Inducción de constituyentes sintácticos en español con técnicas de *clustering* y filtrado por información mutua

Fernando Balbachan
Facultad de Filosofía y Letras
Universidad de Buenos Aires
fernando_balbachan@yahoo.com.ar

Diego Dell'Era
Facultad de Filosofía y Letras
Universidad de Buenos Aires
diego.dellera@gmail.com

Resumen

El Argumento de la Pobreza de los Estímulos (*Argument from the Poverty of Stimulus*, *APS*) se presenta como el gran campo de debate epistemológico entre el paradigma simbólico y el paradigma estadístico en lingüística computacional (Pullum y Scholz 2002). Desde 2000 en adelante aparecieron algunos trabajos dentro del paradigma estadístico que se propusieron atacar el Argumento de la Pobreza de los Estímulos a partir de la postulación de algún algoritmo general no supervisado de adquisición integral del lenguaje. Entre los aportes más importantes, la tesis de doctorado de Clark (2001) recurre a diversas técnicas estadísticas para dar con un algoritmo general no supervisado de inducción del lenguaje, y en particular, de una gramática independiente de contexto para el inglés.

Clark (2001) trabaja con distintas técnicas de inducción para cada fenómeno lingüístico modelizado: morfología mediante modelos markovianos, categorización (*POS-tagging*) mediante *clustering*, etc. Puntualmente, en este trabajo estamos interesados en la inducción de constituyentes sintácticos, dado un *corpus* etiquetado por clase de palabras (*POS-tagged*), como paso previo al procedimiento de inducción de una gramática independiente de contexto. En su propia tesis, el autor reconoce que es necesaria una mayor evidencia translingüística que apoye la plausibilidad psicolingüística de un enfoque como el suyo. Actualmente, no existen trabajos que se hayan propuesto probar el enfoque de Clark (2001) para la inducción de sintaxis en lenguas flexivas y con orden libre de constituyentes, como el español. Así pues, nuestro trabajo se propone contribuir con dicha evidencia translingüística, estudiando la factibilidad de aplicación del algoritmo de inducción de constituyentes de Clark (2001) para el español.

El algoritmo de Clark (2001) que nos ocupa consiste en aplicar técnicas de *clustering K-means* para agrupar secuencias de etiquetas de clase de palabra, según su información distribucional. Luego, se procede a filtrar los resultados para encontrar *clusters* que efectivamente se correspondan con grupos de constituyentes, recurriendo a un criterio de información mutua entre los símbolos inmediatamente anteriores y posteriores a dichas secuencias. Este criterio de filtrado evita el sesgo de un *corpus* escaso, al tiempo que logra distinguir la dependencia buscada entre los límites de las secuencias candidatas a constituyentes por sobre el umbral de la entropía natural de símbolos que co-ocurren a una cierta distancia en el lenguaje (Li 1990).

Nuestra implementación del algoritmo ha sido evaluada en un *corpus* de dimensiones prototípicas, con resultados prometedores. Se obtuvo una cobertura de 74%, una precisión de 58% y una medida F de 65%, en la etapa prototípica. Estos resultados alientan la continuidad del trabajo de investigación a largo plazo, con la meta de lograr un robusto algoritmo de adquisición integral del lenguaje para el español.

1. Introducción

1.1 El debate epistemológico acerca de la adquisición del lenguaje

En el campo transdisciplinario de la lingüística computacional, los enfoques estadísticos, que surgieron principalmente a mediados de la década del '90, vinieron a desplazar al paradigma simbólico, dominante hasta entonces (Moreno Sandoval 1998; Manning y Schütze 1999). El paradigma simbólico, que evolucionó bajo la égida chomskyana de las gramáticas generativas desde la década del '50 (Chomsky 1957, 1965; Moreno Sandoval 1998), se propone manipular categorías sintácticas dadas *a priori* (gramática formal) para deducir o derivar el conjunto de oraciones que constituye una lengua, a partir de la aplicación de reglas, parámetros o principios. El paradigma estadístico, en cambio, echa mano de diversas técnicas probabilísticas, aplicadas a grandes *corpora* de entrenamiento, con vistas a inducir categorías y fenómenos específicos del lenguaje natural a partir de la

detección de patrones estadísticamente significativos en la *tabula rasa* que constituyen los *corpora*. Sin embargo, el paradigma estadístico es más que una mera aplicación de técnicas y modelización matemática: estos enfoques aportan evidencia de plausibilidad psicolingüística a un renovado debate acerca de la naturaleza misma del lenguaje. En efecto, entre el paradigma simbólico y el paradigma estadístico se ha entablado un manifiesto contrapunto de concepciones epistemológicas opuestas en torno al atávico problema de la adquisición del lenguaje, a partir del encolumnamiento de las obras fundacionales del campo detrás de teorías innatistas o teorías empiristas, respectivamente (Piatelli-Palmarini 1980, Cowie 1999, Pullum y Scholz 2002).

> "Probabilistic methods are providing new explanatory approaches to fundamental cognitive science questions of how humans structure, process and acquire language […] Probabilistic models can account for the learning and processing of language, while maintaining the sophistication of symbolic models." [Chater y Manning 2008:335]

Aunque algunos entusiastas de la polémica aseguran que el debate acerca de la adquisición del lenguaje bien podría remontarse al siglo XVII con las posturas filosóficas de Descartes y de Locke (Clark 2001), más recientemente podemos empezar a rastrear esta confrontación en obras primordiales de la lingüística teórica (Chomsky 1957, 1965, 1986) y la psicolingüística (Fodor 1983; Pinker 1994) de la segunda mitad del siglo XX, las cuales defienden un innatismo a ultranza; mientras que las posturas empiristas son esgrimidas por la lingüística cognitiva prototípica de Lakoff y Langacker (Lakoff 1987; Langacker 2000) y filósofos del lenguaje como Quine y otros. Con respecto al problema de la adquisición del lenguaje en el campo transdisciplinario de la lingüística computacional, mientras el paradigma simbólico adscribe a sistemas deductivos que hipotetizan como condición necesaria un estado inicial de conocimiento innato y ricamente estructurado frente a la pobreza de los datos lingüísticos primarios de que dispondrían los niños, los enfoques estadísticos postulan, más bien, sistemas inductivos a partir del aprendizaje de patrones de ocurrencia de eventos en un *corpus* masivo no estructurado, mediante algún algoritmo de aprendizaje de propósitos generales –es decir, no específico de dominio (Clark 2001).

Justamente, el *Argumento de la Pobreza de los Estímulos* (*Argument from the Poverty of Stimulus* o *APS*) se presenta como el gran campo de debate epistemológico entre el paradigma simbólico y el paradigma estadístico. Así pues, el APS se empezó a perfilar como el más robusto adalid de la hipótesis innatista, aunque como bien señalan Pullum y Scholz (2002), ninguna teoría que avale tácita o taxativamente dicha hipótesis deja en claro las propiedades y la estructura de ese conocimiento innato de que dispondríamos durante el proceso de adquisición del lenguaje:

> "The one thing that is clear about the argument from the poverty of the stimulus is what its conclusion is supposed to be: it is supposed to show that human infants are equipped with innate mental mechanisms specifically for assisting in the language acquisition process – in short that the facts about human language acquisition support 'nativist' rather than 'empiricist' epistemological views. What is not clear at all is the structure of the reasoning that is supposed to support this conclusion. Instead of clarifying the reasoning, each successive writer on this topic shakes together an idiosyncratic cocktail of claims about children's learning of language and claims that nativism is thereby supported." [Pullum y Scholz 2002:12]

Por supuesto, desde la otra orilla, las teorías empiristas no deben ser confundidas con un trasnochado conductismo y su concepción del lenguaje como una mera asociación de esquemas estímulo-respuesta. Los empiristas no refutan la existencia de algún mecanismo inicial como condición necesaria para adquirir el lenguaje; simplemente postulan que este mecanismo se trataría de un aspecto más de la inteligencia humana (Piatelli-Palmarini 1980, Clark 2009), un algoritmo de aprendizaje de propósitos generales y no de una habilidad que presupone *a priori* conocimiento de dominio específico (*cf.* concepto de *gramática universal* en Chomsky 1957, 1965, 1986 y concepto de *facultad vertical* en Fodor 1983). Más aún, algunos empiristas no reniegan completamente del procesamiento encapsulado de dominio

específico (Fodor 1983), pero rechazan la idea de que la adquisición del lenguaje sea un proceso llevado a cabo *íntegramente* por este tipo de capacidades cognitivas:

> "There may well be domain-general parts of cognition that are applied to the task of language-acquisition even though the core of it is domain-specific. This sort of research could fruitfully focus the attention of researchers on particular aspects of language where the domain-specificity is more essential; moreover, I think it is clear that at some points in the language acquisition process, even nativists must propose some sort of statistical learning, albeit just for low-level tasks such as word segmentation." [Clark 2001:20]

1.2 Técnicas estadísticas para inducción de sintaxis

Así pues, la confrontación entre el paradigma simbólico y el paradigma estadístico se desató en varios frentes. Por un lado, la supuesta imposibilidad de aprendizaje del lenguaje ante la falta empírica de evidencia negativa (Gold 1967), argumento refutado en Clark (2001). Por otro lado, la renuencia de Chomsky y sus seguidores a dar crédito a las nociones estadísticas de la época como herramienta de análisis:

