Técnicas de clustering para inducción de categorías sintácticas en español.

Autor:
Balbachan, Fernando Ariel
Tutor:
Lion, Carina

2014

Tesis presentada con el fin de cumplimentar con los requisitos finales para la obtención del título de Doctor de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires en Letras

Posgrado

Técnicas de Clustering para Inducción de Categorías Sintácticas en Español

Tesis de Doctorado

Doctorando: Fernando Balbachan
Universidad de Buenos Aires
Facultad de Filosofía y Letras
fernando_balbachan@yahoo.com.ar
Expediente Doctorado en Lingüística: 848.295/08
DNI 23.248.365 Febrero 2014

Directora de tesis: Dra. Zulema Solana (Universidad Nacional de Rosario)
zsolana@arnet.com.ar

Co-Director de tesis: Dr. Carlos Reynoso (Universidad de Buenos Aires)
billyreyno@hotmail.com

Organización de contenidos

Índice de tablas

Índice de figuras

Índice de fórmulas

Agradecimientos

Libre de la metáfora y del mito
labra un arduo cristal: el infinito
mapa de Aquel que es todas Sus estrellas.
(Jorge Luis Borges, Spinoza, *1964)*

El recorrido histórico de la ciencia muchas veces es influenciado por circunstancias mundanas y cotidianas que atañen a los hombres y mujeres de carne y hueso que investigan. Por supuesto, no me refiero a las pintorescas anécdotas que amenizan el relato ficcional de algunos descubrimientos científicos, como el baño de inmersión de Arquímedes o la manzana de Newton. Más bien aludo aquí a las incomensurables circuntancias personales que rodean la actividad del científico. La idealización del científico enclaustrado en su laboratorio, "autoexiliado" en su afán de conocimiento y ajeno a los vaivenes del mundo exterior, como el Spinoza borgeano del epígrafe, es sólo una inverosímil escena de película hollywoodense. Así pues, a lo largo de los cinco años que demandaron la concepción, el desarrollo y la concreción de esta tesis de doctorado, muchas personas desempeñaron un importante rol para llevar el proyecto a buen puerto. Algunos desde lo científico, algunos desde lo académico-institucional, algunos desde el más cotidiano apoyo personal, todos ellos merecen un reconocimiento en las páginas iniciales de esta tesis.

En primer lugar, quisiera agradecer al Profesor Damir Ćavar, quien fuera mi mentor durante mi estadía en Indiana University (IUB) en el programa de maestría en lingüística computacional. Damir fue la primera persona que me formó en Procesamiento del Lenguaje Natural y me transmitió una verdadera pasión por este campo transdisciplinar. A él le debo también los lineamientos iniciales de este proyecto cuando allá por el nevado invierno de 2006, en el medio-oeste norteamericano, debatíamos hasta bien entrada la noche acerca de los patrones estadísticos de información distribucional de naturaleza morfosintáctica en grandes *corpora*.

De regreso al país, logré encaminar este proyecto de investigación bajo la forma de una tesis de doctorado, contando con el apoyo de quien había sido mi profesor en temas relacionados con lingüística formal, Daniel Romero. Daniel se convirtió entonces en mi Director de beca de doctorado (beca UBACyT 2009-2014). Pero Daniel ha sido para mí más que un excelente profesor de lingüística formal, es la persona que más me incitó a llevar la investigación a buen término, sorteando airosamente diversas dificultades que amenazaban con hacer zozobrar el proyecto.

Desde lo institucional, quisiera agradecer también a mi Directora y a mi Co-Director de tesis: Dra. Zulema Solana y Dr. Carlos Reynoso, respectivamente. La lingüística computacional es un área de vacancia académica en el país y son pocos los referentes académicos a nivel nacional capacitados para dirigir una tesis de doctorado. Tal es el caso de Zulema. Ella me brindó un respaldo institucional incondicional en cada instancia del proyecto. A Carlos le debo el

enorme espaldarazo que significó mantener a flote la cátedra de Modelos Formales No Transformacionales (MFNT) de la orientación en Lingüística Formal de la Carrera de Letras (FFyL-UBA), de la cual formo parte desde 2007 y que actuó como ámbito institucional de promoción de actividades académicas de docencia, investigación y formación de recursos humanos en temas relacionados con la lingüística computacional. Quisiera agradecer también a la Dra. Mabel Giammatteo y a la Dra. Hilda Albano, quienes muy ecuánimemente supieron apreciar el valor de la presente investigación en el marco institucional del programa de Doctorado en Lingüística de la Facultad de Filosofía y Letras (FFyL-UBA), como el primer proyecto de Doctorado en Lingüística Computacional radicado en esta Casa de Altos Estudios.

Aunque es injusto mencionar sólo a algunos colegas de entre los muchos que ocasionalmente me han acercado observaciones, opiniones y sugerencias frente a diversos aspectos de mi trabajo, de esa numerosa lista destaco a tres. A la Dra. Yamila Sevilla debo una muy pertinente selección de material bibliográfico de investigaciones psicolingüísticas sobre la adquisición de categorías sintácticas. Al Dr. Rogelio Nazar y a la Dra. Mercè Lorente Casafont, ambos del Instituto Universitario de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF), quisiera agradecerles por sus comentarios acerca de la conformación del corpus de investigación y de la metodología en la investigación lingüística, respectivamente.

A mi compañero de cátedra, Diego Dell'Era, le debo un gran reconocimiento por su ayuda con la programación de mis experimentos en lenguaje Python. Con Diego encaramos una serie de trabajos exploratorios en los comienzos del proyecto, los cuales vieron la luz bajo la forma de publicaciones científicas en revistas especializadas. Dichos trabajos científicos prepararon el terreno para mi experimentación final. También desde el ámbito académico, no quisiera dejar pasar el merecido reconocimiento para los alumnos y adscriptos de la materia Modelos Formales No Transformacionales que colaboraron con anotaciones manuales morfosintácticas de pequeños corpora de referencia, en el marco de un proyecto de investigación grupal con reconocimiento institucional (FFyL-PRIEI 2010-2012) que tuve el honor de dirigir.

En cuanto a los afectos personales, aun aquellos que al día de hoy no están presentes, debo agradecer retroactivamente a mis padres, quienes supieron inculcar en mí muy tempranamente la pasión por el conocimiento y el esfuerzo continuo de autosuperación mediante el estudio, requisitos ambos imprescindibles para todo buen científico. A Alejandra, por su infinito cariño y a Juan, por las extensas charlas de café acerca de la vida misma.

Entones, resulta evidente que el ostracismo científico es una vaga ilusión, una idea tan romántica como la del genio del artista. En muchos proyectos científicos, y una tesis de doctorado lo es, las circunstancias diarias determinadas por las relaciones personales en lo académico, en lo científico y en lo familiar pueden tener tanto o más peso que una manzana cayendo en la cabeza de un somnoliento físico en la apacible campiña inglesa de fines del siglo XVII.

Scientists typically don't study the phenomenal world. That's why they do experiments.
Our phenomenal world is way too complex. If you took videotapes of what's happening outside your window, the physicists and chemists and biologists couldn't do anything with it.

Noam Chomsky

In God we trust, all others bring data.

William Edwards Deming

Organización de la tesis

La siguiente tesis de doctorado se propone como un aporte original al campo de la lingüística computacional, específicamente en la tarea de inducción de gramáticas formales (*grammar inference*) a partir de datos lingüísticos primarios no estructurados. Específicamente, se ofrecerá una modelización plausible al problema de la categorización temprana de palabras durante el proceso de adquisición del lenguaje para el idioma español. Si bien la particularización del modelo sobre un lenguaje puntual resulta fundamental cuando se trabaja en modelización formal estadística, de modo de recrear algorítmicamente las mismas condiciones de posibilidad de inducción de fenómenos sintácticos en comparación con aquellas de que disponen los adquirientes de un lenguaje natural, se espera que el enfoque resulte aplicable a cualquier idioma en virtud de las premisas generales de la hipótesis. En este sentido, debido a la naturaleza transdisciplinaria del enfoque, la tesis releva diversos trabajos con afiliaciones científicas que oscilan entre la psicolingüística, la lingüística formal y la lingüística computacional, en pos de compatibilizar la modelización postulada con la plausibilidad empírica. La hipótesis central de la tesis es, en alguna medida, un argumento indirecto contra el Argumento de la Probreza de los Estímulos (*Argument from the Poverty of the Stimulus* APS) en cuanto a que los Datos Lingüísticos Primarios (*Primary Linguistic Data* PLD) presentarían cierta riqueza factible de ser explotada mediante un mecanismo de aprendizaje general (no específico de dominio), tornando innecasaria la postulación de una Gramática Universal (GU) como requisito para la adquisición del lenguaje. A la luz de este argumento central, la categorización de palabras se presenta como un proceso crucial para la adquisición de una sintaxis rudimentaria. En efecto, esta habilidad temprana es el punto de partida para la construcción de una gramática por parte de los adquirientes de un lenguaje.

La tesis se organiza en nueve capítulos, comenzando por la inserción del proyecto en un paradigma científico específico de investigación lingüística: el paradigma estadístico de la lingüística computacional. El primer capítulo describe entonces los principios epistemológicos de los paradigmas de investigación en lingüística computacional y los distintos enfoques sobre el problema de la adquisición del lenguaje que derivan de la adscripción a cada uno de ellos. En este capítulo inicial también se presenta el debate en torno al Argumento de la Pobreza de los Estímulos, que se configura como nudo gordiano de la discusión sempiterna entre el innatismo y el empirismo.

El capítulo 2 presenta la hipótesis central y la metodología de este trabajo como un aporte a la comprobación empírica de la riqueza estructural de los Datos Lingüísticos Primarios para la adquisición del lenguaje mediante mecanismos generales de aprendizaje no supervisado. En particular, esta tesis se centra sobre la etapa temprana de categorización de palabras como punto de partida para la inducción de sintaxis. El capítulo también incluye una diferenciación operativa

entre las palabras funcionales y las palabras de contenido. La distinción entre palabras funcionales y palabras de contenido resulta de vital importancia para esta tesis, ya que veremos que la evidencia empírica y la modelización estadística contemplan diferencias muy notables para cada una de estas clases de palabras en el proceso ontogenético de adquisición del lenguaje.

El capítulo 3 pasa revista a los modelos formales con motivación psicolingüística que se propusieron para dar cuenta específicamente de la categorización temprana de palabras (Mintz 2002, 2003; Christophe *et al*. 2008). En particular, notaremos cómo estas propuestas adolecen de contradicciones empíricas o teóricas para abarcar el fenómeno.

El capítulo 4 explica en detalle la definición de las técnicas estadísticas de clustering como mecanismo de aprendizaje general no supervisado. Se pasa revista a los distintos algoritmos (clustering jerárquico y no jerárquico), como así también a diversas métricas específicas de la evaluación de la robustez de los clusters.

El capítulo 5 se explaya sobre el estado de la cuestión en torno a las técnicas de clustering para la tarea específica de inducción de categorías sintácticas, ya en el campo del paradigma estadístico de la lingüística computacional. Entre esos trabajos debemos destacar en particular los de Redington *et al.* (1998) y Clark (2002), cuyos lineamientos generales estaremos siguiendo en el diseño de nuestros propios experimentos.

El capítulo 6 analiza en detalle la tesis de doctorado de Wang (2012), un trabajo muy reciente que reproduce, en gran medida, el enfoque transdisciplinario con el que trabajaremos en nuestros experimentos: modelización formal estadística y adecuación explicativa ante la evidencia empírica psicolingüística. Wang (2012) trabaja específicamente con la modelización de la tarea de categorización de palabras funcionales en inglés y en alemán, con premisas de modelización que toman en cuenta la evidencia ontogenética de la adquisición del lenguaje. La tesis de doctorado de Wang (2012) es uno de los pocos trabajos en ofrecer una explicación plausible de la categorización temprana de palabras funcionales, no sólo de la de palabras de contenidos.

El capítulo 7 presenta nuestro propio experimento de categorización de palabras de contenido en español, bajo la premisa del pre-requisito de identificación de palabras funcionales sin tipología diferenciada. El experimento propone, además, algunas modificaciones metodológicas a los trabajos clásicos en técnicas de clustering. Se incluye una exhaustiva evaluación de los datos de salida del experimento.

El capítulo 8 describe otro experimento de inducción de fenómenos sintácticos, conectado con el anterior. Básicamente, se sostiene la plausibilidad algorítmica de aprovechar la información de salida del experimento del capítulo 7 como punto de partida para la construcción de una sintaxis rudimentaria, mediante la inducción de constituyentes sintácticos a partir de la etiquetación morfosintáctica de palabras.

El capítulo 9 retoma el debate en torno al Argumento de la Pobreza de los Estímulos, pero, en esta ocasión, con énfasis en los mecanismos cognitivos que plausiblemente actuarían durante el proceso de adquisición del lenguaje. Se ofrece un exhaustivo relevamiento de las posiciones tradicionales en torno al problema y una relectura de las mismas a la luz de los resultados del experimento central de esta tesis. Este capítulo final también apunta algunas conclusiones generales y traza las líneas de investigación a futuro.

En la parte final de la tesis se adjuntan varios anexos con datos de salida de los dos experimentos propuestos en esta tesis y herramientas de facilitación de la lectura: listado de siglas e índice alfabético de conceptos.

Resumen

El problema de la adquisición del lenguaje se ha presentado tradicionalmente como uno de los campos de estudio por antonomasia de la tradición psicolingüística. No obstante, en las últimas dos décadas, esta "arena epistemológica" ha venido atrayendo la atención de la lingüística computacional o Procesamiento de Lenguaje Natural hacia una plausible modelización de la ontogenésis de la adquisición del lenguaje. Entre el paradigma simbólico y el paradigma estadístico de la lingüística computacional se ha entablado un manifiesto contrapunto de concepciones epistemológicas opuestas, por ejemplo, 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.

Justamente, el Argumento de la Pobreza de los Estímulos (*Argument from the Poverty of the Stimulus* APS) se presenta como el gran campo de debate epistemológico entre el paradigma simbólico y el paradigma estadístico. Mientras el paradigma simbólico adscribe a sistemas deductivos que hipotetizan como condición necesaria un estado inicial de conocimiento innato 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.

En la última década 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 -refutación argumentativa que se conoce como *desafío* (*challenging*) en la bibliografía especializada, estos trabajos enmarcados en el paradigma estadístico abordan el problema desde la misma perspectiva inicial que el paradigma simbólico: la sintaxis como punto de partida para la adquisición del lenguaje y el isomorfismo entre lenguajes formales y lenguajes naturales. El enfoque predominante entre los trabajos del paradigma estadístico de la lingüística computacional de la última década que se abocaron a la inducción integral de una gramática formal a partir de corpora no etiquetado, recurre a las técnicas de clustering como mecanismo de aprendizaje general no supervisado, iterativo y convergente, y a la categorización de palabras como punto de partida del algoritmo. La necesidad de corroborar las hipótesis propuestas con evidencia translingüística se torna imperiosa. Sin embargo, no existen esfuerzos similares de inducción integral de sintaxis para el español.

En definitiva, tal vez sea mucho pedir para una tesis probar la invalidez completa del APS en función de inducir toda una gramática completa de un lenguaje natural a partir de los PLD por medio de métodos no supervisados de aprendizaje de dominio general. Un "atajo argumentativo"

para desafiar la validez del APS como garante de la GU sería demostrar que la etapa temprana de categorización de palabras, punto de partida de los algoritmos integrales de inducción de sintaxis que mencionamos arriba, sí puede ser inducida a partir de los PLD mediante mecanismos no supervisados de aprendizaje general no específicos de dominio. La hipótesis de esta tesis es demostrar que la tarea de categorización temprana puede ser inducida a través de los PLD a partir de indicios facilitadores (palabras funcionales de tipología indiferenciada e información distribucional), con el único pre-requisito del procesamiento fonológico y de la segmentación de palabras y frases. De este modo, el APS como garante último de la GU estaría cayendo parcialmente en cuanto a que los PLD no son tan pobres como se creía.

Si bien en las últimas dos décadas aparecieron bastantes trabajos sobre categorización de palabras, sólo recientemente la adquisición de palabras funcionales ha sido reivindicada por muy pocos trabajos como pre-requisito para el desarrollo lexical temprano, cuando toda la evidencia de producción de lenguaje parece indicar lo contrario. Mientras que las palabras funcionales no están presentes en la producción de lenguaje del niño antes de los 2 años, las palabras de contenido pueden aparecer en producción en el léxico infantil tan prematuramente como desde el año de edad (algunos pocos ítems léxicos) y ciertamente alrededor del año y medio (con medio centenar de ítems léxicos). Esta paradójica inversión lógica de un supuesto pre-requisito evidenciado posteriormente en el tiempo luego de los eductos cuya aparición supuestamente facilitaría, será resuelta en las explicaciones del trabajo de Wang (2012).

Wang (2012) aporta evidencia indirecta (omisión sistemática en producción) de una adquisición muy temprana (antes del año y medio de edad) de las categorías funcionales y evidencia directa de conitinuidad en el repertorio y uso de categorías funcionales entre adultos y niños de alrededor de 2 años de edad. Estas dos observaciones habilitan a considerar a las palabras funcionales como los candidatos ideales para *facilitar* (*boostrapping*) la categorización de palabras de contenido que se evidencia en la *explosión léxica* (*vocabulary spurt*) alrededor de los 2 años. Los indicios prosódicos, que actúan como identificadores de las palabras funcionales, permiten postular en forma muy temprana la representación asbtracta de las mismas, si no la plena adquisición, en niños de edades tan prematuras como los 14 meses.

Existen dos tradiciones experimentales que se corresponden mayormente con los paradigmas científicos dominantes en la investigación de los campos de la psicolingüística y la lingüística computacional: por un lado, las técnicas de clustering como manifestación del paradigma estadístico de la lingüística computacional, ya sea en enfoques puros (Schütze 1993; Redington *et al*. 1998) o combinados con modelos markvianos (Brown *et al*. 1992; Martin *et al*. 1998); por el otro, las teorías de los *marcos frecuentes* (Mintz 2003) y los *protoconstituyentes* sintácticos (Christophe *et al*. 2008). Estos úlitmos trabajos, con una raigambre más simbólica proveniente de la psicolingüística, muestran ciertas falencias al momento de compatibilizar sus postulados teóricos con la evidencia empírica ontogenética de la habilidad temprana de

categorización de palabras. En el terreno del paradigma estadístico, la mayoría de los trabajos de inducción de categorías morfosintácticas a partir de información distribucional mediante técnicas de clustering recurre a una misma premisa: en corpora masivos es de esperar que los ítems lexicales que pertenecen a una misma categoría morfosintáctica tengan una distribución similar, lo cual se traduce en una cercanía en el espacio vectorial.

Nuestra propuesta de modelo de inducción de categorías morfosintácticas del español responde a los siguientes lineamientos. Para el algoritmo de clustering en particular, elegimos el clustering no jerárquico K-means con distancia euclideana sobre los centroides. El espacio vectorial multidimensional quedará definido por un procedimiento de identificación no arbitraria y no apriorística de las *marcas* (*cues*) que habrán de sentar las bases del posterior modelado vectorial de las palabras targets en función de su contexto distribucional inmediato. En cuanto a la escalabilidad del algoritmo, seguiremos a Redington *et al*. (1998) y plantearemos un escenario con un vocabulario reducido de aproximadamente 1000 palabras target. De hecho, esta cantidad de palabras resulta esperable para la finalización de la etapa ontogenética que nos interesa modelizar: la *explosión léxica* (*vocabulary spurt*) que se da en los niños entre los 2 y 3 años de edad.

Aunque algunos otros trabajos clásicos en el campo sortearon con éxito el tratamiento de un espacio vectorial con excesiva dimensionalidad mediante técnicas puramente algebraicas, nos propusimos la desafiante meta de investigar la existencia de alguna propiedad intrínseca en las *marcas* (*cues*) que pudiese ser aprovechada para una reducción de la dimensionalidad más "lingüísticamente motivada". Nuestra intuición apuntaba a una diferenciación en las relaciones bigrámicas a un lado y a otro en las distribuciones cue-targe y target-cue, motivada en la noción de marcación (Lorenzo y Longa 1996) y de expansión lineal de las gramáticas de los lenguajes naturales (Ćavar *et al*. 2004; Ćavar 2010). Esta idea de marcación o informatividad hacia la derecha o hacia la izquierda puede ser matemáticamente representada a partir de la información mutua (Shannon 1948) de una palabra respecto de un corpus o en su métrica conceptualmente inversa, la entropía (Manning y Schütze 1999).

Para la evaluación general de los resultados, implementamos un criterio propio de análisis interno del cluster en función de sus miembros y otro criterio de evaluación de las distribuciones totales de las categorías predominantes. En cuanto al criterio inherente a la composición de cada cluster, tomando como criterio la distribución de frecuencias de los POS-tag en el corpus de referencia, estamos en condiciones de ponderar la incidencia de los POS-tag de cada miembro de un cluster para el *etiquetamiento del cluster* (*cluster_tag*), de modo de disponer de un criterio más adecuado que el mero conteo de la mayoría de los POS-tags presentes en la clase. Para el criterio inherente a la distribución de clusters, el concepto de *hipercluster* o mapeo *many-to-1* resulta muy productivo. Desde un punto de vista metodológico, permite una evaluación que resuelve el problema del mapeo de un número creciente de clusters inducidos en las categorías

del gold standard. Desde un punto de vista algebraico, el hipercluster se ve justificado en gran medida por la ubicación en el espacio vectorial de los centroides de los clusters que los conforman, lo cual, a su vez, refleja particularidades morfosintácticas propias del dominio lingüístico al que pertenecen los datos -por ejemplo, que los sustantivos, ya sean masculinos o femeninos, se parecen más entre sí en contraste con los verbos. En cada ciclo calculamos Precisión, Cobertura y medida F para cada uno de los 16 POS-tags, prevalezcan o no como el cluster_tag, con cada uno de los hiperclusters inducidos. Sobre estas 16 medidas F calculamos el promedio común y el promedio ponderado, obteniendo una efectividad total de 69%, doce puntos porcentuales por arriba del experimento de Redington *et al*. (1998), cuyos lineamientos de diseño nos inspiraron, y algunos puntos por debajo del estado del arte para el inglés 72,4% en el trabajo de Clark (2003). Es de destacar que a partir de los ciclos medios (ciclo 52 en adelante), las medidas F de la mitad de los POS-tag se presentan consolidadas en valores relativamente estables, especialmente para las categorías mayores (sustantivos y verbos).

Los experimentos detallados en esta tesis nos revelan una importante veta de indagación científica que obliga a replantearse cuestiones tan sensibles para la lingüística como la naturaleza del lenguaje y los mecanismos de adquisición del mismo, a la luz de las promisorias técnicas de aprendizaje de máquina y de los procesos de inducción de gramáticas. El experimento central que aquí hemos delineado sostiene, además, la idea de que las marcas distribucionales no sólo capturan patrones morfosintácticos bajo la forma de información estadística, sino que también descubren algunas relaciones semánticas. 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 respectivas agendas de investigación de la lingüística computacional y de la psicolingüística.

Aunque no demostramos necesariamente que sea éste y no otro el mecanismo por el cual se adquiere una gramática de un lenguaje natural, sí demostramos la invalidez del APS en cuanto a que los PLD son suficientmente ricos para inducir una gramática formal (al menos, las categorías POS-tags) únicamente a partir de la información distribucional. 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 última instancia, la evidencia psicolingüística debería ser refrendada por la neurología o incluso por la biolingüística, pero la plausibilidad de dicha evidencia mediante una modelización efectiva es asunto para la agenda actual de la lingüística computacional.

Capítulo 1. El debate epistemológico en torno a un problema recurrente

1.1 Paradigmas de investigación en linguística

Tradicionalmente se han sugerido cuatro metodologías para que los lingüistas comprueben sus hipótesis: la introspección, los experimentos psicolingüísticos, las encuestas de datos y los corpora. Charles Fillmore (1992) describe dos modos aparentemente antagónicos de hacer lingüística: por un lado, la "lingüística de butaca" (*armchair linguistics*) o lingüística teórica, con sus juicios de introspección, y por otro lado, la lingüística de corpus, con la supuesta "trivialidad" de sus resultados observables:

> "[…] A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, «Wow, what a neat fact!», grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. […] A caricature of the corpus linguist is something like this. He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts." [Fillmore 1992:35]

A pesar de la mordaz caricatura de estas metodologías, Fillmore apunta, conciliador, que lo ideal sería que estos dos tipos de lingüistas coexistan "en un mismo cuerpo". La lingüística computacional, una transdisciplina entre la informática y la lingüística, que se ocupa de desarrollar ingenios o mecanismos informáticos para el procesamiento de lenguaje natural con una clara articulación entre investigación aplicada y tecnología, bien podría encarnar este *desideratum* fillmoreano.

En ocasiones se ha simplificado el aporte de la lingüística computacional, reduciéndolo a un mero aspecto ingenieril suplementario de la lingüística teórica:

> "La diferencia entre las tareas y los métodos de la lingüística humanística y de la lingüística computacional se puede comparar con la diferencia existente entre el trabajo de un ornitólogo y un constructor de aviones" [Galicia Haro y Gelbukh 2007:5]

Sin embargo, las investigaciones en lingüística computacional de las últimas dos décadas han aportado un muy fértil sustrato para la elucidación de problemas tradicionalmente considerados por la lingüística teórica, como la adquisición del lenguaje y el procesamiento de textos, trascendiendo así el lugar de mera "técnica aplicada" con que solía ser apartada de las lides epistemológicas de la lingüística y constituyéndose, en cambio, en una pujante transdisciplina científica.

El problema de la adquisición del lenguaje se ha presentado tradicionalmente como uno de los campos de estudio por antonomasia de la tradición psicolingüística. No obstante, en las últimas dos décadas, esta "arena epistemológica" ha venido atrayendo la atención de la lingüística computacional o Procesamiento del Lenguaje Natural hacia una plausible modelización de la ontogenésis de la adquisición del lenguaje. En efecto, la lingüística computacional y la psicolingüística han venido trabajando mancomunadamente en describir en detalle modelos formales plausibles que puedan dar cuenta del proceso específico por el cual un

ser humano desarrolla lenguaje ante una serie de *estímulos* (*input*), en un cierto período de tiempo, utilizando un mecanismo general de aprendizaje y/o conocimiento innato de especie (Clark 2002; Wang 2012) -aunque sobre estos últimos dos recursos existen posturas con marcadas diferencias entre los diversos paradigmas científicos dominantes en lingüística computacional.

Como toda disciplina científica se enmarca en un paradigma o teoría general, es importante entender los distintos compromisos epistemológicos que subyacen a los modelos y algoritmos con los que el lingüista computacional ha de trabajar:

> "Antes de entrar a describir las principales «escuelas» u orientaciones actuales en lingüística, conviene aclarar que debajo de etiquetas denominativas como *gramática*, *modelo gramatical*, *modelo lingüístico*, *mecanismo*, *formalismo* o *teoría lingüística*, residen básicamente tres niveles epistemológicos distintos. El primero, el de la *teoría del lenguaje*, en el que se postulan principios generales sobre la naturaleza y el funcionamiento del lenguaje humano; el segundo, el de los *modelos lingüísticos*, en el que, a partir de las hipótesis principales de la teoría lingüística del primer nivel, se establece una modelización del lenguaje, un simulacro sobre su composicionalidad, estructura y procesamiento; y, finalmente, un tercer nivel epistemológico, el de los *mecanismos*, en el que intervienen aspectos de formalización y de representación de los datos lingüísticos particulares y de los datos generales derivados por generalización." [Cabré y Lorente 2003:1-2]

En el campo de la lingüística computacional existen tres grandes paradigmas de investigación: modelos conexionistas, modelos estadísticos y modelos simbólicos, los cuales no necesariamente comparten principios epistemológicos con los paradigmas dominantes en Lingüística Teórica, aun cuando sí comparten el mismo objeto de estudio.

Más allá del caso puntual de la adquisición del lenguaje, en el campo más general del Procesamiento de Lenguaje Natural, el paradigma conexionista adscribe a un procesamiento paralelo e indiferenciado de datos a través de redes neuronales que refuerzan o debilitan el peso de sus nodos a partir de un mecanismo de aprendizaje convergente denominado *backpropagation* (Schütze 1993). Esto hace que se ubique epistemológicamente muy cerca de las ciencias cognitivas, aunque sus logros en el campo están limitados a ciertas tareas lingüísticas muy compartimentadas (Mc.Murray y Hollich 2009). El paradigma simbólico, que evolucionó bajo la égida chomskyana de las gramáticas generativas de la década del '50 (Chomsky 1957, 1965; Moreno Sandoval 1998), se propone manipular categorías sintácticas *apriorísticas* 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 a diversas técnicas probabilísticas aplicadas a grandes corpora de entrenamiento, los cuales hacen las veces de datos lingüísticos primarios, con vistas a inducir categorías y fenómenos específicos del lenguaje natural a partir de la detección de patrones estadísticamente significativos.

1.2 El problema de la adquisición del lenguaje

No obstante, 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, por ejemplo, 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." [Manning y Charter 2006:335]

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 2002), más recientemente podemos empezar a rastrear esta confrontación en obras primordiales de la lingüística teórica (Chomsky 1957, 1965, 1975, 1986) y de la psicolingüística (Pinker 1979, 1994) de la segunda mitad del siglo XX, las cuales defienden un innatismo a ultranza, en tanto las posturas empiristas puras son esgrimidas por la lingüística cognitiva (Lakoff 1987; Langacker 2000).

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 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 2002).

Justamente, el Argumento de la Pobreza de los Estímulos (*Argument from the Poverty of the Stimulus* 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 (también conocidas como constructivistas) no deben ser confundidas con un trasnochado conductismo y su concepción del lenguaje como una mera asociación de esquemas estímulo-respuesta ni con un ascético fundamentalismo que reniega de cualquier conocimiento inicial que pueda echar a andar el mecanismo de la adquisición del lenguaje, como si acaso el niño se enfrentara a tal problema cual *tabula rasa* (Redington *et al.* 1998). Los empiristas no refutan la existencia de algún mecanismo inicial como condición necesaria para adquirir el lenguaje; simplemente postulan que ese 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 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 2002:20]

Tampoco se pone en duda la necesidad de disponer de cierto conocimiento innato para desarrollar una gramática y usarla tal como los adultos lo hacen (*cf.* concepto de *continuity* en Wang 2012). En este sentido, la principal diferencia entre los innatistas y los empiristas radica en el grado de conocimiento específicamente lingüístico o conocimiento general, respectivamente, que aportarían inicialmente estas estructuras o mecanismos innatos, conocidos en la bibliografía como *sesgos fuertes* o *sesgos débiles* (*strong bias* o *weak bias*), respectivamente (Lappin y Shieber 2007; Clark y Lappin 2013).

1.3 La pobreza de los estímulos y la riqueza de lo innato

La confrontación entre el paradigma simbólico y el paradigma estadístico en torno al problema de la adquisición del lenguaje se desató en dos frentes. Por un lado, desde la teoría de lenguajes formales, la supuesta imposibilidad de aprendizaje del lenguaje natural a través de una gramática formal de jerarquía superior a las Gramáticas Independientes de Contexto (*Context-Free Grammars* o *CFG*) ante la falta empírica de evidencia negativa (Gold 1967). Por otro lado, la renuencia de Chomsky y sus seguidores a dar crédito a las nociones estadísticas de la época como herramientas de análisis:

> "Jonathan Cohen señala que los argumentos que yo uso contra E[*empirismo*] muestran sólo que las «técnicas de simple enumeración» son inadecuadas para el aprendizaje lingüístico […] pero que dichos argumentos no tienen relación con las «técnicas de inducción eliminadora» […] El problema de la

> propuesta de Cohen es que no existen «técnicas de inducción eliminadora» en ningún sentido pertinente." [Chomsky 1975:253]

> "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]

En el capítulo 9 de esta tesis retomaremos este prolífico debate entre empirismo e innatismo. No obstante, es menester introducir en este punto el argumento formal, en torno al cual parece girar la cuestión, y su derrotero epistemológico en el campo de las ideas científicas concurrentes sobre el problema de la adquisición del lenguaje.

El APS como justificación de la postulación de una Gramática Universal (GU) ya había sido sugerido por Chomsky (1965), mucho antes de que los trabajos sobre lenguajes formales pusieran bajo la lupa el isomorfismo chomskyano entre lenguajes naturales y lenguajes formales. Desde su obra fundacional (Chomsky 1957) y hasta bien entrada la década del '70, la gramática generativo-transformacional se asimiló a una *Máquina de Turing* (Peters y Ritchie 1973). Aun con el drástico pasaje de un modelo reglar, como lo era la *Teoría Estándar*, hacia un modelo representacional como *Principios y Parámetros* (Eguren y Soriano 2004), una gramática particular demandaría parametrizaciones en el orden de 10^{15} para un período de aprendizaje que se extiende durante *sólo* 10^{8} segundos (Norvig 2011). Incluso en el caso de contemplar modelos de lenguaje markovianos de n-gramas (mucho menos costosos, pero sin estructura sintáctica), las parametrizaciones para un lenguaje de 20.000 palabras serían $8*10^{12}$ sólo para el caso de trigramas. En uno u otro caso, estas especificaciones formales costosísimas (Grishman 1986) obviamente no podían ser aportadas por los Datos Lingüísticos Primarios (Primary Linguistic Data PLD) en un período crítico de tiempo finito reducido (Pinker 1984), dada su supuesta pobreza estructural (Clark 2002). Así pues, la conclusión de estas premisas es la necesidad de postular la existencia de una Gramática Universal innata disponible para un *Language Acquisition Device* (LAD) (Chomsky 1965, 1975), algo así como un *órgano o facultad de la lengua* con un dominio específico (Fodor 1983):

> "In this chapter we consider several computational learning models that have been applied to the language task. Some of these have yielded results that suggest that the class of natural languages cannot be efficiently learned from the primary linguistic data (PLD) available to children through domain general methods of induction. Several linguists have used these results to motivate the claim that language acquisition requires a strong set of language specific learning biases, encoded in a biologically evolved language faculty that specifies the set of possible languages through a Universal Grammar." [Clark y Lappin 2011:2]

Hasta aquí, el APS como único argumento a favor de la GU podía ser reducido a la siguiente formalización:

> "*a*. Children acquire knowledge of natural language either through domain general learning algorithms or through procedures with strong language specific learning biases that encode the form of a possible grammar. *b*. There are no domain general algorithms that could learn natural languages from the primary linguistic data. *c*. Children do learn natural languages from primary linguistic data. *d*.

