Técnicas de clustering para inducción de categorías sintácticas en un corpus de español

Clustering techniques for induction of syntactic categories onto a Spanish corpus

Fernando Balbachan, Diego Dell'Era
Facultad de Filosofía y Letras, Universidad de Buenos Aires
Buenos Aires, Argentina
fernando_balbachan@yahoo.com.ar , diego.dellera@gmail.com

Abstract

Among statistical models for Natural Language Processing (NLP) approaches, clustering techniques have turned of interest to both computational linguistics and psycholinguistics, as a plausible solution on how a grammar can be acquired completely from scratch (from a *tabula rasa*). We propose this current research as one of the first systematic clustering experiments on large corpora in Spanish, which presents substantial improvements with respect to previous work (Redington et al. 1998). Its short-term goal is to empirically demonstrate that distributional information is a powerful tool for the induction of syntactic categories. Our heuristics includes a *Decreasing Frequency Profile* (Ćavar et al. 2004), mutual information (Shannon 1948), and *K-means* algorithm (Manning and Schütze 1999) in order to find out, in a non-arbitrary non-aprioristic fashion, syntactic cues that will later provide the foundations for the vector space modelization. We will thus obtain clusters of a reasonable purity and evidence for semantic bootstrapping. At long term, we will plan to profit from the results by working with translinguistic evidence and studying the feasibility for inducing more refined syntactic rules.

Keywords: clustering, distributional information, induction of syntactic categories, semantic bootstrapping

Resumen

Como parte de los modelos estadísticos de aproximación al Procesamiento de Lenguaje Natural (PLN), las técnicas de *clustering* han venido atrayendo la atención convergente de la lingüística computacional y de la psicolingüística, como una solución plausible al problema de la adquisición de una gramática a partir de una *tabula rasa*. La presente investigación se enmarca en dicho paradigma y se propone como uno de los primeros intentos sistemáticos de *clustering* para grandes *corpora* en español que incorpora sustanciales mejoras respecto de trabajos anteriores (Redington et al. 1998). La meta a corto plazo es demostrar empíricamente que la información distribucional es una poderosa herramienta para la inducción de juicios de pertenencia de ítems lexicales a categorías sintácticas. En particular, se recurre a una heurística de *Decreasing Frequency Profile* (Ćavar et al. 2004), información mutua (Shannon 1948) y al algoritmo *K-means* (Manning y Schütze 1999) para una identificación no arbitraria y no apriorística de marcas sintácticas que han de sentar las bases del posterior modelado vectorial con *clusters* de razonable pureza y evidencia de facilitación semántica temprana. A más largo plazo se espera sacar provecho de los resultados obtenidos,

proponiendo un modelo cognitivo de adquisición de reglas sintácticas refinadas a partir de evidencia translingüística.

Palabras clave: clustering, información distribucional, inducción de categorías sintácticas, facilitación semántica

1. INTRODUCCIÓN

La lingüística computacional es una novedosa transdisciplina entre la informática y la lingüística que se ocupa de desarrollar ingenios o mecanismos informáticos para el Procesamiento de Lenguaje Natural (o *Natural Language Processing* NLP en inglés) con una clara articulación entre investigación aplicada y tecnología. Desde la década del '90, a la par del salto tecnológico que representaron los procesadores con enorme poder de cálculo, el paradigma de investigación dominante en esta transdisciplina son los enfoques estadísticos, que parten del concepto de *tabula rasa* y de grandes *corpora* para analizar patrones de ocurrencia de eventos estadísticamente significativos, logrando así que la computadora *aprenda* conocimiento inherentemente lingüístico. Esto hace que la agenda de la lingüística computacional desde el paradigma estadístico resulte muy atractiva para la inteligencia artificial y la psicolingüística.