> "It seems to have been demonstrated beyond all reasonable doubt that, quite apart from any question of feasibility, methods of the sort that have been studied in taxonomic linguistics are intrinsically incapable of yielding the systems of grammatical knowledge that must be attributed to the speaker of a language." [Chomsky 1965:54]

> "Dixon speaks freely throughout about the 'probability of a sentence' as though this were an empirically meaningful notion. [...] We might take 'probability' to be an estimate of relative frequency […]. This has the advantages of clarity and objectivity, and the compensating disadvantage that almost no 'normal' sentence can be shown empirically to have a probability distinct from zero. That is, as the size of a real corpus (e.g. the set of sentences in the New York Public Library, or the Congressional Record, or a person's total experience, etc.) grows, the relative frequency of any given sentence diminishes, presumably without limit." [Chomsky 1966:34-35]

Sin embargo, a partir de la denominada revolución bayesiana en lingüística computacional (Manning y Schütze 1999; Clark 2001), las técnicas estadísticas, otrora ineficaces para lidiar con la aceptabilidad de oraciones que requerían los *corpora* reales, se renuevan incorporando la noción de probabilidad en términos de *grado subjetivo* de incertidumbre (*subjective degree of uncertainty*) para ser utilizadas en procesos masivos de inducción que modelan efectivamente distintas áreas del procesamiento del lenguaje natural, demoliendo así el otro bastión de la polémica contra el paradigma estadístico, con lo que queda en pie un último refugio del reinado del innatismo: el Argumento de la Pobreza de los Estímulos. Parafraseando a Klein y Manning (2004), los estímulos 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]

Desde 2000 en adelante aparecieron algunos trabajos dentro del paradigma estadístico que se propusieron atacar el Argumento de la Pobreza de los Estímulos –y consecuentemente, la hipótesis innatista– a partir de la postulación de algún algoritmo general no supervisado de adquisición integral del lenguaje. Pese a que se proponen confrontar con el APS, estos trabajos, enmarcados en el paradigma estadístico, abordan el problema desde la misma perspectiva inicial que Chomsky: la sintaxis como punto de partida para la adquisición del lenguaje y el isomorfismo entre lenguajes formales y naturales (Chomsky 1957).

Entre los trabajos que concitan mayor interés, la tesis de doctorado de Clark (2001) recurre a diversas técnicas estadísticas para dar con un algoritmo general no supervisado de inducción

de sintaxis y, en particular, de una gramática independiente de contexto para el inglés como un modelo formal para la adquisición del lenguaje (Pinker 1979):

> "This question is in one sense thoroughly Chomskyan: I fully accept his characterization of linguistics as, ultimately, a branch of psychology, though for the moment it relies on very different sorts of evidence; I fully accept his argument for complete formality in linguistics, a formality that computer modeling both requires and enforces; I fully accept the idea that one of the central problems of linguistics is how to explain the fact that children manage to learn language in the circumstances that they do. On the other hand, there are many areas in which this work is not so congenial to followers of the Chomskyan paradigm. First, the work here is fully empirical; it is concerned with authentic language, rather than artificial examples. Secondly, it eschews the use of unnecessary hidden entities; far from considering this as the hallmark of a good scientific theory, the unnecessary proliferation of unobservable variables renders the link between theory and surface tenuous and unstable." [Clark 2001:3]

Clark (2001) hace uso de distintas técnicas de inducción para cada fenómeno lingüístico modelizado: morfología mediante modelos markovianos, categorización (*POS-tagging*) mediante *clustering* distribucional, etc. Puntualmente, en este artículo estamos interesados en la inducción de constituyentes sintácticos, dado un *corpus* etiquetado por clase de palabras (*POS-tagged*), como paso previo al procedimiento de inducción de una gramática probabilística independiente de contexto (*Stochastic Context-Free Grammar* o *SCFG* o *Probabilistic Context-Free Grammar* o *PCFG*), que es el fin último de su tesis. Clark mismo reconoce que es necesaria una mayor evidencia translingüística que apoye la plausibilidad psicolingüística de su investigación:

> "There are a number of possible avenues for future research. The most important, in my opinion, is to experiment with other languages, particularly languages with very free word order. There is some evidence that these techniques will work with Chinese (Redington et al. 1995), which has quite fixed word order, but no work has been done in highly inflected languages with free word order." [Clark 2001:148]

Actualmente no existen trabajos que se hayan propuesto probar dicho enfoque para la inducción de sintaxis en español, una lengua flexiva y con orden libre de constituyentes. Así pues, nuestro trabajo se propone contribuir con dicha evidencia translingüística, estudiando, en principio, la factibilidad de aplicación del algoritmo de inducción de constituyentes de Clark (2001) para el español.

1.3 Trabajos previos

La tarea de inducción de constituyentes sintácticos, como primer paso para la inducción de gramáticas, ha venido atrayendo la atención temprana de investigadores:

> "Early work on grammar induction emphasized heuristic structure search, where the primary induction is done by incrementally adding new productions to an initially empty grammar (Olivier 1968; Wolff 1988). In the early 1990s, attempts were made to do grammar induction by parameter search, where the broad structure of the grammar is fixed in advance and only parameters are induced (Lari and Young, 1990; Carroll and Charniak 1992). However, this appeared unpromising and most recent work has returned to using structure search." [Klein y Manning 2002:128]

Más recientemente, con el advenimiento del paradigma estadístico, la tarea adquirió un nuevo vigor científico. Ya sea para rebatir empíricamente el argumento de la pobreza de los estímulos (Clark 2001), para construir bancos masivos de árboles sintácticos –*treebanks*– (van Zaanen 2000) o como parte de modelos de lenguaje (Chen 1995), la inducción no supervisada de estructura sintáctica básica bajo la forma de constituyentes ha probado ser uno de los campos más fértiles de investigación básica en lingüística computacional, a partir de la diversidad de enfoques de los algoritmos de trabajos previos. Clark (2001) releva en detalle los distintos enfoques con que se ha encarado la tarea y agrupa los trabajos más recientes en 3

diferentes aunque en alguna medida superpuestas categorías:

1) Enfoques basados en la probabilidad (Carroll y Charniak 1992 y otros trabajos): Se trata de experimentos basados en el algoritmo *Expectation-Maximization* o *EM* –también conocido como algoritmo *inside-outside* o *IO* (Manning y Schütze 1999)– o en algoritmos genéticos, que han obtenido resultados no muy exitosos:

> "[…] [This approach] produced some rather discouraging research that seemed to indicate that the fact that the IO algorithm converged to a local optimum meant that it would almost always converge to a linguistically implausible grammar." [Clark 2001:124]

2) Enfoques basados en la compresión (Wolff 1988 y otros trabajos): Básicamente recurren a una heurística de compresión de gramáticas inicialmente extensas, a partir del algoritmo *Minimum Description Length* o *MDL* (Manning y Schütze 1999). Si bien estos experimentos obtuvieron cierto éxito en lenguajes artificiales, Clark (2001) observa que este tipo de enfoques fallan al no dar cuenta de las dependencias de larga distancia que caracterizan a los constituyentes sintácticos del lenguaje natural.

3) Enfoques basados en la información distribucional: En este tipo de enfoques se encuadran los trabajos de Finch et al. (1995), Clark (2001) y nuestro propio experimento. La idea subyacente es que las secuencias de palabras o etiquetas morfosintácticas que componen un constituyente sintáctico –símbolo no terminal como, por ejemplo, Sintagma Nominal SN o Sintagma Preposicional SP– aparecerán en similares contextos distribucionales –a izquierda y a derecha– a lo largo de *corpora* masivos.

   Entre los enfoques basados en la información distribucional, encontramos uno de los primeros trabajos de inducción de gramáticas para el español: el algoritmo de inducción de gramática del español de Juárez Gambino y Calvo (2007). Basándose en la noción de sustituibilidad de Harris (2000) para hallar regularidades estructurales, estos investigadores desarrollaron un algoritmo no supervisado para entrenar al sistema de inducción de gramática ABL (*Alignment-Based Learning*) (van Zaanen 2000) con un *corpus* (CAST-3LB) de español de características similares a las de nuestro *corpus* (véanse *Tabla 1* y *Tabla 2*), reportando una medida F de 32,45% en la tarea de inducción de constituyentes sintácticos. No obstante, cabe aclarar que la evaluación del experimento de Juárez Gambino y Calvo (2007) apuntó al *parsing* de oraciones con los constituyentes inducidos, meta más ambiciosa que la evaluación manual de un listado de constituyentes inducidos, como en nuestro caso.

1.4 *Corpora* masivos y *corpora* para implementaciones prototípicas

Los experimentos de aprendizaje de máquina (*machine learning*) nos obligan a reflexionar sobre un aspecto metodológico con profundas incumbencias en el estudio del desarrollo ontogenético del lenguaje. Efectivamente, las técnicas probabilísticas, aplicadas a *corpora* masivos, inducen los fenómenos lingüísticos a partir de la identificación de patrones estadísticamente significativos. De este modo, se busca analogar los datos lingüísticos primarios (*Primary Linguistic Data* o *PLD*) de que dispondría un niño durante el proceso de adquisición del lenguaje a los *corpora* de millones de palabras que son procesados iterativamente por las computadoras.