> Therefore children use learning algorithms with strong language specific learning biases that encode the form of a possible grammar." [Clark y Lappin 2011:2]

Sin embargo, hacia fines de la década del '60 aparece un trabajo fundamental acerca de los límites de la aprendibilidad de los lenguajes formales (Gold 1967) que fascinó a los más acérrimos defensores del innatismo (Johnson 2004; Clark y Lappin 2011), ofreciéndoles, en principio, un asidero argumentativo más robusto, aunque, como veremos más adelante, la lectura que se hizo de este trabajo estaba viciada de errores de interpretación. En efecto, el trabajo de Gold (1967) *Identification In the Limit* (IIL) fue el primero en proponer una teoría computacional de la aprendibilidad de lenguajes formales. Así pues, en virtud del isomorfismo entre lenguajes formales y lenguajes naturales, resultó muy tentador extender las conclusiones del *Teorema de Gold* hacia el problema de la aprendibilidad de lenguajes naturales frente a los Datos Lingüísticos Primarios, si bien este movimiento exegético no estuvo exento de vicios de origen:

> Some linguists and psychologists have invoked learning theoretic considerations to motivate this version of the APS. So Wexler (1999), apparently referring to some of Gold (1967)'s results, states that «*The strongest most central arguments for innateness thus continue to be the arguments from APS and learnability theory...The basic results of the field include the demonstration that without serious constraints on the nature of human grammar, no possible learning mechanism can in fact learn the class of human grammars*.» […] **Gold's results do not entail linguistic nativism. Moreover, his model is highly problematic if taken as a theory of human language learning** […] We suggest that computational learning theory does not motivate strong linguistic nativism, nor it is relevant to the task o funderstanding language acquisition […] **it is not a substitute for a good psycholinguistic account of the facts. However, it can clarify the class of natural language representations that are efficiently learnable from PLD**." [Clark y Lappin 2011:2] (*las negritas y el subrayado son nuestros*)

Gold (1967) propone una teoría del aprendizaje de lenguajes formales como un mecanismo de hipótesis convergente ante la presentación iterativa de *cadenas de símbolos* (*strings*) pertenecientes a un lenguaje de la jerarquía de lenguajes formales de Chomsky (Johnson 2004; Moreno Sandoval 2001; Clark y Lappin 2011). Este mecanismo se denomina *Identificación en el Límite* (*Identification In the Limit* IIL), ya que funciona sobre las bases de un aprendizaje incremental continuo que va restringiendo hipótesis generalizadoras hasta converger en la gramática formal que mejor describe la colección de strings presentados como evidencia. Eventualmente, una clase de lenguajes formales puede ser aprendida en su totalidad si resulta aprendible cada uno de los lenguajes formales pertenecientes a esa clase:

> "In this paradigm a language L consists of a set of strings, and an infinite sequence of these strings is a *presentation* of L. The sequence can be written s_1 , s_2 , … , and every string of a language must appear at least once in the presentation. The learner observes the strings of a presentation one at a time, and on the basis of this evidence, he/she must, at each step, propose a hypothesis for the identity of the language. Given the first string s_1, the learner produces a hypothesis G_1, in response to s_2. He/she will, on the basis of s_1 and s_2, generate G_2, and so on.
> For a language L and a presentation of that language s_1 , s_2 , … the learner identifies in the limit the language L, iff there is some N such that for all $n > N$, $G_n = G_N$, and G_N is a correct representation of L. IIL requires that a learner converge on the correct representation G_L of a language L in a finite but unbounded period of time, on the basis of an unbounded sequence of data samples, and, after constructing G_L, he/she does not depart from it in response to subsequent data. A learner identifies in the limit the class of languages £ iff the learner can identify in the limit every $L \in$ £, for every presentation of strings in the alphabet Σ of L." [Clark y Lappin 2011:6]

Una vez definido formalmente el mecanismo de aprendizaje IIL, Gold (1967) procede a demostrar qué clases de lenguajes formales pueden ser aprendidos ante cierto tipo de evidencia

(*strings*) presentada ante el aprendiz (*learner*). Gold propone dos versiones (modelos) posibles de aprendizaje: el IIL con evidencia positiva únicamente -caso base explicado en la cita anterior- y el IIL con evidencia positiva y evidencia negativa:

> "In Gold's negative evidence (informant) model, a presentation of a language *L* contains the full set of strings Σ^* generated by the alphabet Σ of *L*, and each string is labeled for membership either in *L*, or in its complement *L'*.Therefore, the learner has access to negative evidence for all non-strings of L in Σ^*" [Clark y Lappin 2011:8]

A continuación, Gold demuestra qué clases de lenguajes formales pueden ser aprendidas en uno y otro modelo. Dichos resultados se conocen en la bibliografía especializada como el *Teorema de Gold* y consisten en las siguientes conclusiones:

1) En el modelo IIL de evidencia positiva únicamente puede ser aprendida la totalidad de la clase infinita de lenguajes finitos (un subconjunto de los lenguajes regulares tipo 3) y un subconjunto finito de lenguajes recursivos.

2) En el modelo IIL de evidencia positiva y evidencia negativa puede ser aprendida la totalidad de la clase infinita de lenguajes recursivos, que se compone de un conjunto de las gramáticas dependientes del contexto (tipo 1), la totalidad de las gramáticas independientes de contexto CFG (tipo 2) y la totalidad de los lenguajes regulares (tipo3).

Las clases *aprendibles* (*learnable*) de Gold son sucesivamente inclusivas y se mapean con cada una de las clases de la jerarquía de lenguajes formales de Chomsky (Moreno Sandoval 2001) en función del poder expresivo de sus reglas de producción:

> "Gold's yardstick for measuring which models could learn which classes of languages was the following sequence of mathematically natural classes of languages, which are ordered by the subset relation: (A1) *Gold's classes of languages*: Finite $\subset$ Superfinite $\subset$ Regular $\subset$ Context-free $\subset$ Context-sensitive $\subset$ Primitive Recursive $\subset$ Recursive $\subset$ Recursively Enumerable" [Johnson 2004:588]

Tipo	Lenguaje (Gramática)	Restricciones a la reglas de producción (reglas de rescritura)	Implementación	Limitación en el dominio del lenguaje natural
Tipo-0	Recursivamente enumerable (gramática irrestricta)	Ninguna restricción a ambas partes de la regla. $\alpha \rightarrow \beta$	Máquina de Turing	Demasiado costosas computacionalmente (Grishman 1986) y sobregeneran
Tipo-1	Dependiente de Contexto o propiamente recursivo	La parte derecha contiene como mínimo los símbolos de la parte izquierda en cuanto a longitud $\alpha A\beta \rightarrow \alpha\gamma\beta$ $A \rightarrow \gamma / \alpha\_\beta$	Autómata linealmente finito	Adecuadas en fenómenos fonológicos pero sobregeneran en sintaxis.
Tipo-2	Independiente de Contexto CFG	La parte izquierda sólo puede tener un símbolo no terminal. La parte derecha es irrestricta. $A \rightarrow AB$ $A \rightarrow \gamma$ $A \rightarrow aB$	Autómata push-down	Concordancias externa e interna y subcategorizacion verbal con n-tuplicación de reglas. Problemas en constituyentes discontinuos y construcciones trabadas.
Tipo-3	Regular	La parte derecha sólo puede tener un símbolo terminal o *empty string* o terminal y no terminal $A \rightarrow aB$ $A \rightarrow \gamma$ $A \rightarrow \varepsilon$	Autómata de estados finitos	Cláusulas embebidas. No hay noción de constituyentes sintácticos.

Tabla 1: Jerarquía de lenguajes formales de Chomsky, adaptada de Moreno Sandoval (2001)

El mapeo entre las clases *aprendibles* de Gold y la jerarquía de lenguajes formales de Chomsky resulta de vital importancia para entender la fascinación que ejerció el *Teorema de Gold* entre los (neo)racionalistas chomskyanos. Parecería evidente que los PLD no continenen evidencia negativa (al menos, no directa), por lo que un supuesto aprendiz de un lenguaje natural (el niño), ante la falta de disponibilidad de una GU, se encontraría, en términos de Gold, en el modelo de IIL de evidencia positiva únicamente. En conclusión, el adquiriente de una lengua sólo estaría formalmente capacitado para aprender bajo el "supuesto empirista", cualquier gramática de la clase de lenguajes finitos y sólo algunas (no todas) gramáticas de otra clase de lenguajes más expresiva. A partir de Shieber (1985) se ha demostrado que los lenguajes naturales están formalmente ubicados en la jerarquía de lenguajes formales de Chomsky más allá de los lenguajes independientes de contexto (CFG tipo 2), en una clase intermedia entre los lenguajes de tipo 2 y los de tipo 1, denominada *lenguajes medianamente sensibles al contexto* (*Mildly Context-Sensitive Grammars* MCSG) (Balari *et al*. 2008). En conclusión, el modelo IIL de Gold de evidencia positiva únicamente parecería demostrar la necesidad de postular una GU innata para dar cuenta de la posibilidad de aprender cualquier lenguaje natural ante PLD con evidencia positiva únicamente. Este argumento, concomitante con el APS, que busca demostrar la necesidad de una GU ante los PLD, se conoce en la bibliografía especializada como el *Problema Lógico de la Adquisición del Lenguaje* (*Logical Problem of Language Acquisition* LPLA) (Cowie 1999; Johnson 2004):

> "A common route from LPLA to rationalism goes as follows. If there is no negative information, then there must be some other mechanism that enables the child to learn her language instead of a more expressive language. Such a mechanism would most plausibly be a cognitive ability that somehow prevents the child from entering a situation where negative evidence is needed. Any such cognitive ability would appear to be domain-specific to language and not learned. Thus, the ability must be innate, so rationalists are right about language acquisition and empiricism is false."[Johnson 2004:572]

Bajo esta perspectiva del *Teorema de Gold*, la GU cumple entonces el rol de acotar en forma innata el espacio de hipótesis posibles para una gramática particular ante los PLD particulares de una lengua. A esto se lo conoce como *sesgo innato fuerte* o de dominio específico (Clark y Lappin 2011, 2013), un mecanismo de gran riqueza estructural que especifica los tipos de gramáticas particulares posibles:

> "Universal grammar consists of (i) a mechanism to generate a search space for all candidate mental grammars and (ii) a learning procedure that specifies how to evaluate the sample sentences. Universal grammar is not learned but is required for language learning. It is innate." [Nowak *et al*.2001:114-115]

1.4 El Teorema de Gold revisitado

Las implicancias del Teorema de Gold como aval para la postura innatista derivan de una interpretación falaz y errada del paradigma IIL (Johnson 2004; Clark y Lappin 2011, 2013):

> "Despite it's impressive impact in cognitive science, Gold's Theorem is frequently misinterpreted. All of the authors listed above, for instance, have made false -and in some cases wildly inaccurate- claims about the theorem. Indeed, even rationalists, who might welcome support from the theorem, have made incorrect criticisms of the general assumptions that drive it (Chomsky 1986). The widespread

> confusion about the theorem is especially surprising, since even those who have misunderstood it have claimed that its proof «is quite easy to grasp intuitively».” [Johnson 2004:572]

Las incorrectas derivaciones del *Teorema de Gold* han provenido de diversos campos: desde la neurolingüística (Deacon 1997), desde la psicolingüística (Cowie 1999) y desde la lingüística chomskyana (Chomsky 1975, 1986; Nowak *et al*. 2001).

En primer lugar, resulta fundamental remarcar que Gold (1967) especifica las clases de lenguajes formales que pueden ser aprendidos ante cierta evidencia, no los lenguajes formales particulares que pueden ser aprendidos. La confusión entre el concepto de *complejidad de la clase -complexity between languages*, en términos de Johnson (2004)- y complejidad estructural *dentro de un lenguaje particular -complexity within languages*, en términos de Johnson (2004)- es el craso error en el que incurre Deacon (1997).

En segundo término, desde el punto de vista de la psicolingüística, la aplicabilidad del *Teorema de Gold* –el cual, recordemos, es una teoría del aprendizaje computacional de lenguajes formales- a la aprendibilidad del lenguaje natural (Cowie 1999) debe ser cuestionada con respecto al riguroso requisito del aprendizaje continuo en cada instancia de presentación de sucesivos *strings* de entrada. La propia Cowie (1999) reconoce que el adquiriente de un lenguaje natural no es un aprendiz del estilo de Gold, en el sentido de que un humano no dispone de la capacidad de memoria suficiente para testear cualquier hipótesis válida de gramática contra la presentación de los datos en cada instancia. Esta refutación se basa en los requerimientos que detalla Pinker (1979) como condiciones de aprendibilidad de lenguajes naturales, desgranando así *aprendibilidad* (*learnability*) en sendas interpretaciones como los conceptos de *identificabilidad en el límite* (*identifiability*) para el *Teorema de Gold* y de *asequibilidad* (*aquirability*) para el campo de la psicolingüística (Johnson 2004).

Desde el foro de la lingüística chomskyana (Chomsky 1975, 1986; Nowak *et al*. 2001), la lectura que se hace del *Teorema de Gold* resulta falaz, en cuanto a que demostrar la no *aprendibilidad* del lenguaje natural ante los PLD no es prueba suficiente para invocar la existencia de la GU:

> “Supongamos, por ejemplo, que puede demostrarse que una teoría del aprendizaje particular tiene la siguiente propiedad: un sistema […] puede aproximarse en su límite a cualquier mecanismo de estados finitos que produzca cadenas de izquierda a derecha a medida que pasa de estado a estado, pero nada más que este mecanismo. Puesto que es bien conocido que éste no puede representar ni siquiera la sintaxis de sistemas extremadamente simples (por ejemplo el cálculo proposicional) y con seguridad no los de la sintaxis de la lengua, podemos llegar a la conclusión de que la teoría es inadecuada para explicar el aprendizaje lingüístico.” [Chomsky 1975:198-199]

> “Parametric theories of UG encounter the same complexity issues that other learning models do. Assuming that the hypothesis space of possible grammars is finite does not address the learnabilty issue. In fact, the proofs of the major negative complexity of learning results proceed by defining a series of finitely parameterised sets of grammars, and demonstrating that they are difficult to learn. Therefore, Principles and Parameters (P&P) based models do not solve the complexity problem at the core of the language acquisition task. Some finite hypothesis spaces are efficiently learnable, while others are not. The view that UG consists of a rich set of innate, language specific learning biases that render acquisition tractable contributes nothing of substance to resolving the learning complexity

> problem, unless a detailed learning model is specified for which efficient learning can be shown. To date, no such model has been offered." [Clark y Lappin 2011:18]

Pero aun con toda la fascinación que ejerció Gold (1967) sobre los lingüistas partidarios del innatismo, la principal crítica a la supuesta imposibilidad de aprender un lenguaje natural a partir del modelo *IIL de evidencia positiva únicamente* provino desde el mismo paradigma estadístico y la teoría de la probabilidad. A partir de la denominada *revolución bayesiana* en lingüística computacional (Manning y Schütze 1999; Clark 2002), 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*). De este modo, aparece la noción de *evidencia negativa indirecta* en los PLD, y así, ya no estarían los adquirientes de una lengua alcanzados por el modelo de *IIL de evidencia positiva únicamente*, sino por el modelo de *IIL de evidencia positiva y negativa*, pudiendo aprender entonces cualquier lenguaje formal recursivo (Gold 1967) –un subconjunto de los lenguajes tipo 1 suficiente para abarcar todos los lenguajes naturales (Shieber 1985).

> "This sort of data has traditionally been called «Indirect Negative Evidence». The most natural way to formalise the concept of indirect negative evidence is with probability theory. Under reasonable assumptions, which we discuss below, we can infer from the non-occurrence of a particular sentence in the data that the probability of its being grammatical is very low. It may be that the reason that we have not seen a given example is that we have just been unlucky. The string could actually have quite high probability, but by chance we have not seen it. In fact, it is easy to prove that the likelihood of this situation decreases very rapidly to insignificance." [Clark y Lappin 2011:12]

Los propios Clark y Lappin (Clark 2002; Clark y Lappin 2011) objetan la equivalencia simplista entre probabilidad y gramaticalidad (Norvig 2011), ofreciendo en cambio un incipiente formalismo denominado *Distributional Lattice Grammars* (DLG) para trabajar con evidencia negativa indirecta a partir del recurso del *informante* (*informant*).

De un modo u otro, está claro que la trascendencia del *Teorema de Gold*, desde una mera teoría de la aprendibilidad computacional de lenguajes formales hacia el campo psicolingüístico de la adquisición del lenguaje natural, no está exenta de polémicas interpretaciones. En última instancia, como afirma el propio Clark (Clark 2002; Clark y Lappin 2011), la mejor forma de refutar el APS y la necesidad de una GU es demostrar empíricamente que en los PLD existe cierta riqueza estructural suficiente para inducir una gramática por medio de mecanismos generales (no de dominio específico) de aprendizaje no supervisado:

> "In fact, recent experimental research in unsupervised learning […] indicates that it is possible to achieve accuracy approaching the level of supervised systems. Of course, these results do not show that human language acquisition actually employs these unsupervised algorithms. However, they do provide initial evidence suggesting that weak bias learning methods may well be sufficient to account for language learning. If this is the case, then positing strong biases, rich learning priors, and language specific learning mechanisms requires substantial psychological or neural developmental motivation. The APS does not, in itself, support these devices." [Clark y Lappin 2011:29]

Capítulo 2. La modelización de sintaxis como procesos en cascada

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

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

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

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

Tabla 2: Teorías de adquisición del lenguaje enmarcadas en el innatismo y en el empirismo, adaptado de Clark (2002)

Pese a que se proponen confrontar con el APS -refutación argumentativa que se conoce como desafío (*challenging*) en la bibliografía especializada (Johnson 2004)-, estos trabajos enmarcados en el paradigma estadístico abordan el problema desde la misma perspectiva inicial que el paradigma simbólico: la sintaxis como punto de partida para la adquisición del lenguaje y el isomorfismo entre lenguajes formales y lenguajes naturales (Chomsky 1957, Clark 2002). Así pues, la modelización de la adquisición ontogenética de sintaxis se presenta como un proceso en cascada que toma como punto de partida un corpus de lenguaje escrito cuantitativa y cualitativamente homologable a los PLD (Pullum 1996; Clark 2002) -véase el capítulo 5 de esta tesis para una explicación más detallada acerca del diseño de un corpus para inducción de fenómenos sintácticos.

Algunos trabajos que se focalizan sobre el proceso de categorización de palabras toman en cuenta los indicios fonológicos en su modelización (Popova 1973; Levy 1985; Kelly 1992). En tales casos, será imprescindible que los datos lingüísticos del corpus de entrada al proceso contemplen la especificidad de la oralidad. Si bien dichos trabajos aportan una considerable relevancia al problema de la categorización de palabras, adolecen de un problema insalvable: sus respectivas hipótesis no fueron testeadas en un proceso en cascada para la adquisición integral de sintaxis. En cambio, debido a la naturaleza de la información distribucional que actúa como fuente de información primaria para estos modelos, los trabajos más abarcativos, como los de Clark (2002) y Klein y Manning (2004), optan por experimentar con corpora escritos, asumiendo la habilidad tremprana de procesamiento fonológico y segmentación de palabras y frases que se dan en los niños **en forma previa a la categorización de palabras**, según la abrumadora evidencia proveniente de la psicolingüística (Mehler *et al*. 1998; Jusczyk *et al.* 1999):

> "Taken together, these results (and many others) suggest that when they reach the end of their first year of life, babies have acquired most of the phonology of their mother tongue. In addition, it seems that phonology is acquired before the lexicon contains many items, and in fact helps lexical acquisition (for instance, both phonotactics and typical word pattern may help segmenting sentences into words), rather than the converse, whereby phonology would be acquired by considering a number of lexical items." [Mehler *et al*. 1998:63]

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

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

Clark (2002) recurre a diversas técnicas estadísticas para dar con un algoritmo de aprendizaje general no supervisado de inducción de sintaxis, particularmente una gramática probabilística independiente de contexto (*Probabilistic Context-Free Grammar* PCFG, también conocida como *Stochastic Context-Free Grammar* SCFG) (Roark y Sproat 2007) 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 2002:3]

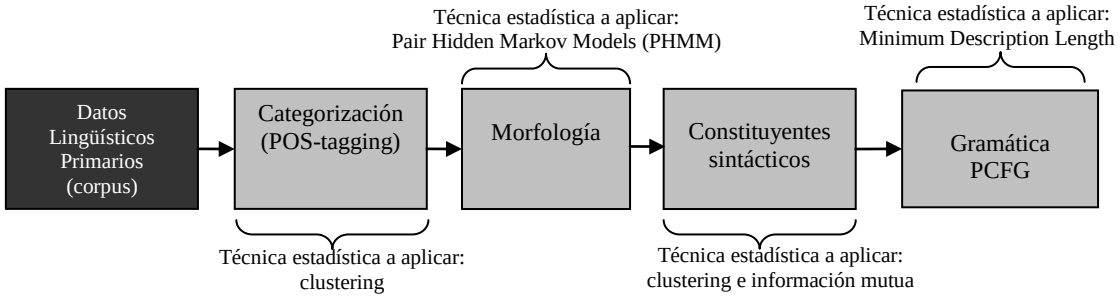

Figura 1: Esquema de algoritmo general de inducción de sintaxis en Clark (2002)

Se ha demostrado que las PCFG tienen mayor poder expresivo que las CFG (Infante-López y de Rijke 2006). A la vez, las PCFG pueden ser aprendidas a patir del modelo IIL de evidencia positiva únicamente (Horning 1969; Manning y Schütze 1999), por lo que serían un candidato plausible para la salida esperada del proceso en cascada, desde el punto de vista de la expresividad de los lenguajes naturales, los cuales se ubican más allá de los lenguajes formales de tipo 1 (CFG) (Shieber 1985), y desde el punto de vista del *Teorema de Gold* (Gold 1967). No obstante, las PCFG también presentan algunos escollos para convertirse en modelos del lenguaje natural:

> "In practice, a PCFG is a worse language model for English than an *n*-gram model (for $n > 1$). An *n*-gram model takes some local lexical context into account, while a PCFG uses none. PCFGs are not good models by themselves, but we could hope to combine the strengths of a PCFG and a trigram model. An early experiment that conditions the rules of a PCFG by word trigrams (and some additional context sensitive knowledge of the tree)." [Manning y Schütze 1999:387]

Por su parte, en Klein y Manning (2004) la salida esperada no es una PCFG sino un *parser de dependencias* (Mel'čuk 1988) a partir de un algoritmo EM (*Expectation Maximization*) (Manning y Schütze 1999) para generar *árboles sintácticos lexicalizados* (*lexicalized trees*). Ellos denominan a su modelo *Modelo de Dependencia con Valencia* (*Dependency Model with Valence* DMV). A pesar de esta importante diferencia entre la salida de Clark (2002) y la de Klein y Manning (2004), en términos de una sintaxis de estructura de frase o de una sintaxis de dependencia, respectivamente, ambos trabajos coinciden en cuanto a una categorización de palabras inducida a partir de técnicas de clustering como punto de partida del algoritmo. Clark (2002) trabaja sobre el *British National Corpus* (BNC) mientras que Klein y Manning (2004) lo hacen sobre una adaptación no POS-etiquetada del *Penn treebank*. Otra divergencia importante entre ambos trabajos es que Klein y Manning (2004) no incorporan inducción o procesamiento diferenciado de información morfológica, mientras que en Clark (2002) la morfología es una etapa a ser inducida inmediatamente después de la inducción de constituyentes sintácticos (véase *Figura 1*) –no obstante, en un experimento posterior, Clark (2003) intercambiará las ubicaciones de las etapas de categorización de palabras y de morfología en el algoritmo, con resultados que serán discuridos en el capítulo 5 de esta tesis.

Con menor trascendencia en la comunidad científica del campo de inducción de gramáticas (*grammar inference*), y sin recurrir a técnicas de clustering, Solan *et al*. (2005) aportan un algoritmo de inducción general de sintaxis denominado ADIOS (*Automatic Distillation Of Structure*), basado en patrones de ocurrencia de palabras, que crea reglas de rescritura del tipo *Context-Free Grammar* (CFG) y hasta del tipo *Context-Sensitive Grammar* (CSG). Ellos proponen su algoritmo para cualquier dominio en donde se trabaje con secuencias simbólicas que presenten estructura jerárquica:

> "We address the problem, fundamental to linguistics, bioinformatics, and certain other disciplines, of using corpora of raw symbolic sequential data to infer underlying rules that govern their production. Given a corpus of strings (such as text, transcribed speech, chromosome or protein sequence data, sheet music, etc.), our unsupervised algorithm recursively distills from it hierarchically structured patterns." [Solan *et al*. 2005:11629]

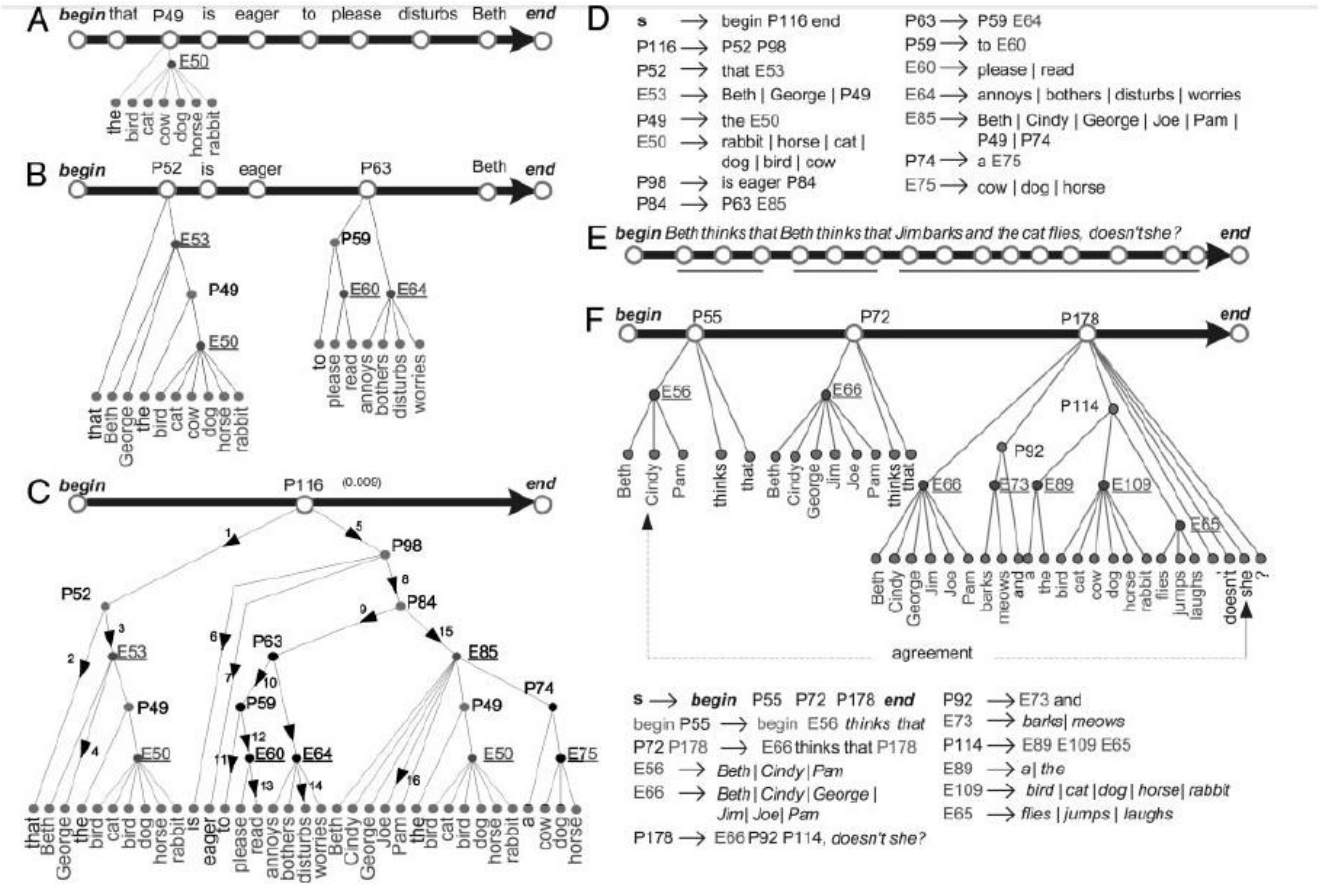

Figura 2: Ejemplo de abstracción progresiva de *patrones* (*patterns*) en Solan *et al*. (2005)

Si bien Solan *et al*. (2005) sostienen que podría ser "instructivo considerar su algoritmo" a la luz de la evidencia psicolingüística de la ontogénesis del lenguaje (Solan *et al*. 2005:11633), la no contemplación de categorías de palabras como pre-requisito para sus propias reglas sintácticas le resta cierta plausibilidad al enfoque ante el problema de la adquisición del lenguaje, aun cuando los autores sostengan lo contrario:

> "[…] the ADIOS representation is inherently safer than grammars that posit globally valid categories (such as «parts of speech» in a natural language)." [Solan et al. 2005:11630]

Como vemos, el enfoque predominante entre los trabajos del paradigma estadístico de la lingüística computacional de la última década, que se abocaron a la inducción integral de una gramática formal a partir de corpora no etiquetado, recurre a las técnicas de clustering como mecanismo de aprendizaje general no supervisado, iterativo y convergente (Christodoulopoulos *et al*. 2010), y a la categorización de palabras como punto de partida del algoritmo (Clark 2002; Klein y Manning 2004). No obstante, en el paradigma conexionista del campo de la lingüística computacional ya habían aparecido en la década del '90 algunos intentos que demostraron limitados resultados a escala. Elman (1991) utiliza una red neuronal de recurrencia simple (*Simple Recurrent Network* SRN) para modelizar reglas de reescritura del tipo CFG. Estas reglas

implican la categorización morfosintáctica de palabras en el mismo aprendizaje de constituyentes y no como paso previo (pre-requisito) de un proceso en cascada. Al no considerar la categorización de palabras como una tarea lingüística temprana previa a la adquisición de reglas sintácticas (Christophe *et al.* 2008), este tipo de experimentos no ofrecen demasiada plausibilidad psicolingüística. El experimento de Elman (1991), muy acotado en sus alcances, abarcaba un lexicón de apenas 22 palabras representadas por vectores de 26 dimensiones (bits). Los resultados muestran un aprendizaje de fenómenos sintácticos como la concordancia, la subcategorización verbal y muy simples instancias de recursividad en subordinadas. Resulta menester aclarar que los trabajos de inducción de gramáticas desde el paradigma conexionista (redes neuronales) han sido escasos y sus resultados, limitados (Reali *et al*. 2003):

> "Both the SRN and Kohonen network have two limitations:
> 1. It has not yet been possible scale up from very small artificial data sets to deal with real linguistic data. For example, in SRNs learning becomes extremely inefficient and slow, if it occurs at all, as vocabulary increases and the language become more complex, since prediction becomes more difficult (Chater & Conkey, 1993).
> 2. The linguistic categories are implicit within these networks, and can only be revealed using a subsequent cluster analysis. Thus, a significant amount of the computational work in approximating syntactic categories is not performed by the network itself." [Redington *et al*. 1998:434]

En español no existen esfuerzos similares de inducción integral de sintaxis desde el paradigma estadístico. Entre los escasos enfoques aplicados a este idioma, uno de los primeros trabajos de inducción de gramáticas es el algoritmo de inducción de Juárez Gambino y Calvo (2007). Basándose en la noción de *sustituibilidad* de Harris (1954) 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 de español escrito (CAST-3LB). No obstante, resulta imperioso destacar que en este caso no se partió de la tarea de categorización de palabras sino de la premisa de información distribucional por sustituibilidad de contextos oracionales, lo cual leresta cierta plausibilidad psicolingúística al experimento. Es más, los propios autores reconocen que el agregado de información morfológica eleva sus propias métricas, por lo que cabría esperar que dicha información morfológica provenga de la etapa de categorización de palabras, la cual está ausente en sus experimentos.

Más recientemente, Graça *et al*. (2011) proponen un muy interesante experimento de inducción de categorías sintácticas a partir del mismo corpus de español no POS-etiquetado (CAST-3LB) mediante un modelo markoviano (*Hidden Markov Model* HMM) enriquecido con parametrización y morfología. Este modelo sólo se enfoca en la etapa de categorización de palabras con evidencia translingüística en búlgaro, danés, portugués, español e inglés. Aun cuando los modelos markovianos escapan del alcance de las técnicas de clustering tradicional objeto de esta tesis, estudiaremos en detalle estos trabajos en el capítulo 5 para evaluar sus resultados.

Justamente, en tanto el campo de inducción de gramáticas (*grammar inference*) trabaja mayormente con enfoques de *aprendizaje de máquina* (*machine learning*) no supervisados, la necesidad de corroborar las hipótesis propuestas con evidencia translingüística se torna imperiosa (Clark 2002). Por ejemplo, Klein y Manning (2004) probaron su modelo DMV en chino y en alemán, además del inglés, con resultados aceptables. Otros trabajos hablan del concepto de *portabilidad* (*portability*) de los algoritmos. Por ejemplo, Christodoulopoulos *et al*. (2010) comparan varias métricas de trabajos canónicos en técnicas de clustering para inducción de categorías sintácticas contra corpora del búlgaro, checo, estoniano, húngaro, rumano, esloveno, serbio e inglés. Así pues, la evidencia translingüística en este tipo de enfoques consolida la plausibilidad psicolingüística de las hipótesis.

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

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

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

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

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

> "A current debate is whether young children possess an abstract representation of functional categories (*e.g*., determiner, auxiliary and preposition) or whether the representation of functional categories is built gradually in an item-by-item fashion. Strong nativist views held that children are innately endowed with a set of grammatical categories including functional categories. They possess abstract knowledge of grammatical categories since the beginning and use that knowledge to learn their first language. Therefore, according to constructivist views, young children do not have abstract knowledge of grammatical categories initially. It is the burden of constructivists to explain how children transform the item-based representation to adult-like grammar." [Wang 2012:3-4]

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

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

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

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

La categorización se vuelve aún más crítica y desafiante en el caso de palabras funcionales, debido a la disimilitud de categorías codificadas en cada gramática particular y al rol fundamental que podrían llevar a cabo las palabras funcionales como facilitadores (*cues*) en la categorización de palabras de contenido (Wang 2012) durante la explosión léxica (*vocabulary*

spurt) (Dromi 1987) que manifiestan los niños alrededor de los 2 años de vida, proceso de facilitación conocido en modelos formales inductivos como *bootstrapping* (Christophe *et al.* 2008):

> "Syntactic categories are interesting in terms of language acquisition because they must be learned before acquiring syntactic structures. […] Because languages use different set of grammatical categories and lexicon is completely different between languages, for both nativist and constructivist views, there are still the problems of how children map words onto grammatical categories and what information in the input and environment is available for children to categorize words. In other words, for functional categories, the problems are how children assign learned word forms to the right categories and whether information in the input could support the task of word categorization. This task is unavoidable even if children are predisposed with notions of functional categories. There must be a learning mechanism that works on the input to categorize words into proto-categories or adult-like categories, which are necessary for acquisition of syntax. […]Therefore, the time when a child possesses the knowledge of syntactic categories would be the earliest possible point in development for his/her knowledge of syntax." [Wang 2012:4-5]

Si bien en las últimas dos décadas aparecieron bastantes trabajos sobre categorización de palabras (Redington *et al.* 1998; Clark 2002; Christodoulopoulos *et al*. 2010), sólo recientemente la adquisición de palabras funcionales ha sido reivindicada por muy pocos trabajos como pre-requisito para el desarrollo lexical temprano (Wang 2012), cuando toda la evidencia de producción de lenguaje parece indicar lo contrario. Mientras que las palabras funcionales no están presentes en la producción de lenguaje del niño antes de los 2 años, las palabras de contenido pueden aparecer en producción en el léxico infantil tan prematuramente como desde el año de edad (algunos pocos ítems léxicos) y ciertamente alrededor del año y medio (con medio centenar de ítems léxicos) (Fenson *et al.* 1993). Esta paradójica inversión lógica de un supuesto pre-requisito evidenciado posteriormente en el tiempo luego de los eductos cuya aparición supuestamente facilitaría, será resuelta en las explicaciones venideras del presente trabajo, a la luz de renovada evidencia de las diferencias ontogenéticas entre producción y comprensión de palabras funcionales en niños y de las diferencias entre adquisición de palabras funcionales como ítems léxicos y procesamiento de sus propiedades distribucionales como clase en un corpus (Elghamry 2004). Así pues, inspirándonos en Redington *et al*. (1998) y en Wang (2012), y retomando nuestra propia investigación (Balbachan y Dell'Era 2008), la hipótesis de este trabajo enmarcado en el paradigma estadístico de la lingüística computacional será demostrar la viabilidad de la identificación temprana de palabras funcionales a partir de indicios de información distribucional, como paso previo y necesario para la categorización de palabras durante la explosión léxica y, consecuentemente, para la adquisición de una rudimentaria sintaxis. La metodología incluirá la técnica de clustering no jerárquico K-means como mecanismo de aprendizaje general no supervisado sobre la información distribucional de un *corpus* que modelice los PLD.

