lázaro y Bès, Gabriel G. 2004 Análisis e implementación de clíticos en una herramienta declarativa de tratamiento automático de corpus. En VI Congreso de Lingüística General, Santiago de Compostela.

Solana, Z. Beltrán, C., Bender, C., Bonino, R., Deco, C., Koza, W., Méndez, B., Rodrigo, A., Tramallino, C. 2009 La interlengua de los aprendientes de español como L2. Aportes de la Lingüística Informática. GRUPO INFOSUR- Ediciones Juglaría.

Una propuesta para el tratamiento de los enclíticos en NooJ

A proposal to analyze the enclitics using NOOJ Tools

Rodolfo Bonino
Grupo INFOSUR
Universidad Nacional de Rosario, Argentina
rodolfobonino@yahoo.com.ar

Abstract

The fact that enclitics form a graphic unit with the verb and, in several cases, may produce graphic alterations (apocopation and chances in the accent) places these sequences between morphology and syntax. The present work tries to solve the empirical problem in the treatment of verb and enclitic sequences using the NooJ tools. In NooJ, the morphological grammars are used to analyze the internal structure of the lexical units and the syntax grammars are used for the relations between the different lexical units. If it is proper to deal with the object of study by means of syntactic grammars, the previous step is to create a productive grammar to analyze that object as a graphical unit formed by two lexical units. Thus, the enclitics are recognized as syntax elements and, consequently, it is possible to elaborate grammatical syntaxes analogous to those employed to treat proclitic and verbal sequences. It is also posed the modifications that must be introduced in the morphological grammar of the verbal system to obtain the graphical forms adopted by verbs when associated with enclitics. Finally, it is presented a brief analysis of the enclitic inserted in the compound forms and the verbal periphrasis.

Key words: NooJ, Spanish language, enclitics, verbs, compound tenses, verbal periphrasis.

Resumen

El hecho de que los enclíticos formen una unidad gráfica con el verbo y, en muchos casos, se produzcan alteraciones gráficas (apócope y cambios de tilde) sitúa a estas secuencias entre la morfología y la sintaxis. En el presente trabajo se intenta resolver el problema empírico que presenta el tratamiento de secuencias de verbos y enclíticos mediante la herramienta NooJ. En NooJ, las gramáticas morfológicas se utilizan para analizar la estructura interna de las unidades léxicas y las gramáticas sintácticas para las relaciones entre distintas unidades léxicas. Si se considera adecuado tratar el objeto de estudio mediante gramáticas sintácticas, el paso previo es crear una gramática productiva que lo analice como una unidad gráfica formada por dos unidades léxicas; así, los enclíticos son reconocidos como elementos de la sintaxis y, consecuentemente, es posible elaborar gramáticas sintácticas análogas a las que se utilizan para tratar secuencias de proclíticos y verbos. También se plantean las modificaciones que se deben introducir en la gramática morfológica del sistema verbal para obtener las formas gráficas que adoptan los verbos cuando se asocian con enclíticos. Finalmente, se presenta brevemente el análisis de los enclíticos insertos en las formas compuestas y las perífrasis verbales.

Palabras claves: NooJ, español, enclíticos, verbos, tiempo compuesto, perífrasis verbales.

1. INTRODUCCIÓN

En trabajos previos se propusieron modelos para el tratamiento automático de la morfología de las formas verbales simples [1] y la sintaxis de las formas compuestas y algunas perífrasis verbales [2]. Estas modelizaciones permiten el análisis de todas las formas simples y compuestas de los verbos incluidos en el diccionario NooJ que hemos confeccionado, excepto las que llevan enclíticos.

A las cuestiones generales que suscitan el análisis automático de los clíticos, los enclíticos suman una serie de problemas específicos derivados del hecho de que forman con el verbo una única palabra escrita, lo que implica:

a- Cambios en la acentuación ortográfica (*amando – amándolo*)

b- La elisión de caracteres (*amad – amaos / amemos –amémonos*)

En este trabajo se propone un procedimiento para lograr el análisis automático de secuencias donde un clíticos sigue a un infinitivo (*amarlo*), un gerundio (*amándolo*), un imperativo afirmativo (*ámalo / amalo*), y a las formas del presente del subjuntivo que se emplean como imperativo (*amémoslo, amémonos, amaos, ámelo, ámense,* etc.). Adicionalmente, se crean gramáticas que reconocen los enclíticos insertos en tiempos compuestos y en perífrasis verbales (*haberlo amado, habiéndolo amado, dejarlo de amar*, etc.).

2. NOOJ [3]

2.1. Generalidades

NooJ es un programa informático desarrollado por Max Silberztein a partir del año 2002, que cuenta con varios útiles para el tratamiento automático de las lenguas naturales[1]:

a) Gramáticas morfológicas y derivacionales (archivos .nof): modelos de flexión y derivación.

b) Diccionarios (archivos .dic): listas de palabras con diversos tipos de información lingüística.

c) Gramáticas productivas (archivos .nom): sistemas regulares o gráficos útiles para el tratamiento cadenas de caracteres con determinadas propiedades formales.

d) Gramáticas sintácticas (archivos .nog): sistemas regulares o gráficos útiles para el tratamiento de cadenas de caracteres formadas por dos o más unidades léxicas, generalmente, separadas por espacios en blanco.

Las gramáticas morfológicas están en la base de los diccionarios, pero estos pueden prescindir de aquellas, o sea, es posible elaborar directamente un diccionario donde se declaran como entradas todas las variantes morfológicas o derivacionales de una palabra, o bien crear gramáticas morfológicas y derivacionales que generen las variaciones a partir de una sola entrada del diccionario. Por ejemplo, se podría hacer un diccionario con las entradas *mesa* y *mesas*, pero también es posible hacer una gramática que produzca la variación de número e indicar al diccionario que determinadas palabras siguen ese modelo flexivo, de este modo, en el diccionario solo será necesario incluir el singular y el modelo de flexión; el plural se generará mediante la gramática. Evidentemente, este procedimiento resulta mucho más eficaz y económico porque el mismo modelo permite flexionar numerosas palabras (*abeja, boda, casa, decena,* etc.), en el caso de

[1] En [1] y [2] se explican las características generales del programa, acá me centraré en los aspectos pertinentes al trabajo desarrollado. Para una descripción detallada de los comandos sugiero consultar el manual del programa y para un entrenamiento en su empleo, el tutorial en español. Este material está disponible en línea en [3].

los verbos en español, que tienen una enorme riqueza flexiva las ventajas de las gramática flexivas son indiscutibles.

Como se mencionó, las gramáticas productivas permiten identificar determinadas secuencias de caracteres a partir de propiedades formales; por ejemplo, una gramática productiva puede reconocer como nombre propio cualquier secuencia de caracteres que comience por una letra mayúscula seguida por un número indefinido de letras minúscula (se reconocerá como nombre propio una secuencia como *Mxnwzyg* o cualquier palabra escrita con mayúscula); pero las gramáticas productivas también se pueden integrar con diccionarios y permiten declarar, por ejemplo que la cadena que se busca debe estar etiquetada en el diccionario como V (verbo) o CL (clítico).

Asimismo, las gramáticas sintácticas pueden operar aisladamente o en interacción con gramáticas productivas y diccionarios que, a su vez, pueden estar integrados con gramáticas flexivas, derivativas y productivas. En el primer caso, las entradas de la gramática serían palabras, por ejemplo, *SN = la →mesa* reconocerá únicamente al sintagma nominal *la mesa*; en el segundo caso las entradas podrían ser categorías declaradas previamente en el diccionario, por ejemplo, *SN = DET→N* (el nombre propio, generado por una gramática productiva, puede ser una subclase de N); esta gramática reconocerá como sintagma nominal cualquier secuencia de determinante y sustantivo, siempre que estos hayan sido previamente reconocidos por un diccionario o una gramática productiva.

En el esquema 1, se muestra cómo pueden interactuar las distintas herramientas del programa:

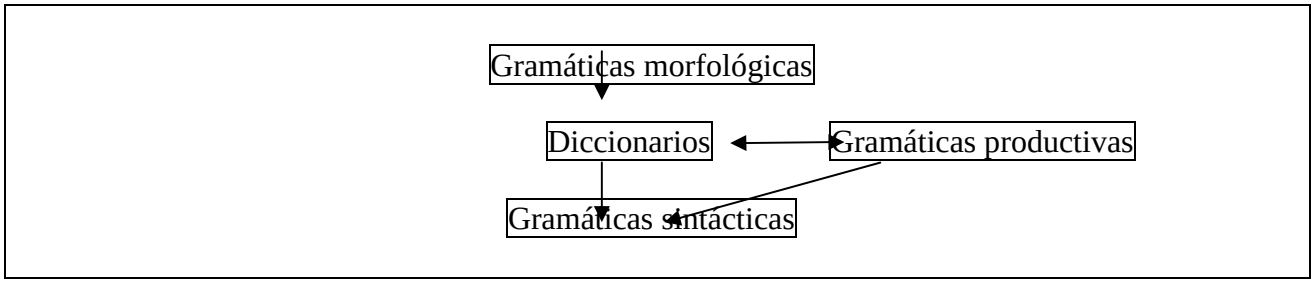

Figura 1: Modo de interacción de las diferentes herramientas del programa

2.2. Uso de variables

Las gramáticas productivas y sintácticas pueden recurrir a variables, que son elementos que presentan dos o más valores en alguno de sus rasgos, cuya variación incide en otro elemento o en la salida de la gramática; por ejemplo:

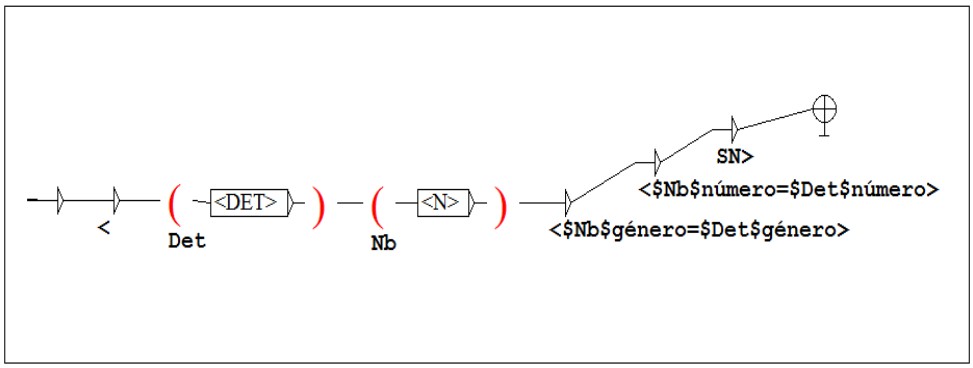

Figura 2: Uso de variables de concordancia

En esta gramática, los elementos entre paréntesis se tratan como variables. De este modo, se puede establecer la concordancia entre los valores de género y número del determinante <DET> y el susantivo <N>: la sentencia <NB$género=$Det$género> indica que, cualquiera sea el valor de género (masculino o femenino) de la segunda variable (Nb), debe ser igual al valor de la primera (Det); la sentencia <Nbnúmero=Detnúmero> expresa lo mismo con respecto al número.