La preocupación concerniente a la plausibilidad de modelización de los PLD es uno de los requerimientos que detalla Pinker (1979) para una teoría formal que se proponga explicar la adquisición del lenguaje:

> "It is instructive to spell out these conditions one by one and examine the progress that has been made in meeting them. First, since all normal children learn the language of their community, a viable theory will have to posit mechanisms powerful enough to acquire a

> natural language. This criterion is doubly stringent: though the rules of language are beyond doubt highly intricate and abstract, children uniformly succeed at learning them nonetheless, unlike chess, calculus and other complex cognitive skills. Let us say that a theory that can account for the fact that languages can be learned in the first place has met the *Learnability Condition*. Second, the theory should not account for the child's success by positing mechanisms narrowly adapted to the acquisition of a particular language. For example, a theory positing an innate grammar for English would fail to meet this criterion, which can be called the *Equipotentiality Condition*. Third, the mechanisms of a viable theory must allow the child to learn his language within the time span normally taken by children, which is in the order of three years for the basic components of language skill. Fourth, the mechanisms must not require as input types of information or amounts of information that are unavailable to the child. Let us call these the *Time and Input Conditions*, respectively. Fifth, the theory should make predictions about the intermediate stages of acquisition that agree with empirical findings in the study of child language. Sixth, the mechanisms described by the theory should not be wildly inconsistent with what is known about the cognitive faculties of the child, such as the perceptual discriminations he can make, his conceptual abilities, his memory, attention, and so forth. These can be called the *Developmental and Cognitive Conditions*, respectively." [Pinker 1979:219]

Pullum (1996) describe bastante bien esta plausibilidad de modelización entre los PLD y los *corpora* masivos del procesamiento computacional:

> "Ideally, what we need to settle the question is a large machine-readable corpus – some tens of millions of words – containing a transcription of most of the utterances used in the presence of some specific infant (less desirably, a number of infants) over a period of years, including particularly the period from about one year (i.e. several months earlier than the age at which two words utterances start to appear in children's speech) to about 4 years." [Pullum 1996:505]

Sin embargo, como Clark (2001) mismo reconoce, resulta difícil emular por completo los datos lingüísticos primarios, toda vez que los recursos de *corpora* de que dispone la comunidad científica están mayormente basados en lenguaje escrito y manifiestan notables diferencias en el registro y grado de formalidad y complejidad de los enunciados en comparación con los que presumiblemente serían los enunciados a los que se ve expuesto un niño entre el año y los 4 años de vida. Es menester mencionar que existen ciertos *corpora* que ofrecen lenguaje de registro especializado, como el *corpus CHILDES* (2,5 millones de palabras organizadas en interacciones orales madre-niño en inglés norteamericano) o el *corpus Wall Street Journal* o *WSJ* (registro periodístico). Aun así, como el mismo Chomsky (1959) concede, se debe tomar en cuenta que los niños en edad de adquirir el lenguaje no sólo se ven expuestos a los enunciados dirigidos específicamente hacia ellos, sino que los medios audiovisuales de comunicación o incluso las conversaciones entre adultos bien podrían funcionar como otros proveedores de datos lingüísticos primarios.

Para su tesis y en particular, para el experimento de inducción de constituyentes sintácticos, Clark (2001) recurre al *British National Corpus* o *BNC* en su primera edición del año 1994, un *corpus* sincrónico de inglés británico que contiene 100 millones de palabras de registro variado (periódicos, obras literarias, etc.), etiquetadas automáticamente según el estándar *C5* (CLAWS5) –un conjunto de 76 etiquetas morfosintácticas al que Clark agrega un símbolo para indicar el fin de oración. Aunque el BNC abarca registros orales en un 10% de la muestra, Clark recorta el *input* del BNC a 12.000.000 de palabras del registro escrito.

Puesto que nuestro objetivo es el estudio de factibilidad del experimento de Clark (2001) acerca de la inducción de constituyentes sintácticos para el español, nuestro trabajo se propuso adaptar su metodología a una implementación prototípica que probara la viabilidad de este enfoque para una lengua flexiva y con constituyentes sintácticos de orden libre.

Entre los recursos gratuitos de lingüística computacional del español, debemos mencionar el *corpus* CRATER, un *corpus* masivo multilingüístico de alineamiento de oraciones entre el inglés, el francés y el español, anotado morfosintácticamente. No obstante, una primera evaluación de la utilidad de este *corpus* para nuestro experimento resultó poco prometedora, ya que CRATER emplea alrededor de 500 etiquetas morfosintácticas y su interfaz de consulta resulta completamente obsoleta.

Entre los recursos de acceso gratuito sólo con fines académicos, analizamos los 2 corpus morfosintácticamente anotados más conocidos del español: CAST-3LB (Civit 2003) y el *Spanish Treebank* (Moreno Sandoval *et al.* 1999).

	CAST-3LB	**Spanish Treebank**
Tamaño en palabras	≈100.000	≈45.000
Tamaño en oraciones	≈3.500	≈1.600
Extensión promedio de oraciones	≈30 palabras	≈28 palabras
Anotación morfosintáctica	≈350 etiquetas	≈200 etiquetas
Criterio de anotación	Anotación semi-manual sintáctica, semántica y pragmática. En el nivel sintáctico, se sigue anotación por constituyentes, con marcaje adicional de funciones sintácticas (Civit 2003)	Automático: *chunking* y *POS-tagging* Validación manual por muestreo aleatorio (Moreno Sandoval *et al.* 1999)

Tabla 1: Comparación entre *corpora* CAST-3LB y *Spanish Treebank*

Sin embargo, la proliferación de etiquetas morfosintácticas en un corpus de reducidas dimensiones podría presentar un problema de dispersión de datos, escollo que debíamos evitar para adaptar el algoritmo de Clark (2001), que trabaja sobre un corpus masivo y con un listado reducido de etiquetas.

Dada la escasez de *corpora* morfosintácticamente anotados para nuestro idioma, nos vimos obligados a encarar la esforzada tarea de generar un *corpus* propio en español, etiquetado según los lineamientos morfosintácticos adaptados del BNC (Leech *et al.* 1994), que alcanzara dimensiones suficientes para trabajar a escala con un prototipo de la implementación del algoritmo adaptado y optimizado para el idioma español.

Debemos agradecer la colaboración de un equipo de entusiastas estudiantes de Lingüística de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires UBA, gracias al cual logramos organizar un *corpus* de aproximadamente 50.000 palabras de registro periodístico escrito, etiquetado morfosintácticamente mediante un riguroso criterio metodológico (véanse *Anexos I*, *II* y *III*). El *corpus* resultante, el instructivo describiendo la metodología utilizada y la implementación prototípica del algoritmo del experimento se encuentran disponibles para la comunidad científica, bajo los alcances de una licencia *Creative Commons* en el sitio web[1].

2. Algoritmo de inducción de constituyentes sintácticos en Clark (2001)

El algoritmo de Clark (2001) que nos ocupa consiste en aplicar técnicas de *clustering K-means* para agrupar secuencias de etiquetas de clase de palabra, según su información distribucional. Luego, se procede a filtrar los resultados para encontrar *clusters* que efectivamente se correspondan con grupos de constituyentes, recurriendo a un criterio de información mutua entre los símbolos inmediatamente anteriores y posteriores a dichas secuencias.

Este criterio de filtrado evita el sesgo de un *corpus* escaso, al tiempo que logra distinguir la dependencia buscada entre los límites de las secuencias candidatas a constituyentes por sobre el umbral de la entropía natural de símbolos que co-ocurren a una cierta distancia en el lenguaje (Li 1990).

[1] http://campus.filo.uba.ar/course/view.php?id=587

	Clark (2001)	**Balbachan-Dell'Era (2010)**
Tamaño en palabras	≈12.000.000	49.925
Tamaño en oraciones	≈700.000	2.108
Extensión promedio de oraciones	16,6 palabras	23,8 palabras
Anotación morfosintáctica	Manual: 77 etiquetas según estándar C5	Manual: 48 etiquetas adaptadas del estándar C5
Criterio de anotación	BNC (Leech *et al.* 1994)	propio, adaptado del BNC

Tabla 2: Comparación entre *corpora* de *input* para ambos experimentos

2.1 Acerca de la naturaleza de un constituyente

La noción de constituyente, en sentido amplio, se aplica a conjuntos de palabras (o etiquetas) que funcionan como unidades sintácticas en la oración. En el sentido estricto en que Clark la usa, la noción de constituyente está restringida a una secuencia continua de etiquetas que reescribe a un nodo no terminal en una derivación sintáctica. El requerimiento de continuidad se debe a que los constituyentes encontrados se usan en etapas subsiguientes del experimento para inducir gramáticas libres de contexto, y esa clase de gramáticas no admite estructuras discontinuas.

La definición de constituyente para el algoritmo admite, por lo tanto, la imbricación de constituyentes en otros constituyentes de mayor extensión, pero excluye explícitamente la oración entera (esto es, se elimina la secuencia delimitada por dos símbolos de fin de oración). Se trata de una simplificación operativa: aunque en sentido amplio la oración se pueda considerar como un constituyente, en el experimento de Clark se intenta hallar los constituyentes más básicos para construir reglas gramaticales. Nótese que en los casos liminares donde un constituyente es en sí mismo una oración, se lo sigue considerando como un constituyente; a la inversa, una oración breve que coincide con un constituyente no se convierte por ello en un constituyente.