Si bien el paradigma estadístico es relativamente reciente en cuanto a investigación propiamente dicha, existen ya diversas *tareas* estándares que marcan los principales lineamientos de trabajo en el campo. Las tareas representan un estándar para evaluar aportes originales frente al mismo problema mediante algoritmos optimizados o enfoques radicalmente novedosos. Una de esas tareas es el *clustering* o agrupamiento de ítems lexicales a partir de grandes corpora lingüísticos en función de la información sintáctica que la máquina va aprendiendo. El presente trabajo aborda las primeras conclusiones surgidas a partir de una investigación llevada a cabo por Fernando Balbachan y Damir Ćavar durante 2006 en Indiana University sobre clustering en un corpus de español mediante un algoritmo optimizado de diseño original.

2. INFORMACIÓN DISTRIBUCIONAL Y ADQUISICIÓN DEL LENGUAJE

2.1. Debate teórico

La adquisición del lenguaje, y en particular de las categorías sintácticas, representa todavía un problema irresuelto. Desde el punto de vista innatista, las reglas gramaticales son innatas, de modo que la adquisición del lenguaje se reduce a establecer una relación entre el léxico y las categorías sintácticas, lo cual deviene, a falta de otras restricciones, en una explosión combinatoria. Desde el punto de vista empirista, el problema es aún más difícil porque ni siquiera se cuenta con dichas reglas innatas.

En principio, las fuentes de información que se pueden considerar relevantes para el problema de la adquisición son las siguientes:

- El contexto comunicativo: En el modelo propuesto por Bruner et al. (1975) la adquisición depende de la comunicación del niño con los adultos que lo crían. En otras palabras, el modelo se basa en la idea de que el uso pragmático del lenguaje determina el proceso de adquisición.

- Los indicios fonológicos: Este modelo sugiere la necesidad de analizar cómo el hablante puede distinguir cuáles indicios fonológicos son relevantes.
- El conocimiento innato: Además de proporcionar información, el conocimiento innato sirve como restricción ante el problema de la explosión combinatoria: o bien provee mecanismos de aprendizaje de la sintaxis, o bien provee las categorías y sus relaciones.
- La prosodia: La prosodia es importante en los modelos donde la mutua predictibilidad entre la sintaxis de una frase y la manera en que se enuncia contribuye al aprendizaje del hablante. La cantidad de datos a analizar para determinar la importancia de la prosodia es tratable.
- La información distribucional: La información distribucional indica el contexto lingüístico en que aparece una palabra. Su utilidad depende de la observación de que las palabras de una misma categoría sintáctica tienen cierta regularidad distribucional.

Debido a las críticas basadas en enfoques innatistas, durante mucho tiempo los estudios de la adquisición del lenguaje evitaron los enfoques alrededor de esta última fuente de información. Sin embargo, estudios relativamente recientes tales como los de Redington et al. (1998) propusieron el uso de análisis distribucionales como una fuente de información relevante (aunque no excluyente). Para ello, intentaron responder en primer lugar a las principales 5 objeciones innatistas, tal como aparecen en Pinker (1984):