Las gramáticas productivas también pueden utilizar variables y transmisión de rasgos de las entradas a las salidas:

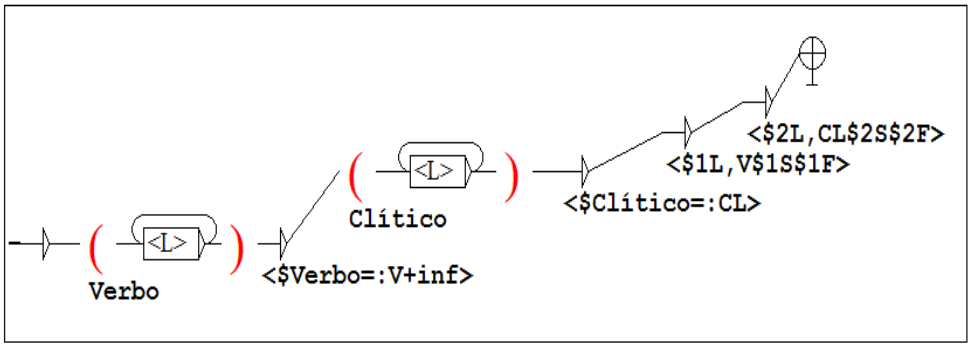

Figura 4: Variables y transmisión de rasgos en gramáticas productivas

En las entradas de esta gramática, <L> indica cualquier letra y el bucle, repetición indefinida; es decir, las entradas son cadenas de caracteres alfabéticos; <$Verbo=:V+inf> indica que la primera secuencia, definida como variable, debe estar previamente etiquetada como verbo infinitivo y <$Clítico=:CL>, que la segunda secuencia debe estar etiquetada como clítico; la sentencia <$1L,V$1S$1F> indica que, en la salida, la primera secuencia debe ser etiquetada como verbo con las mismas propiedades sintáctico semánticas y flexivas que la primera variable y <$2L,V$2S$2F>, que la segunda secuencia debe ser etiquetada como clítico con las mismas propiedades que la segunda variable. En el análisis de un texto, el programa aplica primero el diccionario y luego esta gramática.

3. EL OBJETO LINGÜÍSTICO

Como se ha señalado, el objeto lingüístico que pretendemos analizar son cadenas de verbos y enclíticos.

3.1. Las formas verbales

En español actual las formas verbales simples que admiten enclíticos son:

- el infinitivo,
- el gerundio,
- los imperativos,
- la primera persona del plural del presente del subjuntivo,
- la tercera del singular del presente del subjuntivo,
- la tercera persona del plural del presente del subjuntivo.

El hecho de que el presente del subjuntivo lleve enclíticos solo cuando se utiliza como imperativo no es relevante en esta instancia del análisis.

De las formas compuestas, las únicas que admiten enclíticos son el infinitivo (*haberlo amado*) y el gerundio (*habiéndolo amado*). Las perífrasis de infinitivo pueden llevar enclíticos unidos al infinitivo (*poder amarlo*); estos casos reciben el mismo análisis que los que no forman perífrasis. Cuando el auxiliar también aparece en infinitivo, el enclítico se puede adjuntar a él (*poderlo amar*), como se verá, en tal caso, el procedimiento de análisis será análogo al de los infinitivos compuestos.

3.2. Los clíticos

Si bien este trabajo se circunscribe al análisis formal de la secuencias de verbos y enclíticos, sin tener en cuenta las restricciones que imponen los verbos ni la función de los clíticos; en una etapa posterior se intentará avanzar en esos aspectos. Por lo tanto, los clíticos se clasifican con miras a este objetivo.

En [4], los clíticos se asocian, por un lado con los rasgos morfológicos de persona, género, número y caso, y, por otro, con el de reflexividad. Sin embargo, estas propiedades no tienen una expresión morfológica o léxica sistemática:

Los de tercera persona tienen formas diferenciales para reflexivo y no reflexivo; los no reflexivos diferencian caso acusativo y dativo[2]; los acusativos distinguen género y número y los dativos solo número. El reflexivo no tiene ninguna marca adicional.

Los de primera y segunda persona tienen rasgos diferenciales de persona y número, pero no distinguen género, caso ni reflexividad.

Se, además de ser pronombre reflexivo de tercera persona, puede ser dativo no reflexivo singular o plural cuando precede a un acusativo de tercera persona (*se lo dijo*), y cumple otras funciones que no se asocian al dativo ni al acusativo (índice de cuasi reflejo, marca de pasiva, marca de impersonalidad).

Una alternativa sería darle tantas entradas al diccionario como posibles características tenga; por ejemplo:

se, CL+ac+refl+3era+sg

se, CL+ac+refl+3era+pl

se, CL+dat+refl+3era+sg

se, CL+dat+refl+3era+pl

se, CL+dat+nrefl+3era+sg

se, CL+dat+nrefl+3era+pl

se, CL+crefl

se, CL+pasivo

se, CL+impersonal

Sin embargo, resulta mucho más económico caracterizarlos solo por sus posibilidades funcionales y sus rasgos diferenciales. En una etapa posterior, se intentará precisar los rasgos ambiguos mediante gramáticas sintácticas.

[2] Hay ambigüedades a causa del loísmo y el leísmo, que si bien son fenómenos prácticamente inexistentes en el habla estándar rioplatense, tienen distintos grados de incidencia en otras variedades del español.

En tal sentido, se les asigna la propiedad "Tipo de clítico" que los agrupa por sus posibilidades funcionales:

- Clítico tipo 1: *se* (acusativo reflexivo, dativo reflexivo, dativo no reflexivo, cuasireflejo, pasivo, impersonal).
- Clíticos tipo 2: *me, te, nos, os* (acusativo no reflexivo, dativo no reflexivo, acusativo reflexivo, dativo reflexivo, cuasireflejo).
- Clíticos tipo 3: *lo, los, la, las* (acusativo no reflexivo).
- Clíticos tipo 4: *le, les* (dativo no reflexivo).

El clítico tipo 1 solo tiene rasgo de persona; los tipo 2, persona y número; los tipo 3, persona, género y número; y los tipo 4, persona y número[3].

4. LA IMPLANTACIÓN EN NOOJ

Desde el punto de vista teórico, el tratamiento de las secuencias de verbos con enclíticos mediante una gramática morfológica, que presupone que el clítico es un morfema verbal, no sería ajena a nuestra tradición gramatical; por ejemplo, [5] señala: "Parece, pues, que los pronombres /me, te, etc./ son signos morfológicos que determinan el signo verbal del mismo modo que los signos morfológicos que constituyen sus desinencias" (pág. 149). Sin embargo, desde el punto de vista práctico, surge el inconveniente de que se deberían crear modelos diferentes para verbos que tienen la misma conjugación, pero admiten diferentes clíticos; lo que supone una proliferación de los modelos de conjugación. A esto se suma que, para ser coherente con esta perspectiva teórica, el diccionario también tendría que incluir las formas con proclítico, de modo que, además de la unidad léxica *ama*, tendría que reconocer *me ama, te ama, lo ama,* etc.

El tratamiento de los enclíticos mediante gramáticas sintácticas es la perspectiva teórica sostenida mayoritariamente por las gramáticas del español. La dificultad que presenta es que en NooJ este tipo de gramáticas opera con unidades léxicas que, en general, son palabras formales, es decir, secuencias de caracteres separadas por espacios en blanco. Cuando las unidades léxica no coinciden con palabras, como ocurre con los verbos y los enclíticos, es necesaria una operación previa de reconocimiento de cada una de las unidades que la conforman[4]. En el caso de las contracciones, esta operación se puede efectuar directamente en el diccionario:

al,<a,PREP><el,DET+masc+sg>

del,<de,PREP><el,DET+masc+sg>

En el caso de las secuencias de verbos y clíticos, resultaría poco eficiente la inclusión de todas las secuencias posibles en el diccionario; por lo tanto, será necesario recurrir a una gramática productiva que tome como base el diccionario y la gramática morfológica adecuados para reconocer las dos unidades que componen la palabra gráfica que constituyen el verbo y el clítico.

[3] En el diccionario se agregó el rasgo "caso" a los clíticos 3 y 4 porque se trata de una marca positiva, pero esta etiqueta es, en cierto modo, redundante porque no permite diferenciar elementos del conjunto: todos los elementos del conjunto CL+3 son acusativos y todos los elementos del conjunto CL+4 son dativos. Si se pretendiera analizar casos de leísmo y loísmo, habría que suprimir esta marca y crear gramáticas sintácticas capaces de determinar el valor del rasgo.

[4] También puede darse el caso inverso: que una unidad léxica esté compuesta por más de una unidad formal, pero esto no es pertinente para el objeto que estamos analizando.

4.1. Diccionario y gramática morfológica

Para analizar las secuencias de verbos y enclíticos, el diccionario de entrada de la gramática productiva debe incluir las formas que presentan cambios acentuales o elisión de caracteres y etiquetarlas con sus rasgos morfológicos. NooJ cuenta con la etiqueta especial +NW (non word), que indica que una entrada no ocurre en los textos en forma aislada.

Con la incorporación de un solo clítico, los infinitivos no sufren ninguna alteración gráfica; en todas las demás formas verbales se producen cambios. Por lo tanto, se debe crear una gramática morfológica capaz de generar un diccionario que, además de las formas aisladas, contenga las formas presentes en la combinación con los clíticos, tal como se detalla a continuación.

amando,amar,V+FLX=AMAR+ger

amándo,amar,V+FLX=AMAR+ger+NW (*amándolo*)

ame,amar,V+FLX=AMAR+pte+subj+3era+sg

áme,amar,V+FLX=AMAR+pte+subj+3era+sg+NW (*ámelo*)

amen,amar,V+FLX=AMAR+pte+subj+3era+pl

ámen,amar,V+FLX=AMAR+pte+subj+3era+pl+NW (*ámenlo*)

amemos,amar,V+FLX=AMAR+pte+subj+1era+pl

amémos,amar,V+FLX=AMAR+pte+subj+1era+pl+NW (*amémoslo*)

amémo[5]*,amar,V+FLX=AMAR+pte+subj+1era+pl+1+NW* (*amémonos*)

amá,amar,V+FLX=AMAR+imp+2da+sg+RIOP

ama,amar,V+FLX=AMAR+imp+2da+sg+RIOP+NW (*amalo*)

ama,amar,V+FLX=AMAR+imp+2da+sg

áma,amar,V+FLX=AMAR+imp+2da+sg+NW (*ámalo*)

amad,amar,V+FLX=AMAR+imp+2da+pl

ama,amar,V+FLX=AMAR+imp+2da+sg+RIOP+NW (*amaos*)

4.2. Gramática productiva

La gramática propuesta tiene las características básicas de la que se presenta en la figura 4, pero debe ser perfeccionada con la finalidad de lograr que excluya las secuencias agramaticales (**amadse*) y que reconozca todas las gramaticales, incluso aquellas donde la grafía del verbo se modifica por la presencia del enclítico:

[5] Para la primera persona del plural del presente del subjuntivo, se requieren dos formas NW: *amémo*, que aparece con el clítico *nos* y *amémos*, que ocurre con *lo* y sus variantes y *le* y sus variantes. El rasgo *+1* se agrega a la primera al solo efecto de distinguirla de la otra.