De esta manera, la categorización de palabras se postula como una tarea lingüística temprana con cierta plausibilidad compatible con la evidencia psicolingüística de la ontogénesis del lenguaje. Esta hipótesis central no sólo se relaciona estrechamente con otros trabajos del campo de la psicolingüística que intentan dar un marco teórico para la adquisición temprana del

léxico y los rudimentos de sintaxis, como Christophe *et al.* (2008), sino que también podría arrojar nueva evidencia de plausibilidad sobre modelos teóricos explicativos del lenguaje como la Gramática Chomskyana o sobre ambiciosos trabajos de modelización computacional del complejo proceso de adquisición del lenguaje, como Clark (2002), investigaciones en las cuales el rol temprano que cumplen las palabras funcionales resulta también crítico:

> "In this paper, we focus on the very beginning of language acquisition, and consider processes that may happen during the first two years of life. More specifically, we discuss how infants may start building a lexicon, and how they may start acquiring rudiments of syntax, that is, building the skeleton of a syntactic tree. In particular, we examine the role of two sources of information to which very young infants may plausibly have access: phrasal prosody and function words." [Christophe *et al.* 2008:62]

> "Por otro lado, se cree que las propiedades del léxico (y, en particular, las propiedades de las categorías funcionales) constituyen el *locus* fundamental de la variación sintáctica entre las lenguas." [Eguren y Fernández Soriano 2004:108]

> "I will just say that while the morphology component seems to be very suitable for this sort of model, the category induction algorithm has a serious problem as it stands. The first categories that the algorithm acquires are the closed-class categories – they really leap out of the data, and the open class categories such as noun and so on are acquired later. This is in direct contrast to the order in which children actually produce the words: first, they go through a phase of so-called telegraphic speech, where there speech is composed almost entirely of open-class content words, and only later do they fill in the closed-class function words." [Clark 2002:154-155]

2.3 Palabras funcionales vs. palabras de contenido: una distinción operativa

La distinción entre palabras funcionales y palabras de contenido ha venido siendo marcada por diversos andamiajes teóricos de la lingüística. Según las distintas parafernalias teóricas a las que se apele, existen algunas pequeñas divergencias al caracterizar a los miembros de tales grupos; por ejemplo, la concepción chomskyana de los adverbios como palabras funcionales fue duramente criticada por Jackendoff (1977).

No obstante, mayormente existe un acuerdo generalizado sobre las características generales de las palabras funcionales en contraposición con las palabras de contenido. Las palabras funcionales pertenecen a una clase cerrada de palabras que manifiesta escasos procesos diacrónicos evolutivos y nulos *procesos logogenéticos* (Balbachan 2006), a diferencia de las palabras de contenido, denominadas de clase abierta por su predisposición a incluir nuevos miembros (neologismos).

Desde un punto de vista fonético y fonológico, las palabras funcionales poseen propiedades distintivas en casi todos los idiomas (Shi 1995): menos sílabas y más cortas en duración, no portan acento tónico y suelen ocurrir en los límites de las frases fonológicas. En función de su minimalidad articulatoria (Wang 2012), suelen realizarse a través de morfemas flexivos o partículas enclíticas.

En términos de su probabilidad de ocurrencia, la frecuencia de ocurrencia de las palabras funcionales en cualquier muestreo es altísima (Zipf 1949). Prácticamente todos los miembros de la clase de palabras funcionales presentan una distribución uniforme (Manning y Schütze 1999),

independientemente del tipo de palabras funcional al que pertenezcan (preposición, pronombres, conjunciones, etc.) (Abney 2008).

Por otra parte, existe evidencia neurolingüística de un procesamiento diferenciado para las palabras funcionales (Díaz y McCarthy 2009). Finalmente, en cuanto a la perspectiva psicolingüística, las palabras funcionales también se caracterizan en contraposición a las palabras de contenido por su aparición tardía en el desarrollo ontogenético del lenguaje, especialmente desde el punto de vista de la producción:

> "Regarding the developmental timeline, content words and function words (as well as lexical and functional categories) follow two different paths. The first words produced by children are usually nouns and verbs for most languages. Some functional items (*e.g.*, the present progressive *–ing*, plural *–´s* and prepositions in English) do not appear in production until 2 to 2;6 (although in languages like Turkish grammatical morphemes occur much earlier than other languages), while the production of first recognizable word is around age one. After a slow start, children's vocabulary grows rapidly. They may reach a repertoire of 500-600 words by age of two, while their acquisition of functional words and morphemes follow a relatively fixed schedule and the rate of the inventory growth is in no way comparable to lexical words, although the functional items have much higher frequency than lexical items in the input." [Wang 2012:14]

Adam		Sarah	
I (2;3)		I (2;3)	
II (2;6)	Present progressive *in* *on*, plural	II (2;10)	Plural *in, on* Present progressive, past irregular Possessive
III (2;11)	Uncontractible copula, past irregular	III (3;1)	Uncontractible copula Articles
IV (3;2)	Articles Third person irregular, possessive	IV (3;8)	Third person regular
V(3;6)	Third person regular Past regular Uncontractible auxiliary Contractible copula Contractible auxiliary	V(4;0)	Past regular Uncontractible auxiliary Contractible copula Third person irregular Contractible auxiliary

Tabla 3: Estadios temporales para la adquisición de palabras funcionales del inglés en dos niños (Brown 1973)

Capítulo 3. Estado de la cuestión en categorización: modelos formales con motivación psicolingüística

3.1 La naturaleza de los indicios facilitadores

Las fuentes de información que se pueden considerar relevantes para actuar como facilitadores (*cues*) de la categorización de palabras (Redington *et al.* 1998; Wang 2012) son las siguientes:

- Indicios fonológicos y fonéticos: Kelly (1992) y otros estudios han propuesto que las regularidades entre la estructura fonológica-fonética de las palabras y su pertenencia a categorías sintácticas pueden resultar información útil para incorporar nuevos miembros a dichas clases de palabras durante el proceso de adquisición del lenguaje. Solo por mencionar un ejemplo, las palabras polisilábicas en inglés son predominantemente sustantivos (Kelly 1992). Se presentaron diversos estudios acerca de estos indicios fonológico-fonéticos para varios idiomas: Levy (1983) para el hebreo, Popova (1973) para el ruso, entre otros.
  Si bien esta fuente de información presentaría un aceptable potencial para ser utilizada en modelizaciones del proceso de adquisición del lenguaje (Lafferty y Mercer 1993), es menester reconocer que el proceso de *boostrapping* léxico no puede darse únicamente mediante este tipo de indicios (Redington *et al*. 1998).

- Información semántica: Esta fuente de información plantea que existe una correlación en la forma en que el niño percibe el mundo entre categorías semánticas primitivas (como objeto y acción) y categorías sintácticas (como sustantivo y verbo, respectivamente), de modo que la clasificación inicial de palabras se vería facilitada hasta alcanzar el estadio del lenguaje adulto (Pinker 1984). Este tipo de información para la categorización de palabras resulta difícil de ser evaluada cuantitativamente en experimentos, ya que no es evidente qué aspectos de la información extralingüística del mundo serían procesados para dicha facilitación (Redinton *et al.* 1998). Por otra parte, existe cierto consenso en que los set de clases de palabras morfosintácticas son dependientes de los lenguajes específicos (Nath *et al*. 2008):

  > "It is sometimes claimed that there is a well-defined set that is the same for all languages. This breaks down into two claims; first that the notion of the set of syntactic categories of a particular language is well-defined, and secondly that this set is identical for all human languages. Neither of these assumptions seem to be supported by the evidence." [Clark 2002:56]

  Así pues, existiría un *gap* explicativo para el mapeo de las categorías semánticas primitivas translingüísticas y las categorías sintácticas específicas de cada lenguaje (Clark 2002).

- Conocimiento innato: En este caso es central la idea de los universales lingüísticos sustantivos y formales de Chomsky (1965) como especificaciones innatas para la adquisición del lenguaje (*cf. sesgo fuerte* vs. *sesgo débil* o *strong bias* vs. *weak bias* en Clark y Lappin 2011). Cierta evidencia avala la posición del conocimiento innato; pero como postula la propia Lingüística Chomskyana, este conocimiento innato debe interactuar con los PLD. Por lo tanto, en última instancia, los datos lingüísticos siguen cumpliendo un rol importante en el proceso de adquisición del lenguaje. Así pues, incluso adscribiendo a la tesitura del innatismo, resulta todavía imperioso analizar qué información sería relevante para la interacción con el conocimiento innato, sea éste de dominio específico o general (Fodor 1983).

- Información distribucional: Se basa en el contexto léxico adyacente en que aparece una palabra. La pertinencia de esta fuente de indicios para abordar el problema de la adquisición del lenguaje reside en la observación de que las palabras de una misma categoría sintáctica presentarían cierta regularidad distribucional en la ocurrencia entre contextos inmediatos

(bigramas a derecha y a izquierda en mayor medida, trigramas, tetragramas, etc.). Debido a las críticas provenientes de enfoques innatistas (Pinker 1984), 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. Algunos trabajos en psicolingüística postularon que la posición **absoluta** de las palabras en las oraciones percibidas por los infantes podría influir como facilitación del tipo de categoría de palabra (Maratsos y Chalkley 1981), atendiendo al orden SVO estricto de la sintaxis del inglés. Obviamente, esto debe ser descartado a favor de estudios sobre la posición **relativa** (Schütze 1993) de la palabra *target* respecto de su contexto inmediato, considerando la existencia incontrovertible de lenguajes de orden libre de constituyentes como el español.

Los análisis distribucionales anteriores a la revolución chomskyana en lingüística pretendían relacionar ítems lingüísticos con su contexto mediante el concepto de *sustituibilidad* (Harris 1954). La investigación sobre la adquisición del lenguaje no formaba parte de esos emprendimientos, pues se consideraba al lenguaje como un *constructo cultural* ajeno al punto de vista psicológico o computacional (Redington *et al*. 1998). El legado de Chomsky (1957) dio paso a la crítica de las limitaciones de esos enfoques, y derivó, como efecto secundario, en el relegamiento del análisis distribucional:

> "Chomsky demuestra, probablemente de manera definitiva, que los conductistas y los estructuralistas norteamericanos de la primera mitad del siglo XX […] se equivocaron tanto en la identificación del objeto de estudio, como en los métodos que utilizaban. […] Por otra parte, en lo que a metodología respecta, resulta de todo punto inviable llegar a obtener por medio de la aplicación a un corpus de simples mecanismos de descubrimiento inductivos nociones lingüísticas como las de categoría gramatical o función sintáctica." [Eguren y Fernández Soriano 2004:20]

Sin embargo, hubo quienes mantuvieron viva la llama del análisis distribucional, con suerte dispar. Los avances teóricos del enfoque estadístico a mediados de los años '70 chocaron con la limitación práctica 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, podrían brindar resultados significativos (Manning y Schütze 1999).

3.2 Necesidad o no de facilitadores para la categorización en un lenguaje artificial (Mintz 2002)

Los modelos de categorización basados en información distribucional han recurrido tradicionalmente a algún tipo de indicio (*cue*) para iniciar el proceso de *bootstrapping*. En un trabajo pionero, Smith (1966) trabajó con adultos, apelando a una gramática artificial de cuatro categorías sintácticas (*gramática MN/PQ*) en oraciones de 2 tokens de extensión. Si bien el trabajo de Smith (1966) no arrojó evidencia de categorización basada en información distribucional, inspiró estudios posteriores con lenguajes artificiales en adultos como una forma de estudiar indirectamente la habilidad de categorización en niños durante el proceso de adquisición del lenguaje, habilidad presente tan prematuramente como desde los 18 meses (Shi *et al.* 1999) o incluso desde los 12 meses (Gómez y Maye 2005; Mintz 2006). En este sentido, uno

de los trabajos más interesantes en este tipo de experimentación de categorización con lenguajes artificiales es la investigación de Mintz (2002):

> "The goal of the present study is to take a step in addressing the role of distributional information in children's language acquisition by studying distributional mechanisms in adults. The results will show that adults, when hearing sentences in an artificial language, spontaneously form lexical categories that are based on the distributional regularities in the language. Although this finding does not directly address whether very young children categorize words in this way, it does demonstrate that the cognitive mechanisms needed to carry out such an analysis do occur in the species, and are naturally active during processing of artificial linguistic stimuli. Thus, this study provides a suggestive step forward in investigations of how very young children categorize words." [Mintz 2002:679] (*el subrayado es nuestro*)

Mintz (2002) comienza por preguntarse si la falta de evidencia de categorización distribucional en Smith (1966) puede deberse a la ausencia de indicios facilitadores (*cues*) convergentes. En efecto, trayendo a colación otros trabajos, Mintz (2002) demuestra que aparece evidencia de un análisis distribucional de categorización al enriquecer las palabras que componían las oraciones estímulos de la gramática MN/PQ con indicios convergentes de naturaleza semántica (rasgos masculino/femenino, Braine 1987) y de naturaleza morfológica (prefijos y sufijos con marcada saliencia, Frigo y McDonald 1998).

Mintz considera que el experimento de Smith (1966) adolece de un problema insalvable al considerar oraciones de sólo 2 palabras de extensión sin ningún indicio facilitador. Básicamente, sostiene que en ese escenario una palabra puede funcionar a la vez como *objetivo* (*target*) y como *contexto* (*environment*), dificultando un análisis distribucional efectivo en los aprendientes:

> "One reason that this might be so has to do with the nature of distributional information and distributional analyses. Specifically, a given word can be a word-to-be categorized (target word) while also being a word that functions as a categorizing environment. To perform an effective analysis, learners must track a target word with respect to all of its environments across sentences; likewise, learners must register a word as an environment for all the relevant target words across sentences. While logically it would be possible for an ideal learner to track words simultaneously as targets and environments, without some way to ground a subset of words (*e.g.*, in a category), it might be difficult for human learners to treat a word in a consistent way across sentences, and this might lead to difficulty in tracking the appropriate distributional contingencies. This may be particularly problematic with two word MN/PQ type sentences where there is no basis for making this distinction in distributional role." [Mintz 2002:684]

Por lo tanto, Mintz (2002) se propone recrear el experimento de Smith (1966) recurriendo a un corpus de lenguaje artificial de oraciones con extensión, esta vez, de tres palabras sin sentido (*jabberwocky*) con un paradigma de entrenamiento por sustitución de la palabra media (Harris 1954). Estas oraciones están despojadas de cualquier otro indicio facilitador semántico o morfológico. Contrariando las primeras observaciones de Smith (1966), Mintz (2002) encuentra ahora evidencia de análisis distribucional para categorización de palabras en estímulos sin indicios facilitadores convergentes.

Sin embargo, el propio Mintz reconoce las limitaciones de tales conclusiones al admitir que quizá resulten imprescindibles los indicios facilitadores (*cues*) convergentes, tomando en cuenta la complejidad de las oraciones estímulo de un lenguaje natural (y no una simple combinatoria de tres palabras artifíciales):

"Although in principle the same kinds of mechanisms in evidence here could induce category structure from more complex natural input […], the parameters needed to make them effective might be outside the range of normal human performance." [Mintz 2002:685]

Set A1

Training

Full Paradigm	Partial Paradigm	Alternate Paradigm	No Paradigm
bool nex jiv	sook nex runk	choon pux wug	fimp pux vot
bool kwob jiv	sook kwob runk	choon yult wug	plif daik fimp
bool zich jiv	sook zich runk	choon plif wug	pux ferd daik
bool pren jiv			vot plif ferd
zim nex noof			
zim kwob noof			
zim zich noof			
zim pren noof			
poz nex fen			
poz kwob fen			
poz zich fen			
poz pren fen			

Test

Category Conforming	Control	Repeated	Novel
sook pren runk	choon pren wug	bool pren jiv	daik vot plif
		zim nex noof	ferd fimp pux
		choon plif wug	noof fen poz
		fimp pux vot	jiv bool choon

Set B1

Training

Full Paradigm	Partial Paradigm	Alternate Paradigm	No Paradigm
Same as A1	choon nex wug	sook pux runk	Same as A1
	choon kwob wug	sook yult runk	
	choon zich wug	sook plif runk	

Test

Category Conforming	Control	Repeated	Novel
choon pren wug	sook pren runk	bool pren jiv	Same as A1
		zim nex noof	
		sook plif runk	
		fimp pux vot	

Tabla 4: Materiales de entrenamiento y evaluación para el experimento de Mintz (2002)

Es decir, los resultados de Mintz deben interpretarse como una demostración de la necesidad que evidencian los adquirientes de una lengua de disponer de indicios facilitadores convergentes al momento de encarar la tarea de categorización de palabras basada en la información distribucional. No obstante, estos indicios facilitadores (*cues*) no necesariamente deben estar representados por información temprana de naturaleza semántica o morfológica (postulados improbables de sostener, especialmente en el caso de las palabras funcionales). **El propio Mintz da en la tecla al sugerir que una distinción tajante entre palabras *target* y palabras de contexto (como un contraste fondo-figura) bien podría funcionar como facilitadora de la tarea**:

"In contrast, because of redundant distributional cues, the majority of stimuli that were used here provided a natural distinction: those words that made up a frame and those that occurred in the middle of a frame. This distinction might function like a figure–ground distinction to naturally lead learners to track the patterns of middle words in reference to frames (or frames in reference to middle words), thereby providing a grounding for the distributional analysis. Perhaps what was crucial about the converging cues in prior studies was that they selected a group of words as a target /environment reference point to start distributional learning, not necessarily that they directly (nondistributionally) categorized a set of words. If, here, the initial/final frames played a grounding role, it nevertheless is an

> open question whether natural language input incorporates functionally equivalent framing features to a significant degree." [Mintz 2002:685] (*el subrayado es nuestro*)

Justamente, veremos más adelante cómo esa distinción tajante entre objetivo-contexto bien puede ser mapeada a la separación entre palabras target y palabras cues, respectivamente, actuando estas últimas como los indicios facilitadores respecto de los cuales es posible categorizar las primeras (Balbachan y Dell'Era 2008; Christophe *et al.* 2008; Wang 2012).

3.3 La propuesta de los marcos frecuentes (Mintz 2003; Chemla et al. 2009)

Ya en el terreno de experimentación de los lenguajes naturales, Mintz (2003) encara la problemática de la categorización de palabras de contenido a partir de la información distribucional con una nueva teoría conocida como *marcos frecuentes* (*frequent frames*), basada en la tradición distribucionalista norteamericana (Bloomfield 1933) y en el concepto de *sustituibilidad* de Harris (1954). Este nuevo enfoque, que dominaría sus trabajos posteriores (Mintz 2003 para el inglés; Mintz 2006 y Chemla *et al.* 2009 para el francés), plantea una marcada diferencia en el foco computacional de la modelización respecto de los trabajos de clustering (véase capítulo 5 de esta tesis), también basados en información distribucional, contra los cuales argumenta:

> "Thus, in the frequent-frames approach, the important computational work involves identifying the frequent frames. Once identified, categorization is simply a matter of grouping together the words that intervene in a given frequent frame throughout a corpus. In contrast, in other approaches (Mintz et al., 2002; Redington et al., 1998) the crucial computations involved tracking the statistical profile of each of the most frequent words with respect to all the contexts in which it occurs, and comparing the profiles of each word with all the other words. Thus, an advantage of the frequent-frames categorization process is that, once a set of frequent frames has been identified, a single occurrence of an uncategorized word in a frequent frame would be sufficient for categorization. Moreover, it is computationally simpler, in that fewer total contexts are involved in analysing a corpus." [Chemla *et al.* 2009:397] (*el subrayado es nuestro*)

Justamente, apelando a esta mayor simplicidad de cómputo, Mintz (2003) considera que su teoría ofrece una mayor plausibilidad psicolíngüística para la evidencia ontogenética de la habilidad temprana de categorización de palabras de contenido en infantes, detectable a partir de los 12 a 15 meses (Mintz 2006; Gómez y Maye 2005).

No obstante, como el propio Mintz advierte (Chemla *et al.* 2009), **la simplicidad de la definición de los marcos es también su punto débil. La mera ocurrencia frecuente de diversas palabras target antes o después de las mismas palabras contexto no es garantía para categorizar dichas palabras *target* en un mismo grupo**. Esta excesiva generalización simplista no ocurre en las técnicas de clustering (Redington *et al.* 1998; Balbachan y Dell'Era 2008; Wang 2012), ya que en dichos experimentos **se computa todo el perfil de ocurrencias de una palabra target en función de cada combinatoria de palabras contexto (*cues*) a partir de un espacio vectorial n-dimensional**:

> "One challenge in forming categories from distributional cues is to establish an efficient balance between the detection of the especially informative contexts and the rejection of the potentially

> misleading ones. For example, in (1), that cat and mat both occur after the suggests that the two words belong to the same category. However, applying this very same reasoning to example (2) would lead one to conclude that large and mat belong to the same category […].
> (1) *the cat is on the mat*
> (2) *the large cat is on the mat*
> To address the problem of the variability of informative distributional contexts, the procedures developed by Redington et al . (1998) […] took into account the entire range of contexts a word occurred in, and essentially classified words based on their distributional profiles across entire corpora." [Chemla *et al*. 2009:397]

Desde un punto de vista formal, un *marco frecuente* es un trigrama o co-ocurrencia de una palabra target (la variable 'X' en el siguiente ejemplo) en el medio de dos palabras contexto (las variables 'A' y 'B' en el siguiente ejemplo):

[A X B] , [*you* X *the*] , ...

En Chemla *et al.* (2009), Mintz trabaja con tres experimentos sobre marcos frecuentes "oralizados", extraídos del *corpus* francés de CHILDES (*child directed speech*), reportando evidencia translingüística para sus hallazgos previos en inglés (Mintz 2003). Una gran ventaja de estos experimentos es que su diseño se centra en corpora con estímulos de lenguaje natural, a diferencia de Mintz (2002), y en palabras target acuñadas en dichos lenguajes, a diferencia de las palabras inventadas en Christophe *et al.* (2008). Chemla *et al*. (2009) reportan una precisión altísima para el experimento 1 de *marcos frecuentes continuos*:

Environments		Frames [A x B] (Expt 1)	
Scoring condition		Types	Tokens
French corpus	Accuracy	1.	1.
	(Baseline)	(.13)	(.13)
	Completeness	.33	.34
	(Baseline)	(.16)	(.16)
English corpus	Accuracy	.90	.95
	(Baseline)	(.18)	(.18)
	Completeness	.047	.057
	(Baseline)	(.031)	(.038)

Tabla 5: Precisión (*accuracy*) para el inglés y para el francés en el experimento 1 de Chemla *et al.* (2009)

En experimentos sucesivos, Mintz amplía el concepto de marco frecuente a marcos no necesariamente coincidentes con los límites de las frases fonológicas, como en el caso de [*you* X *the*] (correspondiente al primer árbol sintáctico del grupo (2) de la Figura 3). Esta no coincidencia entre la posible extensión de los marcos frecuentes "*a caballo*" de dos frases fonológicas resulta inadecuada, como veremos más adelante. Esta característica de los marcos frecuentes otorgaría, en términos de Mintz, la ventaja de la posibilidad de dar cuenta con aparente mayor facilidad de los constituyentes sintácticos discontinuos, en tanto resultan un contexto altamente restrictivo en la combinatoria de palabras target, aportando mucha mayor informatividad para la categorización:

> "Thus, the advantage of frames over the alternative contexts examined in this experiment could be explained by the types of syntactic structures that are likely to be involved. These results suggest that discontinuity is a crucial property of frequent frames for the purpose of categorization." [Chemla et al. 2009:403]

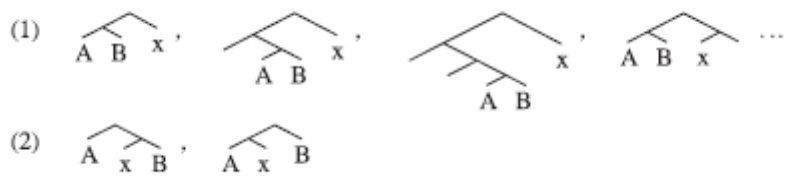

Figura 3: Estructuras sintácticas correspondientes a *marcos frecuentes* continuos (1) y marcos discontinuos (2) en Chemla *et al.* (2009)

Gómez y Maye (2005) reportan la habilidad que tienen los niños de 15 meses para detectar dependencias discontinuas. Sin embargo, estas habilidades han sido verificadas en discontinuidades de palabras, no de estructuras sintácticas. Goodwin (2013) demuestra que a medida que la discontinuidad se agranda en número de palabras, los niños dejan de evidenciar esta habilidad de relacionar las palabras target del marco discontinuo para estímulos que no hayan estado presentes durante el entrenamiento, lo cual sería una prueba de que el procesamiento por estructuras sintácticas ahormacionales (Chomsky 1957) todavía estaría ausente en esta etapa de la ontogénesis (18 meses de edad):

> "Gómez (2002) found that 18-month-old infants could learn nonadjacent dependencies in an artificial grammar (e.g., '*pel__rud*', where the underscore could be several disyllabic words, such as '*waddim'*), but only if there was high variability in the set of possible intervening words. As the set size of possible middle words increased from 2, to 6, to 12, to 24, children improved in their ability to discriminate between trained and untrained word strings (see also Gómez & Maye, 2005, for evidence of nonadjacent dependency learning at younger ages). Because this study did not require children to generalize the pattern to novel audio strings, it is not possible to determine if children learned an abstract rule or a set of specific patterns. However, this does not diminish the importance of the finding that infants were sensitive to the distributional properties of nonadjacent 'words'." [Goodwin 2013:17] (*el subrayado es nuestro*)

Apoyarse en las mayores restricciones combinatorias de un marco **sintáctico** discontinuo para explicar la mejor perfomance del grupo de testeo (experimento 2 de Chemla *et al.* 2009) por sobre el grupo sometido a estímulos de marcos continuos (experimento 1 de Chemla *et al.* 2009) resulta, como mínimo, sospechoso. De ser cierta tal explicación -recordemos que Mintz (2006) postulaba habilidades de categorización a los 12 meses de edad, es decir, 3 meses antes de la remota posibilidad de trabajar con constituyentes discontinuos (Gómez y Maye 2005)-, deberíamos aceptar que los niños enfrentados a la tarea de categorización manifiestan un sesgo de mejora de performance ante estímulos de marcos discontinuos incluso por sobre los estímulos de marcos continuos; pero las discontinuidades en constituyentes sintácticos sólo pueden ser formalizadas a través de reglas de reescritura de frase para gramáticas del tipo *Mildly Context-Sensitive Grammars* MCSG o gramáticas medianamente sensibles al contexto (Russell y Norvig 1995), más expresivas que las *Gramáticas Independientes de Contexto* CFG de tipo 2, según la jerarquía de lenguajes formales de Chomsky de 1957 (Moreno Sandoval 2001). **Es decir, estaríamos postulando que aun antes de que el niño disponga de una "protogramática", ya manifestaría una sensibilidad performativa superior hacia estímulos de constituyentes discontinuos, los cuales, en realidad, formalmente no debería poder distinguir de otros**

***marcos frecuentes* continuos**. Forzosamente, la explicación de tal efecto ha de ser otra para no incurrir en esta flagrante *petitio principii*.

En un tercer experimento (experimento 3 en Chemla *et al.* 2009), Mintz intenta corroborar si, como la intuición lingüística sugiere, el reemplazo de los ítems léxicos en los marcos frecuentes por las "protocategorías" sintácticas en las que van siendo agrupadas sus palabras contextos mejora la performance de categorización de la palabra target intermedia, lo que él denomina "aplicación recursiva del algoritmo":

> "The frequent-frames mechanism as investigated so far could yield initial category knowledge that could serve as a basis for detecting new frame-like contexts with the re-application of the categorization procedure. For instance, if the words *I* and *you* have been categorized together, it may be reasonable to consider them as equivalent in terms of their role in defining frames and to obtain a single group from the frames [*I x it*] and [*you x it*]. This would be a highly desirable outcome, as it could consolidate separate frame-based groups belonging to the same linguistic category (for example, the frames [*I x it*] and [*you x it*] both contain verbs), thus making frame-based categories even more informative linguistically." [Chemla *et al*. 2009:403]

Sin embargo, los resultados observados en este experimento contradicen dicha intución. Pese a que Mintz ensaya una débil justificación para dicha contrariedad, consideramos que éste es otro problema que debe ser subsanado para una mayor adecuación explicativa de la teoría:

> "Second, the recursive analysis presented in Experiment 3 shows that the distributional analysis is maximally efficient when the framing elements *A* and *B* are specific items rather than syntactic categories. Both these principles fit well within a psychologically plausible acquisition model. For instance, infants at the start of the acquisition process already have access to specific items but not yet to established categories. It is then an unexpected bonus that item-specificity leads to better categorization than an analysis in which the framing elements are syntactic categories, even perfect ones. […] After all, if grammars are organized around categories, shouldn't the category of the target word be predicted by the surrounding categories at least as well as by the surrounding words themselves?" [Chemla *et al*. 2009:404]

Finalmente, el último punto débil de la teoría de Mintz es descripto en Chemla *et al.* (2009). La teoría de los marcos frecuentes no ofrece un mecanismo plausible que permita dar cuenta de una estrategia de agrupamiento ulterior (*merging*) de palabras de idéntica categoría morfosintáctica que fueron categorizadas por distintos marcos frecuentes:

> "At this point, infants would possess frame-based categories, containing words that typically «behave the same», in that they belong to the same syntactic category. However, even when accuracy is high, there are typically several frame-based categories for each syntactic category. For instance, several different frames pick out nouns, and several others pick out verbs. Learners would thus need to merge frame-based categories to obtain more comprehensive categories. Several possible strategies can be used to that end, such as grouping together framebased categories that share one of their framing elements as well as some of their categorized words." [Chemla *et al.* 2009:404]

En realidad, Mintz (2006) bosqueja dos posibles estrategias para este agrupamiento ulterior de ítems léxicos (*merging*). Por un lado, propone una estrategia de mapeo entre los ítems léxicos categorizados (*derived categories*) y las categorías innatas mayores apriorísticas, en una manifiesta apelación al paradigma innatista chomskyano mediante un recurso de disponibilidad de información semántica temprana que no se condice con los requisitos iniciales de la teoría:

> "One solution assumes that part of children's innate linguistic knowledge is that there are categories such as noun, verb, and adjective. The problem then becomes one of labeling or associating distributionally derived categories with the innately specified system. Although it is questionable whether verb referents can be identified by learners without access to sentential structural information, […] the referents of concrete nouns have been argued to be recoverable from observations of the

> circumstances in which they are used […] If this is so, then the distributional category that contains nouns could be readily identified based on the concrete nouns that are its members. Note that using a semantic-to-syntactic generalization to label an independently derived category avoids the one-to-many mapping problem encountered when attempting to derive syntactic categories from semantic ones, since the semantic information is simply used to determine a general tendency of a group of words that is independently categorized." [Mintz 2006:25-26]

Alternativamente, propone otra estrategia de agrupamiento ulterior (*merging*), esta vez adscribiendo al paradigma empirista a partir de miembros prototípicos de categorías; aunque en palabras del propio Mintz, esta explicación no zanjaría la discusión sobre este punto:

> "There is an alternative to the view that the distributionally defined categories must be linked to syntactic labels and that infants have innate knowledge of syntactic categories. According to Tomasello (2000a; 2000b), children's early lexical categories are not abstract adult categories, like verb. Rather, initial categories are "item-based" and organized around the specific environments in which words occur. The present findings might appear to mesh well with this view: children's early grammar could be constructed around individual (or consolidated) frame-based categories, and only later would these categories take on the more abstract status as in the adult grammar. In that case, perhaps one need not posit that children have an innate verb category that must be associated with the relevant frequent frames. **But, eventually, children would have to assign words to the adult category, and something akin to the labeling procedures outlined above would be required. Thus, even if children's initial categories turn out to be more restricted than adults, the issue of how they are eventually integrated into an adult grammar remains**." [Mintz 2006:27] (*las negritas son nuestras*)

Estas falencias explicativas debilitan inexorablemente la teoría de los marcos frecuentes. A la vez, observaremos que la adecuación descriptiva de los modelos basados en técnicas de clustering resulta abrumadoramente mayor (véanse capítulos 5, 6 y 7 de esta tesis), sin contrariar la plausibilidad psicolingüística de los mecanismos generales (no de dominio específico) de aprendizaje que dichos modelos postulan como pre-requisitos.

Pese a estas críticas, reconocemos en los marcos frecuentes un germen de plausibilidad psicolingüúistica, en cuanto al rol de facilitadores que cumplen los vecinos de una palabra target en la categorización de la misma. A pesar del esfuerzo de Mintz (2003) en jerarquizar los marcos frecuentes con un mayor poder descriptivo que la **mera conjunción de bigramas** de la palabra target a izquierda y a derecha, esta simple premisa será explotada en todo su potencial estadístico a partir de los modelos basados en técnicas de clustering sobre la información distribucional de las palabras:

> "In the present approach the word '*W*' in the environment '... *X W Y* ...' is stored as «jointly following *X* and preceding *Y*», but such would not be the case if W occurred after *X* and before *Y* on independent occasions. In contrast, […] Redington *et a*l. (1998) studies use «bigram» contexts, which record only independent co-occurrence patterns (*e.g*. «*following X*», «*preceding Y*»). There are potentially important consequences of using frame contexts as opposed to bigram contexts. In particular, the property of joint co-occurrence in the frame contexts involves an additional relationship between the context elements themselves, as well as between context and target word. Hence, it is reasonable to assume a priori that if a given frame occurs *frequently* in a corpus of natural language, it is likely to be caused by some systematic aspect of the language, rather than by accident. Therefore, the target words that occur inside each instance of the frequent frame are likely to have some linguistically pertinent relationship, such as grammatical category membership. This is not a necessary outcome, but is an arguably likely consequence of using frequent frames as contexts. To capitalize on this possibility, not only are frames used as contexts in the present approach, but only words that occur in the most frequent frames are categorized (as opposed to […] Redington *et al*. (1998) studies in which target words were selected based on the frequency of the target words themselves)." [Mintz 2003:94]

3.4 Facilitación mediante frases fonológicas y tipos de palabras funcionales: teoría de los "protoconstituyentes" (Christophe et al. 2008)

Una última visión alternativa a la teoría de los marcos frecuentes para la categorización de palabras de contenido en modelos con motivación psicolingüística es el trabajo de Chistophe *et al.* (2008). Ellos estudian cómo los niños podrían empezar a adquirir un lexicón categorizado a partir de dos tipos de indicios facilitadores: las frases fonológicas (*phonological boostrapping*), cuyos límites actuarían como indicadores de rudimentarios *protoconstituyentes* sintácticos, y los tipos (*prototipos*) de palabras funcionales, que ayudarían a etiquetar dichos protoconstituyentes. Por ejemplo, en la Figura 4, una palabra de contenido o target ('*xxx*') podría ser agrupada en función de la palabra funcional adyacente que la acompaña en el protoconstituyente:

> "Finally, we propose that both infants and adults could perform a first-pass syntactic analysis of incoming speech by putting together these two pieces of information, function words and phrasal prosody: Prosodic boundaries would give syntactic constituent boundaries, while function words would help label these constituents." [Christophe *et al.* 2008:72]

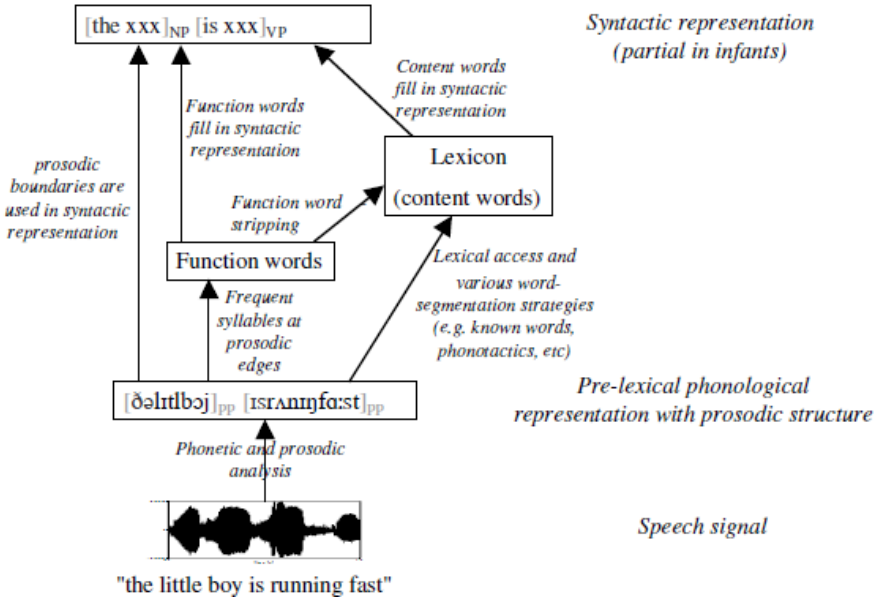

Figura 4: Modelo de *boostrapping* fonológico para adquisición del léxico y sintaxis a partir de palabras funcionales en Christophe *et al.* (2008)

El trabajo de Christophe *et al.* (2008) es, en algún punto, continuador de la motivación distribucional de la teoría de los marcos frecuentes de Mintz (2003), sólo que se propone, por un lado, refinar el concepto de marco frecuente, dotándolo de una entidad sintáctica más robusta, en función de los indicios de unidades fonológicas y, por otro lado, enriquecer la naturaleza de los "nuevos marcos" (*protoconstituyentes*) con información sintáctica rudimentaria, producto de los tipos de palabras funcionales que caracterizan dichos marcos (por ej. el determinante '*the*' etiquetaría un *protoconstituyente* NP, el verbo auxiliar '*is*' etiquetaría un protoconstituyente VP). No obstante, obsérvese que el concepto de *protoconstituyente* sintáctico como elemento facilitador para el proceso de categorización resulta incompatible con los marcos frecuentes (Mintz 2003, 2006; Chemla *et al.* 2009), ya que -recordemos- los marcos frecuentes no coincidían necesariamente con los límites sintácticos de las frases.