1. El análisis distribucional no puede estudiar todas las posibles relaciones estructurales, a riesgo de devenir en explosión combinatoria. Ante esto, Redington observa que no era necesario que los mecanismos de aprendizaje distribucionales estudien todas las relaciones posibles, dado que pueden extraer información útil hasta de las relaciones más simples.
2. Las propiedades abstractas del lenguaje son invisibles para el análisis distribucional, que se limita a la única propiedad visible a su alcance: la adyacencia. La respuesta de Redington consiste en afirmar que el análisis distribucional sirve para descubrir y poner a prueba esas relaciones abstractas.
3. La abundancia de las correlaciones distribucionales es un desperdicio inconducente, porque el lenguaje corrobora sólo ciertas propiedades, y sólo ciertas relaciones entre esas propiedades. Ante esto, la respuesta hace notar que si en realidad había una capacidad lingüística innata, dicha capacidad universal también incurre en el derroche, porque el lenguaje particular que el hablante aprende a partir de ella utilizaba sólo algunos de sus parámetros.
4. El análisis distribucional no puede filtrar relaciones y datos espurios que no son posibles en el lenguaje a aprender. La respuesta obvia es que ese argumento sólo vale contra estudios distribucionales simplistas o mal ejecutados.
5. El análisis distribucional no puede hacer el tipo de pruebas de sustitución de palabras de una misma categoría de las que es capaz alguien que ya ha adquirido el lenguaje, por lo cual resulta inútil para estudiar el problema de la adquisición. Ante esto, Redington sostiene que el modelo distribucional no necesariamente implica que la adquisición proceda de una *tabula rasa*; la información distribucional no era la solución general al problema, sino *una* de las fuentes. De hecho, esto no resulta incompatible con la idea nativista de un *semantic bootstraping*, o facilitamiento semántico, por el cual el niño aprende las categorías sintácticas de sustantivo y verbo en términos de categorías semánticas conocidas (objeto y acción), usando su conocimiento del significado de las palabras para mapear el léxico sobre las categorías sintácticas innatas.

2.2. Historia

Los análisis distribucionales anteriores a la revolución chomskyana pretendían relacionar ítems lingüísticos con su contexto mediante rigurosos métodos de estudio de campo. La investigación sobre la adquisición del lenguaje no formaba parte de ese emprendimiento, pues consideraban al lenguaje como un *constructo* cultural, ajeno al punto de vista psicológico o computacional. El legado de Chomsky dio paso a la crítica de los límites de esos enfoques, y tuvo como efecto secundario el relegar a la oscuridad el análisis distribucional en sí.

Sin embargo, hubo quienes mantuvieron viva la llama, con suerte dispar. Los enfoques de redes neuronales, inicialmente prometedores, resultaron poco efectivos, y mucho menos eficientes. Por su parte, los avances del enfoque estadístico a mediados de los años setenta chocaron con la limitación de los recursos computacionales y de los *corpora* de la época, pero no sin antes demostrar que, si bien eran impracticables a gran escala con *corpora* reales, en principio podían brindar resultados significativos.

3. EXPERIMENTOS PREVIOS CON TÉCNICAS DE CLUSTERING

La propuesta de Redington consiste en el uso de la información distribucional en tres etapas: 1. analizar la distribución del contexto de ocurrencia de cada palabra; 2. comparar la distribución del contexto de ocurrencia de pares de palabras; y 3. agrupar las palabras cuyas distribuciones del contexto de ocurrencia sean similares.

Para analizar la distribución del contexto de ocurrencia de cada palabra se usa una unidad denominada *bigrama*, es decir, la coocurrencia de pares de ítems léxicos en una relación fija. Dicha relación puede ser, por ejemplo, la que existe entre una palabra *target* (es decir, la palabra que se pretende estudiar) y su contexto (la palabra inmediatamente siguiente), relación denominada comúnmente *ventana*. Por ejemplo, si todo el corpus consistiera de una única frase, "*la vaca salta sobre la cerca*", el siguiente fragmento de la tabla representaría el vector de contexto correspondiente a la palabra *salta*:

Target	*Contexto*			
	la	vaca	sobre	cerca
salta	0	0	1	0

Figura 1: Ejemplo de vector de bigramas hacia la derecha para la palabra "*salta*" en la oración "*la vaca salta sobre la cerca*"

Si se considera a cada vector como un punto en el espacio, y se calcula la distancia entre ellos (adoptando como medida de distancia la *Spearman Rank Correlation* en función de su independencia de la estructura del espacio vectorial), es de esperar que los ítems que pertenecen a una misma categoría sintáctica tengan una distribución similar, lo cual se traduce en una cercanía en el espacio vectorial.