Si bien es razonable pensar que las secuencias que forman un constituyente han de ocurrir frecuentemente en un texto, cabe aclarar que la mera frecuencia no garantiza que una secuencia sea una de las estructuras que este experimento se propone encontrar. Por ejemplo, la secuencia AT1 NN1 PRP (artículo singular-sustantivo singular-preposición) es mucho más frecuente que AT1 AV0 AJ1 NN1 (artículo singular-adverbio-adjetivo singular-sustantivo singular) y sin embargo, la secuencia AT1 NN1 PRP no es un constituyente, así como tampoco lo es ninguna de las altamente frecuentes secuencias terminadas en PRP. El caso extremo es la secuencia formada por una única PRP (preposición), que es la etiqueta más frecuente en la mayoría de los textos, pero no un constituyente. A la inversa, un constituyente extenso no deja de serlo por ser muy infrecuente. Es por ello que el corte por frecuencia es sólo el primer umbral del algoritmo de Clark.

Además, la longitud de la secuencia tampoco define el carácter de constituyente: NN1 (sustantivo singular) tiene la misma extensión que PRP y sin embargo es un constituyente.

Aunque existen limitaciones prácticas y teóricas, idealmente un experimento de este tipo ha de encontrar constituyentes continuos de cualquier extensión, particularmente cuando están compuestos de varios niveles imbricados de constituyentes breves, como en la *Figura 1*.

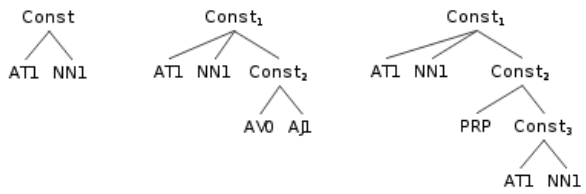


Figura 1: Constituyentes de a) 1 nivel, b) 2 niveles y c) 3 niveles de imbricación

2.2 Paso 1: perfil de frecuencias decrecientes de secuencias candidatas a constituyentes

Para el primer paso de su algoritmo, Clark lista las secuencias de etiquetas con una frecuencia mayor que el número de parámetros de la distribución que modela sus contextos. Utiliza 77 etiquetas, lo cual define una distribución de 77^2 parámetros ≈ 5500, de modo que el piso de frecuencia para las secuencias seleccionadas es de 5000 ocurrencias en su *corpus*.

Si hubiéramos usado este criterio al adaptar el experimento de Clark al español, habríamos debido calcular el siguiente *umbral de ocurrencias u*:

$$u = \left({tags^2}/{size} \right)^{ext} = \left({48^2}/{240} \right)^{1,5} = 31 \text{ ocurrencias (i)}$$

donde distribución de símbolos en contexto *tags* = (48 etiquetas: $48^2 \approx 2300$)

tamaño de *corpus size* = 240 veces menor

extensión de oraciones *ext* = 1,5 veces mayor (promedio de longitud de oraciones)

Aunque el cálculo del umbral arrojaba el valor de 31 ocurrencias, decidimos experimentar con diversos escenarios de corte, entre 10 y 110 ocurrencias. De ese modo, podemos afinar a voluntad la base con la que el resto del algoritmo ha de trabajar, a la vez que nos mantenemos en el orden de valores sugeridos por la adaptación del *corpus* de Clark al nuestro. Con todo, consideramos que estos lineamientos en cuanto al umbral de ocurrencias son comparativamente significativos: Clark obtiene 753 secuencias candidatas, y en nuestro caso obtenemos 198 para el escenario más efectivo de 110 ocurrencias (véase *Tabla 4*).

Una idea importante que opera en el experimento de Clark reside en observar que varias secuencias que forman una misma clase de constituyentes (sintagma nominal, sintagma preposicional, etc.) aparecen en contextos similares, de modo que estudiar los contextos puede brindar información distribucional útil para tratar de detectar constituyentes automáticamente. La idea, en esencia, es la misma que subyace a las pruebas de sustitución para determinar si una secuencia de etiquetas conforma un constituyente o no.

Denominaremos *contexto previo* a la etiqueta que precede a la secuencia y *contexto posterior* a la etiqueta que le sigue. Esta información se puede modelar como dos distribuciones que indican cuántas veces aparece cada posible constituyente (compuesto de una secuencia de etiquetas) en cada contexto (compuesto por los pares posibles combinados de etiquetas anteriores y posteriores). Dado que en nuestro experimento hay 48 etiquetas, cada distribución tiene 48^2 tales pares.

2.3 Paso 2: *Clustering* de secuencias candidatas a constituyentes

Una vez obtenida la anterior tabla de información distribucional, el siguiente paso en el algoritmo de Clark consiste en *arracimar* las secuencias de etiquetas en *clusters* o grupos afines. Para ello, considera que la información de la tabla representa la posición de cada secuencia en un espacio vectorial multidimensional, y que la afinidad entre secuencias se puede medir como la distancia que las separa.

> "If two sequences of tags occur mostly forming the same non-terminal, then we would expect the context that those strings occur in to be similar […] If we clustered sequences according to their distributions we would thus expect to find clusters corresponding to various syntactic constituents" [Clark 2001:132]

Clark sugiere como algoritmo de *clustering* el método iterativo *k-means* y usa la distancia euclidiana entre vectores. Como resultado de esta etapa, espera obtener varios grupos de secuencias que contengan constituyentes válidos, por un lado, y el resto de las secuencias, por el otro. Clark sostiene que es posible determinar automáticamente en cuáles de los *clusters* agrupados por información distribucional hay constituyentes válidos y en cuáles no. Este paso se entiende en el experimento de Clark como instancia previa a la inducción de una SCFG, que es el fin último de su investigación en los procesos de inducción de sintaxis, de modo tal que cada cluster válido resulte el germen para una categoría sintagmática mayor (sintagma nominal, sintagma preposicional, etc.).

	1	2	3	4	…	71	…	73	…	2203	2204	…	2303	2304
	AJ0-AJ0	**AJ0-AJ1**	**AJ0-AJ2**	**AJ0-AJC**	…	**AJ1-NN2**	…	**AJ1-NNP**	…	**VVZ-VVG**	**VVZ-VVI**	…	**$$$-XX0**	**$$$-$$$**
AT1 NN1 PRP AT1 NN1 PRP	0.0	0.0	0.0	0.0	…	1.0	…	0.0	…	1.0	1.0	…	0.0	0.0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
NN1 PRP NNP	0.0	1.0	0.0	0.0	…	0.0	…	4.0	…	0.0	0.0	…	0.0	1.0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Tabla 3: Tabla de información distribucional (secuencias y contextos)

No obstante, en nuestro caso, sólo estamos interesados en el sub-proceso de inducción de constituyentes (véase sección *Modificaciones al experimento original*).

2.4 Paso 3: Criterio de filtrado por información mutua entre etiquetas adyacentes a las secuencias candidatas a constituyentes

Una vez concluido el paso 2, Clark obtiene 100 *clusters*. En nuestro caso, como se ve en el *Anexo IV*, obtuvimos 25 *clusters*. Sin embargo, como Clark observa, esto no significa que todos los *clusters* agrupen constituyentes sintácticos:

> "As expected, the results of the clustering showed clear clusters corresponding to syntactic constituents […] of course, since we are clustering all of the frequent sequences in the corpus, we will also have clusters corresponding to parts of constituents […] we obviously would not want to hypothesize these as constituents: we therefore need some criterion for filtering out these spurious candidates." [Clark 2001:133]

Para determinar cuáles son los grupos de secuencias válidas como constituyentes, Clark propone un filtro basado en el grado de dependencia entre la etiqueta previa y la etiqueta posterior del contexto. La medición de la dependencia entre contextos consiste en estimar su información mutua (*mutual information* o *MI*):

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right) \quad \text{(ii)}$$

Mutual information (información mutua)

donde *I* es la información mutua; *X* e *Y* son las distribuciones de contextos previos y posteriores, respectivamente; *p(x,y)* es la probabilidad conjunta de dos etiquetas dadas de dichos contextos; $p_1(x)$ es la probabilidad marginal de ese contexto previo; y $p_2(y)$ es la probabilidad marginal de ese contexto posterior.

Nótese que en esta fórmula no se mide la interdependencia entre las etiquetas que pertenecen a la secuencia, ni la dependencia entre la secuencia y sus contextos. Esta fórmula de información mutua mide cuán dependientes entre sí son los contextos previo y posterior: una MI de 0 refleja total independencia, mientras que valores altos de MI reflejan cuánto disminuye nuestra perplejidad (Manning y Schütze 1999) cuando, conociendo una etiqueta, encontramos la otra.