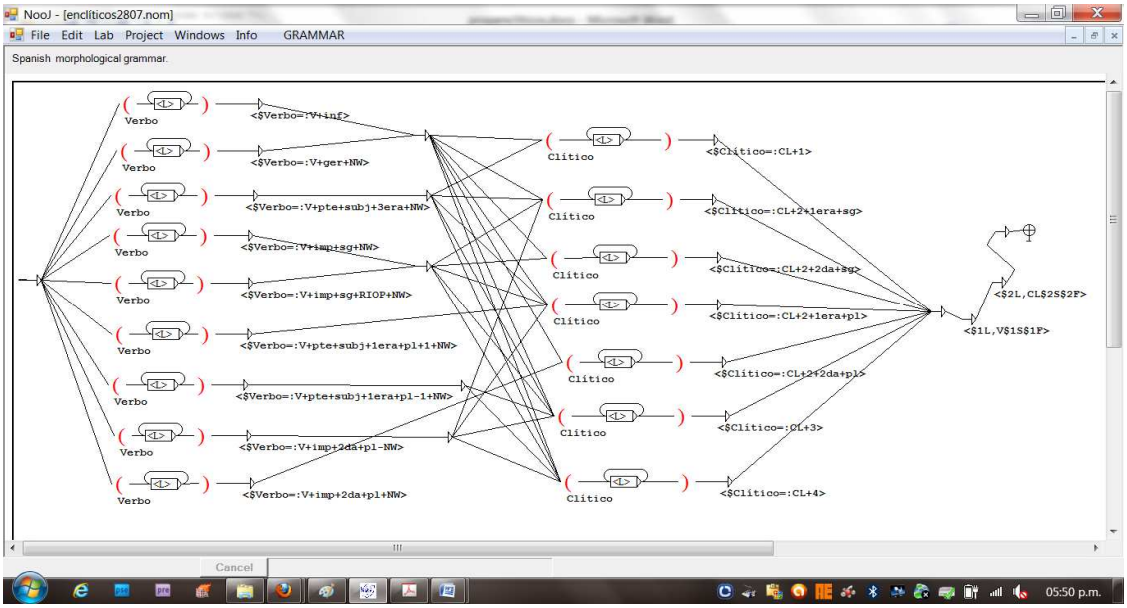

Figura 5: Gramática productiva para el análisis de secuencias de verbos y clíticos

La gramática que se muestra en la figura 5 expresa en el lenguaje formal de NooJ lo siguiente:

Tabla 1: Interpretación de la gramática anterior en lenguaje general

***Verbo* es cualquier secuencia de caracteres que coincide con una entrada del diccionario etiquetada como:**	**Y se puede combinar con *Clítico*, que es cualquier secuencias de caracteres que coincide con una entrada del diccionario etiquetada como:**	**Secuencias lingüísticas**
V+inf (*amar*)	CL+1, CL+2+1era+sg, CL+2+2da+sg, CL+2+1era+pl, CL+2+2da+pl, CL+3, CL+4	*amarse, amarme, amarte, amarnos, amaros, amarlo, amarla, amarlos, amarlas, amarle, amarles*
V+ger+NW (*amándo*)	CL+1, CL+2+1era+sg, CL+2+2da+sg, CL+2+1era+pl, CL+2+2da+pl, CL+3, CL+4	*amándose, amándome, amándote, amándonos, amándoos, amándolo, amándola, amándolos, amándolas, amándole, amándoles*
V+pte+subj+3era+NW (*áme, ámen*)	CL+1, CL+2+1era+sg, CL+2+1era+pl, CL+2+2da+pl, CL+3, CL+4	*ámese, ámeme, ámenos, ámelo, ámela, ámelos, ámelas, ámele, ámeles ámense, ámenme, ámennos, ámenlo, ámenla, ámenlos, ámenlas, ámenle, ámenles*
V+imp+sg+NW (*áma*)	CL+2+1era+sg, CL+2+2da+sg, CL+2+1era+pl, CL+3, CL+4	*ámame, ámate, ámanos, ámalo, ámala, ámalos, ámalas, ámale, ámales*
V+imp+sg+RIOP+NW (ama)	CL+1, CL+2+1era+sg, CL+2+2da+sg, CL+2+1era+pl, CL+3, CL+4	*amase, amame, amate, amanos, amalo, amala, amalos, amalas, amale, amales*
V+pte+subj+1era+pl+1+NW (*amémo*)	CL+2+1era+pl	*amémonos*
V+pte+subj+1era+pl-1+NW (*amémos*)	CL+3, CL+4	*amémoslo, amémosla, amémoslos, amémoslas, amémosle, amémoles*
V+imp+2da+pl-NW (*amad*)	CL+1, CL+2+1era+sg, CL+2+1era+pl, CL+3, CL+4	*amadme, amadnos, amadlo, amadla, amadlos, amadlas, amadle, amadles*
V+imp+2da+pl+NW (*ama*)	CL+2+2da+pl	*amaos*

Los rasgos no especificados incluyen a todos los valores que puede tener; así, dado que en cuanto a la combinatoria con los enclíticos, los verbos en tercera persona de presente del subjuntivo tienen las mismas propiedades tanto en singular como en plural, se utiliza la etiqueta V+pte+subj+3era+NW, que, al no tener especificado el número, incluye singular y plural.

En los clíticos se utilizan etiquetas que subsumen el mayor número de elementos posibles: los clíticos de tercera persona (CL+3 y CL+4) tienen las mismas propiedades combinatorias, independientemente de los otros rasgos diferenciales; por ello, se utilizan etiquetas que no los especifican. En cambio, en el caso de la primera y la segunda persona es necesario utilizar etiquetas específicas porque la primera persona del presente del subjuntivo plural rechaza los clíticos de segunda persona y, además, tiene una forma cuando se concatena con el clítico *nos* (***amémo****nos*) y otra cuando se concatena con clíticos de tercera persona (***amémos****lo*); en la segunda persona del imperativo no son posibles secuencias donde el verbo y el clíticos presentan diferente número (**ámaos,*amadte*).

Entre el verbo y el clítico se introduce un nodo vacío que permite agrupar elementos que tienen las mismas propiedades combinatorias; por ejemplo, el infinitivo y el gerundio se combinan con todos los clíticos, consecuentemente, se introduce un nodo vacío donde convergen ambas formas verbales y todos los clíticos.

5. FORMAS COMPUESTAS Y PERÍFRASIS VERBALES

Como se indicó más arriba, cuando el auxiliar de los tiempos compuestos está en infinitivo o en gerundio, el enclítico lo sigue inmediatamente (*haberlo amado, habiéndolo amado*). En la perífrasis modales y aspectuales de infinitivo, esta construcción alterna con la que presenta el enclítico en posición final (*poderlo amar / poder amarlo, pudiéndolo amar, pudiendo amarlo*).

Para analizar las secuencias donde el enclítico se pospone al auxiliar, la gramática sintáctica propuesta en [2] se debe modificar insertando entre el auxiliar y el verbo principal un nodo alternativo que contenga el clítico:

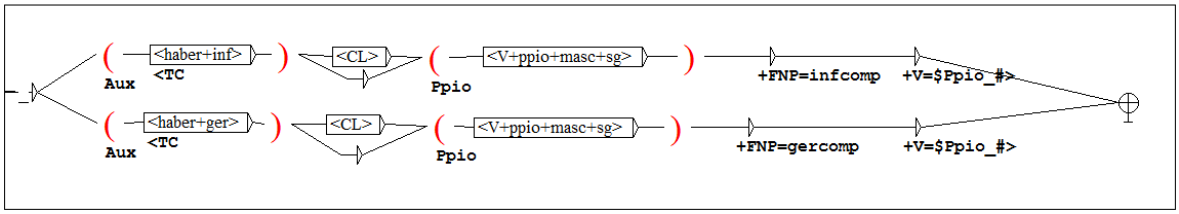

Figura 6: Clíticos insertos entre el auxiliar y el verbo principal

La gramática sintáctica que se muestra en la figura 6 se puede adaptar para el tratamiento de las perífrasis con auxiliar en gerundio e infinitivo. Para el procesamiento, NooJ aplica, en primer lugar, la gramática productiva que reconoce la unidad gráfica formada por el infinitivo y el clítico como dos unidades léxicas diferentes y, luego, las utiliza en las gramáticas sintácticas como si fuesen palabras incluidas en el diccionario. Como se puede observar, después del auxiliar se insertan un nodo <CL> (clítico) y un nodo <E> (nodo vacío, que cuando se conecta se ve representado por una flecha) y se establecen dos conexiones: una que pasa por el clítico y otra que pasa por el nodo vacío; mediante la línea superior NooJ reconoce la secuencia con clítico (*haberlo amado*) y mediante la línea inferior la secuencia sin él (*haber amado*).

6. CONCLUSIONES Y PROYECCIONES

En este trabajo se propone un procedimiento que permite salvar las dificultades que plantea para el tratamiento en NooJ el hecho de que los verbos y los enclíticos, a pesar de su unidad gráfica, sean dos palabras diferentes.

El análisis se limita a los verbos regulares de la primera conjugación; si se pretendiera ampliarlo a todo el sistema, sería necesario modificar las gramáticas con la finalidad de generar en los diccionarios todas las formas no independientes (NW) de cada modelo de flexión.

Para el tratamiento de secuencias con dos clíticos, se puede utilizar un procedimiento similar; pero será necesario determinar cuáles son las combinaciones de los clíticos y cómo inciden en la grafía de la forma verbal.

Si se pretende ir más allá del aspecto formal, será necesario crear diccionarios específicos, que contengan información que permita predecir la selección de los clíticos por parte de los verbos. Por ejemplo, si se tiene un cuenta solo el aspecto formal se obtendrán secuencias como *jactarlo*, que no es gramatical; un diccionario adecuado debe contener información que identifique el grupo de verbos al que pertenece *jactar* y, a partir de él se podrá elaborar una gramática específica para ese grupo, donde el infinitivo, el gerundio y las formas personales de tercera persona se combinan únicamente con CL+1 y las formas personales de primera y segunda, con clíticos que presenten los mismos rasgos de persona y número que el verbo. En tal sentido, resultaría de gran importancia retomar la propuesta de [6], donde se presenta una clasificación de los verbos tomando como criterio las posibilidades combinatorias de los verbos con los clíticos.