Ahora bien, la asignación de categorías sintácticas asume que hay una manera de dividir esas mismas categorías, lo cual plantea un problema. Se han propuesto modelos donde la frontera es discreta, y otros donde es prototípica o basada en similitudes entre ítems individuales. Redington, por su parte, considera que este es un problema previo, de orden teórico; por lo tanto evita tomar posición respecto de cuán estricta sea esta frontera, y decide usar en la práctica un *cluster* (agrupamiento) de tipo jerárquico, en el cual ítems similares se unen para formar una nueva categoría, que a su vez se une con categorías similares. Obtiene así una estructura llamada

dendrograma, donde los miembros de los *clusters* en los nodos de las hojas están más estrechamente vinculados que los miembros de los *clusters* cercanos a la raíz. Si se corta el dendrograma en un nivel dado, se obtiene una cantidad discreta de *clusters* que representan categorías sintácticas.

Redington escoge CHILDES como corpus. Para el experimento, inserta cada palabra en su categoría más frecuente, según la base de datos de Collins Cobuild (es decir, no contempla la ambigüedad de las palabras que pueden pertenecer a más de una categoría). Para obtener una medición de los resultados, corta el dendrograma en distintos niveles, y evalúa su grado de *precision* y *completeness* con fórmulas a tal efecto, respecto de un nivel base. También evalúa la informatividad mutua entre los resultados y el nivel medido inicialmente (es decir, qué porcentaje del total de información de ambos niveles representa la información compartida entre el resultado y el nivel medido inicialmente).

Una serie de experimentos revela cuán importante puede llegar a ser, desde el punto de vista psicológico, la manipulación de parámetros del análisis y de los datos de entrada. A continuación, las conclusiones de cada una de sus experiencias:

- El contexto precedente es más informativo que el contexto siguiente. El uso de contextos más amplios mejora la precisión, pero empeora la completitud (porque a medida que crece el contexto, crece también el número de posibles construcciones sintácticas). El contexto ideal, al nivel del dendrograma elegido, es de dos palabras, antes y después de la palabra *target*. Este contexto, local y pequeño, impone una restricción al tipo de relaciones entre palabras, y constituye una respuesta a la objeción de Pinker según la cual la infinita cantidad de relaciones posibles haría inútil el intento de obtener información válida al usar el enfoque distribucional.
- La efectividad del método de *clustering* varía en forma de campana invertida según el número de palabras *target*: no brinda información cuando hay pocas palabras (porque supone que entre ellas están las más frecuentes, y las más frecuentes pertenecen a categorías cerradas), ni cuando hay muchas (porque su precisión aumenta a la vez que su completitud decrece). El método funciona mejor cuando tanto la cantidad de palabras *target* como la de palabras de contexto es reducida, y se condice con el número de palabras (aproximadamente 1000) que puede llegar a conocer típicamente un niño de tres años.
- La precisión de las categorías descubiertas por el método distribucional es consistente con el orden en que el niño aprende las categorías sintácticas: en primer lugar, la del sustantivo (para la cual la información distribucional es también la más abundante) y luego la del verbo (cuya información distribucional le sigue en importancia).
- La informatividad crece ilimitadamente a medida que aumenta el tamaño del corpus (lo cual es obvio), pero se comprueba que, incluso cuando el corpus se limita a las palabras a las que un niño está expuesto, el método distribucional también brinda información relevante.
- Si se restringen las palabras de contexto a aquellas que deban estar incluidas dentro del límite del mismo enunciado al que pertenece la palabra *target*, existen niveles de *clustering* que brindan información valiosa, pero la diferencia es insignificante. La separación de enunciados, por ende, no es esencial para el método distribucional.
- Agregar la frecuencia de las palabras mejora en mucho cualquier análisis distribucional.
- La omisión de las palabras funcionales deja un corpus de palabras de contenido que sigue siendo informativo distribucionalmente.