Sin embargo, aunque la secuencia propiamente dicha no esté presente en la fórmula, influye en el cálculo. Dado que hay una cierta MI 'natural' entre dos símbolos cercanos cualesquiera de un lenguaje (Li 1990), y que esa MI disminuye a medida que la distancia entre los símbolos crece, la longitud de una secuencia determina la distancia entre sus contextos, de modo que se ha de tomar en cuenta en el cálculo de MI. En la *Figura 2* se puede observar la rápida caída en la curva de MI para distancias crecientes, donde una distancia de 2 símbolos corresponde a una secuencia de 1 etiqueta, una distancia de 3 símbolos, a una de 2 etiquetas, y así sucesivamente:

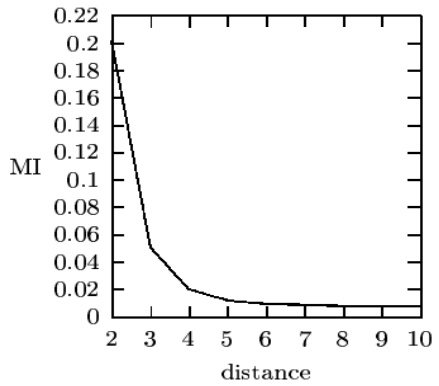


Figura 2: La información mutua entre el contexto previo y el contexto posterior desciende conforme crece la distancia que los separa (medida en símbolos) (Li 1990)

Por esta razón, Clark (2001) toma en cuenta como parámetro la distancia entre las etiquetas del contexto y postula que si en un *cluster* el promedio de la MI de los contextos –ponderado por la longitud de las secuencias– supera el umbral determinado por Li (1990), entonces dicho *cluster* es válido y probablemente agrupe secuencias que son constituyentes.

Cabe aclarar que las conclusiones de Li (1989 y 1990) competen a secuencias de caracteres. Clark (2001) extiende sus conclusiones a secuencias de etiquetas, considerando que una etiqueta funciona como un símbolo; de hecho, Li (1989) mismo ya había contemplado en su artículo la idea de ampliar la unidad de las secuencias de caracteres a palabras.

En la réplica del experimento hasta este punto obtuvimos resultados similares a los reportados por Clark (2001). Mediante una evaluación manual de los *clusters* determinamos una medida F (promedio armónico entre precisión y cobertura) del 65% (véase *Tabla 4*).

2.5 Modificaciones al experimento original

En el experimento original, Clark (2001) agrupa las secuencias en *clusters* con el propósito de inducir símbolos no-terminales automáticamente. Una vez detectados, estos símbolos le son de utilidad para inducir reglas gramaticales. Por esta razón, aplica el filtro de MI sobre el *cluster* entero, ya que la estimación de MI resulta más precisa si se hace sobre la variedad de ocurrencias de etiquetas contenidas en ese grupo.

Es legítimo preguntarse qué pasa si en lugar de aplicar el filtro de MI sobre un *cluster* lo aplicamos sobre cada secuencia. Ello implicaría dejar de lado el objetivo de inducir símbolos no-terminales y reglas gramaticales, pero por otro lado brindaría la posibilidad de determinar en forma más o menos inmediata si una secuencia dada y de ocurrencia frecuente es un constituyente, lo cual reviste utilidad práctica. Para ello, es preciso modificar la manera de estimar su MI: en lugar de un promedio sobre todo el *cluster*, usamos la fórmula para MI punto a punto (*pointwise MI*), calculada sobre un promedio entre secuencias de la misma longitud:

$$MI(X;Y) = \log \frac{p(x,y)}{p(x)p(y)} \qquad \text{(iii)}$$

Pointwise mutual information
(información mutua punto a punto)

En la *Figura 3* se puede apreciar un diagrama que compara el experimento original de Clark y nuestra adaptación al español. Clark agrupa secuencias en *clusters* y luego determina en cuáles hay constituyentes mediante el cálculo de MI de cada *cluster*, mientras que en nuestro caso calculamos MI de cada secuencia y luego las agrupamos en *clusters* de constituyentes de similar distribución. Nuestro paso de *clustering* no es parte esencial del criterio de definición de constituyentes, sino simplemente una forma de asegurarnos la viabilidad del proceso al demostrar convergencia con los resultados del algoritmo original.

3. Evaluación

Resumiendo la descripción de nuestro experimento:

i. Dividimos las 2108 oraciones en dos grupos: 2000 oraciones para entrenamiento y 108 para la aplicación de los constituyentes inducidos (véase *Anexo VI*).
ii. Definimos un umbral de frecuencia de 110 ocurrencias para el paso 1 del algoritmo: obtuvimos 198 secuencias candidatas a constituyentes.
iii. Aplicamos el criterio de MI punto a punto a las secuencias candidatas de (i.) (paso 3 del algoritmo): obtuvimos 107 secuencias validadas como constituyentes.
iv. Arracimamos las 107 secuencias con *clustering* basado en *k-means*: obtuvimos 25 *clusters* de alta pureza (véase *Anexo IV*).
v. Repetimos el experimento con distintos umbrales en (i.), con los resultados de la *Tabla 4*.
vi. Evaluamos manualmente la salida de (iii.) con nuestros propios juicios de gramaticalidad (véase *Anexo V*), de modo de calcular la medida F para cada escenario de (v.): obtuvimos distintos valores para la columna de constituyentes válidos (*n positivos* en la *Tabla 4*).

Las evaluaciones del algoritmo original de Clark y de nuestra implementación revelaron que ambos métodos convergen a resultados similares. Obtuvimos una medida F de alrededor del 65% para el escenario más efectivo de nuestro experimento (véase *Tabla 4*). En cuanto a la extensión y la calidad de los constituyentes, el listado de constituyentes inducidos abarca no sólo casos obvios con etiquetas triviales, sino que, sorprendentemente, en muchos casos se han inducido correctamente constituyentes con etiquetas poco frecuentes (por ejemplo, ocurrencias de CRD –adjetivo cardinal– bien integradas a sintagmas nominales). En otros casos, la extensión de los constituyentes llega a 5 y 6 etiquetas (véase *Anexo V*), lo que demuestra la viabilidad del enfoque para inducción de complejas estructuras de constituyentes.

Precisión

$$P = \frac{n°\,ConstituyentesPositivos}{n°\,ConstituyentesPositivos + FalsosPositivos} \quad \text{(iv)}$$

Cobertura

$$C = \frac{n°\,ConstituyentesPositivos}{n°\,ConstituyentesPositivos + FalsosNegativos} \quad \text{(v)}$$

Medida $F = \frac{(\beta^2 + 1) * P * C}{\beta^2 * P + C}$ (vi)

(Con $\beta = 1$ para asignar igual peso a P y a C)

umbral de ocurrencias	nº Positivos (Paso3)	Candidatos (Paso1)	Precisión %	Cobertura %	medida F %
110	62	198	59	74	66
50	125	464	58	65	61
30	166	786	53	53	53
15	242	1561	49	41	45
10	291	2429	51	32	39

Tabla 4: Evaluación de la medida F para distintos escenarios de experimentación según umbral de ocurrencias

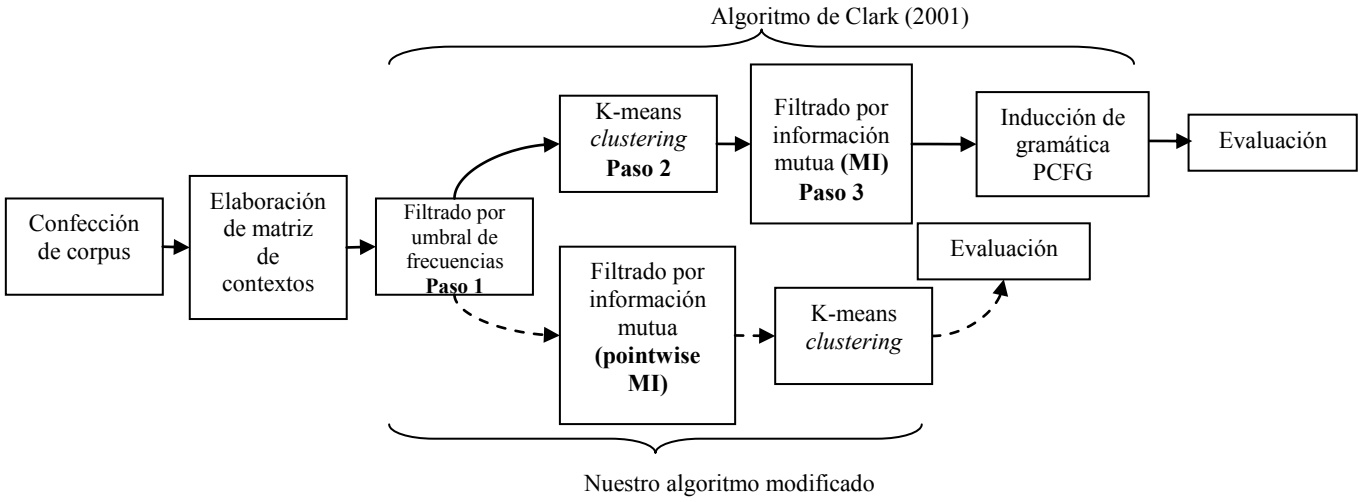


Figura 3: Experimento original de Clark (2001) y nuestra adaptación al español

4. Conclusiones y trabajo a futuro

El presente trabajo verifica la factibilidad de encarar la tarea de inducción de constituyentes sintácticos en español con un enfoque basado en la información distribucional. Además, el experimento ofrece interesantes implicancias acerca de las propiedades formales extrínsecas de los constituyentes sintácticos. Si bien un constituyente es primariamente una secuencia de símbolos más o menos frecuente en la distribución de un *corpus* morfosintácticamente etiquetado, esta condición no es suficiente para definir un constituyente. Más bien, lo que el experimento refleja es que el verdadero filtro entre las secuencias frecuentes de etiquetas, candidatas a ser constituyentes, es la información mutua entre los símbolos que co-ocurren en las adyacencias de dichas secuencias.