En lo sucesivo, adoptaremos el concepto de *protoconstituyente*, a sabiendas de estar alterando el espíritu del trabajo de Christophe *et al.* (2008), porque consideramos de crucial importancia este punto: la idea de que estos constituyentes sintácticos rudimentarios son facilitadores del proceso de categorización de palabras de contenido, según el modelo presentado. Como veremos más adelante, la condición de posibilidad de conformación de un constituyente sintáctico no debería ser lógicamente anterior a la existencia misma de las categoría de palabras morfosintácticas que precede la formación de reglas de reescritura (Clark 2002). De hecho, Clark (2002) y Balbachan y Dell'Era (2010) demuestran que los verdaderos constituyentes sintácticos pueden ser inducidos únicamente a partir de secuencias de clases de palabras morfosintáticas en un corpus de entrenamiento (véanse capítulos 7 y 8 de esta tesis), apelando a técnicas estadísticas de información mutua (*mutual information*). Así pues, utilizaremos el concepto de *protoconstituyentes* sintácticos en lugar de constituyentes sintácticos para lo concerniente al trabajo de Christophe *et al.* (2008).

El experimento principal de Christophe *et al.* (2008) consiste en reproducir en adultos las condiciones ontogenéticas de los niños inmediatas a la aparición de evidencia de categorización de palabras (alrededor de los 18 meses de vida), para categorizar palabras target sin sentido (*jabberwocky*):

> "To test the plausibility of this hypothesis, we presented adult participants with *jabberwocky* sentences, where function words and prosodic information were preserved, but all content words were replaced by nonwords […]. In that way, we simulated the situation of an 18-month-old infant, who may have access to prosodic boundaries and function words, but does not know many content words yet. We created two experimental conditions, one where the target word was immediately preceded by a function word that gave its category (determiner for nouns, pronoun for verbs), and another one where the target word was not immediately preceded by a function word, and a more complex analysis was needed." [Christophe *et al.* 2008:70]

Los investigadores reportan resultados que comprueban la hipótesis del modelo, convalidando el rol crucial que desempeñan las palabras funcionales para categorizar las palabras target adyacentes en un mismo *protoconstituyente* sintáctico. **Es decir, se verifica el papel de las frases fonológicas y de los tipos de palabras funcionales para el proceso de categorización de palabras de contenidos**.

Entonces, puesto que las palabras funcionales cumplen un rol crucial en el modelo propuesto por Christophe *et al.* (2008) para la categorización temprana de palabras de contenido (alrededor del año y medio de edad), resulta de vital importancia compatibilizar los requerimientos de la teoría con la evidencia empírica de la disponibilidad de dichas palabras funcionales. Como veremos más adelante, la ausencia de palabras funcionales en producción no es motivo suficiente para postular su no gravitación en los mecanismos léxicos de la comprensión (Clark 2002; Wang 2012). Christophe *et al.* (2008) explican la disponibilidad temprana de las palabras funcionales a partir de los indicios prosódico-fonológicos que caracterizan esta clase de palabras (Wang 2012) (véase sección *2.3 Palabras funcionales vs. palabras de contenido: una distinción operativa*):

> "A second crucial aspect of the model is the special role played by function words (e.g., determiners, auxiliaries, prepositions, etc.). They are represented within a special lexicon, that is built and accessed from the prelexical representation (paying special attention to prosodic edges) and that directly informs syntactic processing. Infants may be able to discover function words quite early in their acquisition of language because they are extremely frequent syllables that typically occur at prosodic edges (beginning or end depending on the language)." [Christophe *et al.* 2008:63]

Si bien los autores sacan a relucir evidencia empírica de que alrededor del año de vida los niños son capaces de distinguir palabras funcionales en contraposición a palabras de contenido, esta habilidad temprana no alcanzaría para satisfacer el requerimiento del modelo de poder identificar tipos de palabras funcionales, lo cual recién se alcanza en algunos idiomas como el alemán, en el mejor de los casos, a los 16 meses de edad y en otros, como el inglés, recién a los 18 meses:

> "In favor of this hypothesis, several experiments showed that infants around their first birthday already possess some knowledge of the function words of their language […].
> **Identifying a list of functional items would not be sufficient for infants to start doing even a rough syntactic analysis: To that end, infants would need, in addition, to identify categories of function words, such as determiners (signaling nouns) and pronouns (signaling verbs).** […]. In another recent experiment, Kedar, Casasola and Lust (2006) showed that 18- and 24-month-old American infants were better at identifying a known noun depending on whether it was preceded by a correct function word (*the*)
> or an inappropriate one (*and*, as in *"Look at and ball!"* […]).
> **These results suggest that infants within their second year of life are already figuring out what the categories of functional items are in their language.**" [Christophe *et al.* 2008:67] (*las negritas son nuestras*)

Entonces, bajo las premisas del acceso a los tipos de palabras funcionales básicos y de la facilitación prosódica de los límites de *protoconstituyentes* sintácticos, el modelo bien puede ofrecer una explicación plausible para el proceso de categorización de palabras de contenido que constituye la explosión léxica (*vocabulary spurt*) alrededor de los dos años de vida, pero difícilmente puede dar cuenta de la categorización temprana que se da a partir de los 18 meses (o incluso antes) (Shi *et al.* 1999). Es decir, la *teoría de los protoconstituyentes* -como hemos dado en llamar a la teoría de Christophe *et al.* (2008)- ofrece una exploración viable del mecanismo de *boostrapping* léxico una vez que se torna masivamente necesario para la adquisición de cientos de palabras en el vocabulario de los niños, pero los tiempos ontogenéticos para adquirir los tipos de palabras funcionales y para empezar a categorizar palabras de contenidos mediante los mismos se superponen, en el mejor de los casos (por ejemplo para el inglés en alrededor de los 18 meses de edad). El conflicto entre los requerimientos formales de esta teoría y la evidencia empírica se da en torno a la no disponibilidad consolidada de los tipos de palabras funcionales, aunque sí de la noción misma de palabra funcional (en contraposición a la de palabras de contenido). Si bien es lógicamente posible que apenas comienzan a delinearse los tipos de palabras funcionales, éstos sean utilizados como facilitadores de este modelo para la categorización de palabras de contenido, sería deseable que esta habilidad, en tanto pre-requisito, estuviera consolidada antes de ser necesaria. En realidad, como veremos en el capítulo 6, algunos trabajos (Valian 1986; Wang 2012) sostienen que la noción de tipos de palabras funcionales se evidencia antes de los dos años, tan tempranamente como a los 14 meses.

En todo caso, será menester tomar muy en cuenta dicha evidencia empírica en el diseño de nuestro experimento de modelización computacional, tal como se describe en el capítulo 7. *Nuestro experimento: inducción no supervisada de categorías morfosintácticas mediante clustering a partir de palabras funcionales sin tipología diferenciada*). Es decir, como veremos más adelante, podremos contar para nuestro experimento con la disponibilidad de la noción de palabras funcionales pero convenientemente declinaremos disponer de la tipología de la mismas (Redington *et al*. 1998; Elghamry 2004). De ese modo, sortearemos las polémicas alrededor de la posibilidad o no de que los niños dispongan de las nociones de tipos de palabras funcionales antes de evidenciar el proceso de categorización de palabras de contenido (antes de los 18 meses de edad). Adelantamos ahora que el concepto de palabra funcional será el de una palabra "bisagra", que articula la relación entre generalmente dos palabras de contenido o una palabra de contenido y el límite de la frase fonológica (véanse capítulos 6 y 7 de esta tesis), lo cual redunda en un contexto inmediato bastante predecible para una palabra de contenido a ser categorizada:

> "In order to gain some intuition regarding why distributional information is more useful for content words than for function words, consider the kinds of contexts in which each will appear. Content words will tend to have one of a small number of function words as their context. Although content words are typically much less frequent, their context is relatively predictable. Function words, on the other hand, are much more frequent, but will tend to have content words as their context. Because there are many more content words, the context of function words will be relatively amorphous. As the measure of similarity exploits regularities in the distribution of contexts, those words with predictable contexts will be clustered together much more accurately." [Redington *et al.* 1998:456]

Esta función "articulatoria" de las palabras funcionales respecto de las palabras de contenido puede ser asimilada al concepto fondo-figura (Mintz 2002) como plausible mecanismo cognitivo del aprendiente.

Con todo, consideramos que el gran mérito del trabajo de Christophe *et al.* (2008) es haber sido uno de los pioneros desde el campo de la psicolingüistica en reconocerles a las palabras funcionales un rol específico de vital importancia durante el proceso de adquisición temprana del léxico, pese a las escasas evidencias de utilización de dichas palabras en la producción infantil temprana.

Capítulo 4. Técnicas de clustering como mecanismo de aprendizaje general no supervisado

4.1 Representación de objetos en el espacio vectorial multidimensional

Si bien la noción algebraica de clustering como agrupamiento de objetos en un espacio de *n*-dimensiones se remonta bastante atrás en la historia de la ciencia -el concepto de *espacio vectorial*, por ejemplo, puede rastrearse en problemas matemáticos de fines del siglo XIX-, no fue sino hasta entrado el siglo XX que estas técnicas algebraicas se aplicaron a problemas concretos de clasificación o generalización, aunque muy lejanos del dominio del lenguaje natural -véase por ejemplo Cattell (1943). Con el (re)surgimiento del paradigma estadístico en lingüística computacional en la década del '90, a partir de la ampliación del poder de cómputo, las técnicas de clustering atrajeron la atención de los investigadores en NLP.

La idea central es que los objetos se representan como vectores en múltiples dimensiones, de modo tal que la similitud entre los mismos puede ser computada a partir de las distancias respectivas en cada una de esas dimensiones. Por ejemplo, supongamos que queremos representar los siguientes vehículos (*camión*, *automóvil*, *motocicleta*, *cuatriciclo*, *triciclo* y *bicicleta*) en función de tres dimensiones (*cantidad de ruedas*, *precio* y *vida útil*):

Id. de objeto	Descripción	Precio ($)	Vida útil (Kms.)	Ruedas
1	camión	250.000	400.000	8
2	automóvil	150.000	200.000	4
3	motocicleta	80.000	100.000	2
4	cuatriciclo	20.000	10.000	4
5	triciclo	500	100	3
6	bicicleta	3000	5.000	2

Tabla 6: Ejemplo de representación vectorial de objetos en 3 dimensiones

Los vectores que representarían los objetos en un espacio multidimensional serían:

$\vec{V}_1$ = (250000 , 400000 , 8)

$\vec{V}_2$ = (150000 , 200000 , 4)

$\vec{V}_3$ = (80000 , 100000 , 2)

$\vec{V}_4$ = (20000 , 10000 , 4)

$\vec{V}_5$ = (500 , 100 , 3)

$\vec{V}_6$ = (3000 , 5000 , 2)

y los objetos quedarían representados en un espacio vectorial como puntos de coordenadas de 3 dimensiones, tal como se aprecia en la siguiente figura:

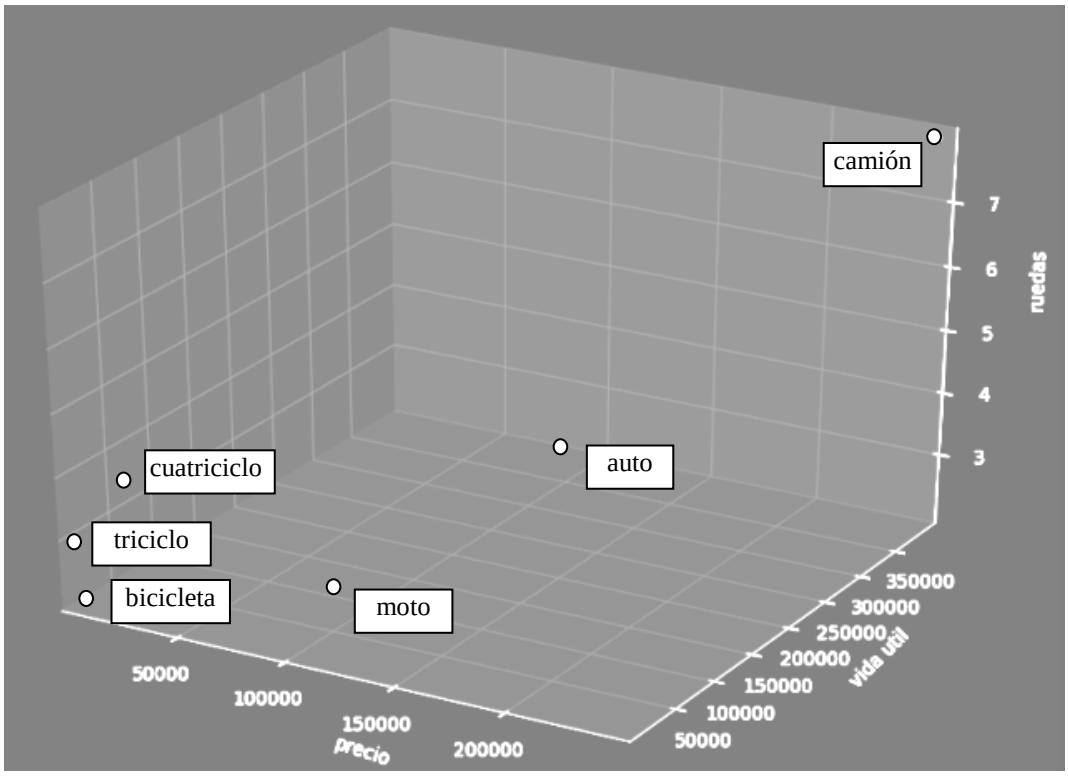

Figura 5: Ubicación de los vectores en un espacio tridimensional adaptado a un gráfico bidimensional

Un cluster es el producto resultante de asignar uno o más objetos (*instances*) a una clase. Como veremos más adelante, se trata de un proceso iterativo y convergente, por el cual se postula la existencia de al menos dos clases (K = 2 clusters) y como máximo tantas clases como objetos a clusterizar (clases de un único miembro). Idealmente, el resultado del proceso de clustering sólo depende de las divisiones naturales ya presentes en los datos (Manning y Schütze 1999).

Intuitivamente, en el ejemplo de los rodados, la representación tridimensional parece predisponer al agrupamiento de los vectores en función de dos clusters. Por un lado, el *triciclo*, la *bicicleta* y el *cuatriciclo* (cluster 1), por otro lado la *moto* y el *automóvil* (cluster 2). Finalmente, el *camión* es el objeto más disímil del conjunto (*outlier*), sin asignación de cluster visible.

Una ventaja del clustering es que se pueden inferir algunas observaciones que *a priori* no resultaban tan evidentes a través de un análisis exploratorio de datos (*Exploratory Data Analysis*) (Manning y Schütze 1999), a saber: que en este ejemplo las dimensiones *precio* y *vida útil* son, en conjunto, buenos criterios separadores de clusters, en tanto el *número de ruedas*, no. Esta apreciación espacial de los clusters podría deberse a un problema de normalización de los valores de distinto orden del número discreto de ruedas vs. los valores de miles de unidades de las otras dos dimensiones (*cf. normalización de vectores* en Manning y Schütze 1999). De hecho, la mejor

forma de agrupar objetos en función de su similitud es calcular una métrica confiable de distancia en un espacio multidimensional.

Estas intuiciones geométricas se plasman en la necesidad de agrupar objetos (*instances*) en clusters a partir de las distancias (o mejor aún, de su proximidad) entre sí. Existen dos tipos de métricas clásicas para las distancias en un espacio vectorial (Abney 2008): la distancia euclideana (también conocida como L_2 *norm*) -definida como la línea recta más corta que une los puntos en el espacio- y la *distancia de Hamming* (también conocida como L_1 *norm* o *distancia Manhattan* o *city-block distance*) -definida como la sumatoria de las diferencias en cada una de las dimensiones de los dos puntos en cuestión:

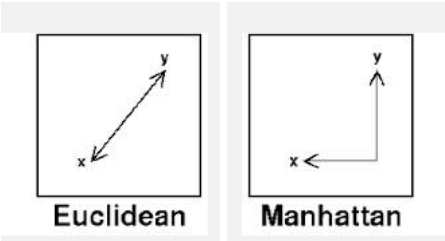

Figura 6: Representación de los dos tipos de distancia vectorial en un espacio bidimensional

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Ecuación 1: Distancia euclideana

$$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{k}|x_i - y_i|$$

Ecuación 2: Distancia Manhattan

El cálculo de las respectivas distancias entre los vectores $\vec{V}_1$ a $\vec{V}_6$ del ejemplo podría ser una buena justificación de la separación en los tres clusters que intuitivamente habíamos observado en el gráfico tridimensional:

> "The geometric intuitions that we appeal to in clustering depend on having a definition of the distance between instances. A good cluster contains many points that are close to each other, and it is well separated (distant) from other clusters." [Abney 2008:133-134]

La representación multidimensional de los datos y el cálculo de distancia vectorial permiten definir otra carácterística de los clusters: el centroide o centro de gravedad de cada cluster. El centroide de un cluster no es más que otro vector $\vec{\mu}$ en el espacio *n*-dimensional que resulta del promedio de la suma vectorial (dimensión por dimensión) de los *M* miembros (vectores $\vec{x}$ que pertenecen a dicho cluster *c*):

$$\vec{\mu} = \frac{1}{M}\sum_{\vec{x}\in c}\vec{x}$$

Ecuación 3: Cálculo del centroide de un cluster

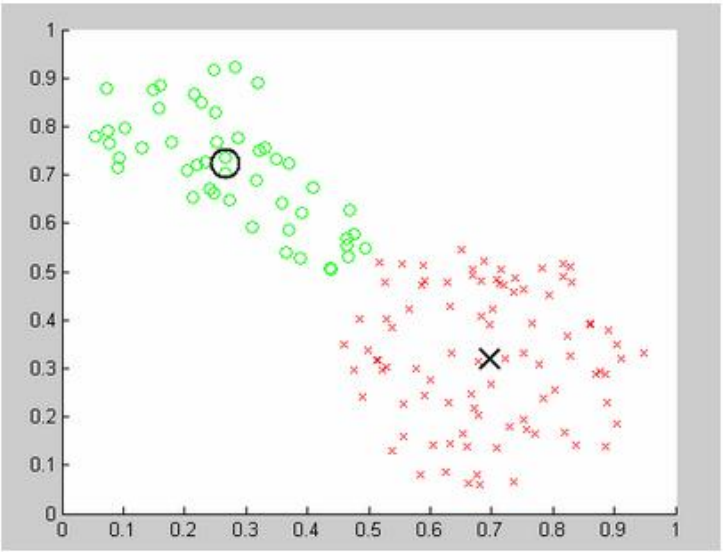

Figura 7: Representación de los centroides en un espacio bidimensional

El centroide es otro vector de *n*-dimensiones que, de algún modo, representa a todo el cluster. La distancia vectorial del centroide de un cluster hacia otros centroides de otros clusters es un buen parámetro para evaluar cuánto difiere un cluster de otro o cuánto se parecen entre sí. Esto nos llevará más adelante al concepto de *hipercluster* como agrupamiento de clusters similares cuyos centroides se encuentren cercanos entre sí (véase *capítulo 7* de esta tesis), como una buena justificación algebraica de la típica evaluación *many-to-1* en trabajos de clustering aplicados a la categorización de palabras (Christoudoulopoulos *et al*. 2010). En algunos casos, suele recurrirse al concepto de *medoide* (Manning y Schütze 1999) como una alternativa del centroide. El medoide es el objeto del cluster más representativo, lo cual implica un enfoque más basado en similitudes prototípicas que en distancias algebraicas.

4.2 Clustering jerárquico o aglomerativo

El clustering jerárquico se basa en la intuición general de que los objetos del espacio vectorial más cercanos están más relacionados entre sí que con los objetos más lejanos. El resultado final se presenta bajo la forma de un *dendrograma* (Manning y Schütze 1999; Abney 2008), un grafo donde los miembros de los clusters en los nodos de las hojas están más estrechamente vinculados entre sí que los miembros de los clusters cercanos a la raíz. Si se corta el dendrograma en un nivel dado (una cierta distancia representada en la figura siguiente por la línea punteada), se obtiene una cantidad discreta de clusters:

El algoritmo iterativo podría detallarse como sigue:

1) La condición de incialización es calcular las distancias entre todos los pares de objetos. Esto es lo mismo que asumir que cada objeto constituye un cluster: $\{C_1, ...,C_N\}$. En el ejemplo de la Figura 8 serían los objetos $X_1, \ldots, X_8$

2) Se buscan los dos clusters más cercanos (C_i, C_j), éstos se *enlazan* y constituyen un único cluster $C_{i\,j}$. En el ejemplo de la Figura 8 serían los pares $X_{1\text{-}2}$, $X_{3\text{-}4}$ y $X_{7\text{-}8}$

3) Se repite el paso 2 hasta que no quedan pares de comparación o hasta que se alcanza una distancia máxima estipulada como parámetro de corte. En el ejemplo esto está representado por la línea punteada. Al alcanzar esta condición final, el algoritmo devuelve los clusters formados con la asignación de miembros correspondiente. En el caso del ejemplo, 3 clusters finales ($X_{1\text{-}2}$, $X_{3\text{-}4\text{-}5}$, $X_{6\text{-}7\text{-}8}$)

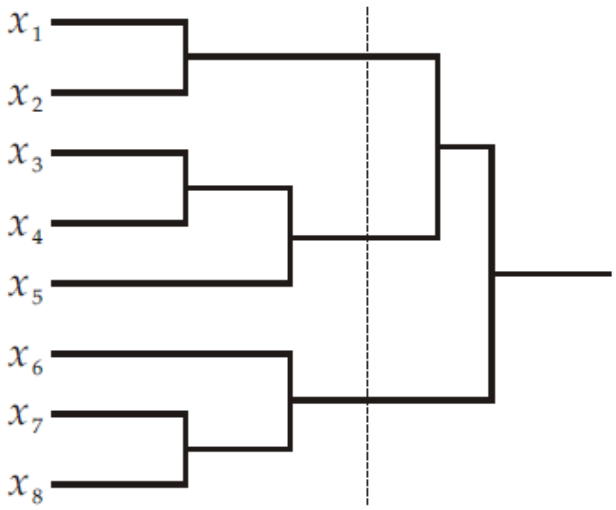

Figura 8: Ejemplo de dendrograma

Como se desprende del algoritmo, la noción de distancia adquiere una importancia crítica para ir *enlazando* objetos acumulativamente. Es decir, cuán lejos un objeto está de otro puede ser fácilmente computado por alguna de las dos métricas de distancia definidas en la sección anterior. Sin embargo, una vez agrupados los objetos en clusters, no es tan sencillo acordar sobre qué tan cerca o lejos están los clusters entre sí. Este parámetro de comparación de distancias de clusters para el modelo de clustering jerárquico es denominado enlace (*linkage*) (Manning y Schütze 1999; Abney 2008). Existen cuatro tipos básicos de enlace:

Single linkage: La distancia entre dos clusters X e Y es la mínima distancia entre dos puntos cualesquiera de sus respectivos miembros $x \in X$ e $y \in Y$.

Complete linkage: La distancia entre dos clusters X e Y es la máxima distancia entre dos puntos cualesquiera de sus respectivos miembros $x \in X$ e $y \in Y$.

Centroid linkage: La distancia entre dos clusters X e Y es la distancia entre sus respectivos centroides.

Average linkage: La distancia entre dos clusters X e Y es la distancia promedio entre todos los puntos de sus respectivos miembros $x \in X$ e $y \in Y$.

Al igual que la distancia de corte, esta parametrización inicial del tipo de enlace resulta crítica, por cuanto los mismos objetos de un espacio vectorial pueden ser sometidos a diferentes agrupamientos en escenarios que contemplen distintas parametrizaciones:

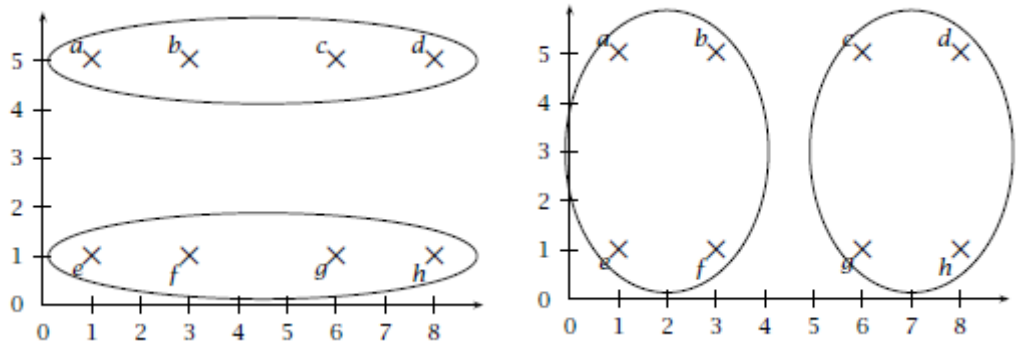

Figura 9: El mismo set de datos agrupados de una u otra manera según el tipo de enlace (*linkage*) de clustering jerárquico: *single linkage* y *complete linkage*, respectivamente

El clustering jerárquico es computacionalmente costoso, ya que la complejidad del algoritmo de *single linkage*, el criterio más ampliamente utilizado en la tarea de categorización morfosintáctica de palabras, es de orden $O(n^2)$.

4.3 Clustering no jerárquico o partitivo

Este tipo de técnica de clustering puede ser entendido como un problema de optimización: hallar los K-centroides (*means*) de los clusters necesarios para que la suma de todas las distancias vectoriales de los miembros asociados a los diversos clusters respecto de dichos centroides sea mínima:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

Ecuación 4: Definición de K-means como optimización de error de ciclo

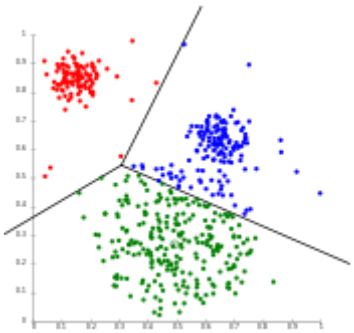

Figura 10: Optimización del espacio vectorial en tres clusters, para el set de datos iniciales

El clustering partitivo es un proceso iterativo y convergente. Esto quiere decir que hay una condición de inicialización a partir de la que se computa la asignación de los objetos (*instances*) a diferentes clusters, lo cual constituye un ciclo. Cada ciclo tiene asociado un error (*cycle error*) en función de la suma de todas las distancias entre los miembros de cada cluster y el centroide de dicho cluster. Si bien lo usual es hacer uso de la distancia euclideana, la distancia Manhattan parece lidiar mejor con los objetos *apartados* (*outliers*) (Manning y Schütze 1999). Con cada iteración se vuelve a computar la asignación en los vectores más alejados de los centroides, intentando minimizar dicho error. El algoritmo más usado es conocido como *K-means* o *algoritmo de Lloyd* (Manning y Schütze 1999; Abney 2008) y puede ser expresado como:

1) Comenzar por ubicar 2 centroides al azar (ciclo K=2).
2) Calcular la distancia euclideana de cada uno de los objetos del espacio vectorial a dichos centroides y asignarlos a uno u otro en función de la distancia mínima.
3) Computar el error de ciclo como la sumatoria de las distancias euclideanas a sus respectivos centroides de todos los vectores de cada cluster.
4) Comenzar una nueva iteración con un nuevo cluster (K = n+1), inicializar los correspondientes centroides al azar y reasignar los vectores a los nuevos centroides.
5) Recalcular centroides para los nuevos clusters y el nuevo error de ciclo.
6) Iterar el algoritmo desde el paso 2) hasta que el error de ciclo de una nueva asignación sea mayor que el de la iteración actual o hasta que se alcanza el ciclo K= máximo de clusters (parámetro).

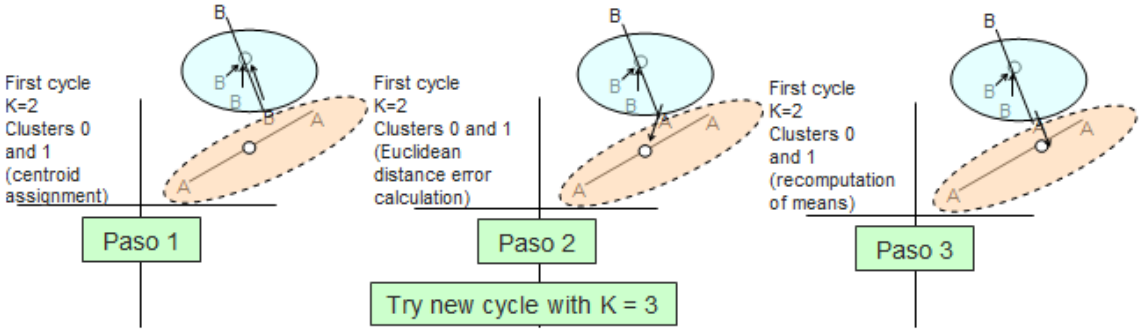

Figura 11: Ciclo de iteración con el algoritmo de *clustering* K-means. Esquema de 2 clusters (K=2) de vectores con sus centroides.

Obsérvese la reasignación del *outlier* B del cluster *b* del paso 1 al cluster *a* del paso 2 y correspondientemente el cambio de ubicación del centroide del cluster *a* en el paso 3.

En realidad, el algoritmo descripto arriba es nuestra propia implementación del algoritmo K-means con ciertas variaciones. Introdujimos la variación de iterar sobre diversos escenarios (desde K=2 clusters hasta un parámetro definido K=max) en función de "historizar" la conformación de los clusters y hallar un escenario que optimice dichas asignaciones según el criterio (*gold standard*) de categoría de palabra morfosintáctica que estaremos investigando en el capítulo 7 de esta tesis. De hecho, ante la ausencia de un gold standard, es sabido que el problema de hallar el número de clusters que optmiza una distribución de objetos en el espacio vectorial mediante técnicas de clustering no es nada sencillo. No obstante, algunos trabajos que analizaremos luego se han centrado en métricas de evaluación de la robustez de los diversos escenarios en función de la partición (*partioning*) en nuevos clusters o reagrupamiento en

hiperclusters (*merging*) que puedan optimizar el clustering de datos (Manning y Schütze 1999; Böhm *et al*. 2006).

El clustering no jerárquico es computacionalmente menos costoso que el jerárquico, ya que la complejidad del algoritmo es de orden $O(n)$ para un número constante de iteraciones.

4.4 Consideraciones acerca de la pertinencia de las técnicas de clustering para la categorización de palabras

En la mayoría de los trabajos de inducción de categorías morfosintácticas a partir de información distribucional mediante técnicas de clustering se recurre a una misma premisa: para analizar la distribución del contexto de ocurrencia de cada palabra (*target*) usaremos una unidad denominada bigrama: co-ocurrencia de pares de ítems léxicos en una relación fija contigua. Dicha relación puede ser, por ejemplo, la contigüidad que existe entre una palabra *target* (es decir, la palabra que se pretende estudiar) y su contexto inmediato (la palabra inmediatamente siguiente o anterior), relación denominada comúnmente *ventana de análisis* y en particular, bigrama hacia la derecha o bigrama hacia la izquierda, respectivamente. Por ejemplo, si todo el *corpus* consistiera en una única frase "*la vaca salta sobre la cerca*", la siguiente tabla representaría el vector de ocho dimensiones del contexto correspondiente a la palabra *salta* (Manning y Schütze 1999; Zhitomirsky-Geffet y Dagan 2009):

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

Tabla 7: Ejemplo de vector de bigramas hacia la derecha y hacia la izquierda para la palabra "*salta*" en la oración "*la vaca salta sobre la cerca*"

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

En corpora masivos es de esperar que los ítems lexicales que pertenecen a una misma categoría morfosintáctica tengan una distribución similar, lo cual se traduce en una cercanía en el espacio vectorial (Manning y Schütze 1999) susceptible de ser descubierta a partir de técnicas de

clustering. Esta premisa básica era la que subyacía también a la teoría de los marcos frecuentes (Mintz 2003; Chemla *et al*. 2009) que analizamos en el capítulo anterior de esta tesis. El mapeo de categorías sintácticas sobre un espacio vectorial multidimensional asume que hay una manera de dividir esas mismas categorías bajo un criterio geométrico: tradicionalmente se han propuesto modelos donde la frontera es discreta, y otros donde es prototípica o basada en similitudes entre ítems lexicales individuales. Algunos de estos criterios serán descriptos en detalle en el capítulo 7 de esta tesis.

Esta premisa básica compartida por el clustering sobre información distribucional y por la teoría de los marcos frecuentes presenta, no obstante, sustanciales diferencias. Por un lado, en el caso de computar vectores sobre un corpus masivo se trabaja con un enfoque estadístico sistemático. Un vector de bigramas cubriría así todos los marcos de ocurrencia de una palabra target en cuestión y no sólo los más frecuentes. Es de esperar que la palabra '*niño*' en español presente no sólo un elevado número para la frecuencia absoluta de bigramas como *el-niño* y *niño-es* sino también para otro tipo de combinaciones sistemáticas. Por otro lado, la ausencia de determinados bigramas (el número cero ocupando la dimensión correspondiente en el vector) también es significativa para el agrupamiento mediante clustering en el espacio vectorial. Así pues, la no ocurrencia de los bigramas *niño-son* o *la-niño* posiblemente sea una característica de todos los vectores que representan a los sustantivos comunes singulares masculinos en español. Mientras que esta información no era tomada en cuenta por la teoría de los marcos frecuentes, en las técnicas de clustering es naturalmente incorporada a los vectores en función de una buena determinación de las palabras de contexto que actúan como facilitadores (*cues*).