- La información sobre la frecuencia de elementos lexicales particulares sirve para clasificar mejor las palabras de otras categorías sintácticas.
- El estudio distribucional sobre un corpus del lenguaje que las madres usan con sus niños y sobre un corpus de lenguaje adulto indica que la información distribucional no cambia de uno a otro, de lo cual se puede inferir que el lenguaje maternal (*motherese*) no ayuda particularmente al aprendizaje de las categorías sintácticas en el niño. Ello a su vez se condice con la evidencia externa de que los niños cuyas madres no les hablan en el lenguaje maternal no experimentan retrasos en el aprendizaje del lenguaje.

4. CRÍTICA A REDINGTON – APORTE ORIGINAL DEL EXPERIMENTO BALBACHAN-ĆAVAR (2006)

Si bien el trabajo de Redington fue uno de los primeros intentos sistemáticos en trabajar con clustering sobre grandes *corpora*, el experimento adolece de ciertas fallas de diseño, que podrían resultar incompatibles con los lineamientos epistemológicos del paradigma estadístico. En particular, la crítica al algoritmo de Redington se centra en dos aspectos claves del diseño del experimento:

1) Redington recurre a una modelización apriorística del espacio vectorial en función de definir arbitrariamente el número de marcas sintácticas (*cues*) vs. palabras blanco (*target*), tomando como único parámetro la frecuencia relativa de ocurrencia de los tipos (los primeros 150 tipos y los siguientes 1000 tipos, respectivamente sobre un corte arbitrario de los 1150 tipos más frecuentes, lo cual en un corpus de 1.500.000 tokens y decenas de miles de tipos parece algo escaso). Es decir, no se invocan propiedades inherentes a la información distribucional a la hora de distinguir entre *marcas* y *palabras blanco*.
2) Redington trabaja con clustering **jerárquico** (*Hierarchical Average Link*) del tipo dendrograma, pero estipula un nivel de corte de los clusters en 0,8. Este coeficiente es otro parámetro de diseño arbitrario con sensibles consecuencias en el modelado lingüístico del espacio vectorial.

La arbitrariedad en el diseño del algoritmo socava los mismos principios epistemológicos que el paradigma estadístico se propone defender. La idea de la *tabula rasa* es que la máquina misma tome las decisiones de naturaleza lingüística exclusivamente en función de la información distribucional de los ítems léxicos, y no que éstas sean estipuladas *a priori* por la intuición lingüística o el arbitrio del investigador. No obstante, sí es menester reconocerle a Redington un gran aporte en cuanto al estudio del papel concreto que juega la ventana (palabras contexto alrededor de las palabras *target*) en los modelos de n-gramas en cuanto a la informatividad del análisis distribucional.

Tomando esta crítica 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 es dar con un procedimiento de identificación no arbitraria y no apriorística de las marcas sintácticas (*cues*) que habrán de sentar las bases del posterior modelado vectorial con *clusters*, demostrando que la información distribucional es una poderosa herramienta para la inducción de juicios de pertenencia de ítems lexicales a categorías sintácticas.

4.1. Corpus Balanceado

Se comenzó por armar un corpus de entrenamiento con 2 millones de *tokens*, organizados en oraciones bien formadas del español. Debido a la masiva necesidad de oraciones gramaticales y a los requerimientos de procesamiento (*tokenización*) se optó por la incorporación de voluminosos textos en formato electrónico (libros electrónicos) y artículos periodísticos, de modo de balancear el registro textual. El corpus final abarcó 2.000.000 de palabras y 71.467 tipos.

4.2. Primera etapa del algoritmo: Identificación de marcas

Para la identificación de cues o marcas se trabajó con un *Perfil de Frecuencia Decreciente* (Ćavar et al. 2004): una lista de bigramas ordenados por frecuencia y una bolsa de palabras conteniendo todos los tipos. La heurística consiste en anotar las ocurrencias a la derecha y a la izquierda de las palabras más frecuentes hasta abarcar el 90% de la bolsa de palabras totales. En ese punto es dable pensar en un corte entre las marcas sintácticas versus las palabras de contexto. Nuestra intuición algorítmica se justifica plenamente en que esa lista de 107 *cues* sintácticas coincide aproximadamente con la clase de palabras funcionales del español.