Ahora bien, en nuestro experimento prototípico también tropezamos con algunos obstáculos. Por un lado, nos encontramos con el consabido problema de la sobreestimación del rol de la información mutua (Li 1990) y de la dispersión de datos en los modelos estadísticos (Fong y Berwick 2008), situación que se ve agravada al trabajar con un *corpus* de implementación prototípica que implica dimensiones no tan masivas. Esto explica por qué la medida F decrece tanto cuando el umbral de aceptación de candidatos a constituyentes baja a apenas 15 ocurrencias (véase *Tabla 4*).

Por otro lado, como el mismo Clark (2001) reconoce, el modelo falla en capturar como constituyentes secuencias de etiquetas de muy rara ocurrencia. Esto se condice con la extensión y composición de los constituyentes inducidos (véase *Anexo V*). Los constituyentes extensos (5 o más etiquetas), en general, pueden describirse como constituyentes cortos de etiquetas frecuentes imbricados en otros. Es decir, existe la tendencia a que los constituyentes más extensos se compongan de las etiquetas más frecuentes. Esto se verifica, por ejemplo, en la dificultad que encuentra el experimento para modelar proposiciones subordinadas o constituyentes en los que entran en juego etiquetas menos frecuentes.

El experimento nos revela 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.

En cuanto al trabajo a futuro, nos trazamos los siguientes ejes de organización. Primero, resulta fundamental continuar con los trabajos de revisión y ampliación del *corpus*, hasta alcanzar dimensiones apropiadas para una experimentación con contundencia científica, más allá de la etapa exploratoria del prototipo. A su vez, debemos refinar y homologar aún más los criterios de anotación para los eventuales colaboradores del proyecto, a fin de alcanzar dicha masividad en el *corpus*.

En segundo término, notamos la necesidad de refinar y validar el algoritmo de inducción, experimentando con otras métricas derivadas de la teoría de la información, tales como la distancia de divergencia Kullback-Leibler (Manning y Schütze 1999). Asimismo, se hace imprescindible considerar otros enfoques propuestos para el problema de la inducción de constituyentes sintácticos y evaluar la factibilidad de los mismos en un *corpus* en español. Los trabajos más conocidos para dicha tarea son los modelos de Klein y Manning (2002 y 2004), como así también trabajos provenientes de otros paradigmas de investigación como el conexionismo (Reali *et al.* 2003). En particular, una muy fructífera línea de investigación podría ser la combinación de modelos de inducción de constituyentes lineales, como el que Clark (2001) propone, con modelos de dependencias sintácticas (Mel'čuk 1988; Paskin 2002; Klein y Manning 2004).

Finalmente, en lo que atañe al objetivo más ambicioso de trabajo a futuro, nos proponemos continuar con una investigación integral para analizar la factibilidad de inducir una gramática independiente de contexto completa del español en relación con la toma de una postura en el debate epistemológico actual entre sesgos fuertes *versus* débiles (*weak bias* y *strong bias*, respectivamente) como componente innato en la adquisición del lenguaje (Lappin y Schieber 2007). Para encarar dicho objetivo, es importante segmentar el proceso total de inducción en tareas parciales, una de las cuales, en principio,

bien podría ser la inducción de constituyentes sintácticos que estamos describiendo en este artículo. De hecho, en investigaciones previas (Balbachan y Dell'Era 2008) probamos la plausibilidad de la inducción de categorías sintácticas en español a partir de un algoritmo no supervisado, tarea cuya salida podría ser aprovechada para que sea automáticamente procesada como los datos de entrada del proceso de inducción de constituyentes sintácticos.

Consideramos que el mérito de la presente implementación prototípica es experimentar con modelos de inducción de fenómenos sintácticos que puedan aportar renovada evidencia al debate acerca de la adquisición del lenguaje; en especial, si consideramos que la investigación de 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 o incluso la biolingüística, pero la plausibilidad de dicha evidencia mediante una modelización efectiva es asunto para la agenda actual de la lingüística computacional.

Anexo I Listado completo de etiquetas morfosintácticas para anotación de corpus

El conjunto de etiquetas que usaremos se basa en el llamado *C5 tagset*, un estándar que se aplicó al etiquetado del *British National Corpus* (*BNC*). Como etiquetamos texto en español, prescindimos de algunas etiquetas específicas del idioma inglés y agregamos otras más apropiadas.

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*, *así que*, *luego*)
12	CJS	conjunción subordinante (excepto *que*)
13	CJT	conjunción subordinante *que* (en "Dijo que...")
14	CRD	adjetivo numeral cardinal (*tres*)
15	DAT	fecha (*7 de noviembre*)
16	DPS	determinante posesivo (*su*, *mi*)
17	DT1	determinante definido singular (*aquel* hombre)
18	DT2	determinante definido plural (*aquellos* hombres, *todos* los hombres)
19	EX0	existencial *hay*
20	ITJ	interjección (*ah*, *ehmm*)
21	NN0	sustantivo neutro en número (*virus*)
22	NN1	sustantivo singular (*lápiz*)
23	NN2	sustantivo plural (*lápices*)
24	NNP	sustantivo propio (*Buenos_Aires*)
25	ORD	adjetivo numeral ordinal (*sexto*, *3ro.*, *último*)
26	PND	pronombre demostrativo (¿Cuál querés? *Éste*, ¿Cuál querés? *Esto*)
27	PNI	pronombre indefinido (*ninguno*, *todo*)
28	PNP	pronombre personal (*tú*)
29	PNQ	pronombre interrogativo (*quién*)
30	POS	pronombre posesivo (*mío*)
31	PPE	pronombre personal enclítico (dar\ *lo*, *se* cuasi-reflejo (*morirse*, él *se* cayó)
32	PRP	preposición (excepto *de*) (*sin*)
33	REL	pronombre relativo (*quien* en "el presidente, quien avisó...")
34	SEP	*se* pasivo ("se venden casas") e impersonal ("se reprimió a los manifestantes")
35	VBG	gerundio de verbo cópula (*siendo*)
36	VBI	infinitivo de verbo cópula (*ser*)
37	VBN	participio de verbo cópula (*sido*)
38	VBZ	verbo cópula conjugado (*es*)

39	VM0	infinitivo de verbo modal (*soler*)
40	VMZ	verbo modal conjugado (*debía*)
41	VMG	gerundio de verbo modal (*pudiendo*)
42	VMN	participio de verbo modal
43	VVG	gerundio de verbo léxico (*obrando*)
44	VVI	infinitivo de verbo léxico (*vivir*)
45	VVN	participio de verbo léxico (*cifrado*)
46	VVZ	verbo léxico conjugado (*vive*)
47	XX0	adverbio de negación (*no*)
48	$$$	fin de oración

Anexo II Composición del corpus

longitud promedio de oraciones: 23,68
cantidad de oraciones: 2.108
cantidad de palabras etiquetadas: 49.925

etiqueta	n	%
PRP	9613	19.25
NN1	8280	16.58
AT1	6484	12.99
VVZ	3857	7.73
NN2	3409	6.83
NNP	2405	4.82
AJ1	1897	3.80
AV0	1610	3.22
CJC	1574	3.15
AT2	1501	3.01
CRD	1056	2.12
VVI	902	1.81
AJ2	851	1.70
VVN	847	1.70
REL	835	1.67
CJT	730	1.46
PPE	621	1.24
VBZ	571	1.14
DPS	424	0.85
SEP	347	0.70
CJS	250	0.50
DT1	250	0.50
XX0	226	0.45
ORD	153	0.31
PNP	135	0.27
AJ0	127	0.25
VMZ	118	0.24
PNI	118	0.24
DAT	113	0.23
VVG	112	0.22
DT2	104	0.21
PND	103	0.21
AT0	64	0.13
EX0	58	0.12
NN0	54	0.11
VBI	47	0.09
AJC	46	0.09
PNQ	11	0.02
AJS	8	0.02
VBG	7	0.01
AVQ	5	0.01
VMN	1	0.00
ITJ	0	0.00
POS	0	0.00
VBN	0	0.00
VM0	0	0.00
VMG	0	0.00