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

especializada da en llamar *cue* (Redington *et al*. 1998; Clark 2002) o *feature words* (Nath *et al*. 2008)-, o bien a partir de la simplificación de la matriz resultante de los vectores, desestimando las submatrices *anuladas* en cero – a partir de técnicas como *Single Value Decomposition* (SVD) (Deerwester *et al*. 1990; Schütze 1993) o *Principal Component Analysis* (PCA) (Böhm *et al*. 2006).

Este procedimiento algebraico de reducción de la dimensionalidad del espacio vectorial a partir de la identificación de palabras marcas (*cues*) tiene su perfecto correlato en la evidencia psicolingüística ontogenética de la adquisición de la habilidad temprana de categorización de palabras que estudiamos en el capítulo anterior de esta tesis: aprendemos a categorizar palabras en función de cierta información facilitadora (*cues*), la cual bien puede estar representada por ciertos *descriptores* preferenciales (Redington *et al.* 1998; Clark 2002) para todos los tipos de palabras. Como mencionamos en el capítulo 2 de esta tesis, la hipótesis central de este trabajo sostiene que dicho papel sería desempeñado mayormente por las palabras funcionales de un idioma, en virtud de su ocurrencia masiva y de sus propiedades distribucionales y articulatorias (actúan como bisagras) respecto de las restantes palabras. Dos grandes desafíos se derivan de esta hipótesis central: demostrar que estas *cues* están disponibles para el adquiriente de un lenguaje en forma previa a los tipos de palabras morfosintácticas a inducir -si no como palabras plenamente adquiridas, al menos como marcas formales en los PLD- y demostrar que esta inducción puede ser llevada a cabo mediante mecanismos generales (no de dominio específico) de aprendizaje no supervisado.

Justamente, todas estas consideraciones nos llevan a contemplar algunos aspectos de modelización que deben ser cuidadosamente analizados para este tipo de enfoques en experimentos de clustering. Algunas consideraciones son inherentes a la naturaleza del problema de la categorización de palabras y otras, en cambio, atañen a las técnicas de clustering empleadas como metodología para la presente investigación. Dichas consideraciones serán analizadas en detalle en el capítulo 7 de esta tesis, al momento de presentar el diseño de nuestro experimento de clustering para categorización de palabras en español.

Capítulo 5. Estado de la cuestión en categorización: modelos formales basados en clustering

5.1 Dos décadas de inducción no supervisada de categorías de palabras mediante clustering

A esta altura de la tesis estamos en condiciones de historizar los diversos trabajos canónicos que han aparecido a lo largo de las últimas dos décadas para la inducción de categorías morfosintácticas de palabras mediante técnicas de clustering (Christodoulopoulos *et al*. 2010). Algunos de estos trabajos difieren en cuanto a las métricas de evaluación. Algunos otros difieren en cuanto al objetivo final del experimento: demostrar cierta plausibilidad psicolingüística (Redington *et al*. 1998; Clark 2002; Nath *et al*. 2008) o mejorar el proceso de entrenamiento de etiquetadores morfosintácticos (POS-*taggers*) (Graça *et al*. 2011). Ocasionalmente, algunos trabajos incluyen recursos de otro paradigma científico más allá del paradigma estadístico, como es el caso de la apelación adicional a una red neuronal en Schütze (1993), pero en todos los casos existe una metodología básica compartida y una misma premisa lingüística formal:

> "The general methodology […] for inducing word class information can be outlined as follows, (a) collect global context vectors of target words by counting how often feature words appear in the neighboring positions, and, (b) apply a clustering algorithm on these vectors to obtain word classes." [Nath et al. 2008:1221]

5.2 Brown et al. (1992)

Brown *et al*. (1992) fue el primer trabajo de clustering para inducción no supervisada de categorías de palabras en inglés. Básicamente el algoritmo recurre a clustering jerárquico de un vocabulario en *V* clusters con un criterio de agrupamiento (*merging*) iterativo que maximice la probabilidad de un modelo markoviano (*perplejidad* mínima) de primer orden (bigramas a izquierda), criterio de calidad que es esquivalente a encontrar la partición que maximice la *mutual information* entre los clusters resultantes o, en otros términos, la *pérdida de información mutua MI-loss* debe ser mínima (Manning y Schütze 1999):

$$\text{MI-loss}(c_i, c_j) = \sum_{c_k \in C \setminus \{c_i, c_j\}} I(c_k;\ c_i) + I(c_k;\ c_j) - I(c_k;\ c_i \cup c_j)$$

Ecuación 5: Criterio de calidad de agrupamiento de clusters basado en minimización de pérdida de información mutua *I* entre los clusters a ser agrupados C_i y C_j

El algoritmo se detiene cuando se ha alcanzado un número K de clusters fijado como parámetro inicial en 1000 (Brown *et al*. 1992).

Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays
June March July April January December October November September August
people guys folks fellows CEOs chaps doubters commies unfortunates blokes
down backwards ashore sideways southward northward overboard aloft downwards adrift
water gas coal liquid acid sand carbon steam shale iron
great big vast sudden mere sheer gigantic lifelong scant colossal
man woman boy girl lawyer doctor guy farmer teacher citizen
American Indian European Japanese German African Catholic Israeli Italian Arab
pressure temperature permeability density porosity stress velocity viscosity gravity tension
mother wife father son husband brother daughter sister boss uncle
machine device controller processor CPU printer spindle subsystem compiler plotter
John George James Bob Robert Paul William Jim David Mike
anyone someone anybody somebody
feet miles pounds degrees inches barrels tons acres meters bytes
director chief professor commissioner commander treasurer founder superintendent dean custodian
liberal conservative parliamentary royal progressive Tory provisional separatist federalist PQ
had hadn't hath would've could've should've must've might've
asking telling wondering instructing informing kidding reminding bothering thanking deposing
that tha theat
head body hands eyes voice arm seat eye hair mouth

Tabla 8: Ejemplos de los 10 miembros más frecuentes de algunos de los 1000 clusters inducidos

La bibliografía especializada se ocupó de relevar alguna falencias de este trabajo fundacional:

- El algoritmo resultante es *voraz* (*greedy*), computacionalmente costoso y de complejidad $O(V^3)$ (Christodoulopoulos *et al*. 2010). Inicialmente Brown *et al*. (1992) trabajaron con un corpus de 365.893.263 tokens y 260.741 palabras (*types*) de vocabulario de diversas fuentes en una partición final de 1000 clusters.

- No hay justificación alguna del parámetro de 1000 clusters finales (Manning y Schütze 1999).

- No hay un tratamiento adecuado de la ambigüedad del tipo de palabra (*POS-tag ambiguity*) (Clark 2002). Cada palabra sólo puede ser asignada a un único cluster (*hard clustering*). Esta falencia resulta bastante invalidante para un idioma como el inglés, en el que una misma forma léxica puede actuar como verbo o sustantivo tan sólo en virtud de su posición sintáctica absoluta.

- El modelo de lenguaje subyacente al criterio de agrupamiento de clusters se basa en la premisa de que la inclusión de una palabra target W_i a un cluster C_i depende solamente de a qué cluster C_3 se asignó la palabra predecesora W_3. Claramente esto es un reduccionismo del determinismo que juegan los tipos de palabra de los vecinos de una palabra target (Redington *et al*. 1998).

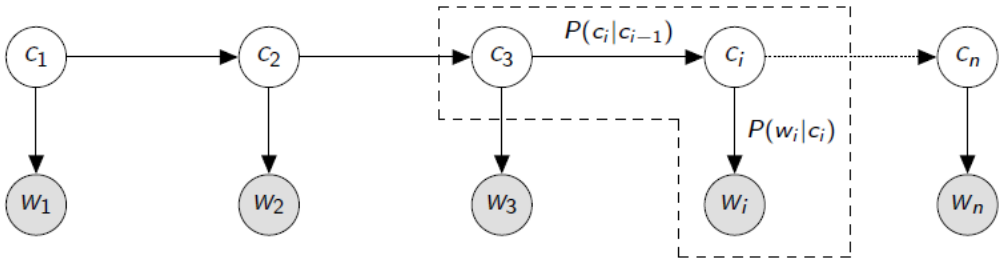

Figura 12: Modelo bayesiano de lenguaje subyacente a criterio de agrupamiento de clusters

Pese a estas críticas, es menester destacar que el algoritmo de Brown *et al*. (1992) demostró ser uno de los más efectivos en comparación con los avances posteriores, constituyéndose en un baseline para los trabajos siguientes -en Graça *et al*. (2011) se realiza un exhaustivo trabajo de comparación de diversos experimentos clásicos bajo las mismas condiciones y el algoritmo de Brown *et al.* (1992) resulta favorecido en varios escenarios con medidas F que alcanzan 68,7% en la evaluación *many-to-1* (Christodoulopoulos *et al*. 2010):

> "The degree to which the classes capture both syntactic and semantic aspects of English is quite surprising given that they were constructed from nothing more than counts of bigrams." [Brown *et al.* 1992:475]

Los lineamientos básicos de este trabajo fundacional continuarán en Martin *et al*. (1998).

5.3 Schütze (1993)

El trabajo de Schütze (1993) presenta una arquitectura algorítmica bastante sofisticada: dos iteraciones de clustering seguidas de un procesamiento a través de una red neuronal bi-recurrente (Elman 1991). A diferencia de Brown *et al*. (1992) , las *cues* que actúan en la premisa de la información distribucional se extienden más allá de la palabra predecesora (bigramas a izquierda <W_{-1},W_{target}>) a bigramas a derecha <W_{target},W_1>) y a los bigramas con palabras a distancia de una palabra de la *target* <W_{-2},W_{target}>) y <W_{target},W_2>).

W_{-2} W_{-1} W_{target} W_1 W_2

El corpus está definido por texto escrito con artículos del *New York Times News Service* desde junio de 1990 hasta octubre de 1990, identificando las 5.000 palabras (*types*) más frecuentes como palabras target a categorizar. Es decir, cada una de las 5.000 palabras *target* está caracterizada por la posible co-ocurrencia con la combinatoria de estas 5.000 palabras en las posiciones W_{-2} W_{-1} W_{target} W_1 W_2 , lo que significa una matriz inicial de 5.000 filas o vectores (cada W_{target}) por 20.000 columnas (dimensiones con las frecuencias absolutas de los bigramas <W_{-1},W_{target}><W_{target},W_1><W_{-2},W_{target}><W_{target},W_2> para cada palabra combinada en las posiciones W_{-2} W_{-1} W_1 W_2.

La primera intución de Schütze (1993) es que bastaría con modelizar los 5.000 vectores de 20.000 dimensiones en el espacio para encontrar directamente la similitud entre palabras,

formando clases mediante, por ejemplo, la métrica del coseno del ángulo vectorial entre las palabras *target*:

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Ecuación 6: Similitud de dos vectores X e Y basada en el coseno del ángulo que los separa en el espacio vectorial

Sin embargo, esto resulta inviable debido a la excesiva dimensionalidad de los vectores. Por lo tanto, se recurre a una técnica de reducción de la dimensionalidad denominada *Single Value Decomposition* (SVD) (Deerwester *et al*. 1990), obteniendo así una matriz equivalente reducida de sólo 15 dimensiones y luego sí se aplica el criterio de similitud por coseno del ángulo vectorial para los 5.000 vectores de palabras target:

word	nearest neighbors
accompanied	submitted banned financed developed authorized headed canceled awarded barred
almost	virtually merely formally fully quite officially just nearly only less
causing	reflecting forcing providing creating producing becoming carrying particularly
classes	elections courses payments losses computers performances violations levels pictures
directors	professionals investigations materials competitors agreements papers transactions
goal	mood roof eye image tool song pool scene gap voice
japanese	chinese iraqi american western arab foreign european federal soviet indian
represent	reveal attend deliver reflect choose contain impose manage establish retain
think	believe wish know realize wonder assume feel say mean bet
york	angeles francisco sox rouge kong diego zone vegas inning layer
on	through in at over into with from for by across
must	might would could cannot will should can may does helps
they	we you i he she nobody who it everybody there

Tabla 9: Ejemplos de los 10 miembros más cercanos a cada una de las 10 palabras target seleccionadas

Como Schütze (1993) sostiene, el coseno del ángulo vectorial como criterio de similitud no es una solución que pueda escalar. A la vez, si bien los vecinos cercanos resultantes tienden a pertenecer a la misma categoría sintáctica (véase *Tabla 9*), la premisa de las dimensiones como bigramas de ocurrencia distribucional de palabras no resulta tan eficaz, por lo que Schütze (1993) decide formar nuevos vectores ya no con los conteos de ocurrencias de palabras individuales sino con las ocurrencias de las palabras target con clases de palabras resultantes de un primer ciclo de clustering.

Para el primer ciclo de clustering, Schütze toma como entrada los 5.000 vectores de 15 dimensiones reducidas luego de haber aplicado SVD. El algoritmo de clustering utilizado es el de Buckshot clustering (Cutting *et al*. 1992), una versión más simple del algoritmo jerárquico aglomerativo que tiene como objetivo reducir el tiempo de cómputo desde $O(n^2)$ a $O(k*n)$.

> "Instead of counting the number of occurrences of individual words, we would now count classes. The space was clustered with Buckshot, a linear-time clustering algorithm described in (Cutting *et al.* 1992). Buckshort applies a high-quality quadratic clustering algorithm to a random sample of size $\sqrt{k*n}$, where k is the number of desired cluster centers and n is the number of vectors to be clustered.

> Each of the remaining $n - \sqrt{k*n}$ vectors is assigned to the nearest cluster center. The high-quality quadratic clustering algorithm used was truncated group average agglomeration." [Schütze 1993:252]

Para este ciclo inicial de clustering, Schütze (1993) toma la precaución de dejar a un lado las palabras funcionales -el corte arbitrario deja afuera las primeras 278 palabras más frecuentes- que no serán clusterizadas pero sí actuarán como cues o features en el segundo ciclo de clustering. El resultado de este primer ciclo son 4.722 palabras target (5.000 menos las 278 palabras funcionales) agrupadas en 222 clases, las cuales actuarán como features más reducidas en dimensionalidad que los miembros individuales de las clases para el segundo ciclo, posibilitando así la cobertura de un número masivo de vocabulario.

El segundo ciclo de clustering toma como entrada un número mucho mayor de palabras: 22.771 palabras (*types*) que superaron las 100 ocurrencias en 18 meses de corpus del *New York Times News Service* (mayo de 1989 a octubre de 1990) en las 500 dimensiones (que surgen de las 278 palabras funcionales y los 222 clases del primer ciclo) en las cuatro posiciones combinatorias distribucionales W_{-2} W_{-1} W_1 W_2 respecto de la palabra target. Esto determina una matriz de 22.771 filas (vectores) y 2.000 columnas (dimensiones), sobre la cual vuelve a aplicar SVD hasta reducirla a 10 dimensiones y luego vuelve a clusterizar utilizando Buckshot. Schütze (1993) no especifica en su trabajo el parámetro *k* de clusters deseados para la aplicación de Buckshot, pero deducimos que está en el orden de los 500 clusters, en función del valor de *n*=22.771 vectores a clusterizar. La siguiente tabla muestra 20 ejemplos de clusters finales y los vecinos más cercanos a la palabra (*word*) que actúa como etiqueta del cluster.

word	nearest neighbors
armaments	turmoil weaponry landmarks coordination prejudices secrecy brutality unrest harassment |
athlete	virus scenario | event audience disorder organism candidate procedure epidemic
b'nai	| suffolk sri allegheny cosmopolitan berkshire cuny broward multimedia bovine nytimes
bowers	jacobs levine carr hahn schwartz adams buckley dershowitz fitzpatrick peterson |
clerk	salesman | psychologist photographer preacher mechanic dancer lawyer trooper trainer
cliches	pests wrinkles outbursts streams icons endorsements | friction unease appraisals lifestyles
cruz	antonio | clara pont saud monica paulo rosa mae attorney palma
declaration	sequence mood profession marketplace concept facade populace downturn moratorium |
desirable	recognizable | frightening loyal devastating exciting troublesome awkward palpable
dome	blackout furnace temblor quartet citation chain countdown thermometer shaft |
equally	| somewhat progressively acutely enormously excessively unnecessarily largely scattered
financings	| endeavors monopolies raids patrols stalls offerings occupations philosophies religions
gibbs	adler reid webb jenkins stevens carr laurent dempsey hayes farrell |
luxuries	volatility insight hostility dissatisfaction stereotypes competence unease animosity residues |
northwestern	| baja rancho harvard westchester ubs humboldt laguna guinness vero granada
oh	gee gosh ah hey | appleton ashton dolly boldface baskin lo
sole	| lengthy vast monumental rudimentary nonviolent extramarital lingering meager gruesome
transports	| spokesman copyboy staffer barrios comptroller alloy stalks spokeswoman dal spokesperson
vividly	skillfully frantically calmly confidently streaming relentlessly discreetly spontaneously
walks	floats | jumps collapsed sticks stares crumbled peaked disapproved runs crashed
claims	credits promises | forecasts shifts searches trades practices processes supplements controls
on	through from in | at by within with under against for
must	will might would cannot could can should won't | doesn't may
they	we | i you who nobody he it she everybody there

Tabla 10: Ejemplos de los 10 miembros más cercanos a cada uno de las 20 clusters seleccionados. Obsérvese el signo ´|´ delimitando a izquierda la densidad del cluster.

El signo ´|´ delimita a izquierda aquellos vecinos que tienen una correlación mayor a 0,978 con la palabra que etiqueta al cluster en cuestión. Esta notación permite distinguir fácilmente aquellos clusters densos (algo deseable) de aquellos cuyos miembros están más dispersos:

> "As can be seen from the table, the regions occupied by nouns and proper names are dense, whereas adverbs and adjectives have more distant nearest neighbors. One could attempt to find a fixed threshold that would separate neighbors of the same category from syntactically different ones. For instance, the neighbors of oh with a correlation higher than 0.978 are all interjections and the neighbors of cliches within the threshold region are all plural nouns. However, since the density in the space is different for different regions, it is unlikely that a general threshold for all syntactic categories can be found." [Schütze 1993:254]

El análisis pormenorizado de los clusters seleccionados hace que Schütze se tope por primera vez con el escollo de la ambigüedad del tipo de palabra morfosintáctica del inglés y una definición formal del problema en el espacio vectorial, en virtud de cómo la consideración de contextos distribucionales alternativos, propios de la ambigüedad, interfieren en el perfil que representa a dicha palabra ambigüa, haciendo que en el espacio vectorial termine agrupada con otras que claramente no son de ninguna de las dos (o más clases) "*puras*" a las que pertenecería en cada una de sus interpretaciones individuales de Part-Of-Speech (POS):

> "The neighborhoods of *transports* and *walks* are not very homogeneous. These two words are ambiguous between third person singular present tense and plural noun. Ambiguity is a problem for the vector representation scheme used here, because the two components of an ambiguous vector can add up in a way that makes it by chance similar to an unambiguous word of a different syntactic category. If we call the distributional vector $\vec{v}_c$ of words of category *c* the *profile* of category *c*, and if a word w_l is used with frequency α in category c_1 and with frequency β in category c_2, then the weighted sum of the profiles […] may turn out to be the same as the profile of an unrelated third category c_3 :
> $\alpha * \vec{v}_{c1} + \beta * \vec{v}_{c2} = \vec{v}_{c3}$ " [Schütze 1993:254]

El problema de la ambigüedad de una misma forma léxica como distintos tipos de palabra morfosintáctica es claramente un escollo para los enfoques de clustering. Una forma de lidiar con ese escollo es recurrir a un *soft clustering* (Manning y Schütze 1999) que permite que un miembro sea asignado a más de una clase a la vez. Éste es el camino por el que optan algunos trabajos posteriores como el de Clark (2002). No obstante, Schütze (1993) elige un diseño alternativo para lidiar con palabras ambiguas, el cual apela incluso a conocimiento lingüístico específico y supervisado, socavando así los propios principios epistemológicos que deberían ser respetados en pos de la plausibilidad psicolingüística de este tipo de experimentos. Schütze (1993) busca la forma de desambiguar los posibles vectores ambigüos al intercalar una red neuronal bi-recurrente entre los vectores de dimensionalidad reducida que se obtienen de la aplicación de SVD en el segundo ciclo de clustering (22.771 vectores de 10 dimensiones) y la aplicación de Buckshot clustering.

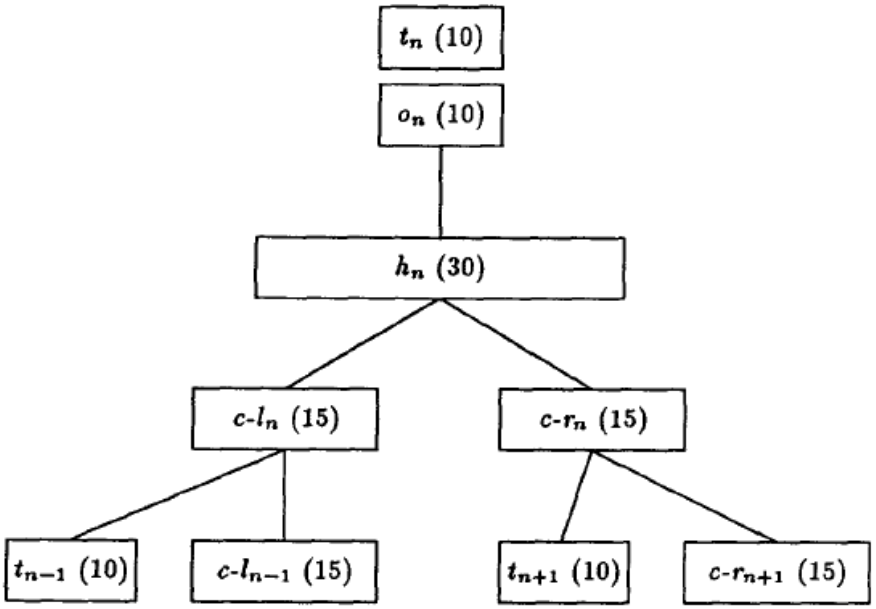

Figura 13: Arquitectura de la red neuronal bi-recurrente para el tratamiento de palabras ambiguas en Schütze (1993)

Schütze (1993) reconoce que el algoritmo de desambiguación falla en aprender la desambiguación de algunos POS y por eso recurre a dos reglas adicionales de dominio específico, tornando su experimento en un mecanismo de aprendizaje semisupervisado:

> "The network failed to learn the distinctions between adjectives, intransitive present participles and past participles in the frame '*to-be* + [] + non-NP'. For this reason, the adjective close, the present participle *beginning*, and the past participle *shot* are all classified as belonging to the category STRUGGLING_TRAVELING. (Present Participles are successfully discriminated in the frame '*to-be* + [] + NP': see *winning* in the table, which is classified as the progressive form of a transitive verb: HOLDING_PROMISING.) This is the place where linguistic knowledge has to be injected in form of the following two rules:
> 1) If a word in STRUGGLING_TRAVELING is a morphological present participle or past participle assign it to that category, otherwise to the category ADJECTIVE_PREDICATIVE.
> 2) If a word in a noun category is a morphological plural assign it to NOUN_PLURAL, to NOUN_SINGULAR otherwise. With these two rules, all major" [Schütze 1993:256]

Las críticas al trabajo de Schütze se centran en los siguientes dos aspectos:

- El experimento no ofrece una métrica final de evaluación para comparar la efectividad del enfoque.
- Algunos parámetros fueron definidos arbitrariamente sin ninguna justificación de tal elección (frecuencias de corte).
- La incorpopración de reglas específicas de dominio en la salida de la red neuronal socavan el principio epistemológico de mecanismos generales de aprendizaje no supervisado.

Con todo, es menester reconocerle al trabajo de Schütze (1993) ciertos méritos:

- Fue el primero en lidiar formalmente con el problema de la ambigüedad del POS-tag, problema típico del inglés. Veremos más adelante que este problema se da con menor dramatismo en lenguajes con morfología más rica (Clark 2002; Graça *et al*. 2011).

- Fue el primero en introducir con éxito técnicas de reducción de dimensionalidad del espacio vectorial (por ejemplo, SVD).

- Fue el primero en proponer el estudio de una ventana de análisis más extensa que no sólo involucre a la palabra predecesora de la palabra target. Esta perspectiva se analizará con más detalle en Redington *et al*. (1998).

5.4 Redington et al. (1998)

Redington *et al.* (1998), por su parte, representa la primera investigación exhaustiva sobre clustering para categorización de palabras con un enfoque específicamente orientado hacia la naturaleza lingüística del problema. En efecto, Redington *et al*. (1998) proponen una serie de nueve experimentos en los que no sólo ofrecen un algoritmo computacional para un mecanismo de aprendizaje no supervisado de inducción de categorías sintácticas, sino que también se preocupan por brindar adecuación explicativa para el fenómeno elucidado en virtud de la plausibilidad psicolingüítica de la metodología y de los hallazgos.

Desde le punto de vista del diseño del algoritmo, Redington *et al*. (1998) retoma el trabajo de Finch y Chater (1992) en cuanto a recurrir a un algoritmo de *hard clustering* (no da cuenta de la ambigüedad POS-tag, ya que asigna un miembro a una única clase), jerárquico aglomerativo de *enlace promedio* (*average linkage*) -más apropiado para lidiar mediante el debido tratamiento de los casos apartados (*outliers*) con clusters elongados y no tan esféricos como los que se espera en la modelización vectorial de la categorización de palabras.

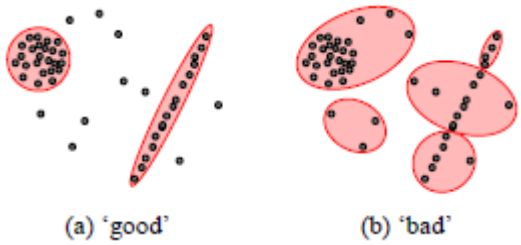

Figura 14: Tratamiento adecuado de *outliers* en clusters elongados densos (a) versus clusters esféricos inadecuados (b) para un mismo set de datos

Los datos de entrada están basados en la sección transcripta de discurso adulto del corpus CHILDES (*Child Language Data Exchange System*) en inglés, un *corpus* con emisiones de padres durante el proceso de adquisición del lenguaje de sus hijos, establecido en 1985 con 2,5 millones de tokens de aproximadamente 3.000 hablantes (2/3 de los cuales son mujeres):

> "This large and noisy corpus of adult speech provides a full-scale and realistic test of the usefulness of distributional information as a potential cue to linguistic categories. Indeed, in some ways, the corpus presents a greater challenge than that faced by children, because the number of speakers. dialects, constructions, topics, and vocabulary items is large. The language to which a single child, interacting with a small number of adults, is exposed will tend to be much more homogeneous." [Redington *et al*. 1998:439]

Los propios autores admiten que el *corpus* elegido no necesariamente incluye lenguaje únicamente dirigido a niños (*child directed speech*), como el *maternés* (*motherese*) o *baby talk*. 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 PLD.

A partir de dicho corpus, los autores trazan un *Perfil de Frecuencia Decreciente* (*Decreasing Frequency Profile* DFP) con las ocurrencias de palabras (*types*). El DFP ofrece información fundamental para identificar las palabras que actuarán como cues o features y las que serán objeto (*target*) del clustering. Efectivamente, Redington *et al*. (1998) apartan las 150 palabras más frecuentes como cues del contexto y seleccionan las restantes 1000 palabras más frecuentes como palabras target para sus experimentos de clustering. Esta decisión de diseño del experimento los diferencia de los enfoques anteriores (Brown *et al*. 1992; Schütze 1993) que buscaban cubrir con técnicas de clustering vocabulario de tamaño masivo. Esta modelización de un subset de palabras target restringido tiene su justificación desde el punto de vista de la plausibilidad psicolingüística:

> "It is not necessary (or even desirable) to record these statistics for every word in the input in order to provide useful information. From a psychological perspective, in the early stages of syntactic category acquisition, it seems unlikely that a syntactic category will be assigned to every word in the child's input, particularly given that the child's vocabulary is very limited." [Redington *et al*. 1998:436]

y su justificación algorítmica e implementativa:

> "It may also be computationally appropriate to focus on a small number of target words in order to provide more reliable distributional information and to avoid unnecessarily complex computation. Moreover, it may be appropriate to be even more restrictive with respect to the set of context words (over which frequency distributions are observed). This is because each target word may occur in a relatively small number of contexts, and only the most frequent words in these contexts will provide reliable frequency information." [Redington *et al*. 1998:436]

> "Because of the Zipfian distribution of words, cutting out low frequency items will greatly reduce the parameter space (and the memory requirements of the system being built), while not appreciably affecting the model quality." [Manning y Schütze 1999:199]

La explicación de por qué las palabras cues no son a su vez sometidas al proceso de clustering aparece en Martin et al. (1998):

> "As a result, the clustering process tries to distribute the frequent words uniformly over the word classes. Consequently, there will never be a homogeneous word class of numbers or function words at the end of a clustering process, because of the fact that these words appear quite often. Instead, the frequent words will be spread over all word classes, regardless of the initialization method." [Martin et al. 1998:34]

En nuestros experimentos de clustering observaremos esta misma decisión de diseño en cuanto a la separación algorítmica de cues y palabras target y la clusterización de estas últimas. Sin embargo, como observaremos en el capítulo 7 de esta tesis, la identificación algorítmica y no apriorística de cues (Elghamry 2004) en función de sus propiedades distribucionales en un corpus, resultará fundamental para otorgar una mayor plausibilidad psicolingüística a nuestro

modelo en comparación con la identificación arbitraria de cues en la que incurren Redington *et al*. (1998).

Como ya mencionamos, el modelo de Redington *et al*. (1998) no da cuenta de la posible ambigüedad del tipo de palabras (*hard clustering*). Esta carencia obliga a los autores a disponer de una referencia (*benchmarking)* según el tipo de palabra más probable (en el caso de ambigüedad) para cada forma léxica de las 1000 palabras target. Para ello recurren a la base de datos *Collins Cobuild Lexical Database*. Finalmente obtienen 956 palabras target con un único POS-tag de referencia:

Category	Example	n
noun	truck, card, hand	407
adjective	little, favorite, white	81
numeral	two, ten, twelve	10
verb	could, hope, empty	239
article	the, a, an	3
pronoun	you, whose, more	52
adverb	rather, always, softly	60
preposition	in, around, between	21
conjunction	cos, while, and	9
interjection	oh, huh, wow	16
simple contraction		0
complex contraction	I'll, can't, there's	58

Note. 44 words remained unclassified.

Tabla 11: POS-etiquetamiento de las 956 palabras finales del subset de palabras target a clusterizar

Este procedimiento de *benchmarking* para desambigüación puede ser considerado el gold standard contra el cual se evaluará la efectividad de los clusters. En nuestro experimento adoptaremos esta misma decisión de diseño. Los autores aclaran que de las 1000 palabras target, 44 no pudieron ser POS-etiquetadas y que unas 100 fueron etiquetadas a mano (mayormente, nombres propios). Esto significa un subset final de palabras *target* de 956 palabras.

Recordemos que el clustering jerárquico devuelve dendrogramas en función de un criterio de similitud, distancia o corte entre clusters (véase capítulo 4 de esta tesis). El criterio de corte por el que se inclinan los autores es el coeficiente ρ de la *correlación Spearman* (*Spearman rank correlation*):

> "The Rank correlation measure may be the most successful because it is a robust measure which makes no assumptions about the underlying structure of the set of points in the space […]. This distribution is non-normal and the absolute differences between points on some dimensions can be very large, which may potentially swamp all other differences if parametric measures are used (e.g., Euclidean distance). In fact, these large differences are inevitable, as bigram frequencies, like word frequencies, have an extremely skewed distribution (specifically, they follow Zipf's law). Intuitively, in linguistic terms, the distribution is non-normal, since the items tend to be clumped within distinct regions of the space (corresponding, to some extent, to syntactic categories). Again, it is intuitively apparent that some elements of the vectors will be orders of magnitude larger than others, reflecting the fact that some words appear in almost stereotyped relationships (e.g., *of the, in the, of a*)." [Redington *et al.* 1998:437]

Para obtener una medición de los resultados se procede a cortar el *dendrograma* resultante en distintos niveles (*similarity level*) y a evaluar para los miembros de cada cluster el grado de

precisión (*precision*, a la que denominan erróneamente *accuracy*) -donde gravitan los falsos positivos (*false alarmas*)- y cobertura (*completeness* o *recall*) -donde gravitan los falsos negativos (*misses*). Estas métricas de evaluación son para cada cluster. Sin embargo, los autores no especifican el procedimiento matemático (presumiblemente el promedio o promedio ponderado en función de cantidad de miembros entre los clusters resultantes) que aplican en los análisis globales de cada escenario de sus experimentos. Como *Precision* -mal denominada *accuracy* en Redington *et al*. (1998)- y *Recall* -denominada *completeness* en Redington *et al*. (1998)- son métricas generalmente complementarias, hubiese sido deseable que los autores trabajen con la medida F (*F-score*) que resulta del promedio armónico de ambas, tal como se especifica en nuestro experimento del capítulo 7 de esta tesis.

$$\text{Accuracy} = \frac{\text{hits}}{\text{hits + false alarms}}$$

Ecuación 7: Precisión (*accuracy*) en los clusters

$$\text{Completeness} = \frac{\text{hits}}{\text{hits + misses}}$$

Ecuación 8: Cobertura (*completeness*) en los clusters

Hits son los aciertos para las palabras asignadas a un cluster que representa al POS correspondiente en el *benchmark.*

Para solucionar esta tendencia a un *trade-off* (situación de equilibrio complementario) entre *Accuracy* y *Completeness*, los autores proponen otra métrica de evaluación de los clusters basada en la *teoría de la información* (Shannon 1948), denominada *informatividad* (*informativeness*):

> "However, having two measures for each classification makes comparing the scores for two different classifications difficult. Specifically, when one classification has a higher accuracy, but a lower completeness than a second, or vice versa, it is unclear how accuracy and completeness should be traded off. This second information-theoretic kind of scoring produces only one measure, avoiding this problem. The measure of goodness is the mutual information between the classification and the benchmark (the information that they share), as a percentage of their joint information (the information conveyed by the classification and the benchmark together). This measure reflects both accuracy and completeness. Groupings in either classification or benchmark that are not reflected in the other (that is, both false alarms and misses) will increase the joint information, and penalize the measure." [Redington *et al*. 1998:441]

Este criterio basado en la información mutua entre los clusters resultantes y los clusters del *benchmark* o gold standard dará origen a métricas más sofisticadas como Variación de la Información (*Variation of Infomation* VI) (Meilă 2003) y medida V (*V-measure*) (Rosenberg y Hirschberg 2007):

$$\text{Informativeness} = \frac{I_i + I_j - I_{ij}}{I_{ij}}.$$

where I_i is the amount of information in the classification, as given by:

$$I_i = -\sum_i p(i) \log_2 p(i).$$

$$I_{ij} = -\sum_{ij} p(ij) \log_2 p(ij)$$

Ecuación 9: Informatividad (*informativeness*) en Redington *et al*. (1998)

I_i es la cantidad de información de la clasificación en función de la probabilidad de que un ítem sea asignado al cluster *i* de los resultados. Similarmente, I_{ij} es la cantidad de información en función de la probabilidad de que un ítem sea asignado al cluster *i* cuando el cluster *i* correponde a la categoría *j* del *benchmark*.

Este tipo de métricas basadas en la teoría de la información, ha sido ampliamente utilizado en otros trabajos (Christodoulopoulos *et al*. 2010) como una indicación confiable de la partición de clusters o su agrupamiento en hiperclusters (véase capítulo 7 de esta tesis).

Una vez establecidas la forma de leer los resultados y la referencia (*benchmark*) y las métricas con las que se evaluarán los mismos, los autores proponen una serie de nueve experimentos, analizando detalladamente las variaciones paramétricas del escenario de clustering.