Referencias

[1] BONINO, Rodolfo (2011). "Una propuesta para la implantación de la morfología verbal del español en NooJ" en Revista Infosur N° 5. Octubre de 2011. En línea <http://www.infosurrevista.com.ar>. Última consulta 14 de mayo de 2013.

[2] BONINO, Rodolfo y Andrea Rodrigo (en prensa). "Análisis automático del sistema verbal en "La muchacha del atado" de Roberto Arlt". Ponencia presentada en el II Congreso Internacional de Profesores de Lenguas Oficiales del MERCOSUR.

[3] <http://www.nooj4nlp.net>. Última consulta 14 de mayo de 2013.

[4] Real Academia Española (2010). Nueva gramática de la lengua española (Manual). Espasa Libros, S.L., Buenos Aires.

[5] ALARCOS LLORACH, Emilio (1973). "Pronombres personales" en Gramática funcional del español. Gredos, S.A., Madrid.

[6] SOLANA, Zulema (2008) "Clíticos como clasificadores de verbos" en Revista Infosur N° 2. Agosto de 2008. En línea <http://www.infosurrevista.com.ar>. Última consulta 28 de julio de 2013.

Análisis Estadístico de Datos Aplicados al Estudio de Calidad en Servicios de Traducción

Statistical Data Analysis Applied to the Study of Quality in Translation Services

Analía Marta Pogliano
Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Argentina
analiampoglino@gmail.com

Abstract

This research studies the existing linguistic quality problems in translation of texts, from a leading company in translation services and its Client . The overall objective is to evaluate and estimate the values of the metrics required by the Client to ensure the expected quality.

Using logistic regression models to analyze data to find models that describe the audited linguistic errors in the forms of quality control for both the Client and the Company.

Also, using analysis of Canonical Correlations, the measurements between the two groups are studied, identifying the factors influencing the behavior of the errors.

The results indicate that the form of linguistic control between the Customer and the Company is not the same. The structures of correlation between the variables under study differ in each set of measurements , as well as the models adjusted for each group of measurements do not show the same significant variables.

Keywords: logistic regression, canonical correlation, translation services, language errors, linguistic quality control

Resumen

En esta investigación se estudian los problemas de calidad lingüística existentes en traducciones de textos, entre una Empresa líder en servicio de traducción y su Cliente. El objetivo general es evaluar y estimar los valores de los parámetros de medición exigidos por el Cliente para garantizar la calidad esperada.

Mediante los modelos de Regresión Logística se analizan los datos para encontrar los modelos que mejor describan a los errores lingüísticos auditados en los formularios de control de calidad, tanto para el Cliente como para la Empresa.

Asimismo, empleando el análisis de Correlaciones Canónicas se estudian las asociaciones existentes entre las mediciones de los dos grupos, identificando los factores influyentes en el comportamiento de los errores.

Los resultados obtenidos indican que la modalidad de control lingüístico entre el Cliente y la Empresa no es la misma. Las estructuras de correlación entre las variables bajo estudio difieren en cada grupo de mediciones, como así también, los modelos ajustados para cada grupo de mediciones no presentan las mismas variables significativas.

Palabras claves: regresión logística, correlaciones canónicas, servicios de traducción, errores lingüísticos, control de calidad lingüístico.

1. INTRODUCCION

La Empresa ha presentado serios problemas en cuanto a la calidad de ciertos pedidos entregados a un Cliente en particular. Al tratarse de traducciones de textos, el control de calidad está basado de acuerdo a las normas estándares internacionales de LISA (Localization Industry Standards Association). Debido a esto, la calidad de las entregas por parte de la Empresa deben alcanzar los valores establecidos como “aceptables”, según las normas internacionales.

Cliente emplea un proceso de control de calidad interno, con sus propios revisores y especialistas lingüísticos, quienes van a estar encargados de decidir si el pedido recibido presenta buena o mala calidad. De la misma manera, la Empresa cuenta con su departamento de control de calidad lingüística formado por revisores nativos de cada idioma localizado. Y ellos son los encargados de decidir si la traducción alcanza o no los niveles estándares de calidad.

La Empresa comenzó a recibir mal feedback del Cliente, haciendo referencia a la baja calidad en los pedidos entregados para ciertos idiomas, y que éstos no alcanzaban los valores estándares requeridos.

Esta noticia provocó gran incertidumbre y preocupación dentro de la Empresa, ya que todas las medidas de control lingüístico estaban siendo cumplidas, obteniéndose en la mayoría de los casos resultados positivos, garantizándole al Cliente una buena calidad.

Pero los resultados que la Empresa recibió en los reportes mensuales enviados por el Cliente no coincidían con los que ésta contaba. A pesar de los intentos de mejora por parte de la empresa, ésta no pudo alcanzar los parámetros de calidad exigido por el Cliente.

Por tal motivo se propuso investigar con sumo detalle, mediante la aplicación de métodos estadísticos el porqué de esta inconsistencia en los resultados. Una consideración importante a tener en cuenta para esta investigación es el cumplimiento del siguiente lema: “el Cliente siempre tiene la razón”, por más que los datos y resultados demuestren lo contrario. Debido a esto, el feedback recibido y los resultados obtenidos por sus revisores deben ser aceptados y considerados como “lo correcto, lo ideal”, es decir, en términos estadísticos, los datos del Cliente se deben tomar como grupo control.

2. MATERIALES Y METODOS

2.1. Materiales

La información recopilada para la creación de la base de datos proviene de formularios de control de calidad lingüísticos llamados LQA (“Lingusitic Quality Assurance”), pertenecientes a dos grupos: del Cliente y de la Empresa.

Estos formularios son completados por los revisores linguisticos, que, automáticamente, mediante fórmulas de cálculos que ponderan la cantidad de errores con sus correspondientes pesos de error, se determina el resultado final del control de calidad, esto es, si el resultado es aceptable (“Pass”) o si es inaceptable (“Fail”). La metodología de evaluación es prácticamente la misma entre el Cliente y la Empresa, pero las variables medidas, en algunos casos, son diferentes. Esto se debe a la continua actualización de los formularios con el fin de mejorar las mediciones y obtener resultados más confiables.

2.2. Métodos Estadísticos

2.2.1. Análisis de Correlaciones Canónicas

El análisis de correlaciones canónicas estudia las relaciones existentes entre dos grupos de variables. Investiga en detalle las interdependencias lineales entre dichos conjuntos, que pueden ser tratados simétricamente, o bien desempeñar un rol diferente en el análisis: un grupo de variables predictoras y otro de variables respuesta. Ambos conjuntos no deben ser independientes, se trata de descubrir "relaciones complejas" que reflejan la estructura existente entre ambos grupos de variables. El objetivo de esta técnica es resumir las asociaciones entre estos dos grupos de variables mediante la creación de nuevas variables a partir de las variables de cada grupo.

Sean las variables del primer grupo identificadas por *X* y las del segundo grupo identificadas por *Y*. Estas variables pueden pensarse distribuidas conjuntamente con esperanza nula (sin pérdida de generalidad), y matriz de covariancias, particionadas en cuatro submatrices (2.2.1).

$\underline{Z}' = (\underline{X}' \mid \underline{Y}') = (X_1 X_2 ... X_p \mid Y_1 Y_2 ... Y_q)$ distribuida conjuntamente con $E(\underline{Z}') = \underline{0}$ y $\underline{\Sigma} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$ (1)

Se desean construir dos combinaciones lineales U y V (2), determinando el conjunto de coeficientes de forma que la correlación entre U y V sea máxima. Por lo tanto deberá expresarse dicha correlación como función de los coeficientes.

Esto es:

$$\begin{aligned} U &= \alpha' X = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p \\ V &= \gamma' Y = \gamma_1 y_1 + \gamma_2 y_2 + ... + \gamma_q y_q \end{aligned} \qquad (2)$$

tales que:

$$\rho_{U,V} = \frac{E(U \cdot V)}{\sigma_U \sigma_V} = \frac{E(\alpha' X \cdot \gamma' Y)}{\sqrt{E(\alpha' X)^2} \cdot \sqrt{E(\gamma' Y)^2}} = \frac{\alpha' \Sigma_{XY}\, \gamma}{(\alpha' \Sigma_{XX}\, \alpha)(\gamma' \Sigma_{YY}\, \gamma)} \quad (3)$$

Suponiendo que ρ_1 es la correlación máxima entre U y V, es decir: $\rho_1 = \max_{\underline{\alpha} \neq \underline{0}, \underline{\gamma} \neq \underline{0}} [corr(a'x, \gamma' y)]$ (4)

Luego, la primera correlación canónica entre X e Y se define por ρ_1.

Además, $U_1 = \underline{\alpha_1}' \underline{X}$ y $V_1 = \underline{\gamma_1}' \underline{Y}$ en donde α_1 y γ_1 son los valores de α y γ que producen esta correlación máxima, se conocen como las primeras variables canónicas.

Sin pérdida de generalidad, se pueden elegir α_1 y γ_1 de modo que Var (U_1) = Var (V_1) = 1

Sean ahora $U_2 = \underline{\alpha_2}' \underline{X}$ y $V_2 = \underline{\gamma_2}' \underline{Y}$, en donde se eligen α_2 y γ_2 de modo que:

1. U_2 y V_2 no están correlacionadas con U_1 y V_1.
2. Var (U_2) = Var (V_2) = 1 y
3. la correlación entre $\underline{\alpha_2}' \underline{X}$ y $\underline{\gamma_2}' \underline{Y}$, denotada por ρ_2 es un máximo sobre todos los α_2 y γ_2.

Entonces ρ_2 es la segunda correlación canónica y $U_2 = \underline{\alpha_2}' \underline{X}$ y $V_2 = \underline{\gamma_2}' \underline{Y}$ reciben el nombre de segundas variables canónicas. La cantidad real de correlaciones canónicas posibles es igual al

mínimo de *q* y *p-q*. La cantidad de correlaciones canónicas diferentes de cero es igual al rango de la matriz Σ_{12}.

2.2.2. Modelos de Regresión Logística Múltiple

Los métodos de regresión son una componente integral de cualquier análisis de datos asociado con la descripción de la relación entre una variable respuesta y una o más variables explicativas, cuyo objetivo es encontrar el modelo que mejor ajuste los datos y que sea el más parsimonioso.

En modelos de regresión logística, a diferencia de los modelos de regresión lineal, la variable respuesta es binaria o dicotómica. Se desea conocer la relación entre:

- Una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial).
- Una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas.

Por sus características, los modelos de regresión logística permiten dos finalidades:

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente.
- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

Considerando un conjunto de p variables independientes las cuales serán denotadas con el vector $x' = (x_1, x_2, \ldots, x_p)$, la probabilidad condicional de que y tome el valor 1 (presencia de la característica estudiada), en presencia de las covariables X:

$$P(y=1/X)=\pi(x)=\frac{e^{g(x)}}{1+e^{g(x)}} \qquad (5)$$

Siendo $g(x)=\beta_0+\beta_1 x_1+\ldots+\beta_p x_p$ donde

- β_0 es la constante del modelo o término independiente
- p el número de covariables
- β_i los coeficientes de las covariables
- x_i las covariables que forman parte del modelo.