Anexo III Ejemplo de anotación para el corpus

CONTRA LOS ACUSADOS
AMIA: piden que se amplíe un embargo
El/AT1 fiscal/NN1 Alberto_Nisman/NNP le/PNP pidió/VVZ a/PRP el/AT1 juez/NN1 federal/AJ1 Rodolfo_Canicoba_Corral/NNP ampliar/VVI a/PRP 540/CRD millones/NN2 de/PRP dólares/NN2 el/AT1 embargo/NN1 contra/PRP los/AT2 iraníes/NN2 acusados/VVN de/PRP haber_planeado/VVI y/CJC
ordenado/VVI la/AT1 ejecución/NN1 de/PRP el/AT1 atentado/NN1 terrorista/AJ1 contra/PRP la/AT1 AMIA/PNP ./$
El/AT1 monto/NN1 --expresado en pesos significan 1.843 millones-- surge/VVZ a/PRP el/AT1 considerar/VVI todos/DT2 los/AT2 daños/NN2 provocados/VVN por/PRP el/AT1 atentado/NN1 de/PRP 1994/NN0 , incluido/VVN un/AT1 resarcimiento/NN1 para/PRP los/AT2 familiares/NN2 de/PRP los/AT2 85/CRD muertos/NN1 y/CJC para/PRP los/AT2 más/AV0 de/PRP 200/CRD heridos/NN2 que/REL dejó/VVZ el/AT1 ataque/NN1 ./$

Anexo IV Paso 2 del algoritmo de Clark (2001) aplicado a nuestro corpus: clustering de secuencias candidatas

Anexo IV: Paso 2 del algoritmo Clark (2001): clustering de secuencias candidatas a constituyentes

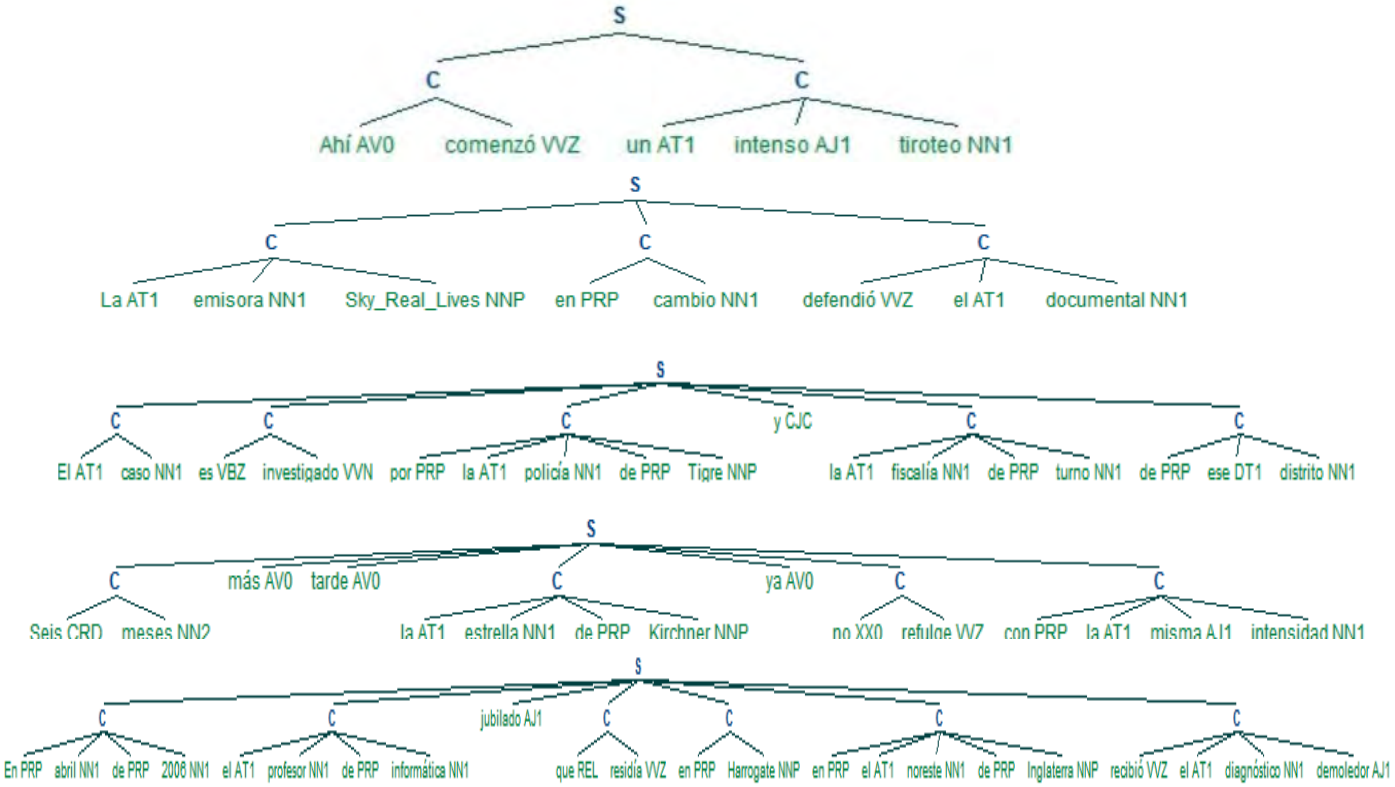
0	1	2	3	4	5	6	7	8
PRP DT1	AV0 AT1NN1	NNP NNP	PRP AT1NN1PRP AT1NN1	XX0	NN2 PRP AT1	REL VVZ	NN1AT1NN1	VVZ
AV0 PRP AT1	PRP VVI	AT1NN1CJC	PRP AT1NN1PRP NN1	SEP	AT1AJ1	AJ1CJC	AJ1NN1	VVZ PRP AT1NN1
PRP DPS	AV0	NNP CJC	VVN	PPE	AT1	CJC	NN1AV0	VBZ VVN
PRP NN1PRP AT1	CJT VVZ		PRP NN2 AJ2	PNP	CJT AT1	PRP AT1NN1CJC	NN1	VVZ AV0
PRP VVI AT1	CJT AT1NN1		PRP AT1NN1NNP		AT1NN1PRP AT1	CJC VVZ	NN1PRP AT1NN1AJ1	VVZ NN2
AV0 AT1	CJT		PRP AT1NN1PRP NNP		DT1		NN1AJ1	VVZ PRP VVI
PRP AT1			PRP CRD NN2		DPS		NN1PRP AT1NN1	PPE VVZ
PRP AT1NN1PRP AT1			CJC NN2		NNP PRP AT1		NN1PRP CRD NN2	VVZ AT1NN1
PRP AT1AJ1			PRP AT1NN1		NNP AT1		DAT	VVZ VVN
			CJC AT1NN1		AT2 NN2 PRP AT1		NN1PRP NNP	
			AV0 PRP AT1NN1		VVI AT1		NN1PRP NN1	
			PRP AT1NNP				NN1NNP	
			PRP AT2 NN2					
			PRP NN1PRP AT1NN1					
			AJ1PRP AT1NN1					
			PRP NN1AJ1					
			PRP AT1NN1AJ1					
			PRP NNP					
			VVN PRP AT1NN1					
			PRP NN1					
			PRP AT1AJ1NN1					
			PRP DPS NN1					
			PRP AV0					
			PRP NN2					

13	14	15	16	17	18	19	20	21
NN2 PRP	REL VVZ PRP	AT1NN1PRP AT2	NN2 PRP AT2	NN2 CJC	NN2 AJ2	VVZ AT1	NN1PRP AT2 NN2	NN1PRP CRD
AT1NN1PRP	VBZ VVN PRP	AT2		NN2 REL	AJ2 NN2	CJC AT1	NN1PRP NN2	NN1REL VVZ
AT2 NN2 PRP	VVZ PRP				NN2 PRP AT1NN1	VVZ AT1NN1PRP AT1	NN1VVN	NN1VBZ
NNP PRP	AT1NN1VVZ PRP				NN2	AJ1PRP AT1		NN1CJC
CRD NN2 PRP	AJ2 PRP				NN2 PRP NN2	VVN PRP AT1		NN1PRP AT2
NN2 PRP AT1NN1PRP	VVZ PRP AT1NN1PRP				NN2 PRP NN1	VVZ PRP AT1		NN1VVZ
NN2 AJ2 PRP	NNP VVZ PRP				NN2 AV0			
AT1NN1PRP NN1PRP	PRP NN1PRP				NN2 VVZ			
VVI PRP	PRP NNP PRP							
AT1NN1AJ1PRP	CRD PRP							
AT1NN1PRP AT1NN1PRP	PRP AT2 NN2 PRP							
	PRP NN2 PRP							
	PRP AT1NN1PRP							
	VVN PRP							
	PRP VVI PRP							
	PRP							
	CJC PRP							
	SEP VVZ PRP							
	AV0 PRP							
	PRP AT1NN1AJ1PRP							
	VVZ AT1NN1PRP							
	PPE VVZ PRP							
	AJ1PRP							