5.4.0 Experimento 0 (inicial): Parámetros por default

Corpus: corpus original de 2,5 millones de tokens
Cues: 150 types más frecuentes
Palabras target: 956 types siguientes en el DFP
Contexto: bigramas y trigramas a derecha y a izquierda (600 dimensiones para cada target)
Corte del dendrograma: 0,8
Criterio de similitud: correlación de Spearman
Evaluación: precisión y cobertura
Salida: 37 clusters en total, 12 clusters conteniendo más de 10 miembros (Figura 15), 25 clusters indecidibles por miembros escasos
Comentarios: clusters no homogéneos, por ejemplo aparecen bastantes nombres propios en el cluster de conjunciones e interjecciones.

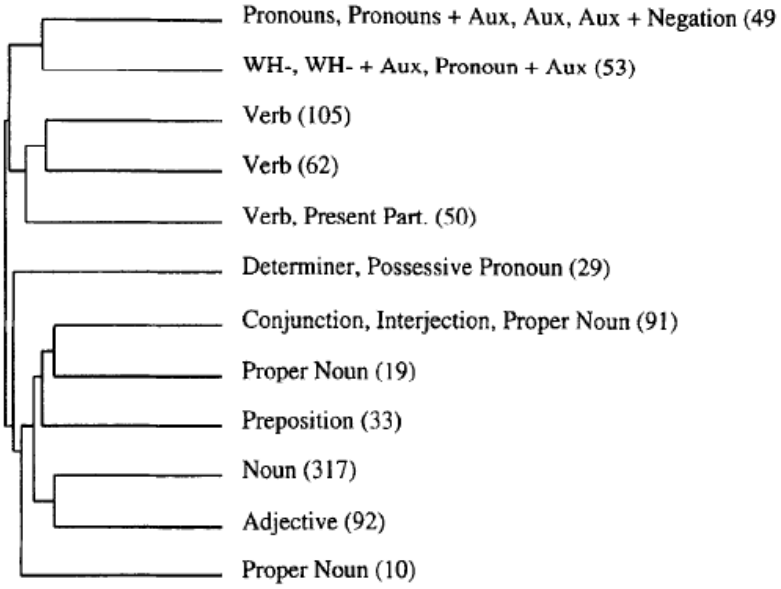

Figura 15: Dendrograma de salida del experimento 0

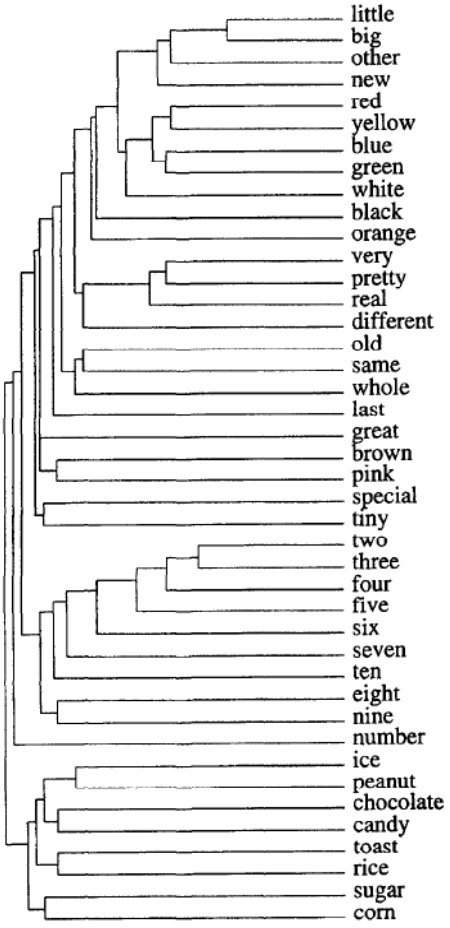

Figura 16: Parte de la estructura interna del cluster *Adjetivos*

5.4.1 Experimento 1: Diferentes contextos y diferentes coeficientes de corte

Contexto: se amplía la ventana de análisis: bigramas, trigramas, tetragramas y pentagramas siguientes a derecha (Figura 17) y precedentes a izquierda (Figura 18), cada uno de ellos **por separado.**

Corte del dendrograma: variable entre 0 y 0,9

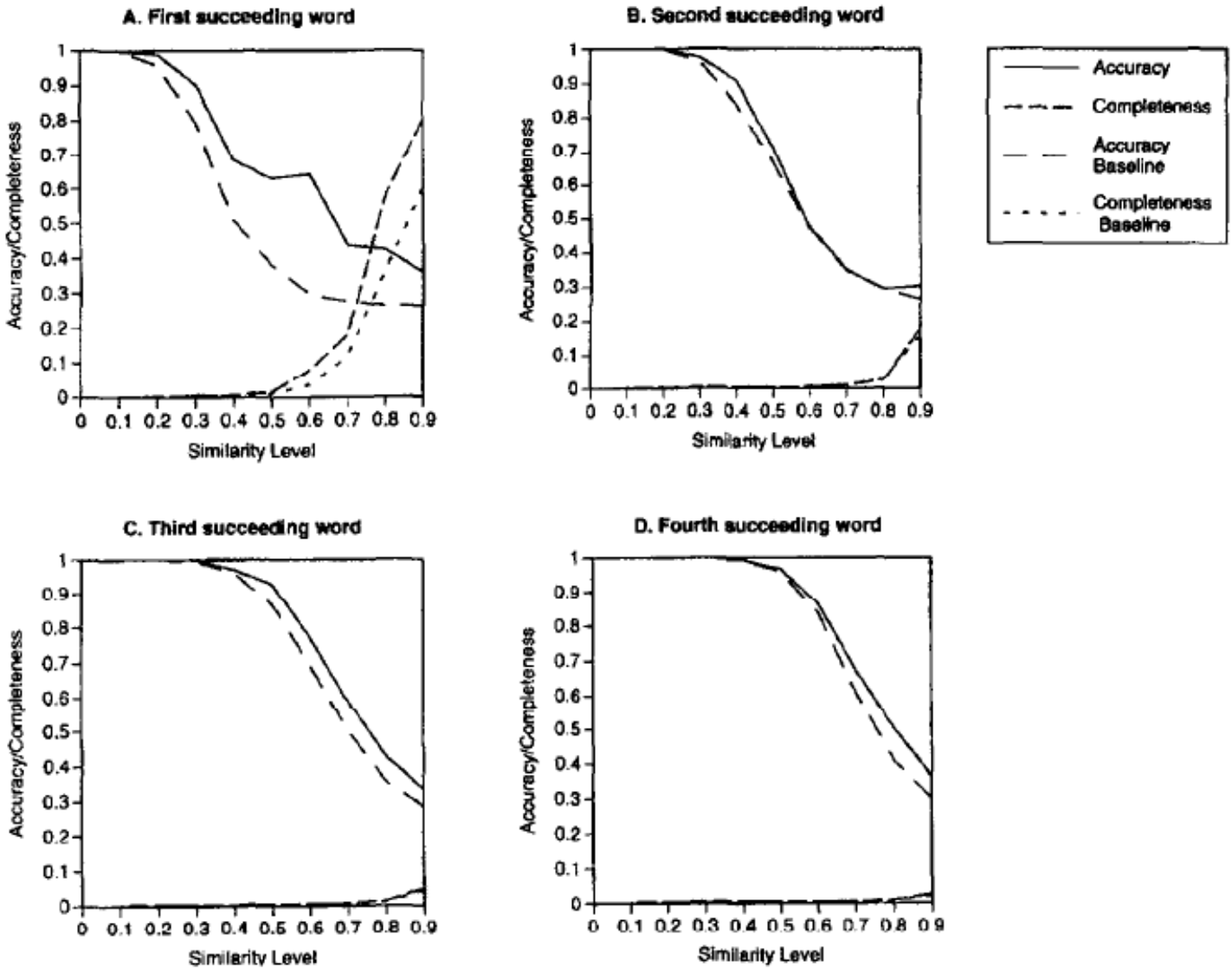

Figura 17: Salida del experimento 1 con contexto siguiente (Redington *et al*. 1998)

Comentarios: El contexto precedente es más informativo que el contexto siguiente. El uso de contextos más amplios (tetragramas y pentagramas) mejora la precisión, pero empeora la cobertura (porque a medida que crece el contexto crece también el número de posibles construcciones sintácticas). El contexto ideal, al nivel del dendrograma elegido, es la combinación de bigramas y trigramas, antes y después de la palabra *target* (parámetro por default del experimento inicial 0), con una precisión de 0,79 y una cobertura de 0,45. Este contexto local y pequeño impone una restricción al tipo de relaciones entre palabras y constituye una respuesta a la objeción de Pinker (1984), según la cual la infinita cantidad de relaciones posibles haría inútil el intento de obtener información válida al usar el enfoque distribucional. Las métricas de *baseline* surgen de corridas con asignación al azar.

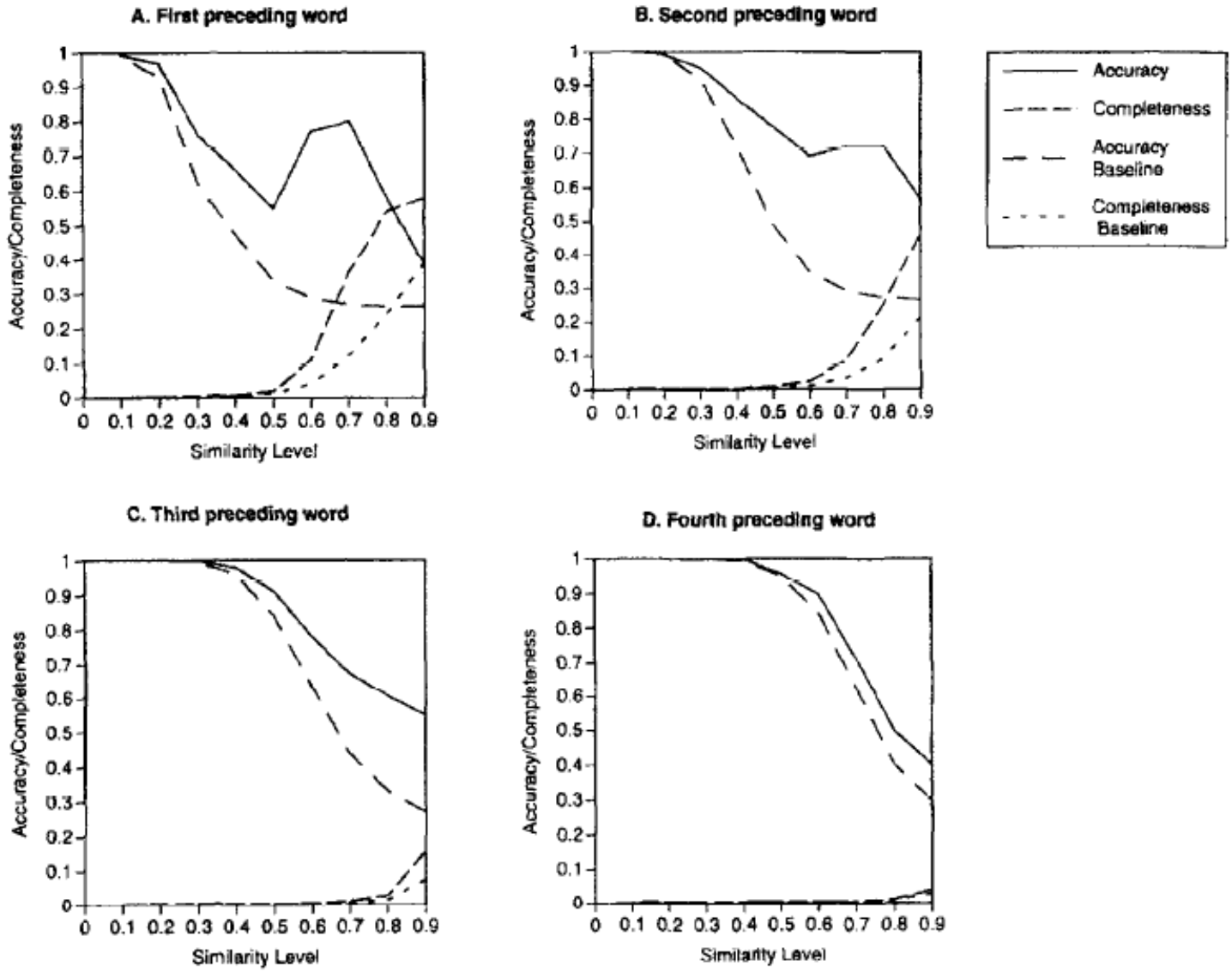

Figura 18: Salida del experimento 1 con contexto precedente (Redington *et al*. 1998)

5.4.2 Experimento 2: Variación en el número de palabras target

Palabras target: desde los 10 types siguientes a las 150 cues en el DFP hasta 2000 palabras en total

Comentarios: 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, que pertenecen a categorías cerradas), ni cuando hay muchas (porque la precisión del modelo aumenta a la vez que su cobertura decrece). El método funciona mejor cuando tanto la cantidad de palabras target como la de palabras contexto son reducidas y se condice con el número de palabras (aproximadamente 1000) que puede llegar a conocer típicamente un niño de tres años (Bates *et al*. 1994).

5.4.3 Experimento 3: Discrimanción de resultados del experimento inicial 0 según POS-tag

		Observed		Baseline	
Class	n	Accuracy	Completeness	Accuracy	Completeness
noun	407	.90	.53	.43	.14
adjective	81	.38	.45	.09	.16
numeral	10	.09	.82	.02	.27
verb	239	.72	.24	.25	.14
article	3	.10	1.00	.01	.51
pronoun	52	.25	.24	.06	.14
adverb	60	.17	.18	.07	.16
preposition	21	.33	.53	.03	.16
conjunction	9	.06	.33	.02	.24
interjection	16	.18	.67	.02	.20
complex contraction	58	.55	.47	.07	.17
Overall	956	.72	.47	.27	.17

Tabla 12: Precisión y Cobertura para cada POS-tag en el experimento 3 de Redington *et al*. (1998)

Comentarios: Será muy importante analizar esta tabla con la salida de nuestro experimento de clustering (véase *capítulo 7* de esta tesis).

5.4.4 Experimento 4: Variación del tamaño del corpus

Corpus: Variación entre 100.000, 500.000, 1 millón y 2 millones de tokens
Comentarios: Se observa una marcada mejora en correlación el aumento del tamaño del corpus a partir de 1 millón de tokens.

5.4.5 Experimento 5: Agregado de información de límite de oraciones en el corpus

Explicación: El experimento original de Redington *et al*. (1998) no contemplaba los límites de oraciones ni de frases. Sin embargo, existe numerosa evidencia empírica de que los niños acceden a esta habilidad fonética de segmentación de frases en forma muy temprana (Christophe *et al*. 2008). A su vez, plantear relaciones de contexto-target-contexto que atraviesen los límites de oraciones (y hasta de frases) representa un error de modelización . De este modo, en este experimento 5 los autores incorporan información de límite de oraciones como un símbolo más en la distribución de ocurrencias de bigramas y trigramas y luego vuelven a correr los parámetros del experimento inicial 0 a distintos niveles de corte.
Corte del dendrograma: variable entre 0 y 0,9
Evaluación: informatividad
Comentarios: Se observan mejoras en la métrica de informatividad con picos que se dan con un coeficiente de corte menor entre clusters en comparación con el experimento 0.

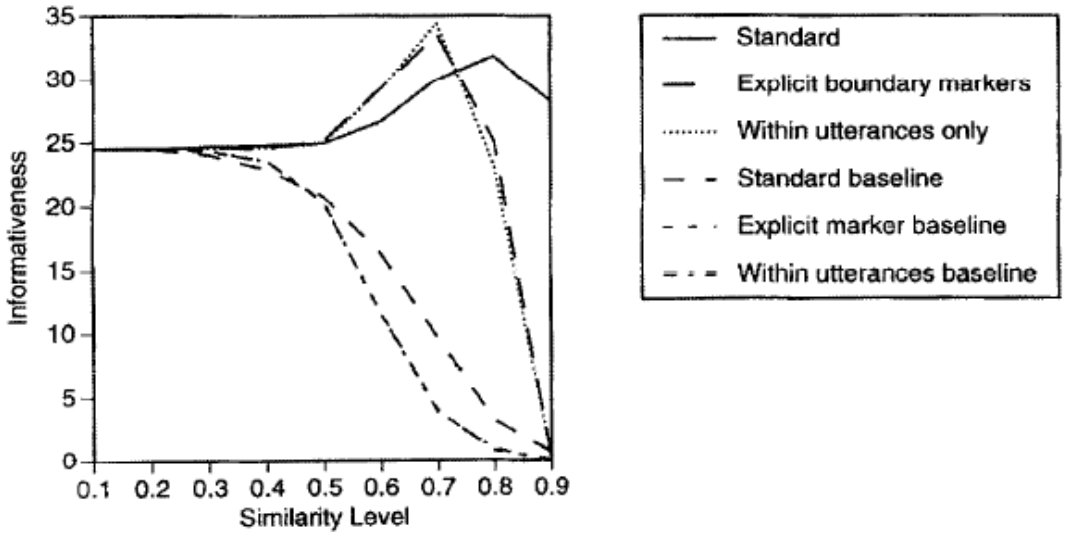

Figura 19: Salida del experimento 5 con información de límite de oraciones en Redington *et al*. (1998)

5.4.6 Experimento 6: Cambio en el criterio de similitud entre clusters

Corte del dendrograma: variable entre 0 y 0,9
Criterio de similitud: distancia Manhattan o city-block
Evaluación: informatividad
Comentarios: Se observa un descenso en las métricas de evaluación entre el experimento inicial con criterio basado en la correlación de Spearman y el experimento 6 con la distancia Manhattan o city-block (véase *Ecuación 2*). Esta sensibilidad de la evaluación a los criterios de corte es moderada.

5.4.7 Experimento 7: Remoción de las palabras funcionales del corpus

Comentarios: Se observa un marcado descenso en las métricas de evaluación entre el experimento inicial 0 y el experimento 7 con las palabras funcionales removidas del corpus. Como mencionamos anteriormente, las técnicas de clustering son particularmente efectivas en escenarios con una distribución de eventos muy frecuentes (Martin *et al*. 1998). Como las palabras funcionales en cualquier idioma suelen ser las palabras más frecuentes en cualquier corpus, la conclusión a que nos lleva este experimento no es inesperada.

5.4.8 Experimento 8: Cambios en la naturaleza del corpus

Corpus: Comparación de CHILDES con una muestra de igual tamaño del *British National Corpus* (BNC) con lenguaje entre adultos (no dirigido a niños).
Comentarios: No se observan cambios significativos en los valores de las métricas de evaluación.

5.4.9 Valoración general del trabajo de Redington et al. (1998)

Si bien el trabajo de Redington *et al*. (1998) fue uno de los primeros intentos sistemáticos en investigar técnicas de clustering sobre grandes corpora (Klein y Manning 2004), el experimento adolece de ciertas fallas de diseño que podrían resultar incompatibles con los lineamientos

epistemológicos del paradigma estadístico y podrían hacer mella en la plausibilidad psicolingüística de esta modelización. En particular, la crítica al algoritmo se centra en la identificación apriorística y arbitraria de las 150 palabras *cues*. La arbitrariedad en el diseño del algoritmo socava los mismos principios epistemológicos que el paradigma estadístico se propone defender. Idealmente, se esperaría que en un enfoque no supervisado fuera el propio algoritmo el que 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 (Headen *et al.* 2008). Justamente, a partir de esta crítica, presentaremos nuestro propio experimento de clustering en el capítulo 7 de esta tesis.

No obstante, es menester reconocerle al trabajo de Redington *et al*. (1998) un gran aporte en cuanto al estudio exhaustivo del papel concreto que juega la ventana de análisis (palabras contexto alrededor de las palabras target), como así también su rol fundacional en las modelizaciones computacionales del proceso de categorización de palabras de contenido con plausibilidad psicolingüística.

5.5 Martin et al. (1998)

Martin *et al*. (1998) proponen la continuación del experimento de Brown *et al*. (1992) en cuanto a un enfoque basado en categorías morfosintácticas que minimicen la perplejidad de modelos markovianos. Los aportes de Martin *et al*. (1998) se basan en extender la teoría inicial desde los bigramas a izquierda a los bigramas y trigramas precedentes y siguientes a la palabra target, como así también en definir mejores prácticas de análisis estadístico.

El algoritmo de Martin *et al*. (1998) es conoccido como *algoritmo de intercambio* (*exchange algorithm*) y es una variación del algoritmo *bottom-up* de clustering jerárquico que utilizaba Brown *et al*. (1992), en la que el proceso de agrupamiento de clusters era llevado a cabo iterativamente por pares. El *algoritmo de intercambio* de Martin *et al*. (1998) puede resumirse a :

1) Inicialización: Se elige el valor del parámetro inicial *G* (cantidad de clusters deseados) y se asigna cada una de las *G-1* palabras más frecuentes del corpus a un cluster distinto.

2) Las restantes palabras del vocabulario *V* son asignadas en su totalidad a otro cluster C_v.

3) Se computa la perplejidad del modelo markoviano de bigramas y de trigramas para iterativamente ir removiendo, insertando o intercambiando cada una de las palabras de C_v a las distintas clases de inicialización, de modo de obtener valores de perplejidad menores.

4) El criterio de finalización es: o bien se alcanza el número de iteraciones previstas (parámetro inicial) o bien no hay más palabras en C_v para intercambiar.

El corpus elegido para los experimentos consiste en tres diferentes cortes del corpus *Wall Street Journal* (WSJ) corpus con tamaños de 1 millón (1M), 4 millones (4M) y 39 millones

(39M) de tokens y 20.000 types (19.979+2 símbolos especiales de palabra desconocida y de límite de oración). Para la parametrización del modelo markoviano de bigramas se eligieron los valores de 50, 100, 200, 500, 1.000 y 2.000 clusters mientras que el modelo de trigramas se inicializó sucesivamente con 50, 100 y 200 clases.

$g = 2$	THE, JAPAN'S, YESTERDAY'S, BRITAIN'S, TODAY'S, CANADA'S, CHINA'S, FRANCE'S, MEXICO'S, ITALY'S, AUSTRALIA'S, ISRAEL'S, CALIFORNIA'S, TOKYO'S, TAIWAN'S, NICARAGUA'S, SWEDEN'S, POLAND'S, NASDAQ'S, TOMORROW'S, ...
$g = 12$	SAID, SAYS, ADDS, SUCCEEDS, CONTENDS, RECALLS, EXPLAINS, ASKS, PREDICTS, CONCEDES, SUCCEEDING, INSISTS, ASSERTS, WARNS, ADMITS, COMPLAINS, REPLIED, CONCLUDES, DECLARES, OBSERVES, ...
$g = 22$	BY, THEREBY
$g = 32$	PLANS, AGREED, EXPECTS, BEGAN, DID, MAKES, CAME, TOOK, GOT, DOES, CONTINUED, CALLS, HELPED, WANTS, DECIDED, WENT, MEANS, OWNS, FAILED, HOLDS, ...
$g = 42$	NEW, MAJOR, BIG, OLD, FULL, ADDITIONAL, SINGLE, NON, JOINT, LEADING, WIDE, DOUBLE, LEVERAGED, PRE, PARTICULAR, CONVENTIONAL, TRIPLE, COMPARABLE, FORT, GRAMM, ...
$g = 52$	U., JONES, BROTHERS, LYNCH, LEHMAN, STANLEY, HUTTON, SACHS, REYNOLDS, BACHE, PEABODY, INDUSTRIALS, STEARNS, HANOVER, WITTER, GENERALE, KRAVIS, LUFKIN, GUARANTY, GRENFELL, ...
$g = 62$	THAN, QUARTER, HALF, EIGHTHS, QUARTERS, EIGHTH, SIXTEENTHS, INTERSTATES
$g = 72$	BUSINESS, INTEREST, TAX, TRADE, DEBT, MONEY, CAPITAL, MANAGEMENT, WORK, CASH, GROWTH, PRODUCTION, POLICY, POWER, NEWS, CREDIT, TAKEOVER, SUPPORT, BUDGET, INFORMATION, ...
$g = 82$	INCORPORATED, CORPORATION, GROUP, UNIT, LIMITED, MAKER, INDUSTRIES, DIVISION, UNIVERSITY, HOLDINGS, SUBSIDIARY, PARTNERSHIP, CORPORATION'S, ASSOCIATES, INCORPORATED'S, OPERATOR, BANCORP, AFFILIATE, SUPPLIER, LABORATORIES, ...
$g = 92$	OFFICIALS, IT'S, ANALYSTS, TRADERS, EXECUTIVES,THAT'S, WE'RE, SOURCES, THERE'S, DEALERS, BANKERS, I'M, THEY'RE, HE'S, ECONOMISTS, BROKERS, STATISTICS, YOU'RE, EXPERTS, CRITICS, ...

Tabla 13: Ejemplos de miembros de los clusters resultantes por el modelo de trigramas para 100 clases (G = 100) sobre el corpus 39M

La evaluación de los clases inducidas se realiza sobre un corpus de evaluación (*test corpus*) con 324.655 tokens de texto no incluido en el corpus de entrenamiento (etapa de clustering). Se calcula la perplejidad para los modelos markovianos basados en bigramas y en trigramas basados en las clases inducidas, aplicando dichos modelos sobre este corpus de prueba. A menor perplejidad, mayor validación de las clases de palabras inducidas.

Classes	Class bigram			Class trigram		
	1 M	4 M	39 M	1 M	4 M	39 M
50	454.5	427.8	421.3	396.5	347.1	343.4
100	396.9	361.1	352.7	415.9	285.7	264.9
200	360.4	310.8	301.8	479.9	258.9	206.2
500	326.8	259.9	244.2		–	
1000	318.1	233.5	211.0		–	
2000	305.9	218.6	187.1		–	

Tabla 14: Perplejidad en corpus de evaluación para modelos markovianos según bigramas o trigramas de clases inducidas

Uno de los mayores problemas en la evaluación de los enfoques basados en modelos markovianos es la dispersión de datos, lo que resultaría en probabilidad 0 para algunos de los bigramas o trigramas de clases inducidas durante el entrenamiento que no se hallaran en el corpus de evaluación:

> "The dependence of the conditional probability of observing a word w_n at a position n is assumed to be restricted to its immediate (m-1) predecessor words $w_{n-m}+1 \ldots w_{n-1}$. The resulting model is that of a Markov chain and is referred to as m-gram model. For m=2 and m=3, we obtain the widely used bigram and trigram models, respectively. These bigram and trigram models are estimated from a text corpus during a training phase. But even for these restricted models, most of the possible events, i.e., word pairs and word triples, are never seen in training because there are so many of them. Therefore in

> order to allow for events not seen in training, the probability distributions obtained in these m-gram approaches are smoothed with more general distributions." [Martin *et al*. 1998:20]

El trabajo propone, entonces, un método de suavizado (*smoothing*) para lidiar con el problema de la dispersión de datos, basado en el método de interpolación absoluta con una distribución generalizada de instancia única (*absolute interpolation with a singleton generalized distribution*) (Ney *et al*. 1995).

Sin mucha sorpresa, ya que un mayor número de clases tiende a reducir la perplejidad de un modelo markoviano, notamos que la mejor validación en la Tabla 14 se da con bigramas, con parámetros de mayor granularidad de clases esperadas y con el mayor corpus de entrenamiento: es decir, el mejor modelo son las 2.000 clases inducidas actuando como bigramas en el corpus de 39 millones de tokens. Sin embargo, este resultado de validación dista mucho de ser verosímil, ya que es dudoso que el mejor escenario de categorización morfosintáctica de palabras implique miles de categorías para un vocabulario de apenas unas decenas de miles de palabras (*types*).

5.6 Clark (2000, 2002, 2003)

Alexander Clark (2000, 2002, 2003) es otro ejemplo de un trabajo exhaustivo sobre el problema de la inducción de categorías sintácticas a partir de técnicas de clustering. En particular, Clark (2002) incluye sus experimentos de induccción de categorías sintácticas en un esquema general de inducción de sintaxis (véase Figura 1) en su tesis de doctorado, como una reelaboración de Clark (2000).

El enfoque general de estos trabajos es tomar como datos una secuencia de fonemas o caracteres segmentados en palabras y oraciones y obtener como salida diversas clases morfosintácticas de palabras. La técnica utilizada, al igual que en todos los trabajos presentados hasta ahora, es considerar la información distribucional local de un palabra (su contexto inmediato) como fuente de información suficiente para posibilitar la inducción de clases de palabras morfosintácticas en inglés.

Clark identifica los trabajos previos de Brown *et al*. (1992), de Schütze (1993) y de Finch y Chater (1992) –similar a Redington *et. al* (1998) como los antecedentes más cercanos a su propio experimento. Sin embargo, se ocupa de marcar tajantes diferencias respecto al tratamiento de palabras ambiguas en cuanto a su POS-tag, algo que ni Brown *et al*. (1992) ni Finch y Chater (1992) habían trabajado y que Schütze (1993) había resuelto a medias, y respecto a la cobertura de palabras raras (*rare words*, incluso *hapax legomena*) para dar cuenta de todos los types target de un corpus. A pesar de su declamada cobertura exhaustiva, los modelos de Brown *et al*. (1992) y Schütze (1993) recortaban palabras *target* por arriba de un cierto umbral de ocurrencias. Finch y Chater (1992), al igual que Redington *et al*. (1998) directamente se enfocaba en clusterizar sólo palabras frecuentes.

Una innovación de Clark (2002) respecto de estos trabajos es la elección de un criterio de similitud entre palabras a ser clusterizadas diferente a los trabajos anteriores: la divergencia Kullback-Leibler (Kullback-Leibler Divergence KLD). A pesar de que KLD no es una métrica de distancia, debido a que no es simétrica y no satisface el criterio de desigualdad triangular (*triangle inequality*) (Abney 2008), Clark (2002) defiende su elección.

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -H(p(x)) - \frac{1}{n} \log(\prod_x q(x)^{c(x)})$$

Ecuación 10: Kullback-Leibler Divergence y redefinición en una entropía constante y una probabilidad máxima (*maximum likehood*) para un cluster

El corpus de Clark es una selección de 12 millones de tokens del *British National Corpus* (BNC). El algoritmo de Clark (2000, 2002), denominado Clustering de Distribución de Contexto (*Context Distribution Clustering* CDC), puede resumirse a un procedimiento general y otros dos específicos para palabras ambiguas y para palabras raras (de escasa ocurrencia). El procedimiento general consta de los siguientes pasos:

0) Preparación del corpus: se agrega información de límite de oración &SENTENCE y de elipsis &HELLIP como POS-tag relevantes a ser clusterizados.

1) Seleccionar las K^{th} palabras más frecuentes del corpus y asignarlas a K clusters (parámetro inicial instanciado en tres escenarios de 77, 100 y 150 clusters, respectivamente). No obstante, Clark (2002) prefiere el escenario de 77 clusters por ser el más semejante al set de etiquetas morfosintácticas con que cuenta el BNC (estándar CLAWS-4), lo que facilitará la comparación final en la evaluación.

2) En cada iteración calcular la distribución de contextos (bigramas a cada lado de la palabra target) de cada cluster (KLD), que es el promedio ponderado con respecto a los *K* clusters del ciclo y un cluster "reservorio" en donde quedan las palabras no clusterizadas por debajo de las K^{th} iniciales. En este reservorio inicial, que iterativamente irá incluyendo menos y menos términos, se encuentran las palabras "generales" (aquellas que superan un umbral de 50 ocurrencias en el corpus).

3) Para una palabra target del reservorio se ordenan los valores de KLD de cada cluster en forma creciente y se asigna al cluster con el valor de KLD más bajo.

4) Se repite iterativamente el ciclo desde el paso 2) hasta cubrir un porcentaje del reservorio (80%) o hasta que la KLD entre dos clusters caiga por debajo de un umbral, en cuyo caso ambos clusters son agrupados y se forma un úlitmo cluster con el resto de las palabras del reservorio.

HOWEVER OF_COURSE FOR_EXAMPLE INDEED
NEVERTHELESS
FAR INFINITELY
LATER AGO EARLIER THEREAFTER
ENOUGH
IMPORTANT POSSIBLE CLEAR HARD CLOSE
NEW OTHER FIRST OWN GOOD
LAST NEXT GOLDEN FT-SE
BETTER WORSE LONGER BIGGER STRONGER
YEARS PER_CENT DAYS TIMES MONTHS
GROUP NUMBER SYSTEM OFFICE CENTRE
PEOPLE WORK LIFE RIGHT END
WORLD GOVERNMENT PARTY FAMILY WEST
US BRITAIN LONDON GOD LABOUR
TIME WAY YEAR DAY MAN
PART SORT THINKING LACK NONE
NEED NEEDS SEEM ATTEMPT OPPORTUNITY
FACT IMPRESSION ASSUMPTION IMPLICATION
MR MRS DR HONG MR.
CHARLES MARK PHILIP HENRY MARY
JOHN SIR DAVID ST DE
KLERK CLOWES HOWE COLI GAULLE
ARE WERE
BE HAVE DO MAKE GET
WANT WANTED TRIED WISH WANTS
MADE USED FOUND LEFT PUT
BASED RESPONSIBLE COMPARED INTERESTED ASSOCIATED
SAID SAYS WROTE EXPLAINED REPLIED
THOUGHT FELT KNEW DECIDED HOPE
ASKED LIKED WATCHED SMILED INVITED
CAME WENT LOOKED SEEMED BEGAN
TOOK TOLD SAW GAVE MAKES
LOOK RUN LIVE MOVE TALK
USE HELP FORM CHANGE SUPPORT
THINK BELIEVE SUPPOSE INSIST RECKON
GO COME TRY CONTINUE APPEAR
SEE SAY FEEL MEAN REMEMBER
KNOW UNDERSTAND REALISE
GOING ABLE LOOKING TRYING COMING
COPE DEPEND CONCENTRATE SUCCEED COMPETE
SUCH USING PROVIDING DEVELOPING WINNING

Figura 20: Ejemplo de clusters de palabras de contenido (cada línea es un cluster) con los 5 miembros más frecuentes

Para el caso de palabras ambiguas, Clark (2002) sostiene que se identifican naturalmente a partir de una combinación lineal de las distribuciones de contexto para varios clusters, como el coeficiente α de la siguiente fórmula:

$$\alpha_i^{new} = \sum_x p(x) p^{old}(i|x) = \sum_x p(x) \frac{\alpha_i^{old} q_i(x)}{\sum_i \alpha_i^{old} q_i(x)}$$

Ecuación 11: Cálculo de coeficiente de pertenencia de una palabra ambigua a diversos clusters

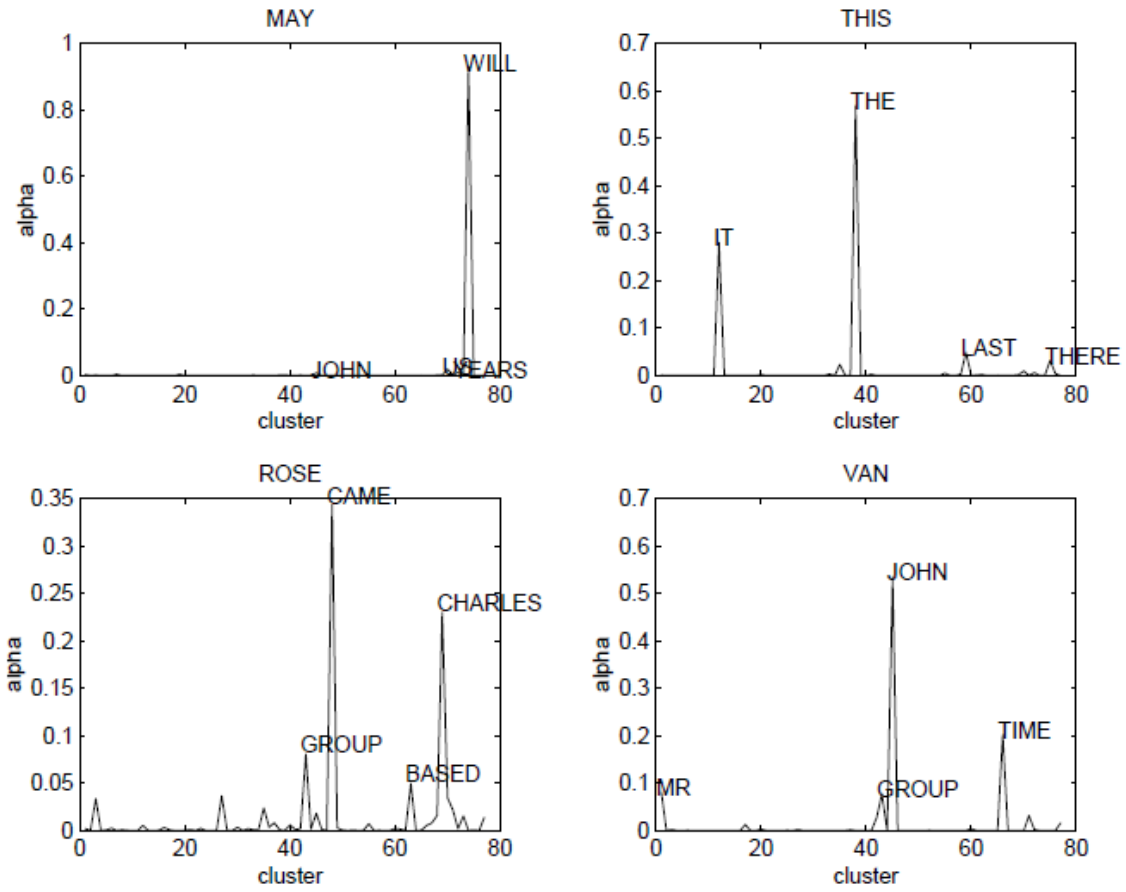

Figura 21: Ejemplo de palabras ambiguas. Cálculo de coeficiente α de pertenencia de palabras ambiguas a diversos clusters.