Si se divide la expresión (5) por su complemento, es decir, si se construye su odds se obtiene una expresión de más fácil manejo matemático: $\frac{P(y=1/X)}{1-P(y=1/X)}=\frac{\pi(x)}{1-\pi(x)}=e^{g(x)}$ (6)

Si ahora se realiza su transformación logarítmica con el logaritmo natural, se obtiene una ecuación lineal que es lógicamente de manejo matemático aún más fácil y de mayor comprensión:

$$\log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right)=\log(e^{g(x)})=g(x) \quad (7)$$

En la expresión (7) la primera igualdad es el llamado logit, es decir, el logaritmo natural de la odds de la variable dependiente (esto es, el logaritmo de la razón de proporciones de cometer un error al

traducir). El término a la derecha de la igualdad es la expresión lineal, idéntica a la del modelo general de regresión lineal: $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$ (8)

La importancia de esta transformación es que *g(x)* tiene muchas propiedades deseables de un modelo de regresión lineal. El logit, *g(x),* es lineal en sus parámetros, puede ser continuo, variando entre -∞ y +∞, según del rango de variación de *x*.

3. RESULTADOS

3.1. Análisis de Correlaciones Canónicas entre las mediciones del Cliente y de la Empresa

Es de interés estudiar si existen relaciones complejas entre los dos grupos de variables medidas por el Cliente y por la Empresa. Para ello se desea investigar las interdependencias lineales entre dichos conjuntos, que serán tratados como un grupo de variables predictoras (variables medidas por la Empresa) y otro de variables respuesta (variables medidas por el Cliente). Así, se podrá ver cómo estos dos conjuntos de mediciones se relacionan.

Las variables que se analizan para el estudio de correlaciones canónicas son: Cumplimiento, Significado, Puntuación, Terminología y Ediciones. Cada una de ellas mide la frecuencia de errores de cada tipo encontrados en los documentos traducidos. Los valores que pueden asumir estas variables son: 0, 1, 2,…etc.

Previo a este análisis, fue necesaria una transformación logarítmica en las variables para lograr que los datos sigan una distribución normal, supuesto requerido para esta técnica. Pero como las variables pueden tomar el valor 0, se recurre a la utilización de la transformación logarítmica: $\ln(x + 0.05)$

Una vez realizadas las transformaciones, se probó la normalidad de las variables, en forma individual, mediante el Test de Shapiro-Wilks y se concluyó que, trabajando con un nivel de significación del 5%, se cumple la normalidad de cada una de las variables transformadas, lo cual es una condición necesaria pero no suficiente de normalidad conjunta.

Se presenta el análisis de los resultados obtenidos mediante el procedimiento Proc Cancorr de SAS para el análisis de correlaciones canónicas:

Según la matriz de correlaciones entre las variables del Cliente, $\boldsymbol{R_{11}}$ (Tabla 1), se observa que existe asociación entre las variables. Es decir, los errores lingüísticos encontrados por el Cliente están correlacionados entre sí.

Tabla 1: Matriz de correlaciones entre las variables del Cliente $\boldsymbol{R_{11}}$

Correlaciones entre Variables Cliente						
	cumplimiento	**gramática**	**significado**	**puntuación**	**terminología**	**edición**
cumplimiento	1.0000	**0.3221**	**0.3434**	**0.2287**	**0.3304**	**0.2051**
gramática	0.3221	1.0000	**0.4395**	**0.3927**	**0.3005**	**0.2519**
significado	0.3434	0.4395	1.0000	**0.3046**	**0.3550**	**0.2185**
puntuación	0.2287	0.3927	0.3046	1.0000	**0.3151**	**0.2527**
terminología	0.3304	0.3005	0.3550	0.3151	1.0000	**0.1952**
edición	0.2051	0.2519	0.2185	0.2527	0.1952	1.0000

En cambio, analizando la matriz de correlaciones entre las variables de la Empresa, $\boldsymbol{R_{22}}$ (Tabla 2), se observa los errores lingüísticos encontrados por la Empresa se podrían pensar que son independientes entre ellos (no hay asociacion).

Tabla 2: Matriz de correlaciones entre las variables de la Empresa $\boldsymbol{R_{22}}$

	cumplimiento	**gramática**	**significado**	**puntuación**	**terminología**	**edición**
cumplimiento	1.0000	**0.1279**	**0.0214**	**0.0969**	**0.1147**	**0.0707**
gramática	0.1279	1.0000	**0.2318**	**0.1859**	**0.1526**	**0.0108**
significado	0.0214	0.2318	1.0000	**0.0764**	**0.1688**	**-0.0298**
puntuación	0.0969	0.1859	0.0764	1.0000	**0.1733**	**0.0391**
terminología	0.1147	0.1526	0.1688	0.1733	1.0000	**0.1775**
edición	0.0707	0.0108	-0.0298	0.0391	0.1775	1.0000

Si se analiza la matriz de correlaciones entre las variables del Cliente y de la Empresa, $\boldsymbol{R_{12}}$, (Tabla 3) se distingue que la variable Gramática de la Empresa está altamente correlacionada con las variables Significado, Puntuación y Terminología del Cliente. Es decir, los errores gramaticales encontrados por la Empresa están asociados con los errores de significado, de puntuación y de terminología encontrados por el Cliente.

Estos valores indican que las estructuras de las variables entre el Cliente y la Empresa son diferentes, a pesar que ambos grupos de variables estén midiendo lo mismo, sobre la misma muestra. Es por ello que los resultados obtenidos por uno u otro grupo son tan diferentes.

Tabla 3: Matriz de correlaciones entre las variables del Cliente $\boldsymbol{R_{12}}$

	Cumplimiento_E	**Gramática_E**	**Significado_E**	**Puntuación_E**	**Terminología_E**	**Edicion_E**
Cumplimiento_C	0.0760	0.2421	0.1462	0.0974	0.2511	0.0507
Gramática_C	0.1162	0.2395	0.1118	0.1743	0.1024	0.1164
Significado_C	0.1645	**0.2993**	0.1427	0.0730	0.0824	0.0766
Puntuación_C	0.1438	**0.3272**	0.1260	0.1369	0.0445	0.1812
Terminología_C	0.0995	**0.2971**	0.0770	0.1531	0.1174	0.1300
Ediciones_C	0.1241	0.1386	0.0337	0.1169	-0.0068	0.0924

En base al Test de significación de correlaciones canónicas, presentado en la tabla 4, mediante la estadística de Wilks, se observa que la primera correlación canónica es la única significativa, a un nivel de significación $\alpha = 0.05$. Todas las restantes correlaciones, a partir de la segunda, no fueron significativas, con un valor de p-asociado = 0.3692. Es por ello que se cuenta con un solo par de variables canónicas, cuya interpretación es relevante al estudio.

Tabla 4: Test para la significación de las correlaciones canónicas

	Razón de verosimilitud	**F**	**Gl numerador**	**Gl denominador**	**Pr > F**
1	**0.68367539**	**2.75**	**36**	**1096.2**	**<.0001**
2	0.89989808	1.07	25	930.21	0.3692
3	0.96634623	0.54	16	767.46	0.9261
4	0.98958609	0.29	9	613.45	0.9765
5	0.99669622	0.21	4	506	0.9332
6	0.99950544	0.13	1	254	0.7232

El primer y único par de variables canónicas significativas está dado por las siguientes combinaciones lineales:

$$U_1 = Cliente1 = 0.0866\ \text{Cumplimiento} + 0.0704\ \text{Gramática} + 0.1367\ \text{Significado} + 0.2514\ \text{Puntuación} + 0.1740\ \text{Terminología} + 0.0352\ \text{Ediciones}$$

$$V_1 = Empresa1 = 0.1692\ \text{Cumplimiento} + 0.3856\ \text{Gramática} + 0.0864\ \text{Significado} + 0.1294\ \text{Puntuación} + 0.0093\ \text{Terminología} + 0.2297\ \text{Ediciones}$$

Observando las cargas canónicas entre las variables del Cliente y sus variables canónicas Cliente1 en la tabla 5, éstas indican que del grupo de variables del Cliente, P*untuación*, T*erminología* y S*ignificado* son las más importantes dentro de la combinación lineal U_1.

Tabla 5: Correlaciones entre las variables del Cliente y Empresa y sus respectivas variables canónicas.

Cliente/Empresa	Cliente1 (U_1)	Empresa1 (V_1)
cumplimiento	0.5375	0.3800
gramática	0.6123	**0.8631**
significado	**0.6705**	0.3431
puntuación	**0.7950**	0.3775
terminología	**0.6841**	0.2786
ediciones	0.3934	0.3756

De igual manera, se presentan las cargas canónicas entre las variables de la Empresa y su variable canónica Empresa1 (V_1). En este caso, para el grupo de las variables de la Empresa, la única variable que aporta mucha importancia dentro de la combinación lineal es la variable *Gramática*.

Para ver la relación entre una variable de un conjunto, con la variable canónica del otro conjunto, se analizan las cargas cruzadas entre variables originales y variables canónicas (Tabla 6). Nuevamente se observa que, para el grupo de las variables del Cliente, las variables con más peso sobre el grupo de variables canónicas de la Empresa son P*untuación*, T*erminología* y S*ignificado;* y para el grupo de las variables de la Empresa, la variable con más importancia sobre el grupo de las variables canónicas del Cliente es *Gramática.*

Tabla 6: Cargas canónicas cruzadas para Cliente y para Empresa

Correlaciones entre las variables del Cliente y la variable canónica de la Empresa	
Cliente	**Empresa1**
Cumplimiento_C	0.2635
Gramática_C	0.3001
Significado_C	0.3287
Puntuación_C	0.3897
Terminología_C	0.3353
Edicion_C	0.1928

Correlaciones entre las variables de la Empresa y la variable canónica del Cliente	
Empresa	**Cliente1**
Cumplimiento_E	0.1863
Gramática_E	0.4231
Significado_E	0.1682
Puntuación_E	0.1850
Terminología_E	0.1366
Edicion_E	0.1841

Se determina que sí existe una relación entre las variables del Cliente y la Empresa; la misma está dada entre las variables Puntuación, Terminología y Significado, medidas por el Cliente, y la variable Gramática, medida por la Empresa. Éstas han demostrado ser variables que hacen un aporte mayor a las interdependencias lineales de cada grupo. Así, en los documentos en los que el Cliente distingue más errores de puntuación, terminología y significado, la Empresa encuentra errores gramaticales.

Otro aspecto que deja en evidencia este análisis es que las estructuras de las variables bajo estudio difieren en cada grupo de mediciones. Para el caso de las variables del Cliente, las mediciones de los errores lingüísticos presentan cierta asociación entre ellas. En cambio, para el caso de la Empresa, no existe asociación entre los distintos tipos de errores. Es por ello que los resultados de los controles de calidad lingüística pertenecientes a un mismo documento analizado por la Empresa y el Cliente a menudo difieren sustancialmente.

