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:

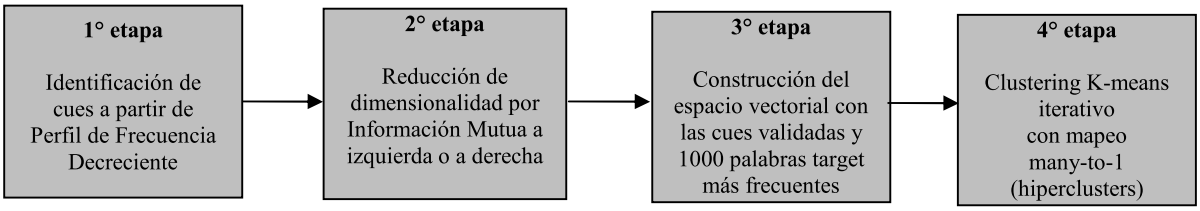

Figura 1: Esquema del algoritmo de categorización de palabras propuesto

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