Para el caso de palabras raras (menos de 50 ocurrencias en el corpus), Clark (2002) propone calcular la probabilidad posterior de que la palabra esté en cada cluster en función del número de ocurrencias. El autor encuentra particularmente exitosa la comparación de la evaluación de esta asignación de palabras raras a clusters para el caso de types de hasta 5 ocurrencias en el corpus.

Para la evaluación general del experimento de Clark, el autor descarta el recurso de un gold standard (*benchmarking*), como en el caso de Redington *et al*. (1998) y, en cambio, apela al cálculo de la perplejidad (a menor perplejidad, mejor inducción de clusters) de un modelo markoviano de trigramas entrenado con los 77 clusters inducidos y evaluado contra cuatro sets de 100.000 tokens del BNC (texto no contenido en el corpus original). Clark (2002) ofrece sus métricas de evaluación (*CDC* en la siguiente tabla) contra el mismo modelo markoviano

entrenado según los clusters del algoritmo de Brown *et al*. (1992) y según el listado de etiquetas del BNC (*CLAWS* en la siguiente tabla) disponible para cada uno de los cuatro sets de evaluación (*test set* en la siguiente tabla):

Test set	1	2	3	4	Mean
CLAWS	411	301	478	413	395
Brown et al.	380	252	444	369	354
CDC	372	255	427	354	346

Tabla 15: Perplejidad de modelos markovianos como métricas de evaluación para 3 modelos y para cada set con la respectiva media geométrica (*mean*)

En un trabajo posterior, Clark (2003) propone utilizar información morfológica en forma previa a su algoritmo CDC, bajo la forma de un modelo markoviano de bigramas de caracteres, considerando también información de frecuencia de la palabra target a clusterizar. Clark (2003) invierte así el orden de las etapas iniciales del algoritmo de inducción de sintaxis de su propio trabajo (Clark 2002) (véase Figura 1), obteniendo resultados levemente más prometedores. En inglés, esta modificación redundará en una leve mejora (Graça *et al*. 2011), pero presumiblemente puede resultar más beneficiosa para lenguajes ricos en morfología como el español:

> "Languages with comparatively rich morphology tend to have rather free word order, which might cause problems with distributional induction techniques. However these languages tend to signal the part of speech in the surface form of words, so it would be possible to use that information to learn. What would create serious problems is a very free word order language with very limited morphology: fortunately such languages seem not to exist." [Clark 2002:76]

En dos interesantes trabajos de homologación de la evaluación de diversos modelos bajo las mismas condiciones, Christoudoulopoulos *et al*. (2010) determinan que el algoritmo de Clark (2002) es uno de los mejores con las métricas de evaluación basadas en los mapeos *many-to-1* y *1-to-1*, alcanzando medidas F de 71.2%. A su vez, Graça *et al*. (2011) comprueban una medida F de 72.4% para el algoritmo de Clark (2003), que incluía información morfológica previa, en la evaluación *many-to-1*. Estas métricas de evaluación sitúan el trabajo de Clark como uno de los principales aportes al problema de inducción de categorías sintácticas a partir de técnicas de clustering sobre información distribucional.

5.7 Investigaciones actuales a partir de los trabajos fundacionales

A partir de los relevamientos de los trabajos fundacionales podemos observar que básicamente existen dos tradiciones: por un lado, clases de palabras inducidas a partir de modelos markovianos (*Hidden Markovian Models* HMM) en conjunción con técnicas de clustering (Brown *et al*. 1992; Martin *et al*. 1998), y por el otro, enfoques puros basados en

técnicas de clustering (Finch y Chater 1992; Schütze 1993; Redington *et al*. 1995, 1998; Clark 2000, 2002, 2003).

En los últimos años se han presentado algunos trabajos que adhieren a una u otra tradición y que han validado las premisas de los modelos con evidencia translingüística (Christodoulopoulos *et al*. 2010; Graça *et al*. 2011). En algunos casos puntuales, ciertos trabajos ofrecieron extensiones metodológicas respecto de los trabajos fundacionales que merecen ser mencionadas sucintamente, tales son los casos de Berg-Kirkpatrick *et al*. (2010) y de Nath *et al*. (2008).

Berg-Kirkpatrick *et al*. (2010) proponen un modelo basado en HMM tradicional pero asumiendo distribuciones de emisión logísticas antes que multinomiales, lo cual posibilita el agregado de features discriminativas como morfología (trigramas de caracteres) y capitalización (distinción mayúsculas/minúsculas). El modelo reporta altísimas métricas de evaluación contra el corpus del *Wall Street Journal* (WSJ), con una medida F en el mapeo *many-to-1* de 75.5%, superando incluso la performance del trabajo de Clark (2002, 2003) y elevando la barra en el estado del arte para la cuestión. Sin embargo, cabe mencionar que la consideración de features tales como capitalización definitivamente aleja estos modelos de la plausibilidad psicolingüística de las investigaciones originales, puesto que no es viable suponer que tal feature discriminativa esté disponible para los adquirientes de una lengua en la modelización de los PLD.

Por su parte, Nath *et al*. (2008) presentan un innovador trabajo para la inducción de categorías en bengalí que adhiere más bien a la tradición purista de la técnicas de clustering, aunque con el agregado de un exhaustivo análisis de las propiedades topológicas emergentes de un corpus, organizado como un red (*network*) de palabras:

> "A key to the identification of these natural word classes is to understand the syntactic structure of a language, which is captured through the complex interaction of the words. This is arguably an outcome of a self-organizing process governing the dynamics of language and grounded in the cognitive abilities of human beings […] In this context, language can be viewed as a network of words and formation of lexical categories an emergent property of this network. Thus, understanding the structure and function of this network will help us in procuring deeper insight into the nature of word classes in a given language." [Nath *et al*. 2008:1220]

Los investigadores plantean en principio un enfoque basado en dos técnicas de clustering: el clustering aglomerativo jerárquico tradicional y el algoritmo *Chinese Whispers* (Biemann 2006a):

> "The Chinese Whispers algorithm (Biemann 2006a) is a non-parametric random-walk based clustering algorithm, where initially each node is in a separate cluster. In every iteration, the nodes propagate information about their current cluster to all the neighbors, and in turn, decide upon their own cluster labels based on a weighted majority voting of the cluster information received from the neighbors. The algorithm terminates when the labels do not change considerably over successive iterations." [Nath *et al*. 2008:1223]

Una vez organizada la red de palabras como clusters con estructura de comunidad, los autores estudian diversas propiedades topológicas de la red, tales como la *distribución de grados* (*Cumulative Degree Distribution* CDD) -la cantidad de conexiones que tiene un nodo- y el *coeficiente de clustering* de un nodo de la red -la probabilidad de que los miembros vecinos a una ocurrencia elegida al azar de la palabra nodo sean, a su vez, vecinos entre sí. Los autores

sostienen que las propiedades topológicas de la red reflejan ciertas propiedades sintácticas de las palabras que la conforman:

> "Power-law networks are believed to have a self-similar hierarchical structure. In this case, the hierarchy is a reflection of syntactic ambiguities. Highly ambiguous words that belong to several lexical categories have the highest degrees. The next level of hierarchy is manifested by words that belong to a few lexical categories, whereas the last level of hierarchy is represented by the words that are unambiguous in nature. The power-law indicates that there are few words that belong to a large number of lexical categories, while the most of the words belong to only one lexical category." [Nath *et al*. 2008:1223]

Si bien el experimento central de esta tesis (véase *capítulo 7*) está enmarcado en la tradición iniciada por Redington *et al*. (1998) en cuanto a un enfoque tradicional y purista basado en las técnicas de clustering, es importante tener presentes también lineamientos alternativos del estado actual de la cuestión, tal como los hemos relevado a lo largo de este extenso capítulo. Oportunamente, entonces, retomaremos estos temas con la justificación de las decisiones de diseño y de las técnicas de evaluación para nuestro propio experimento de clustering.

Capítulo 6. Una propuesta conciliatoria entre la psicolingüística y la lingüística computacional (Wang 2012)

6.1 Categorización temprana de palabras funcionales

Hasta ahora hemos repasado los trabajos principales en categorización de palabras de contenidos basada en la información distribucional del contexto de la palabra target. Apuntamos que existen dos tradiciones experimentales que se corresponden mayormente con los paradigmas científicos dominantes en la investigación de los campos de la psicolingüística y la lingüística computacional. Por un lado, las técnicas de clustering como manifestación del paradigma estadístico de la lingüística computacional, ya sea en enfoques puros (Schütze 1993; Redington *et al.* 1998) o combinados con modelos markvianos (Brown *et al*. 1992; Martin *et al*. 1998); por el otro, las teorías de los marcos frecuentes (Mintz 2003) y los protoconstituyentes sintácticos (Christophe *et al.* 2008), con una raigambre simbólica proveniente de la psicolingüística. Estos trabajos, como se ha argumentado en los capítulos anteriores, muestran, en mayor o menor medida, ciertas falencias al momento de compatibilizar sus postulados teóricos con la evidencia empírica ontogenética de la habilidad temprana de categorización de palabras (alrededor del año y medio). Por lo menos en dos de los modelos detallados (Redington *et al.* 1998 y Christophe *et al.* 2008), las palabras funcionales cumplen un rol crucial; pero mientras que en Redington *et al.* (1998) el problema es la arbitrariedad apriorística de una defínción de clase que abarque a las palabras funcionales (dónde se encuentra el corte en el *Perfil de Frecuencia Decreciente*), en Christophe *et al.* (2008) el problema se traduce en la incosistencia temporal de una tipología granular de las categorías funcionales para el momento en que despunta la habilidad temprana de categorización de palabras (no es viable asumir la habilidad en el adquiriente de manipular frases verbales y nominales para el año y medio de vida).

Con este estado de la cuestión en las investigaciones empíricas y teóricas de la tarea lingüística de categorización temprana de palabras, entra en escena un trabajo innovador por sus objetivos de investigación. Enmarcada tanto en la psicolingüística como en la lingüística computacional, la tesis de doctorado de Wang (2012) se propone como uno de los trabajos pioneros en investigar específicamente la habilidad de categorización de palabras funcionales como pre-requisito para la categorización de palabras de contenido:

> "Word categorization is necessary to map words onto categories for language acquisition. There have been many research on categorizing content words (*e.g.*, nouns and verbs). Currently there are three major theoretical approaches to the categorization problem: semantic bootstrapping, phonological bootstrapping and distributional bootstrapping. This dissertation focuses on the distributional approach. Several distributional cues (including bigrams, frequent frames and preceding function words) were shown to be informative for categorizing nouns, verbs and some adjectives in typologically different languages. Although the above-mentioned distributional patterns were able to capture a few groups of function words or morphemes, most of the groups or frames were dominated by lexical items. No research to date has specifically studied the categorization of functional items using distributional information." [Wang 2012:62]

Como se explicó anteriormente en la sección *3.1 La naturaleza de los indicios facilitadores*, las tareas de modelización de categorización de palabras típicamente consideraban tres fuentes de indicios facilitadores: información fonética-fonológica, información semántica e información distribucional. El trabajo de Wang (2012) descarta la posibilidad de que la información semántica cumpla algún rol en el caso de la categorización temprana de palabras funcionales y, por lo tanto, se focaliza en la posibilidad de que los indicios prosódicos ayuden muy tempranamente a distinguir palabras funcionales en contraposición con las de contenido (Shi 1995) para que, luego, la información distribucional actúe como facilitadora del agrupamiento más refinado de categorías de palabras funcionales. Esto evita la circularidad de tener que definir palabras funcionales en función de la información distribucional de palabras de contenido, las que, a su vez son definidas a partir de las palabras funcionales:

> "Previous research has shown that there are some acoustic, phonological and prosodic cues that distinguish function words from content words (Shi *et al.* 1998). However, function words belong to a number of different functional categories (such as determiner, auxiliary, conjunction, preposition and postposition). Those cues may not be able to make finer distinctions between functional categories. Semantics are not so helpful here because 1) function words usually carry less meaning than content words. It would be difficult to form functional categories based on their meanings; 2) children may not have learned the meaning of function words when they start to categorize words because some researchers claim that the reason for their initial omission in production is that no meaning is attached to the form and they only use it when they have learned both form and meaning […] The hypothesis is that distributional information in the input is a very informative cue for categorizing function words if they could be differentiated from lexical items using acoustic cues." [Wang 2012:62-63]

6.2 Omisión sistemática de categorías funcionales en el "discurso telegráfico" de los niños

Al comenzar a investigar su hipótesis, Wang (2012) se topa con un problema metodológico amenazante. En efecto, su objeto de estudio raramente se manifiesta en el habla temprana de sus informantes:

> Young children's early speech often lacks functional elements such as function words (e.g., determiners in English) and inflectional morphemes (*e.g.*, past tense *-d* and plural -s morphemes in English). Early speech was characterized as 'telegraphic speech' […] The process that is most relevant to the current discussion is the first process, imitation and reduction, in which children's imitation of mothers' speech often omits the functional elements. ." [Wang 2012:18-19]

Como las omisiones resultan siempre muy sistemáticas, Wang (2012) logra explicar sastisfactoriamente esta tendencia a la imitación telegráfica del lenguaje materno a partir del concepto de Longitud Promedio de los Enunciados (*Mean Length of Utterance* MLU) y su consiguiente reformulación a Longitud Promedio de los Enunciados medido en morfemas (*Mean Length of Utterance in morphemes* MLUm) como parámetros de la evolución ontogénetica que podrían restringir las posibilidades combinatorias de estos enunciados en producción:

> "Children omitted words so their production is usually between two to four morphemes. In other words, some words were omitted in order to keep the length constraint. The underlying mechanism that is responsible for the length constraint and the initial omission is still uncertain. One can explain that the length constraint originates from certain domain-general cognitive processing or working memory limitations." [Wang 2012:21]

Algunos estudios tratan de explicar esta omisión de palabras funcionales en producción a partir de postular que las palabras de contenido son adquiridas mucho antes que las palabras funcionales:

> "Researchers have been trying to understand why those functional elements are often omitted in early speech. Brown (1973) evaluated three possibilities: 1) content words with semantic meaning are easy to learn; 2) content words carry more information; 3) content words often have heavier stress which is easier to notice by children." [Wang 2012:22]

Por lo tanto, el discurso telegráfico de los niños de 2 años en la etapa I de la adquisición de palabras funcionales (Brown 1973) (véanse *Tabla 16* y *Tabla 17*) podría explicarse como una saturación por parte de las palabras de contenido de la extensión máxima disponible en MLUm como recurso cognitivo para esa etapa de desarrollo. El MLUm de la etapa I es de apenas 1,75 morfemas, con lo que bastarían dos morfemas para saturar la capacidad productiva de los enunciados del niño.

> "In Brown's (1973) classical study of the development of 14 grammatical morphemes, the development was divided in to five stages according to the children's MLUm with Stage I has an MLUm of 1.75, Stage II 2.25, Stage III 2.75, Stage IV 3.50, Stage V 4.00. Brown suggested that when MLUm is over 4.0 it is not an accurate measure of linguistic knowledge anymore." [Wang 2012:16]

> "Children typically begin this stage by juxtaposing two words with equal intonation on both and a pause between them, as if each were a word being pronounced in isolation. *Mommy . . . Sit.* Following that, children combine words into what appear to be rudimentary sentences, with no pauses between the words and falling intonation at the end. Some examples are given in Tager-Flusberg (1997): *more car*, *more read*, *no pee*, *bye-bye Papa*, *there potty*, *Mommy stair*. You will notice that these utterances tend to be dominated by content words, often nouns, adjectives, and verbs. For this reason, these forerunners of adult sentences have been labeled telegraphic, because they exhibit the same economy of expression that telegrams did when they served as a form of urgent and expensive communication. Most function words, such as prepositions and helping verbs, are missing from children's initial two-word utterances, but some of the more salient ones do occur, such as more, no, and off." [Barry 2002:173]

Parent	Child
There goes one	There go one
Daddy's brief case	Daddy brief case
Fraser will be unhappy	Fraser unhappy
He's going out	He go out
That's an old time train	Old time train
It's not the same dog as Pepper	Dog Pepper
No, you can't write on Mr. Cromer's Shoe	Write Cromer shoe

Tabla 16: Ejemplos de discurso telegráfico de niños en etapa I de adquisición de palabras funcionales (Wang 2012)

Es decir, nuevamente nos topamos con la idea de que las palabras funcionales presentan una función articulatoria (*pivot*) o de *pegamento (glue*) de las palabras de contenido (Elghamry 2004) y, por lo tanto, tienden a ocurrir en la sintaxis naturalmente entre ellas. Salvo casos excepcionales como "*más cereal*" o "*sin manos*", etc. será imposible esperar que se den en

producción enunciados de dos palabras durante esta etapa I del proceso de adquisición de palabras funcionales (Brown 1973):

> "There is no recognizable stage that marks the transition from two-word to multiple-word utterances. Once children get the idea of syntax, they may combine more than two words at a time, as in Goodluck's examples: *clock on there*, *kitty down there*, *other cover down there*, *up on there some more* (Goodluck 1991). Children's syntactic growth during this period is measured by the mean length of utterance (MLU), calculated according to the average number of morphemes per utterance. **Although children may develop at very different rates, when their utterances approach a MLU of about 2.0, they begin to add the grammatical «glue» that holds together adult sentences, such as tense and number markers, possessive markers, helping verbs, and certain prepositions. This marks the transition to the next stage of development, what we might term the grammatical morpheme stage**." [Barry 2002:174] (*las negritas y el subrayado son nuestros*)

	Adam		Sarah
I (2;3)		I (2;3)	
II (2;6)	Present progressive *in* *on*, plural	II (2;10)	Plural *in, on* Present progressive, past irregular Possessive
III (2;11)	Uncontractible copula, past irregular	III (3;1)	Uncontractible copula Articles
IV (3;2)	Articles Third person irregular, possessive	IV (3;8)	Third person regular
V(3;6)	Third person regular Past regular Uncontractible auxiliary Contractible copula Contractible auxiliary	V(4;0)	Past regular Uncontractible auxiliary Contractible copula Third person irregular Contractible auxiliary

Tabla 17: Estadios temporales para la adquisición de palabras funcionales del inglés en dos niños (Brown 1973)

No obstante, para Wang (2012) el que las palabras omitidas en la reproducción del discurso materno sean sistemáticamente las palabras funcionales es una evidencia indirecta de que al menos en comprensión ya hay una representación rudimentaria de dicho tipo de palabras mucho antes de los 2 años de edad (Hochmann *et al*. 2010), en el momento indicado para propiciar la categorización temprana de palabras de contenidos y también la posterior explosión de vocabulario. Lo anterior no significa que las palabras de contenido iniciales no sean más fáciles

de aprender, como bien sostiene Brown (1973), sino que simplemente se sugiere que la ausencia sistemática de palabras funcionales en producción no necesariamente implica la ausencia de dicho tipo de palabras en los procesos activos de comprensión:

> "Shi and Melançon (2010) tested young children's knowledge of determiners and whether they can generalize between different determiners. […] Therefore, the result indicates that **children as young as 14 months old** are treating some determiners as a group so they could generalize and transfer knowledge of co-occurrence statistics between determiners. **It further suggests that there is a primitive, if not completely abstract, determiner category in the grammar of young children**." [Wang 2012:28] (*las negritas y el subrayado son nuestros*)

Así pues, Wang (2012) se pregunta acerca de la naturaleza de esa presunta representación rudimentaria temprana de las categorías funcionales a la luz del paradigma innatista y del paradigma empirista (constructivista):

> "One of the much debated problems is whether children's early syntactic representation resembles adults' syntactic representation (*i.e.*, the issue of continuity), which would also have consequences on linguistic theories. In the case of functional categories, the question is whether functional categories in adults' grammar are available to young children in early stage of development. Nativists argue that children are predisposed with notions of some abstract syntactic categories (such as noun, verb and determiner), though the contents of those categories are yet to be determined from the language experience […]. From nativists' perspective, young children would possess the same repertoire of abstract syntactic categories as adult speakers. Constructivists argue that children build the abstract categories by learning limited-scope formulae and generalize them to adult like syntactic categories […] Hence, children's early knowledge of syntactic categories (if any) is qualitatively different from adult speakers." [Wang 2012:5-6]

Esta disyuntiva acerca de la naturaleza del conocimineto rudimentario temprano sobre las categorías funcionales no puede ser zanjada fácilmente, aunque Wang (2012) aporta cierta evidencia que apoyaría la postura innatista a partir de la cuestión de la continuidad (*continuity*) en el repertorio y uso de categorías sintácticas entre niños y adultos. En efecto, para cuantificar dicho conocimiento inicial presuntamente innato, los investigadores recurrieron típicamente al concepto de *solapamiento* (*overlap*) en el uso de las categorías funcionales, comparando el discurso de los niños con el de sus progenitores (Pine y Martindale 1996, Valian *et al*. 2009):

> "If a child have abstract knowledge of functional categories, it is expected to observe that he or she will use both determiners a and the before nouns to a degree that resembles adults' usage. Overlap of *a* and *the* is computed by dividing the number of noun types that were used with both *a* and *the* by the number of noun types that were used with either *a* or *the* [Ecuación 12]. The computation of overlap can be easily generalized to all determiners [Ecuación 13]. […] The overlap test, proposed in Pine & Martindale (1996), compares children's overlap to adults' overlap (usually their mothers) to determine the time point when children have possessed the knowledge of abstract categories […]." [Wang 2012:36]

$$Overlap_{a\_the} = \frac{\#\,nouns\ occurring\ with\ \textbf{both}\ a\ and\ the}{\#\,nouns\ occurring\ with\ \textbf{either}\ a\ or\ the}$$

Ecuación 12: Solapamiento (*overlap*) entre '*a*' y '*the'*

$$Overlap_{all\_det} = \frac{\#\,nouns\ occurring\ with\ \textbf{any two determiners}}{\#\,nouns\ occurring\ with\ \textbf{any determiner}}$$

Ecuación 13: Solapamiento (*overlap*) para toda la clase de determinantes

No obstante, como bien apunta Wang (2012), el concepto tradicional de *overlap* es muy sensible al tamaño de la muestra, por lo que Wang reformula la métrica como *solapamiento esperado* (*expected overlap*), analizando la desviación estándar entre el *solapamiento esperado* y el *solapamiento observado* como un indicador de cuánto se parecen entre sí los usos de las categorías funcionales en el discurso de los niños y en el de sus progenitores:

$$Expected\ Overlap = \left\lfloor 0.5 + \sum_{All\ nouns} p(overlap) \right\rfloor$$

$$Deviation = \frac{Expected\ Overlap - Actual\ Overlap}{Expected\ Overlap}$$

Ecuación 14: Reformulación de *solapamiento esperado* (*expected overlap*) y desviación estándar. Ejemplo para sustantivos.

Con esta métrica Wang (2012) analiza ocho sets del corpus CHILDES para el inglés y corpora espontáneos del alemán, concluyendo que los resultados sugieren que los aprendientes de aquellos dos idiomas manifiestan una desviación estándar en el uso de los determinantes que replica la de sus progenitores:

> "This analysis demonstrates that children's actual overlap deviates from expected overlap to the same degree as their mothers. The results suggest that English-learning young children use determiners in the same way as adults do." [Wang 2012:49]

En resumen, Wang (2012) aporta evidencia indirecta (omisión sistemática en producción) de una adquisición muy temprana (antes del año y medio de edad) de las categorías funcionales y evidencia directa de conitinuidad en el repertorio y uso de categorías funcionales entre adultos y niños de alrededor de 2 años de edad. Estas dos observaciones habilitan a considerar a las palabras funcionales como candidatos ideales para facilitar (*boostrapping*) la categorización de palabras de contenido que se evidencia en la explosión léxica (*vocabulary spurt*) alrededor de los 2 años:

> "The analyses showed that young children at the first half of their second year of life already possess abstract knowledge of some functional categories like determiner. Even before the second birthday, they already process function words/morphemes as abstract categories (Shi & Melançon, 2010). When they start producing combinatorial speech, they are able to quickly generalize nouns to different determiners. These evidence strongly suggests that they are actively using abstract knowledge in language processing and production […].
> Their early sensitivity to function words/morphemes and abstract knowledge of functional categories indicate that they can use functional items to categorize nouns and verbs." [Wang 2012:60-61]

Resta, pues, modelizar plausiblemente cómo podrían ser categorizadas esas palabras funcionales inicialmente a partir de información distribucional. A ello se aboca Wang (2012) en la preparación de sus dos experimentos principales, uno de naturaleza más estadística (clustering jerárquico en el experimento 1) y otro de naturaleza más simbólica (marcos frecuentes en el experimento 2).

6.3 Experimento 1 de Wang (2012): clustering jerárquico sobre categorías funcionales

Wang (2012) adapta un experimento anterior de Mintz *et al.* (2002) considerando como palabras target las palabras funcionales de cuatro categorías: determinantes, pronombres, preposiciones y verbos auxiliares/modales. Para ello, toma el mismo corpus de 8 informantes en inglés de CHILDES que Mintz *et al.* (2002) y filtra un listado de 82 palabras target en función de un Perfil de Frecuencia Decreciente. Luego, forma dos listas de bigramas a derecha y a izquierda de la palabra target con de las 200 palabras de contexto más frecuentes.

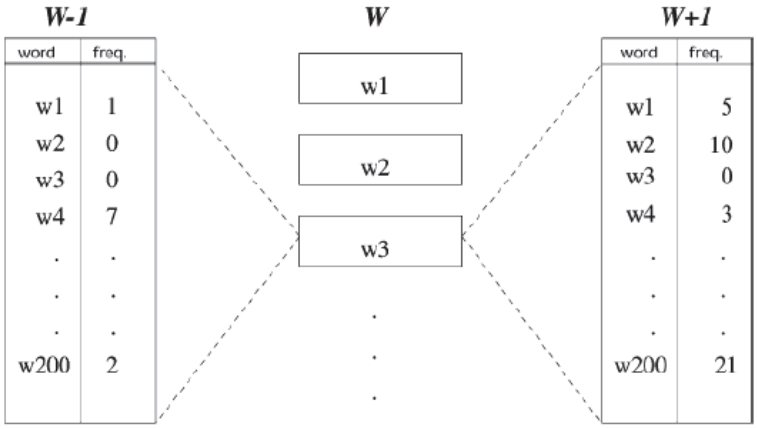

Figura 22: Bigramas a derecha y a izquierda de las 82 palabras funcionales target en el experimento 1 de Wang (2012)

Se generan entonces 82 vectores de 400 dimensiones que se proceden a clusterizar con un clustering jerárquico -Wang (2012) no especifica el tipo de enlace (*linkage*) implemementado-, recurriendo como criterio de comparación entre ellos a una distancia Manhattan ponderada, conocida como distancia *Canberra*.

$$d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Ecuación 15: Distancia *Canberra* entre dos vectores $\vec{p}$ y $\vec{q}$

A su vez, Wang (2012) establece coeficientes de cortes para 3 escenarios de análisis: 8 clusters, 12 clusters y 16 clusters, reportando altos valores de precisión pero baja cobertura (*recall*), con medida F en torno a 0,55 -aunque los niveles de pureza sí parecen ser significativos (en torno a 0,8 para el escenario de 16 clusters.

Cabe desatacar las enormes semejanzas de este experimento con los enfoques tradicionales de clustering como el de Redington (Redington *et al.* 1998) o nuestro propio experimento (véase capítulo 7). No obstante, en el próximo capítulo de esta tesis remarcaremos cómo las distintas decisiones de diseño en el algoritmo repercuten en la plausibilidad de la adecuación explicativa

de estos tres experimentos semejantes -ya hemos puntualizado algunas de esas críticas en el capítulo anterior.

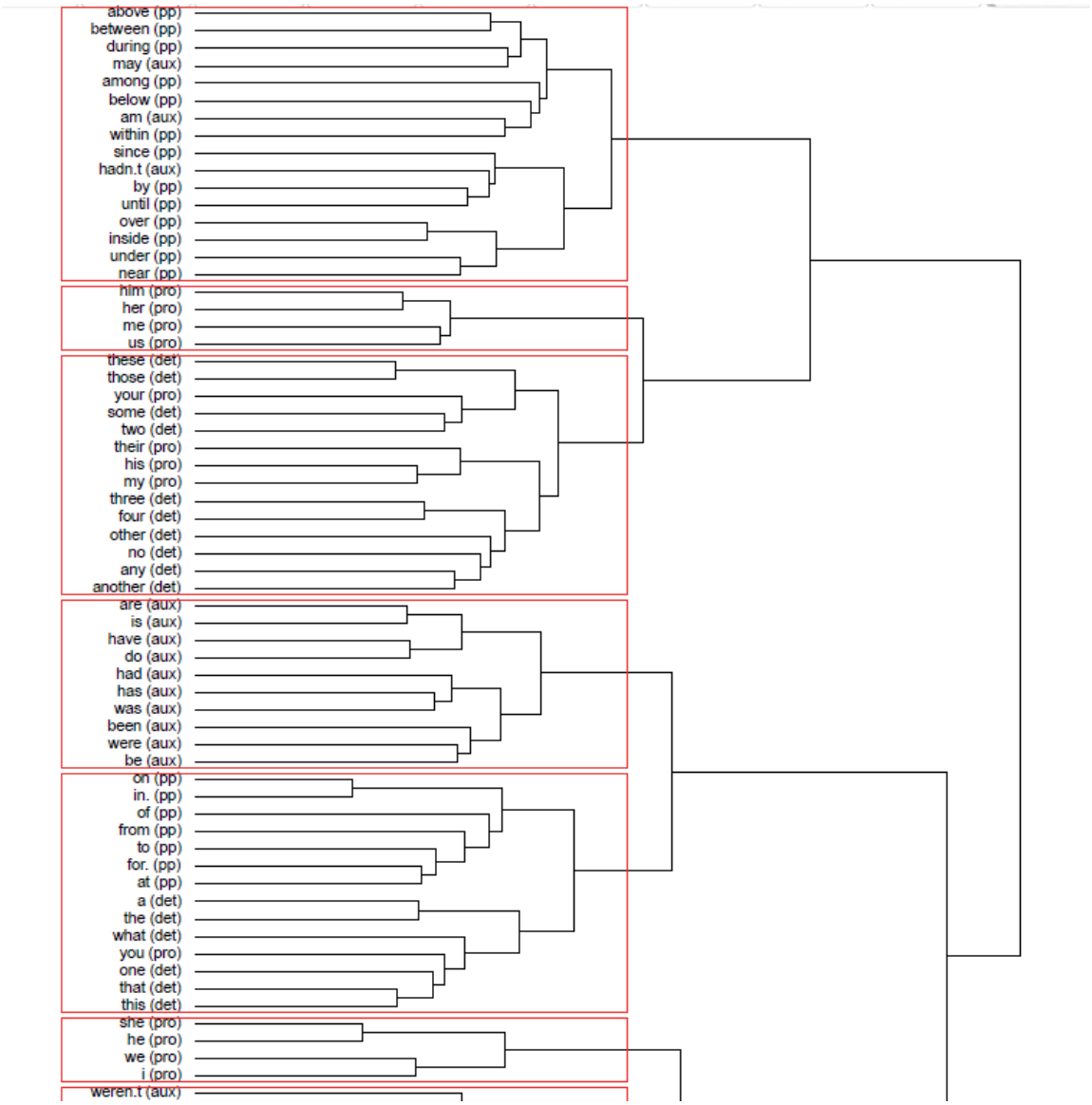

Figura 23: Dendrograma con categorización de 82 palabras funcionales en el experimento 1 de Wang (2012)

Para dimensionar los resultados del experimento 1 de Wang (2012), hay que tomar en cuenta que la categorización de palabras funcionales (de clase cerrada) resulta un escollo algebraico importante para las técnicas de clustering, debido a que en todos los idiomas, las categorías de palabras funcionales tienden a presentar clases mucho más reducidas en el número de miembros que las categorías de palabras de contenido (de clase abierta):

> "However, those statistical regularities are not enough to overcome several challenges. […] category size distribution tends to be uneven. For example, the vast majority of the word types are open class (nouns, verbs, adjectives), and even among open class categories, there are many more nouns than adjectives. This runs contrary to the learning biases in commonly-used statistical models. A common failure of those models is to clump several rare categories together and split common categories." [Graça *et al*. 2011:528]

Al analizar los resultados de este experimento 1, Wang (2012) reporta que dos de las cuatro categorías funcionales originales fueron clusterizadas mejor que el resto: pronombres y verbos auxiliares/modales. Sin embargo, no sugiere ninguna explicación al respecto. La conclusión de este experimento 1 es que las técnicas de clustering bien pueden inducir las categorías de palabras funcionales a partir de la información distribucional.

6.4 Experimento 2 de Wang (2012): marcos frecuentes para categorías funcionales

Para su segundo experimento, Wang (2012) apela a la teoría de los marcos frecuentes (Mintz 2003). Adapta el mismo corpus que en el experimento anterior para capturar los 45 marcos frecuentes de mayor ocurrencia que involucran palabras funcionales target.