3.2. Modelo de Regresión Logística Múltiple

Mediante el análisis de regresión logística se desea encontrar el modelo que mejor ajuste los datos, cuantificando la importancia de la relación existente entre cada una de las covariables y la variable dependiente.

En base a los datos de la investigación, se cuenta con dos grupos de variables: las medidas por el Cliente y las medidas por la Empresa. Entonces es necesario encontrar dos modelos de regresión logística que mejor describan la variable respuesta dicotómica “Resultado”, para luego compararlos y ver cómo difieren entre ellos y qué tan distintas son las estimaciones de los coeficientes de los modelos.

Asimismo, es de interés analizar cuáles son las mediciones realizadas por la Empresa que tienen efecto sobre el resultado del Cliente. Es decir, modelar la probabilidad de rechazo del documento por parte del Cliente (“Fail”) en base a las mediciones de la Empresa. La variable respuesta “Resultado” toma valores 0 = “Pass” y 1 = “Fail”.

3.2.2. Ajuste del modelo para los Datos del Cliente

Utilizando el procedimiento Logistic de SAS, mediante el método de Selección de variables hacia atrás “Backward”, el modelo resultante presenta los efectos de las variables referidas a los errores de terminología, cumplimiento, gramática, significado, año, cantidad de palabras analizadas (conteo) y las interacciones de esta última con los errores de gramática, cumplimiento y significado. (Tabla 6)

Tabla 6: Efectos incluidos en el modelo final y sus significancias

Efectos	GL	Wald Chi-Cuadrado	Pr > ChiSq
cumplimiento	1	7.0214	0.0081
gramática	1	0.6989	0.4032
significado	1	3.6531	0.0560
terminología	1	5.4174	0.0199
conteo	1	0.8033	0.3701
año	1	4.8186	0.0282
cumpli*conteo	1	4.8533	0.0276
grama*conteo	1	7.7957	0.0052
sig*conteo	1	9.1734	0.0025

Además, el valor de la estadística Chi-Cuadrado en la Test de Hipótesis Global, mediante el Test de Razón de Verosimilitud, fue de 109.83, con 9 grados de libertad, la cual resultó ser significativo a un nivel α=1%. Esta prueba indica que las variables predictoras que se están usando son variables estadísticamente significativas del resultado de los LQAs analizados.

Las estimaciones de los parámetros del modelo obtenidas por el método de máxima verosimilitud nos permiten obtener la función logit g (x) estimada como:

$$\hat{g}(x) = -4.4352 + 0.2931\,cumplimien\,to + 0.1051\,gramática + 0.1563\,significad\,o + 0.2357\,terminilog\,ía + 0.8389\,conteo + 1.2169\,año + 0.9359\,cumplimien\,to * conteo + 2.9701\,gramática * conteo + 1.4724\,significad\,o * conteo$$

Los valores de dichas estimaciones se ven resumidos en la Tabla 7.

Tabla 7: Estimaciones de los coeficientes del modelo del Cliente

Parámetro		GL	Estimador	Error Estándar	Chi-Cuad	Pr > ChiSq
Intercepto		1	-4.4352	0.6477	46.8840	<.0001
cumplimiento		1	0.2931	0.1106	7.0214	0.0081
gramática		1	0.1051	0.1257	0.6989	0.4032
significado		1	0.1563	0.0818	3.6531	0.0560
terminología		1	0.2357	0.1013	5.4174	0.0199
conteo	**A**	1	0.8389	0.9360	0.8033	0.3701
año	**2008**	1	1.2169	0.5544	4.8186	0.0282
cumpli*conteo	**A**	1	0.9359	0.4248	4.8533	0.0276
grama*conteo	**A**	1	2.9701	1.0637	7.7957	0.0052
sig*conteo	**A**	1	1.4724	0.4861	9.1734	0.0025

La interpretación de los coeficientes del modelo de regresión logística se hace en términos de la razón de odds, cuya expresión es

$$\hat{\theta} = \frac{\left[\frac{\pi(x+1)}{1-\pi(x+1)}\right]}{\left[\frac{\pi(x)}{1-\pi(x)}\right]} \Rightarrow \ln\hat{\theta} = \hat{g}(x+1) - \hat{g}(x)$$

A continuación se presentan las razones de odds correspondientes a cada tipo de error (Tabla 8):

Tabla8: Razones de Odds para Cliente

Variables	Conteo <1000	Conteo > 1000
Cumplimiento	3.41	1.34
Gramática	21.65	1.11
Significado	5.10	1.17
Terminología	1.266	
Año	3.377	

Un aspecto relevante que se observa en este análisis es la gran influencia sobre los resultados finales que aporta la cantidad de palabras revisadas (variables cumplimiento, gramática y significado). Esto significa que la importancia de los errores de significado, gramática o cumplimiento se consideran más “severos” cuando se analizan menos palabras.

También se observa que el año marca una gran diferencia en el resultado. Depende el año de realización del QA, la chance aumenta 3 veces de un año a otro. Esto responde a uno de los interrogantes planteados previamente en esta investigación.

3.2.3. Ajuste del modelo para los Datos de la Empresa

Utilizando el procedimiento Logistic de SAS, mediante el método de Selección de variables hacia atrás "Backward", el modelo resultante presenta los efectos de las variables referidas a los errores de terminología, cumplimiento, gramática, significado, conteo, año y la interacción entre significado y la cantidad de palabras (conteo). (Tabla 9)

Tabla 9: Efectos incluidos en el modelo final y sus significancias

Efectos	GL	Chi-Cuadrado	Pr > ChiSq
cumplimiento	1	4.8451	0.0277
gramática	1	16.0914	<.0001
significado	1	1.5443	0.2140
terminología	1	16.7958	<.0001
conteo	1	17.0856	<.0001
año	1	4.9855	0.0256
sig*conteo	1	8.7700	0.0031

La medida global de Bondad de Ajuste del modelo, mediante el Test de Hipótesis Global indica un buen ajuste del modelo con las variables que fueron seleccionadas, obteniéndose un valor de la estadística Chi-Cuadrada de 99.15, con 7 grados de libertad, cuya probabilidad asociada es menor a 0.001, siendo significativa a un nivel de α del 1%.

Las estimaciones de los parámetros del modelo (ver tabla 10) obtenidas por el método de máxima verosimilitud nos permiten obtener la función logit g (x) estimada como:

$$\hat{g}(x) = -3.6931 + 0.5062 cumplimiento + 0.3000 gramática + 0.1541 significado + 0.6332 terminilogía + 2.1761 conteo + 1.0399 año + 1.1364 significado * conteo$$

Tabla 10: Estimaciones de los coeficientes del modelo de la Empresa

Parámetro		DF	Estimador	Error Estándar	Chi-Cuad	Pr > ChiSq
Intercepto		1	-3.6931	0.4968	55.2590	<.0001
cumplimiento		1	0.5062	0.2300	4.8451	0.0277
gramática		1	0.3000	0.0748	16.0914	<.0001
significado		1	0.1541	0.1240	1.5443	0.2140
terminología		1	0.6332	0.1545	16.7958	<.0001
conteo	**A**	1	2.1761	0.5264	17.0856	<.0001
año	**2008**	1	1.0399	0.4657	4.9855	0.0256
sig*conteo	**A**	1	1.1364	0.3837	8.7700	0.0031

De la misma manera que se analizara anteriormente para el modelo del Cliente, la interpretación de los coeficientes del modelo en regresión logística para las mediciones realizadas por la Empresa, se hace en términos de la razón de odds (Tabla11).

Tabla11: Razones de Odds para Empresa

Variables	Conteo <1000	Conteo > 1000
Cumplimiento	1.659	
Gramática	1.350	
Significado	3.634	1.17
Terminología	1.884	
Año	2.829	

A diferencia del modelo, si bien se observa el mismo tipo de efecto, el conteo no influye con la misma intensidad. Si se comparan las razones de Odds correspondientes a los errores de significado, para el Cliente se observa que la chance de Fail es 5 veces mayor al incrementarse en un error, mientras que para la Empresa es 4 veces mayor. Asimismo, en las mediciones del Cliente la cantidad de palabras interactúa también con los errores de gramática y cumplimiento.Con respecto al Año, se observa lo mismo para ambos grupos. La chance de obtener un rechazo aumenta 3 veces de un año a otro.

3.2.4. Análisis de la respuesta del Cliente en función de las mediciones de la Empresa

Es de interés analizar cuáles son las mediciones realizadas por la Empresa que tienen efecto sobre el resultado del Cliente. El objetivo es modelar la probabilidad de rechazo del documento por parte del Cliente (“Fail”) en base a las mediciones de la Empresa realizadas sobre los documentos.

Se lleva a cabo el mismo procedimiento que en los análisis precedentes obteniéndose un modelo final que incluye los efectos de los errores de gramática, terminología, la cantidad de palabras evaluadas (conteo) y la interacción de estos últimos dos. (Tabla 12).

Tabla 12: Efectos incluidos en el modelo final y sus significancias

Efectos	GL	Chi-Cuadrado	Pr > ChiSq
gramática	1	7.4872	0.0062
terminología	1	0.2827	0.5949
conteo	1	1.7142	0.1904
term*conteo	1	4.3206	0.0377

Además, el valor de la estadística Chi-Cuadrado en la Test de Hipótesis Global, mediante el Test de Razón de Verosimilitud, fue de 16.21, con 4 grados de libertad, la cual resultó ser significativa a un nivel α=1%. Esta prueba indica que las variables predictoras que se están usando son variables estadísticamente significativas del resultado de los LQAs analizados.

Las estimaciones de los parámetros del modelo (Tabla 13) obtenidas por el método de máxima verosimilitud nos permiten obtener la función logit g (x) estimada como:

$$\hat{g}(x) = -2.2167 + 0.1551 gramática - 0.0891 terminolog ía + 0.5851 conteo + 0.7605 terminolog ía * conteo$$

Tabla 13: Estimaciones de los coeficientes del modelo Cliente-Empresa

Parámetro	DF	Estimador	Error Estándar	Chi-Cuad	Pr > ChiSq
Intercepto	1	-2.2167	0.3375	43.1390	<.0001
conteo	1	0.5851	0.4469	1.7142	0.1904
gramática	1	0.1551	0.0567	7.4872	0.0062
terminología	1	-0.0891	0.1675	0.2827	0.5949
term*conteo	1	0.7605	0.3659	4.3206	0.0377

La interpretación de los coeficientes del modelo en regresión logística para el resultado del Cliente y las mediciones realizadas por la Empresa, se hace en términos de la razón de odds (Tabla 14).