De este modo, a diferencia del de Redington, nuestro experimento se propone como un método no apriorístico y no arbitrario para identificar marcas sintácticas sobre las cuales se trabajará con técnicas de *clustering*. Este enfoque, además, presenta interesantes implicancias para un abordaje desde la perspectiva psicolingüística en cuanto a la hipótesis de adquisición del lenguaje a partir de la identificación de las palabras funcionales de una lengua.

4.3. Segunda etapa del algoritmo: Modelización del espacio vectorial

Una vez identificadas las marcas (cues), el resto de los tipos (71.360) pasa a ser considerado palabras *target*. De este modo, se procede a una modelización de un espacio vectorial bajo la forma de una matriz *cue* vs. *target*, conteniendo cada vector 107 dimensiones. Es decir, cada palabra contexto es caracterizada en función de su frecuencia de aparición en bigramas *cue-target* respecto de cada una de las 107 palabras *marca*. Se incorpora un análisis de Información Mutua (Shannon 1948) para limitar la ventana de análisis por bigramas únicamente hacia la derecha de la palabra *marca*, lo cual coincide con la intuición de gramática creciente hacia la derecha para las categorías nucleares.

Este espacio vectorial requiere un poder de cálculo 12,5 veces mayor que el del experimento de Redington, pero confiamos en que pueda sentar las bases de un estudio sistemático de mayor alcance respecto del aprendizaje de máquina para la inducción de juicios gramaticales.

4.4. Tercera etapa del algoritmo: Clustering con algoritmo K-means

En el paradigma estadístico de la lingüística computacional, las técnicas de clustering se han posicionado como una tarea estándar a la hora de explorar datos. Para nuestro experimento recurrimos a un algoritmo no jerárquico denominado K-means que consiste en:

1) Comenzar por asignar 2 clusters (K=2), tomando los dos vectores más alejados entre sí, y calcular los respectivos centroides de esos clusters.
2) Calcular el error como la sumatoria de la distancia Euclideana al centroide de todos los vectores de cada cluster.
3) Iniciar una nueva iteración con un nuevo cluster (K=n+1) y reasignar los vectores más alejados del centroide al nuevo cluster
4) Recalcular centroides para los nuevos clusters y nuevo error

5) Repetir el algoritmo desde el paso 2) hasta que el error de nueva asignación sea mayor que el de la iteración actual. Idealmente, el número máximo de clusters es Kmax = nº de *cues*

Si bien el clustering o agrupamiento es un procedimiento aplicable a cualquier conjunto de datos, en el dominio específico del Procesamiento de Lenguaje Natural la aplicación de clustering sobre los tipos de palabras conlleva una ventaja adicional: la inducción de categorías sintácticas o clases de palabras como categorías lingüísticas no apriorísticos con un claro correlato matemático-estadístico. En particular, en el experimento Balbachan- Ćavar (2006), manejamos la hipótesis de que los distintos clusters estarán definidos por un comportamiento estadístico diferencial en cuanto a la ocurrencia sintáctica de sus miembros respecto de las palabras *marcas* (palabras funcionales).