Anexo V Muestra de la salida final del experimento con constituyentes filtrados

	secuencia	longitud	MI max	MI argmax	MI promedio
	AJ1 NN1	2	8.345	AT1--CJC	0.053
error	AJ1 PRP	2	10.122	NN1--AT1	0.050
error	AJ1 PRP AT1 NN1	4	8.645	NN1--AJ1	0.014
error	AT1	1	11.372	PRP--REL	0.242
	AT1 AJ1 NN1	3	8.129	PRP--PRP	0.040
	AT1 NN1	2	11.325	PRP--CJC	0.188
	AT1 NN1 AJ1	3	9.972	PRP--PRP	0.052
error	AT1 NN1 AJ1 PRP	4	8.536	PRP--AT2	0.072
error	AT1 NN1 PRP	3	10.391	PRP--NN2	0.080
error	AT1 NN1 PRP AT1	4	10.411	PRP--NN1	0.102
	AT1 NN1 PRP AT1 NN1	5	8.203	PRP--AV0	0.077
error	AT1 NN1 PRP AT1 NN1 PRP	6	8.009	PRP--AT1	0.022
	AT1 NN1 PRP NN1	4	9.223	PRP--AJ1	0.055
	AT1 NN1 PRP NN2	4	6.180	PRP--VVZ	0.017
	AT1 NN1 PRP NNP	4	8.917	PRP--VVZ	0.047
	AT1 NN1 VVN	3	8.387	PRP--PRP	0.021
	AT1 NN1 VVZ	3	7.885	$$$--PRP	0.050
error	AT1 NN1 VVZ PRP	4	8.151	PRP--VVI	0.064
	AT2 NN2	2	9.409	PRP--CJC	0.060
error	AT2 NN2 PRP	3	8.630	PRP--ORD	0.032
error	AT2 NN2 PRP AT1	4	8.277	PRP--NN1	0.043
	AT2 NN2 PRP AT1 NN1	5	5.741	PRP--$$$	0.037
	AT2 NN2 VVZ	3	5.964	PRP--NN2	0.029
error	AV0	1	10.030	VVZ--PNI	0.313
error	AV0 AT1 NN1	3	5.991	$$$--AJ1	0.032
error	AV0 PRP	2	9.209	VVZ--DPS	0.081
	AV0 PRP AT1 NN1	4	6.107	PRP--AJ1	0.032
	AV0 VVZ	2	7.798	PRP--DPS	0.049
error	CJC	1	9.969	NN1--PNI	0.270
error	CJC AT1 NN1	3	8.002	NN2--ORD	0.036
	CJC VVZ	2	8.537	NN1--DPS	0.046
error	CJT AT1 NN1	3	6.951	VVZ--PRP	0.022
	CJT VVZ	2	7.798	PRP--DPS	0.045
	CRD NN2	2	8.979	PRP--CJC	0.061
	NN1	1	12.421	AT1--REL	0.344
	NN1 AJ1	2	11.061	AT1--CJC	0.071
error	NN1 AJ1 PRP AT1	4	8.852	AT1--NN1	0.023
error	NN1 AT1 NN1	3	8.630	PRP--ORD	0.035
error	NN1 PRP	2	11.943	AT1--NN0	0.129
error	NN1 PRP AT1	3	11.137	AT1--ORD	0.049
	NN1 PRP AT1 NN1	4	10.639	AT1--AJ1	0.082
	NN1 PRP AT1 NN1 AJ1	5	6.965	AT1--$$$	0.036
error	NN1 PRP AT1 NN1 PRP AT1	6	5.970	AT1--REL	0.010
error	NN1 PRP AT2	3	10.419	AT1--NN2	0.046
	NN1 PRP AT2 NN2	4	6.807	AT1--VVZ	0.021
	NN1 PRP CRD	3	10.185	AT1--NN2	0.023
	NN1 PRP CRD NN2	4	7.544	AT1--VVZ	0.013
	NN1 PRP NN1	3	9.554	AT1--PRP	0.049

Anexo VI Muestra de constituyentes inducidos sobre algunas oraciones de prueba



Bibliografía

Balbachan, Fernando y Diego Dell'Era. 2008. Técnicas de clustering para inducción de categorías sintácticas en un corpus de español. En *Infosur* (2):95-104.

Carroll, Glenn y Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. En C. Weir, S. Abney, R. Grishman y R. Weischedel (eds.). *Working notes of the workshop statistically-based NLP techniques*. AAAI Press.

Chater, Nick y Christopher Manning. 2006. Probabilistic models of language processing and acquisition. En *Trends in Cognitive Sciences* (10):335-344.

Chen, Stanley. 1995. Bayesian grammar induction for language modeling. En *ACL* (33):228–235.

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

Chomsky, Noam. 1959. A review of B. F. Skinner's 'verbal behavior'. En *Language* (35):26–58.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA. MIT Press.

Chomsky, Noam. 1966. *Topics in the Theory of Generative Grammar*. París. Mouton.

Chomsky, Noam. 1986. *El conocimiento del lenguaje*. Madrid. Alianza.

Civit i Torruella, Montserrat. 2003. *Criterios de etiquetación morfosintáctica de corpus en español*. Tesis de doctorado, Universidad de Barcelona.

Clark, Alexander. 2001. Unsupervised language acquisition: theory and practice. Sussex. School of Cognitive and Computing Sciences, University of Sussex Press.

Clark, Eve. 2009. *First Language Acquisition*. Cambridge, Inglaterra. Cambridge University Press.

Cowie, Fiona. 1999. *What's Within? Nativism Reconsidered*. Oxford. Oxford University Press.

Finch, Steve, Nick Chater y Martin Redington. 1995. Acquiring syntactic information from distributional statistics. En J. P. Levy, D. Bairaktaris, J. A. Bullinaria y P. Cairns (eds.). *Connectionist Models of Memory and Language*. Londres. UCL Press.

Fodor, Jerry. 1983. *La modularidad de la mente*. Madrid. Morata.

Fong, Sandiway y Robert Berwick. 2008. Treebank parsing and knowledge of language: a cognitive perspective. En *Proceedings of the 30th annual conference of the Cognitive Science Society*:539-544. Austin, Texas.

Gambino, Omar J. y Hiram Calvo. 2007. On the usage of morphological tags for grammar induction. En Alexander Gelbukh y A. F. Kuri

Morales (eds.). *MICAI 2007, LNAI 4827*: 912–921. Berlín. Springer-Verlag.

Gold, E. Mark. 1967. Language identification in the limit. En *Information and control* (10):447-474.

Harris, Zellig S. 2000 (1951). *Structural Linguistics*. University of Chicago Press.

Klein, Dan y Christopher Manning. 2001. Distributional phrase structure induction. En *Proceedings of CoNLL 2001*:113-121.

Klein, Dan y Christopher Manning. 2002. A generative constituent-context model for improved grammar induction. En *Proceedings of ACL 2002*:128-135. Philadelphia.

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

Lakoff, George. 1987. *Women, Fire, and Dangerous Things*. The University of Chicago Press, Chicago.

Langacker, Ronald. 2000. Estructura de la cláusula en la gramática cognoscitiva. En *Volumen monográfico 2000:*19-65. Universidad de California, San Diego.

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

Lari, Karim y Steven Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. Computer Speech and Language (4):35–56.

Leech, Geoffrey, Roger Garside y Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. Reporte técnico, Lancaster University.

Li, Wentian. 1989. Mutual information functions of natural language texts. Santa Fe Institute preprint.

Li, Wentian. 1990. Mutual information functions versus correlation functions. En *Journal of statistical physics* (60):823-837.

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

Mel'čuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany, NY.

Meyer, Charles. 2002. *English Corpus Linguistics*. Cambridge, Inglaterra. Cambridge University Press.

Moreno Sandoval, Antonio. 1998. *Lingüística computacional*. Madrid. Síntesis.

Moreno Sandoval, Antonio, Susana López Ruesga y Fernando Sánchez León. 1999. Spanish Treebank. Reporte técnico, versión 5. Universidad Autónoma de Madrid.

Olivier, Donald. 1968. *Stochastic grammars and language acquisition mechanisms*. Tesis de doctorado, Harvard University.

Paskin, Mark A. 2002. Grammatical bigrams. En T. G. Dietterich, S. Becker y Z. Ghahramani (eds.) *Advances in Neural Information Processing Systems 14*. Cambridge, Massachusetts. MIT Press.

Piattelli-Palmarini, Massimo (ed.). 1980. *Language and Learning: the Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA. Harvard University Press.

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

Pinker, Steven. 1994. *El instinto del lenguaje*. Madrid. Alianza.

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*. Berkeley, California.

Pullum, Geoffrey y Barbara Scholz. 2002. Empirical assessment of stimulus poverty arguments. En *The Linguistic Review* (19):9-50.

Reali, Florencia, Morten Christiansen y Padraic Monaghan. 2003. Phonological and distributional cues in syntax acquisition: scaling-up the connectionist approach to multiple-cue integration. En *Proceedings of the 25th annual conference of the Cognitive Science Society Lawrence Erlbaum Associates, Inc*.:970-975. Mahwah, New Jersey.

Redington, Martin, Nick Chater, Chu-Ren Huang, Li-Pin Chang, Steve Finch y Keh-jiann Chen. 1995. The universality of simple distributional methods: identifying syntactic categories in Chinese. En *Proceedings of the Cognitive Science of Natural Language Processing*. Dublín.

Sánchez León, Fernando. 1994. Spanish tagset for the CRATER project. Reporte técnico, Universidad Autónoma de Madrid.

Santorini, Beatrice. 1991. Part-of-speech tagging guidelines for the Penn treebank project. Reporte técnico MS-CIS-90-47, University of Pennsylvania.

Van Zaanen, Menno. 2000. ABL: Alignment-based learning. En *COLING* (18):961–967.

Wolff, J. Gerard. 1988. Learning syntax and meanings through optimization and distributional analysis. En Y. Levy, I. M. Schlesinger y M. D. S. Braine (eds.). *Categories and Processes in Language Acquisition*. Lawrence Erlbaum, Hillsdale, NJ.