Tabla 14: Razones de Odds Cliente/Empresa

Variable	Conteo <1000	Conteo >1000
Gramática	1.17	
Terminología	1.92	0.92

Este modelo pone en evidencia que la aceptación o rechazo de la traducción de un documento por parte del Cliente sólo se relaciona con los errores de gramática y terminología hallados por la Empresa. Por cada error de gramática que se encuentra en el texto evaluado, la chance de ser rechazado por el Cliente se incrementa un 17%. El efecto de un error de terminología es importante sólo cuando el número de palabras evaluadas es inferior a 1000.

4. CONCLUSIONES

El análisis de Correlaciones Canónicas determina que existe una relación entre las variables del Cliente y la Empresa; la misma está dada entre las variables Puntuación, Terminología y Significado, medidas por el Cliente, y la variable Gramática, medida por la Empresa.

Éstas han demostrado ser variables que hacen un aporte mayor a las interdependencias lineales de cada grupo. Así, en los documentos en los que el Cliente distingue más errores de puntuación, terminología y significado, la Empresa encuentra errores gramaticales. Otro aspecto que deja en evidencia este análisis es que las estructuras de correlación entre las variables bajo estudio difieren en cada grupo de mediciones. Es por ello que los resultados de los controles de calidad lingüística pertenecientes a un mismo documento analizado por la Empresa y el Cliente a menudo difieren sustancialmente.

Mediante el análisis de Regresión Logística, para los datos del Cliente, un aspecto relevante que se observa es la gran influencia sobre los resultados finales que aporta la cantidad de palabras revisadas.

A diferencia del modelo anterior, para los datos del Empresa, si bien se observa el mismo tipo de efecto, el conteo no influye con la misma intensidad.

Analizando cuáles son las mediciones realizadas por la Empresa que tienen efecto sobre el resultado del Cliente, el modelo obtenido pone en evidencia que la aceptación o rechazo de la traducción de un documento por parte del Cliente sólo se relaciona con los errores de gramática y terminología hallados por la Empresa.

Referencias

[1] Beltrán, Celina, “Modelización lingüística e información estadística”- 1ra ed.- Rosario-Juglaría, 2009.

[2] Dallas E. Johnson, “Métodos Multivariados aplicados al análisis de datos”. 1998.

[3] Eriksson L., Johansson E., Kettanen N. –Wold and S. Wold, “Multi and Megavariate Data Analysis, Principles and Applications”. 1991-2001.

[4] Grupo de Investigación Traducción, literatura y sociedad, “Ética y política de la traducción literaria”. Volumen de colección Itaca. Miguel Gómez Ediciones 2004.

[5] Hosmer, David W. y Lemeshow, Stanley. “Applied Logistic Regression”. Wisley Series in Probability and Mathematical Statistics.

[6] Khattree, Ravindra y Dayanand N. Naik, “Multivariate Data Reduction and Discrimination with SAS Software, Cary, NC: SAS Institute Inc. 2000.

[7] Lomprecht, James L, “Applied Data Analysis for Process Improvement, A practical Guide to Six Sigma Block Belt Statistics”. 2005.

[8] Mauly, Bryan F.J., “Multivariate Statistical Methods”. 1986-2004.

[9] Waddignton, Christopher, “Estudio Comparativo de Diferentes Métodos de Evaluación de traducción general”. 1999-, Madrid.

Comparación de métodos de clasificación aplicados a textos Científicos y No Científicos

Comparison among classification methods applied to Scientific and Non Scientific texts.

Ivana Barbona
Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina.
ivanabarbona@gmail.com

Abstract

The aim of this work is to compare the performance of the classification methods Support Vector Machine, Sequential Minimal Optimization, Logistic Regression, Lineal and Quadratic Discriminant Analysis by applying them to Scientific and Non Scientific Texts. This comparison among classification methods is carried out taking into account the measure of the misclassification error percentage calculated by (total of misclassified texts/total of texts)x100 . The software used for this analysis are Weka and JMP programs.

Keywords: Support Vector Machine - Sequential Minimal Optimization - Logistic Regression – Lineal Discriminant Analysis - Quadratic Discriminant Analysis – Learning Machine.

Resumen

En este trabajo se comparan el desempeño de los métodos de clasificación Support Vector Machine, Sequential Minimal Optimization, Regresión Logística, Análisis Discriminante Lineal y Cuadrático mediante su aplicación a textos Científicos y No Científicos. La comparación entre los métodos se realiza teniendo en cuenta la medida del porcentaje de error de mala clasificación calculado como (total de textos mal clasificados/total de textos)x100. El software utilizado para realizar este análisis son los programas Weka y JMP.

Palabras claves: Support Vector Machine, Sequential Minimal Optimization, Regresión Logística, Análisis Discriminante Lineal, Análisis Discriminante Cuadrático, Métodos de Clasificación, Aprendizaje de Máquina.

1. Introducción

La clasificación automática de textos ha tomado gran importancia en los últimos años debido al aumento de información disponible en formato electrónico. Su objetivo es categorizar documentos dentro de una cantidad fija de categorías predefinidas en función de su contenido.

En el presente trabajo se compara el desempeño de varios métodos de clasificación. Para esto se cuenta con una base de datos de textos clasificados en Científicos y No Científicos a los cuales se les midieron una determinada cantidad de características. La comparación entre los métodos se realiza teniendo en cuenta la medida del porcentaje de error de mala clasificación calculado como (total de textos mal clasificados/total de textos)x100.

Los métodos que se van a utilizar son: Support Vector Machine (SVM), Sequential Minimal Optimization (SMO), Regresión Logística, Análisis Discriminante Lineal (ADL) y Cuadrático (ADC).

2. Material y Métodos.

Los datos a utilizar provienen de un proyecto de investigación que se lleva a cabo en la Facultad de Ciencias Agrarias de la UNR denominado “Modelización Estadística en la Clasificación de Textos: Científicos y No Científicos”. En este proyecto se pretende evaluar el desempeño de las técnicas multivariadas para clasificar/agrupar textos según el género utilizando la frecuencia de aparición de distintas clases de palabras, entre otras características.

La base de datos corresponde a información de 150 textos. Está compuesta por una variable de clasificación GENERO que se refiere al tipo de texto, cuyas categorías son CIENTÍFICO y NO CIENTÍFICO, así como también 12 variables que representan características propias de estos textos.

GENERO	Género al que pertenece el texto
TEXTO	Identificador del texto dentro del corpus
adj	cantidad de adjetivos del texto
adv	cantidad de adverbios del texto
cl	cantidad de clíticos del texto
cop	cantidad de copulativos del texto
det	cantidad de determinantes del texto
nom	cantidad de nombres (sustantivos) del texto
prep	cantidad de preposiciones del texto
v	cantidad de verbos del texto
otro	cantidad de otras etiquetas del texto
total_pal	cantidad total de palabras del texto

Tabla 1: Variables en la base de datos.

Dado que el tamaño en cuanto a la cantidad de palabras es distinto para cada texto, se decide considerar la proporción de cada clase de palabras en lugar de la frecuencia. Además, mediante una selección de variables se descarta del análisis la variable "otro", lo cual permite eliminar la restricción de que la suma de los valores de las variables para cada texto es igual a uno. Restricción que proviene de tener en cuenta la proporción en lugar de la frecuencia, y que podrían causar problemas en la aplicación de algunos métodos.

Por lo tanto, la base definitiva a que se utiliza está compuesta por la variable de clasificación GENERO Y 9 variables que caracterizan a los textos. Éstas son la proporción de adjetivos, adverbios, clícticos, copulativos, determinantes, sustantivos, preposiciones y verbos.

Los Métodos de clasificación utilizados, que van a ser comparados son los siguientes:

2.1. Support Vector Machine:

Es un método de clasificación supervisada lineal que determina la frontera óptima entre dos grupos que pueden ser linealmente separables o no. Encuentra un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad que puede llegar a ser infinita, mediante la utilización de vectores soportes. Los vectores soportes son los datos que caen más próximos al hiperplano. Luego, mediante una trasformación inversa se obtiene la frontera no lineal que separa a esos dos grupos en el espacio original.

En el caso de clasificar en 2 categorías, se busca el hiperplano que tenga la máxima distancia o margen con los puntos más cercanos a él. Luego, los puntos pertenecientes a una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado.

Mediante la Función Kernel se puede proyectar la información a un espacio de características de mayor dimensionalidad, las funciones kernel que se utilizan en este trabajo son:

- Lineal
- Polinomio de segundo grado
- Radial Basis Function.

Se realizó una grilla en la cual se aplicó método de SVM variando la constante de penalización C y otros parámetros dentro de los distintos kernel considerados.

2.2. Sequential Minimal Optimization (SMO):

Es un algoritmo que resuelve un problema, que surge en SVM, de optimización de una función cuadrática de varias variables, pero sujetas a una restricción lineal de esas variables.

2.3. Regresión Logística:

Es un modelo estadístico que sirve para describir la relación entre una variable respuesta categórica y un conjunto de variables explicativas, mediante el uso de la función de enlace logit.

Sea un conjunto de p variables independientes denotadas por el vector $x' = (x_1, x_2, \dots, x_p)$. Y sea la probabilidad condicional de que la variable respuesta Y sea igual a 1 $P(Y = 1/x) = \pi(x)$. El modelo de regresión logística está dado por la ecuación:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

donde,

$$\pi(x) = \frac{e^{g(x)}}{1 - e^{g(x)}}$$

2.4. Análisis Discriminante:

El Análisis Discriminante es una técnica Multivariada exploratoria utilizada para describir si existen diferencias entre k grupos de unidades o poblaciones (individuos, objetos, etc.) respecto a un conjunto de p variables medidas sobre estas unidades. Mediante éste análisis se obtiene una regla de clasificación basada en una función discriminante que puede ser utilizada con el fin de asignar futuras unidades a una de las k poblaciones según sus valores observados.

Existen varios métodos para obtener la función discriminante. En el caso del Discriminador Lineal se supone la estructura de covariancias de las p variables es la misma para todas las poblaciones. En cambio, para el Discriminador Cuadrático, se supone normalidad multivariada pero estructuras de covariancias distintas para las distintas poblaciones.

Sea ω un individuo que puede provenir de k poblaciones $\pi_1, \pi_2, \dots, \pi_k$, con k ≥ 3. Se quiere encontrar una regla de clasificación para asignar a ω a una de las k poblaciones basándose en el vector observado $x = (x_1, x_2, \dots, x_p)'$ de p variables para este individuo en particular. Sea μ_i el vector de medias y Σ_i la matriz de variancias y covariancias de las p variables en la i-ésima población.

2.4.1. Función Lineal Discriminante.

Se supone matriz de variancias y covariancias Σ común para las k poblaciones.