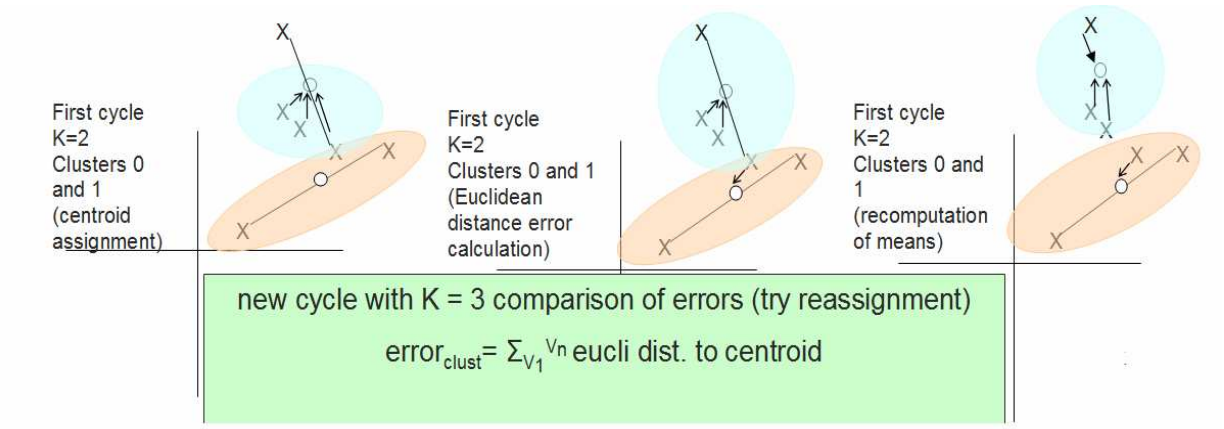

Figura 2: Ciclo de iteración con el algoritmo de clustering K-means. Esquema de 2 clusters (K=2) de vectores con centroides

5. ANÁLISIS DE LOS RESULTADOS

El experimento Balbachan-Ćavar (2006) consistió en aplicar el algoritmo descrito en los parágrafos 4.2, 4.3 y 4.4 al corpus de entrenamiento mencionado en el parágrafo 4.1. Como ya se ha mencionado anteriormente, el poder de cálculo requerido es elevado, por lo que la implementación del experimento resultó un aspecto no trivial. Al cabo de varios días de procesamiento disponíamos de los primeros resultados del proceso de clustering. Algunas observaciones a destacar en el análisis de los resultados son:

1) Todas las categorías sintácticas mayores habían sido inducidas con razonable grado de pureza, y refinadas en rasgos de género y número (para sustantivos) y en otras caracterizaciones (verbos finitos vs. verbos en infinitivo).
2) En algunos casos han sido inducidos clusters de alto grado de refinamiento categorial como pronombres demostrativos y posesivos.
3) Se observa evidencia de facilitación semántica (*semantic bootstrapping*). En efecto, uno de los clusters agrupó relaciones familiares como *madre*, *padre*, *marido*, *mujer*, *hermano*, aun cuando estos sustantivos son de género diverso.
4) Incluso entre los verbos se evidencia cierto grado de especificidad subcategorial al agrupar en un cluster aparte verbos de actitud verbal como *contestó*, *exclamó*, *repuso*, *preguntó*, *añadió*.

5) La pureza y refinamiento de los clusters es mayor para los grupos más numerosos (sustantivos y verbos con una pureza cercana a 100%), pero tiende a valores estadísticamente inaceptables en el caso de grupos de pocos miembros (preposiciones con pureza de 25%).

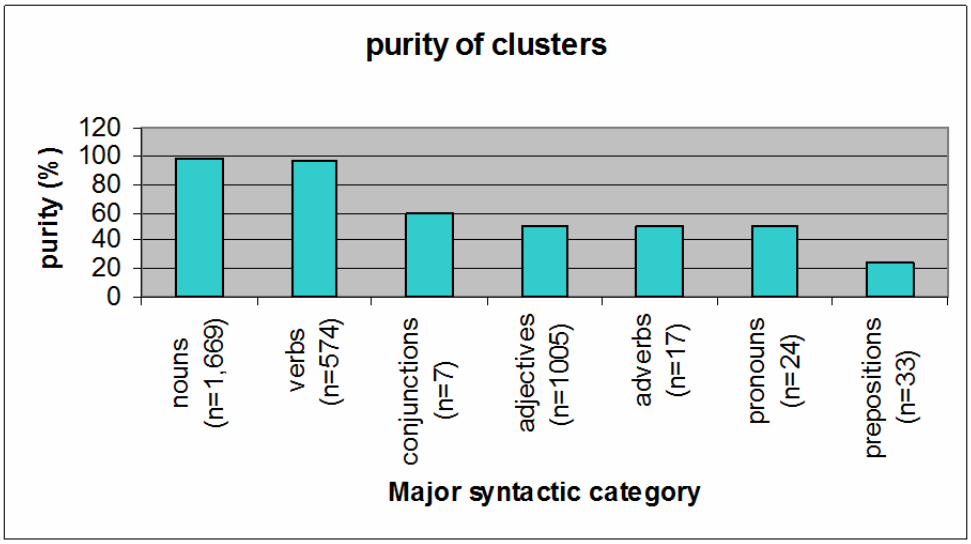

Figura 3: Gráfico de pureza de clusters para categorías sintácticas mayores en experimento Balbachan-Ćavar (2006)

Si bien el experimento ha suministrado evidencia contundente de la viabilidad de enfoques que contemplen las técnicas de clustering en el marco del paradigma estadístico de investigación de la lingüística computacional, el objetivo a largo plazo de estos experimentos no es el agrupamiento en sí mismo sino la posibilidad de inducir una gramática completa a partir del mismo. Para ello es importante tomar una perspectiva interlingüística (trabajar con corpora en varios idiomas) que avale tales directrices de diseño de experimentos. No obstante, reconocemos que esta meta a largo plazo es por demás muy ambiciosa para el estado de arte actual de la disciplina, tomando en cuenta las limitaciones tecnológicas de poder de cálculo. Con todo, confiamos en que experimentos como los descritos en este artículo puedan ir sentando las bases de un campo de estudio todavía fértil y muy promisorio.

6. CONCLUSIONES

El progreso de las tecnologías de la información 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 sintácticas y la composición de las categorías sintácticas mismas. Los experimentos que aquí hemos delineado sostienen además la idea de que las marcas distribucionales no solo capturan patrones sintácticos bajo la forma de información estadística, sino que también descubren algunas relaciones semánticas. Además, hemos demostrado empíricamente la estrecha correlación entre palabras *cue* vs. palabras *target*, en función de la distinción lingüística 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 perspectivas computacional y psicolingüistica.

En este sentido, y sin menoscabo de otros mecanismos de aprendizaje que actúan simultáneamente, se puede concluir que la información distribucional se perfila como un enfoque enriquecedor. De este modo, el paradigma estadístico se propone como un promisorio *framework* de investigación que requerirá una amplia gama de herramientas y experimentos para explorar cabalmente todo su potencial.

Referencias bibliográficas

[1] Manning, C. y Schütze H. (1999). Foundations of Statistical Natural Language Processing. The MIT Press. Cambridge, Massachusetts.

[2[Redington, M., Charter N. y Finch S. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. Cognitive Science vol. 22(4) pp.425-469.

[3] Elghamry, Khaled y Ćavar, Damir (2004). Bootstrapping Cues for Cue-based Bootstrapping. Tesis de doctorado. Departamento de Lingüística, Indiana University.

[4] Shannon, Claude (1948). A mathematical theory of communication. Bell System Technical Journal, XXVII: 379-423.

[5] Damir Ćavar, Paul Rodrigues y Giancarlo Schrementi (2004). Syntactic Parsing Using Mutual Information and Relative Entropy. En Indiana University Linguistics Club (IULC) Online Working papers, proceedings of Midwest Computational Linguistics Colloquium 2004 (MCLC).

[6] Bruner, J. (1975). The ontogenesis of speech acts. Journal of Child Language, 2, 1 - 19.

[7] Pinker, S. (1984). Language learnability and language development. Cambridge, MA: Harvard University Press.