Sean $M^2(x,\mu_i) = (x-\mu_i)'\Sigma^{-1}(x-\mu_i)$, con i=1,…,k la distancia de Mahalanobis de ω a las poblaciones. Entonces se puede pensar en un criterio de clasificación que asigne a ω a la población más próxima de la siguiente forma:

Si $M^2(x,\mu_i) = \min\{M^2(x,\mu_1), M^2(x,\mu_2), \dots, M^2(x,\mu_k)\}$ se asigna ω a π_i

Utilizando las funciones lineales discriminantes se obtiene la expresión

$$L_{ij}(x) = (\mu_i - \mu_j)'\Sigma^{-1}x - \frac{1}{2}(\mu_i - \mu_j)'\Sigma^{-1}(\mu_i + \mu_j)$$

y trabajando algebraicamente se puede probar que:

Si $L_{ij}(x) > 0$ para todo $j \neq i$, entonces se asigna ω a π_i

Como las funciones $L_{ij}(x)$ verifican:

$$L_{ij}(x) = \frac{1}{2}\left[M^2(x,\mu_j) - M^2(x,\mu_i)\right]$$

$$L_{ij}(x) = -L_{ji}(x)$$

$$L_{rs}(x) = L_{is}(x) - L_{ir}(x)$$

Entonces sólo se necesitan k-1 funciones discriminantes.

2.4.2. Función Cuadrática Discriminante.

Se puede deducir en base a la regla de máxima verosimilitud de la siguiente manera:

Sea la función de densidad de x $f_i(x)$ en la i-ésima población π_i.

Si $f_i(x) = \max\{f_1(x), f_2(x), \ldots, f_k(x)\}$ entonces se asigna ω a π_i

Este criterio se relaciona con las funciones discriminantes

$$V_{ij}(x) = \ln f_i(x) - \ln f_j(x)$$

Si se cumple el supuesto de normalidad multivariante y las matrices de covariancias $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ difieren para las distintas poblaciones entonces se obtiene el siguiente discriminador cuadrático:

$$Q_{ij}(x) = \frac{1}{2}x'\left(\Sigma_j^{-1} - \Sigma_i^{-1}\right)x + x'\left(\Sigma_i^{-1}\mu_i - \Sigma_j^{-1}\mu_j\right) + \frac{1}{2}\mu_j'\Sigma_j^{-1}\mu_j - \frac{1}{2}\mu_i'\Sigma_i^{-1}\mu_i + \frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}\ln|\Sigma_i|$$

Todos los métodos se implementaron con el programa WEKA, excepto los Análisis DL y DC que fueron aplicados mediante el programa JMP.

Para compararlos se considera la medida del error de mala clasificación calculado como total de textos mal clasificados/total de textos.

3. Resultados

En la tabla 2 se observa una grilla con los resultados del porcentaje de error de mala clasificación para los métodos SVM y SMO variando el kernel y valores de parámetros.

Con respecto al método Support Vector Machine, se consideran distintos valores del parámetro C para los kernel lineal, polinomio de grado 2 y radial basis function. Lo mismo para el parámetro γ del kernel RBF.

		SVM									SMO
		Kernel: Lineal	Kernel: Polinomio grado 2	Kernel: Radial Basis Function.							
				γ = 0.0	γ= 0.1	γ= 0.2	γ= 0.3	γ= 0.5	γ= 1	γ= 5	γ= 10	
C	**0.1**	***19.33***	21.33	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00
	0.2	***19.33***	22.67	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	34.00
	0.3	20.67	24.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	30.00
	0.4	20.67	24.67	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	28.00
	0.5	20.67	24.67	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	25.33
	0.6	20.00	25.33	39.33	39.33	40.00	40.00	40.00	40.00	40.00	40.00	25.33
	0.7	20.00	26.00	39.33	38.67	40.00	40.00	40.00	40.00	40.00	40.00	23.33
	0.8	20.00	26.00	38.00	38.00	40.00	40.00	40.00	40.00	40.00	40.00	21.33
	0.9	20.00	27.33	38.67	36.67	39.33	40.00	40.00	40.00	40.00	40.00	20.67
	1	20.00	26.00	36.00	31.33	40.00	40.67	40.00	40.00	40.00	40.00	18.67
	1.5	20.00	26.67	32.67	31.33	40.67	41.33	40.67	40.00	40.00	40.00	18.67
	2	20.00	26.67	33.33	32.00	40.67	41.33	40.67	40.00	40.00	40.00	19.33
	2.5	20.00	30.00	33.33	32.67	40.67	41.33	40.67	40.00	40.00	40.00	18.67
	3	20.00	29.33	33.33	32.67	40.67	41.33	40.67	40.00	40.00	40.00	18.67
	3.5	20.00	29.33	34.00	33.33	40.67	41.33	40.67	40.00	40.00	40.00	18.67
	5	20.00	28.00	34.67	34.67	40.67	41.33	40.67	40.00	40.00	40.00	***18.00***
	10	20.67	26.67	34.00	34.00	40.67	41.33	40.67	40.00	40.00	40.00	***18.00***
	20	20.67	27.33	34.00	34.00	40.67	41.33	40.67	40.00	40.00	40.00	18.67

Tabla 2: Valores del Error de Mala Clasificación al aplicar SVM y SMO para distintos valores de parámetros.

Método		**Porcentaje de Mala Clasificación**
Regresión Logística		20.67
Análisis Discriminante	**Lineal**	***18***
	Cuadrático	***16.67***

Tabla 3: Resultados de error de mala clasificación al aplicar Análisis Discriminante y Regresión Logística.

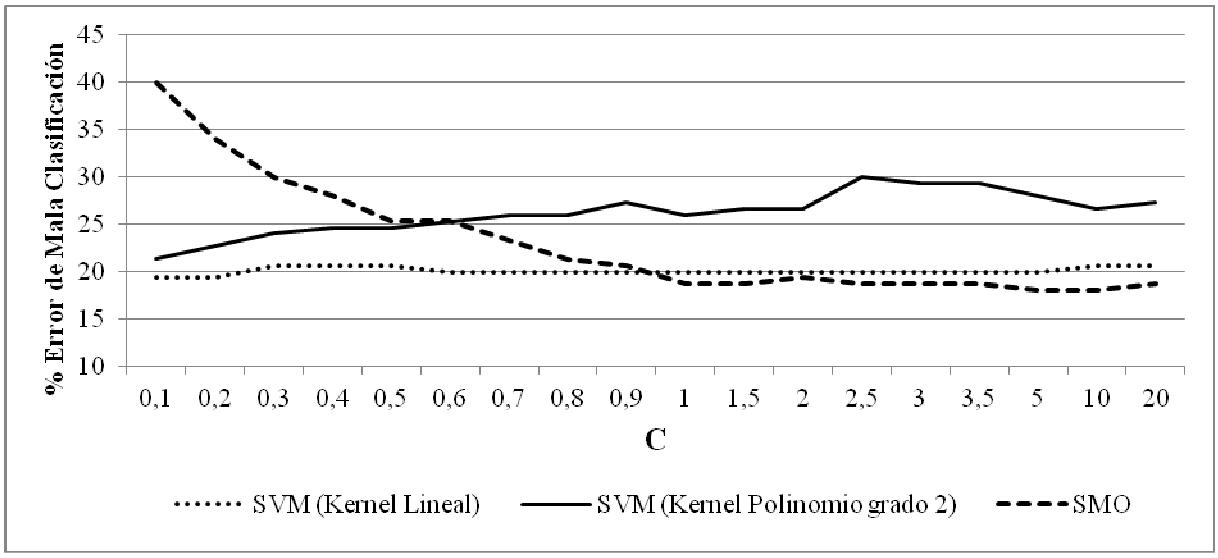

Figura 1: Comparación entre SVM (Kernel Lineal), SVM (Kernel Polinomio grado 2) y SMO.

Se observa que los métodos con porcentajes de mala clasificación menores son el Análisis Discriminante Lineal (18%) y Cuadrático (16.67%), SMO con C=5 y 10 (18% para ambos casos) y SVM con kernel lineal y C= 0.1 y 0.2 (19.33%).

El método SMO presentó porcentajes de mala clasificación bajos, del aproximadamente el orden del 18%. No obstante, se visualiza cierta variabilidad en cuanto a estos porcentajes para diferentes valores del parámetro C (Figura 1), con valores que van del 18% al 40%. Esto estaría indicando cierta inestabilidad del método para clasificar, representando una desventaja. En cambio el método SVM con kernel lineal, obtuvo valores un poco más altos de porcentajes de mala clasificación que SMO, pero al variar el parámetro C, sus resultados fueron más estables, con valores que fueron del 19.33% al 20.67%.

En cuanto al tiempo de ejecución, SVM y SMO no presentaron diferencias. Pero al comprar los métodos observando las demás medidas de error que proporciona WEKA en sus resultados, SVM con Kernel lineal y C= 0.1 o 0.2 resultan mejores.

La tabla 3 muestra los resultados de los errores de mala clasificación para los métodos de Regresión Logística y Análisis Discriminante. Los resultados observados son de esperarse en cuanto al cumplimiento de los supuestos a los cuales están sujetos los ADL y ADC. Dado que los datos con los que se trabajó presentan estructuras de covariancias distintas para los grupos, resulta lógico que el ADC (16.67% de mala clasificación) funcione mejor que el ADL (18% de mala clasificación).

Por otro lado, cabe destacar que no es posible saber cómo afectan las estructuras de covariancias distintas entre los grupos para los métodos SMO y SVM, lo cual representa una ventaja del AD frente métodos de aprendizaje de máquina como el SVM.

4. Conclusiones

De todos los métodos de clasificación considerados, el que presentó el menor porcentaje de mala clasificación fue el ADC (16.67%).

Tanto el ADL como el ADC dieron buenos resultados al clasificar los textos en Científicos y No Científicos, presentando un 18% y 16.67% de mala clasificación respectivamente.

En cuanto a los métodos de aprendizaje de máquina, el que presenta mejores resultados es el SVM con kernel lineal y constante de penalización C= 0.1 o 0.2 (19.33%).

Del resto de los métodos aplicados, el que presenta peores resultados es SVM con kernel RBF, arrojando valores de porcentaje de error de mala clasificación que van del 34% al 40%.

Si bien el método SMO presentó porcentajes bajos de mala clasificación para valores altos de C (18%), no es considerado uno de los mejores debido a la variabilidad que presenta en sus resultados al considerar distintos valores de la constante C, dando indicios de cierta inestabilidad del método para clasificar bien.

Referencias

Beltrán, C. 2010. Análisis discriminante aplicado a textos académicos: Biometría y Filosofía. Revista Infosur. N° 4.

Cherkassky, V., Mulier, F. 2007. Learning From Data. Concepts, Theory, and Methods. John Wiley & Sons.

Cuadras, C. 2012. Nuevos Métodos de Análisis Multivariante. CMC Editions.

Hastie, T., Tibshirani, R., Friedman, J. 2009. The Elements of Statistical Learning. Data Mining, Inference, and Predicion. Springer Series in Statistics.

Hosmer, D., Lemeshow, S., Sturdivant, R. 2013. Applied Logistic Regression. John Wiley & Sons.

Witten, I., Frank, E. 2005. Data Mining. Practical Machine Learning Tools and Techniques. Elsevier.