Frame	Token	Type	Primary category*	% primary category	Category (Token)
what__you	2291	13	aux	100%	aux (2291)
there__are	1865	3	pro	99%	pro (1864) aux (1)
would__like	1822	7	pro	99%	pro (1820) aux (1) pp (1)
are__going	1183	2	pro	100%	pro (1183)
can__see	1170	4	pro	100%	pro (1170)
I__think	975	10	aux	99%	pro (3) aux (972)
do__think	965	2	pro	100%	pro (965)
on__floor	815	1	det	100%	det (815)
back__the	678	8	pp	100%	pp (678)
I__know	659	5	aux	99%	pro (1) aux (658)
out__the	641	11	pp	99%	det (2) pp (639)
you__a	616	21	aux	91%	det (2) aux (565) pp (49)
do__remember	606	2	pro	99%	det (1) pro (605)
it__the	592	20	pp	83%	det (3) aux (97) pp (492)
have__look	554	3	det	97%	det (540) pp (14)
go__the	520	12	pp	99%	aux (2) pp (518)
is__what	511	2	det	100%	det (511)
what__we	463	11	aux	100%	aux (463)
it__a	462	13	aux	82%	det (1) aux (382) pp (79)
to__a	453	8	aux	99%	pro (2) aux (449) pp (2)
do__want	446	1	pro	100%	pro (446)
have__got	418	4	pro	99%	pro (417) pp (1)
you__have	408	17	aux	98%	pro (2) aux (402) pp (4)
look__the	405	11	pp	99%	det (2) pp (403)

** det = determiner, pp = preposition, pro = pronoun, aux = auxiliary*

Tabla 18: 45 marcos frecuentes de un informante en el experimento 2 de Wang (2012)

Los números entre paréntesis indican la mayor probabilidad de que el *gap* en el marco frecuente sea llenado por determinada categoría de palabra funcional en frecuencia absoluta.

Si bien el enfoque de los marcos frecuentes podría resultar particularmente inapropiado para la categorización de palabras funcionales, ya que los contextos distribucionales de determinadas

palabras funcionales experimentan una distribución completamente disjunta -por ejemplo, los determimantes '*a*' y '*an'* (Schütze 1993)-, Wang (2012) reporta elevadas medidas de precisión (*accuracy* por sobre 0,95) para los 8 sets, pero la cobertura (*recall*) es realmente insignificante en todos los casos (en torno a 0,10). Otra de las desventajas de este método de categorización de palabras funcionales, como el mismo Wang reconoce, es que la teoría de los marcos frecuentes no toma en cuenta los límites de las oraciones. Como las palabras funcionales (especialmente los determinantes) suelen ocurrir en los inicios de oraciones (e incluso, en los límites de las frases fonológicas), esta información distribucional no puede ser explotada en el experimento 2. Por último, otra observación crítica se centra en la escasa variedad de los tipos de categorías de palabras funcionales involucrados y en la necesidad de un mecanismo posterior para inducir la categoría funcional propiamente dicha a partir de los diversos marcos frecuentes de ocurrencia de las palabras miembros de la clase. Todo esto lleva a considerar el experimento 1 como más exitoso que el experimento 2 a los fines de probar la hipótesis central de Wang (2012) de categorización de palabras funcionales.

6.5 Evaluación general de Wang (2012)

Los experimentos propuestos por Wang (2012) demuestran que la categorización temprana de palabras funcionales puede ser modelizada exitosamente con técnicas de clustering, una vez que se han identificado las palabras funcionales a partir de los indicios prosódicos. A su vez, y en función de una mayor adecuación explicativa, se demuestra que este mecanismo de categorización inicial de palabras funcionales puede ser el mismo que potencia la categorización de palabras de contenido: la información distribucional.

> "An implication of the results from the two analyses is that even if semantic information is not available to children, acoustic and distributional cues in the input would be enough to bootstrap children from scratch to some kind of prototypical categories. **At last, this is favorable in terms of simplicity of theories because the bootstrapping of both lexical and functional categories can be accounted for under the distributional approach**." [Wang 2012:89] (*las negritas y el subrayado son nuestros*)

El gran aporte del trabajo de Wang (2012) es demostrar la plausibilidad de un estadio previo en un trabajo de modelización tendiente a inducir sintaxis en forma integral. Por ejemplo, el proceso de categorización de palabras funcionales del experimento 1 bien podría ser considerado una etapa previa de procesamiento para ofrecer las condiciones de posibilidad de un modelo como el de los protoconstituyentes (Christophe *et al.* 2008), que ya analizamos en la sección 3.4. Es decir, el experimento 1 de Wang (2012) podría dar las condiciones de posibilidad a nivel del desarrollo ontogenético para que el adquiriente plausiblemente apele a nociones de protoconstituyentes NP y VP (caracterizados por determinantes y verbo auxiliar, respectivamente) en forma previa a la categorización de palabras de contenido, tal como sugería el modelo (Christophe *et al.* 2008):

> "Functional categories are available in children's grammar since very early on. Function words are late in production is not because of lack of knowledge of functional categories but probably due to performance-related factors. Functional categories are used by young children in the same way as adults do. The results are in favor of nativist views or a strong learning mechanism. With the knowledge of functional categories, children can start acquire syntactic constituents like NP and VP (because functional categories are the heads of the projections) and more complex syntactic structures. Distributional information in the input was able to accurately categorize function words with the help of acoustic cues. Therefore, it could be a primary source for initial categorization of function words and bootstrapping of functional categories." [Wang 2012:90-91]

Pero lo más importante a destacar del trabajo de Wang (2012) para nuestra propia hipótesis es haber demostrado que los indicios prosódicos, que actúan como identificadores de las palabras funcionales, permiten postular en forma muy temprana la representación asbtracta de las mismas, si no la plena adquisición, en niños de edades tan prematuras como los 14 meses. Si bien existe bastante evidencia empírica de tal facilitación prosódica en inglés y algunos otros idiomas, cabe preguntarse si para el caso del español la hipótesis de Wang (2012) resultará igualmente validada. Como veremos en el próximo capítulo de esta tesis, las palabras funcionales del español son bastante diversas en cuanto a sus propiedades fonéticas y fonológicas –mientras que los pronombres enclíticos parecen adecuarse a la característica de minimalidad prosódica (Wang 2012), ése claramente no es el caso de los pronombres demostrativos, los pronombres personales nominativos o los pronombres indefinidos. De todos modos, como veremos en el próximo capítulo, la justificación de la identificación de las palabras funcionales como cues para la categorización de palabras en nuestro experimento no descansa tanto en el perfil prosódico de dichas palabras funcionales como más bien en sus propiedades distribucionales en cualquier corpus masivo (Elghmary 2004).

Capítulo 7. Nuestro experimento: Inducción no supervisada de categorías morfosintácticas mediante clustering a partir de palabras funcionales sin tipología diferenciada

7.1 Motivación de las decisiones de diseño

Habiendo analizado en detalle los diversos enfoques para el problema de la categorización temprana de palabras, estamos en condiciones de proceder a delinear nuestro propio experimento. En función de las fortalezas y las críticas relevadas para los trabajos que durante las últimas dos décadas atacaron el problema de cómo los adquirientes de una lengua conforman clases morfosintácticas de palabras, nuestro experimento se propone como un enfoque computacional compatible con la evidencia empírica de la psicolingüística, con mayor una adecuación explicativa. Así pues, nuestra propuesta de modelo de adquisición de categorías morfosintácticas del español responde a los siguientes lineamientos, algunos de los cuales exploraremos en detalle en sucesivas secciones del presente capítulo:

1) Para el marco epistemológico general, optamos por el paradigma estadístico de la lingüística computacional, en detrimento del paradigma simbólico. A pesar de que algunos modelos enmarcados en el paradigma simbólico, tal como explicamos en el capítulo 3 de esta tesis, son compatibles con nuestra hipótesis de un sesgo débil (Lappin y Shieber 2007; Clark y Lappin 2013) para inducir sintaxis a partir de un mecanismo de aprendizaje general, consideramos que los modelos disponibles de marcos frecuentes (Mintz 2003; Chemla *et al*. 2009) y de protoconstituyentes (Christophe *et al*. 2008) presentan insalvables cuestionamientos a la adecuación descriptiva y a la adecuación explicativa, respectivamente.

2) Desde el paradigma estadístico de la lingüística computacional, nos inclinamos hacia las técnicas de clustering con un enfoque tradicional, sin el agregado de técnicas avanzadas de *machine learning* (véase capítulo 5 de esta tesis). Esto nos garantiza una aceptable cobertura del fenómeno a elucidar, sin contradecir la hipótesis de un mecanismo general de aprendizaje, ya que como explicamos en la sección *5.7 Investigaciones actuales a partir de los trabajos fundacionales*, algunos modelos actuales logran una mayor efectividad en inducir categorías sintácticas a partir de considerar features como la distinción mayúscula/minúscula (Berg-Kirkpatrick *et al*. 2010) , algo que obviamente nos está vedado en función de mantener las condiciones de aprendibilidad de una teoría formal de inducción de sintaxis (Pinker 1979).

3) Para el algoritmo de clustering en particular, elegimos el clustering no jerárquico K-means con distancia euclideana sobre los centroides. Como ya explicamos en el capítulo 4 de esta tesis, nos proponemos "historizar" el proceso iterativo de inducción de categorías hasta hallar una distribución óptima en función del conjunto de datos iniciales y una parametrización creciente de los números de clusters desde *K=2* hasta *K=nº máximo de cues*. Esta historización sería inviable con un algoritmo de clustering jerárquico. Además, K-means ofrece otra ventaja: la menor complejidad de poder de cómputo. La distancia euclideana (véase *Ecuación 1*) como criterio de similitud de objetos en el espacio vectorial se nos presenta más intuitivamente correcta que la distancia Manhattan (véase *Ecuación 2*) para garantizar la plausibilidad de un mecanismo de aprendizaje general, a pesar de que se considera que esta última resulta menos sensible que la primera a la influencia de los objetos apartados (*outliers*) en el espacio vectorial (Manning y Schütze 1999).

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

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

6) En cuanto a la escalabilidad del algoritmo, seguiremos a Redington *et al*. (1998) y plantearemos un escenario con un vocabulario reducido de aproximadamente 1000 palabras target. De hecho, esa cantidad de palabras resulta esperable para la finalización de la etapa ontogenética que nos interesa modelizar: la explosión léxica (*vocabulary spurt*) (Dromi 1987) que se da en los niños entre los 2 y 3 años de edad. Por supuesto, este corte en las palabras target nos aleja de enfoques exhaustivos como los de Clark (2002). Sin embargo, como ya se ha explicado en el capítulo 5 de esta tesis, consideramos que el aprnedizaje no supervisado basado en técnicas de clustering es especialmente eficaz en agrupar eventos con una cierta ocurrencia frecuente en el espacio vectorial (Martin *et al*. 1998). A su vez, esta decisión de diseño se condice con la plausibilidad de la evidencia empírica psicolingüística y con la robustez de los modelos matemáticos postulados en dichas técnicas de clustering, reduciendo los costos implementativos:

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

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

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

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

9) Para la evaluación de nuestro experimento exploraremos diversas alternativas, pero podemos adelantar que nos basaremos principalmente en la métrica *many-to-1* (Christodoulopoulos *et al*. 2010). También seguiremos a Redington *et al*. (1998) en una evaluación discriminada para cada tipo de categoría inducida (véase *Tabla 12*) y postularemos nuestra propia justificación algebraica del agrupamiento de clusters (*cluster merging*) (Böhm *et al*. 2006) en *hiperclusters* a partir del mapeo *many-to-1*.

Básicamente el algoritmo propuesto en esta tesis se muestra en el siguiente esquema:

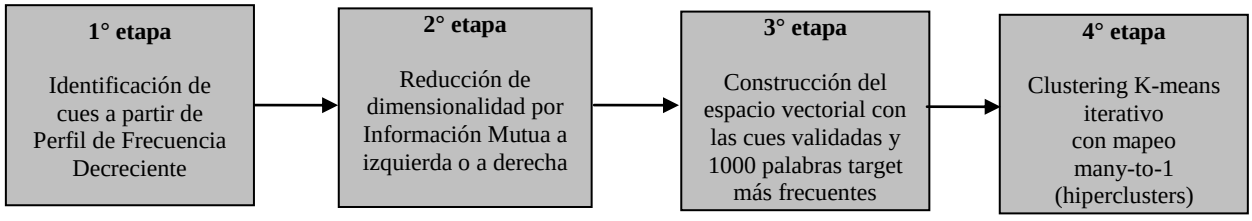

Figura 24: Esquema del algoritmo de categorización de palabras propuesto para la tesis

En las póximas secciones del presente capítulo de esta tesis procederemos a extendernos en detalle sobre las decisiones de diseño delineadas en esta sección.

7.2 Corpus de PLD

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* PLD) de que dispondría un niño durante el proceso de adquisición del lenguaje a los *corpora* de millones de palabras (*tokens*) que son procesados por las computadoras.

Una ingenua objeción a nuestra conformación de los PLD podría poner la lupa en la diferencia entre un proceso de aprendizaje incremental a lo largo de años (como sucede efectivamente en los niños) en contraste con un acceso inmediato a todos los PLD por parte de la computadora. Análogamente, desde un punto de vista neurológico se podría aducir que no es lo mismo la experimentación con los procesos orgánicos involucrados en nuestra mente/cerebro que la experimentación con modelos psicolingüísticos. Consideramos que estas cuestiones atañen desde un punto de vista epistemológico a la idea misma de modelo. En toda disciplina científica la postulación de un modelo representa la formalización de propiedades isomórficas de los fenómenos observables en la realidad, pero en ningún caso un modelo es la realidad misma a estudiar.

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] (*las negritas y el subrayado son nuestros*)

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 mismo reconoce (Clark 2002), resulta difícil emular por completo los PLD, 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 para la elucidación del problema de la adquisición del lenguaje, como el corpus multilingüe CHILDES (30 millones de palabras en 20 idiomas organizadas en interacciones orales madre-niño). Desafortunadamente, CHILDES no está disponible en español.

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. Por ejemplo, para su tesis de inducción integral de sintaxis Clark (2002) recurre al *British National Corpus* (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 C4 (CLAWS-4) –un conjunto de 76 etiquetas morfosintácticas al que Clark (2002) agrega un símbolo para indicar el fin de oración. Aunque el BNC abarca registros orales en un 10% de la muestra, Clark (2002) recorta el input del BNC a 12 millones de palabras del registro escrito.

Tomando como guía el trabajo de Redington *et al.* (1998), en nuestro experimento comenzamos por armar un corpus no anotado morfosintácticamente de 1,8 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 1,8 millones de palabras (*tokens*) y 71.467 *tipos* (*types*, entendidos como cualquier cadena de caracteres alfanuméricos entre signos de puntuación o espacios en blanco sin distinción mayúsculas/minúsculas). Una única instancia de type puede manifestarse en un corpus con la ocurrencia concreta de dicha palabra en numerosos tokens.

7.3 Primera etapa del algoritmo: Identificación de cues

7.3.1 Intuición distribucional acerca de las palabras funcionales vs. palabras de contenido

A esta altura estamos familiarizados con la idea de que la información distribucional en grandes corpora resulta de gran valor para los juicios de pertenencia de palabras a ciertas clases morfosintácticas. Hemos definido dicha información distribucional como la relación de las palabras target con sus vecinos inmediatos (bigramas a derecha y a izquierda). Resta, pues, justificar la selección no arbitraria de las palabras que actuarán como marcadoras de dichas relaciones bigrámicas (*cues* o *palabras marca* o *palabras indicio*) y la consiguiente selección de las palabras target.

Hemos observado que las técnicas de clustering resultan particularmente sensibles a la frecuencia de los eventos que ocurren en el espacio vectorial. Es decir, los resultados de clustering tienden a ser mejores cuanto más frecuentes son los eventos a clusterizar, ya que naturalmente es de esperar que los patrones estadísticos tiendan a consolidarse con la ley de los grandes números de la teoría de la probabililidad (Martin *et al*. 1998). A su vez, en la relación de palabras funcionales vs. palabras de contenido en cualquier idioma se verifica el criterio de frecuencias diferenciadas (Zipf 1949). Justamente, desde un punto de vista teórico es sabido que las palabras funcionales tienden a ocurrir en los contextos de las palabras de contenido con cierta predicibilidad (Redington *et al*. 1998), articulando su significado bajo la luz de diversas presentaciones lingüísticas (por ejemplo, '*el hombre con auto*', '*el hombre sin auto*', '*el hombre detrás del auto*', '*el auto del hombre*', etc.).

> "In order to gain some intuition regarding why distributional information is more useful for content words than for function words, consider the kinds of contexts in which each will appear. Content words will tend to have one of a small number of function words as their context. Although content words are typically much less frequent, their context is relatively predictable. Function words, on the other hand, are much more frequent, but will tend to have content words as their context. Because there are many more content words, the context of function words will be relatively amorphous. As the measure of similarity exploits regularities in the distribution of contexts, those words with predictable contexts will be clustered together much more accurately." [Redington *et al.* 1998:456]

Esta convergencia de propiedades nos obliga a dirigir nuestra atención hacia la distinción entre palabras funcionales versus palabras de contenido como un criterio de selección "natural" de cues y palabras target, respectivamente. Ya en el capítulo 2 y en la conclusión del capítulo 3 de esta tesis habíamos adelantado esta distinción operativa entre palabras funcionales y palabras de contenido, pero ha llegado el momento de demostrar empíricamente ciertas propiedades distribucionales que se verifican en corpora masivos, más allá de los etiquetamientos gramaticales para dichas clases de palabras.

7.3.2 Ley de Zipf

Zipf (1949) fue uno de los primeros en estudiar las distribuciones de palabras en corpora masivos. A partir de sus investigaciones, Zipf (1949) sostiene la existencia de un principio unificador, el *Principio del Menor Esfuerzo*, el cual subyace a la naturaleza humana. La

evidencia a favor de esta teoría consiste en ciertas leyes empíricas que Zipf expone. La presentación de las mismas comienza con su propia investigación, mostrando ciertas distribuciones estadísticas en el lenguaje. No se comentará aquí su teoría general, sino algunas de sus leyes empíricas acerca del lenguaje.

Si contamos cuántas veces figura cada palabra (tipo o *type*) de una lengua en un corpus extenso, y después ordenamos las palabras en función de su frecuencia de aparición, podemos explorar la relación entre la frecuencia de una palabra *f* y su posición en la lista, conocida como su *ranking r*.

$$f \propto \frac{1}{r}$$

Ecuación 16: Ley de Zipf para las frecuencias de palabras en corpora masivos

En otras palabras:
Hay una constante *k* tal que $f \cdot r = k$

En el ejemplo de la Tabla 19 esto significa que la 50º palabra más común del ranking (‘*esto*’) debería ocurrir con el cuádruple de frecuencia que la 200º palabra más común del ranking (‘*Rocinante*’). No consideramos este resultado como una ley propiamente, sino como una caracterización aproximada de ciertos hechos empíricos:

Palabra (type)	Frecuencia (*f*)	Ranking (*r*)	$f \cdot r$
que	20.549	1	20.549
de	17.998	2	35.996
a	9.532	5	47.660
los	4.681	10	46.810
como	2.244	20	44.880
esto	849	50	42.450
tiene	385	100	38.500
Rocinante	203	200	40.600
libertad	80	500	40.000
deja	37	1.000	37.000
león	16	2.000	32.000
ocio	5	5.000	25.000
raudal	2	10.000	20.000
sartas	1	20.000	20.000

Tabla 19: Muestra de verificación empírica de la Ley de Zipf en el corpus del texto *El ingenioso hidalgo don Quijote de la Mancha* (Cervantes 1604)

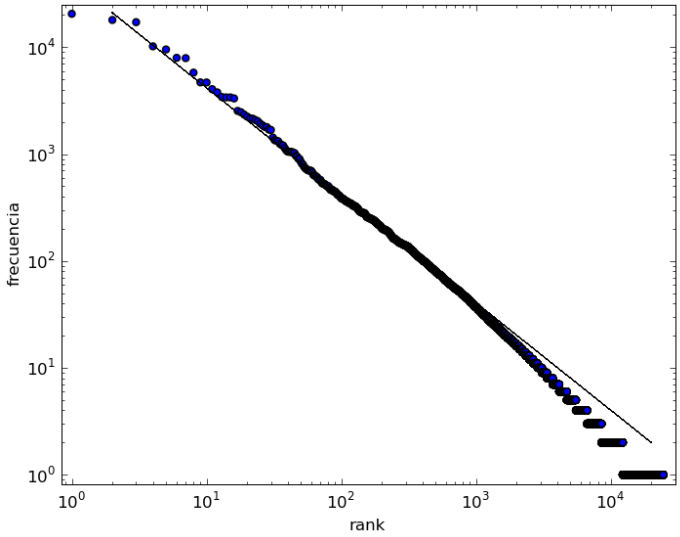

Figura 25: Gráfico de escala logarítmica entre las frecuencias y los rankings para el corpus del texto *El ingenioso hidalgo don Quijote de la Mancha* (Cervantes 1604)

El gráfico muestra el *ranking* en el eje X y la *frecuencia* en el eje Y, utilizando escala logarítimica. Los puntos corresponden a los rankings y frecuencias de las palabras. La línea recta es la predicción hecha por la *Ley de Zipf*, con $k = 40.000$, es decir que $f \cdot r = 40.000$.

La Tabla 19 y la Figura 25 muestran una verificación empírica de la *Ley de Zipf* tomando a *Don Quijote* como texto base. La *Ley de Zipf* parece confirmarse aproximadamente, con un valor de k alrededor de 40.000, pero notamos que para las palabras de frecuencia muy alta (>10.000) o muy baja (< 20) los resultados se alejan bastante de la predicción.

De acuerdo con esta teoría, tanto el hablante como el oyente están tratando de minimizar su esfuerzo. El esfuerzo del hablante se conservaría utilizando un vocabulario reducido de palabras comunes y el esfuerzo del oyente se reduciría disponiendo de un extenso vocabulario de palabras poco comunes, de manera que los mensajes resulten menos ambiguos. Zipf (1949) arguye que el acuerdo entre estas necesidades opuestas que resulta más económico para ambas es el tipo de relación recíproca entre frecuencia y ranking que figura en los datos a favor de la *Ley de Zipf*.

7.3.3 Perfil de Frecuencia Decreciente (Dreceasing Frequency Profile DFP)

La *Ley de Zipf* resulta útil como una descripción básica de la frecuencia de distribución de las palabras en los lenguajes naturales: hay unas pocas palabras muy comunes (generalmente, palabras funcionales), una cantidad media de palabras de frecuencia intermedia y un gran número de palabras de baja frecuencia (generalmente, palabras de contenido). De este modo, es dable suponer que la principal característica de las palabras funcionales es su mayor frecuencia en corpora masivos. La base léxica común de cualquier texto, independientemente de la temática o de la estilística, termina aislando las palabras funcionales a lo largo de corpora masivos balanceados. Esto se verificará en cualquier muestra masiva de texto suficientemente representativa.

Para demostrar la utilidad de un *Perfil de Frecuencia Decreciente* (Elghamry 2004) en la identificación de palabras funcionales, nos propusimos analizar una muestra de apenas unas

decenas de miles de tokens del corpus Brown. El corpus Brown es probablemente el corpus más conocido. Es un corpus de aproximadamente un millón de palabras etiquetadas, recopiladas entre 1960 y 1970 en Brown University, con una composición balanceada, es decir, se intentó que el corpus representara una muestra significativa del inglés norteamericano de esa época. Cubre los géneros periodístico, ficción, texto científico y legal, entre otros. Para el reporte de salida de la Tabla 20 sólo consideramos la mitad de la sección *editoriales* del genéro periodístico (apenas una vigésima parte del total del corpus). Notaremos cómo el Perfil de Frecuencia Decreciente verifica la intuición de que las palabras funcionales son las más frecuentes en cualquier corpus masivo.

En particular, en este registro de salida observamos la interferencia de algunas palabras de contenido en las primeras posiciones del ranking: '*Mr.*', '*state*', '*year*', '*home*', '*city*', '*president*'. Claramente, esto se debe al sesgo de ciertas expresiones y temáticas periodísticas en un corpus de dimensiones aún pequeñas. Sin embargo, este escenario exploratorio de dimensiones reducidas bien puede funcionar como una prueba de concepto para la demostración de que las palabras funcionales son más frecuentes que las palabras de contenido en cualquier corpus masivo balanceado.

7.3.4 Punto de corte entre palabras funcionales y palabras de contenido en el DFP

A modo de observación empírica de las distribuciones predichas por la *Ley de Zipf*, nótese cómo las palabras pertenecientes a los primeros 50 rankings del DFP cubren el 40% del corpus (véase *cobertura* en la *Tabla 20*) con apenas 55 types (palabras), que son apenas el 0,7% de la diversidad léxica del corpus. En el otro extremo, las palabras que ocupan los últimos rankings de frecuencia (del 100 al 113) tienen la misma cobertura aproximadamente (39%) pero representan el 94,6% de la diversidad léxica del corpus. Unas 5000 palabras (types), desde el type ubicado en la posición 2398 hasta el type 7559, son *hapax legomena* (una única ocurrencia, en el ranking 113) o *dis legomena* (dos ocurrencias, en el ranking 112), con lo cual pueden ser consideradas eventos dispersos. Resulta evidente que, si es que existe un punto de corte entre el ordenamiento por frecuencia en primer lugar de las palabras funcionales y luego las palabras de contenido, dicho punto de quiebre puede ser calculado en base a las propiedades distribucionales del corpus. En este caso, tal punto de quiebre se ubicaría en algún lugar de los rankings 50 a 100. No obstante, ese tramo del ranking muestra una ocurrencia indiferenciada de palabras funcionales y palabras de contenido. En este experimento exploratorio, tal coexistencia indiferenciada se debe, como veremos, a las dimensiones aún reducidas del corpus.

#	type	fre %	freq	rank	size	freq of freq
1	the	0,07283	3198	1	3	1
2	of	0,03131	1375	2	2	1
3	a	0,02535	1113	3	1	1
4	and	0,02414	1060	4	3	1
5	to	0,02412	1059	5	2	1
6	in	0,02280	1001	6	2	1
7	for	0,01223	537	7	3	1
8	that	0,00877	385	8	4	1
9	was	0,00831	365	9	3	1
10	he	0,00820	360	10	2	1
11	on	0,00795	349	11	2	1
12	is	0,00752	330	12	2	1
13	be	0,00674	296	13	2	1
14	said	0,00660	290	14	4	2
15	at	0,00660	290	14	2	2
16	by	0,00588	258	15	2	1
17	as	0,00569	250	16	2	2
18	with	0,00569	250	16	4	2
19	will	0,00560	246	17	4	1
20	his	0,00540	237	18	3	1
21	it	0,00501	220	19	2	1
22	mrs	0,00492	216	20	3	1
23	would	0,00383	168	21	5	1
24	has	0,00376	165	22	3	1
25	who	0,00353	155	23	3	1
26	an	0,00337	148	24	2	1
27	from	0,00332	146	25	4	1
28	this	0,00330	145	26	4	1
29	are	0,00305	134	27	3	1
30	have	0,00287	126	28	4	1
31	but	0,00280	123	29	3	1
32	were	0,00271	119	30	4	1
33	been	0,00257	113	31	4	2
34	not	0,00257	113	31	3	2
35	new	0,00255	112	32	3	1
36	mr	0,00253	111	33	2	1
37	which	0,00251	110	34	5	2
38	had	0,00251	110	34	3	2
39	two	0,00246	108	35	3	1
40	state	0,00244	107	36	5	1
41	year	0,00241	106	37	4	2
42	their	0,00241	106	37	5	2
43	one	0,00239	105	38	3	2
44	they	0,00239	105	38	4	2
45	after	0,00219	96	39	5	1
46	last	0,00212	93	40	4	1
47	there	0,00207	91	41	5	1
48	when	0,00203	89	42	4	1
49	first	0,00198	87	43	5	1
50	out	0,00196	86	44	3	1
51	i	0,00194	85	45	1	1
52	home	0,00191	84	46	4	1
53	more	0,00187	82	47	4	1
54	its	0,00180	79	48	3	1
55	all	0,00175	77	49	3	1
56	city	0,00173	76	50	4	1
57	or	0,00171	75	51	2	2
58	other	0,00171	75	51	5	2
59	also	0,00169	74	52	4	2
60	up	0,00169	74	52	2	2
61	her	0,00164	72	53	3	1
62	over	0,00162	71	54	4	1
63	no	0,00157	69	55	2	1
64	some	0,00143	63	56	4	1
65	three	0,00141	62	57	5	2
66	president	0,00141	62	57	9	2
67	only	0,00134	59	58	4	3
68	into	0,00134	59	58	4	3
69	made	0,00134	59	58	4	3
70	than	0,00130	57	59	4	3
71	years	0,00130	57	59	5	3

tokens	43911
types	7559

ranking		distribución		cobertura	
1 to 50	55 types	0,70%	17369 tokens	40%
50 to 100	355 types	4,70%	9437 tokens	21%
100 to 113	7149 types	94,60%	17105 tokens	39%

1 to 50	3,22	letter per word
50 to 100	5,18	letter per word
100 to 113	7,10	letter per word

size average vs. ranking

Tabla 20: Perfil de Frecuencia Decreciente para una sección del corpus Brown

Es esperable que en corpora masivos, la frecuencia ordene naturalmente la distinción de palabras funcionales claramente por encima de las palabras de contenido, estabilizando los puntos de quiebre de todos los corpora masivos para un idioma aproximadamente en torno al mismo ranking, e identificando así las palabras funcionales -base léxica común para todos esos corpora masivos, en términos de Zipf (1949). Para explorar las propiedades distribucionales de las palabras funcionales y su importancia como cues identificatorias del contexto de uso de las palabras de contenido recurriremos al algorimo de identificación de cues de Elghamry (2004):

> "A first approximation to this procedure can make use of the frequency of certain elements or features in the input. Accordingly, a cue can be any member of the set of the highly frequent elements in the input. Consequently, function words, stress, and silence as indicated by utterance boundaries can be possible cues. Utterance boundaries are cues by definition since they indicate the beginning and end of some constituents. Function words are highly frequent in the input, which, among other features, makes them stand out in the input. For that reason, some of the learning methods discussed in previous chapters have used these words as cues." [Elghamry 2004:81-82]

> "The approximate method proposed here makes a direct use of the highly frequent words in a corpus on the assumption that these words would provide information about the distributional properties of other words in the corpus. […] The core of this method is to find the smallest subset of words in the corpus that co-occur with a number of words that converges to an order of the number of word types in that corpus. […] It is expected that higher orders of approximation should give more fine-grained information about the distributional properties of the words in the corpus." [Elghamry 2004:83-86]

Para implementar este algoritmo de identificación de cues o marcas en función de las propiedades distribucionales de las palabras trabajamos incialmente con un Perfil de Frecuencia Decreciente (DFP) sobre el corpus de experimentación descripto en la sección *7.2 Corpus de PLD*. A su vez, inicializamos una bolsa de palabras vacía. Nuestra heurística propone una sutil modificación de la heurística de corte de Elghamary (2004): consiste en recorrer el DFP y anotar las ocurrencias de types a la derecha y a la izquierda de cada palabra del DFP hasta que en un cierto punto una nueva palabra en el ranking ya no incorpora ningún nuevo type a la bolsa de palabras ni a izquierda ni a derecha para ninguna de sus ocurrencias (tokens) en el corpus. En ese punto es dable pensar que tal palabra en el DFP tiende a comportarse más como una palabra de contenido (con contexto predecible) que como una palabra funcional (con contexto impredecible).

En nuestro corpus de 71.467 types y 1,8 millones de tokens la aplicación del algoritmo de identificación automática de cues nos devolvió una lista de 106 palabras (en su enorme mayoría, palabras funcionales). No obstante, es menester reconocer que esta primera etapa de nuestro algoritmo de catgorización ya nos está dando algunas señales de alerta sobre la posibilidad de un corpus no enteramente balanceado. La aparición como cues de las palabras '*trabajo*', '*producción*' y '*capital*' es la consecuencia de la incorporación al corpus de la obra *El Capital* de Marx (1867) (sesgo temático), mientras que se da una situación análoga con los nombres propios '*levin*' y '*ana*' por haber incorporado la voluminosa novela *La guerra y la Paz* de Tolstói (1869). A pesar de estas excepciones, debemos resaltar la prevalencia de palabras funcionales en el listado de cues. Para ilustar el poder empírico de la heurística de corte, considerando hasta el

punto de corte en la palabra de orden 106 del DFP, la cobertura es en types de un 85,5% del corpus (61.095 types en la bolsa de palabras sobre 71.467), lo cual valida por completo el enfoque: estas 106 cues detectadas son vecinas inmediatas (a uno u otro lado) para al menos un 85,5% de las palabras (types) del corpus.

Cues identificadas:

#	Cue	Ocurrencias	#	Cue	Ocurrencias	#	Cue	Ocurrencias
1	de	110854	37	sobre	3355	73	sino	1845
2	la	68582	38	cuando	3279	74	puede	1844
3	que	57630	39	qué	3202	75	bien	1828
4	el	48327	40	ha	3001	76	cada	1814
5	y	47851	41	mi	2838	77	sí	1773
6	en	45480	42	también	2816	78	después	1771
7	a	38686	43	este	2815	79	levin	1763
8	los	29176	44	sólo	2756	80	hay	1718
9	se	23499	45	vez	2715	81	menos	1704
10	no	21342	46	dijo	2654	82	está	1702
11	las	19149	47	hasta	2645	83	casa	1694
12	un	19125	48	fue	2563	84	general	1690
13	del	18828	49	mismo	2534	85	antes	1672
14	por	18070	50	dos	2527	86	decir	1665
15	una	16650	51	entre	2507	87	nada	1637
16	con	16255	52	ser	2498	88	otra	1605
17	su	16238	53	esta	2486	89	hombre	1593
18	lo	11741	54	tiempo	2343	90	día	1590
19	es	11510	55	todos	2342	91	siempre	1581
20	para	11113	56	ni	2332	92	otro	1552
21	al	11017	57	valor	2305	93	algo	1538
22	más	9044	58	tan	2278	94	capital	1535
23	como	8893	59	muy	2258	95	uno	1486
24	pero	7657	60	ella	2243	96	producción	1484
25	le	7206	61	son	2092	97	tiene	1418
26	sus	6159	62	desde	2079	98	modo	1404
27	o	5869	63	años	2034	99	forma	1404
28	había	5700	64	ahora	1974	100	te	1395
29	me	5671	65	yo	1971	101	mundo	1386
30	trabajo	5122	66	tanto	1946	102	tenía	1353
31	sin	4544	67	nos	1930	103	pues	1353
32	era	4503	68	estaba	1907	104	esto	1353
33	si	3953	69	porque	1896	105	ana	1328
34	todo	3906	70	así	1880	106	eso	1326
35	ya	3537	71	vida	1876			
36	él	3403	72	parte	1857	palabras_context_hasta_este_corte (types):		61095

Tabla 21: Identificación de cues por punto de corte en el DFP de nuestro experimento central (1º Etapa)

Otra interesante conclusión provisoria a remarcar es que la identificación de cues se basa más en las propiedades distribucionales de las palabras más frecuentes de un corpus que en su pertenencia a clases cerradas (funcionales) o abiertas (de contenido). Esto se ve refrendado en ciertas observaciones empíricas de estudios de frecuencia. Si bien las palabras funcionales tienden a ocurrir con mayor frecuencia que las de contenido, esto no siempre es valido para cada uno de los miembros de ambas clases. Por ejemplo, Clark (2002) hace notar que entre los 12 millones de tokens extraídos del BNC que componen su corpus no figura ninguna ocurrencia del pronombre personal reflexivo '*ourselves'*.

De este modo, a diferencia del experimento de Redington *et al.* (1998), nuestro experimento se propone como un método no apriorístico y no arbitrario para identificar marcas sintácticas que

habrán de constituir las relaciones bigrámicas de las dimensiones de los vectores sobre los cuales se trabajará con técnicas de clustering. Este enfoque, además, presenta interesantes implicancias 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 en forma temprana como pre-requisito para la categorización de palabras de contenido (Christophe *et al*. 2008; Wang 2012), punto que retomaremos en la sección *7.11 Discusión de los resultados y conclusiones*.

7.4 Segunda etapa del algoritmo: Reducción de dimensionalidad

En esta etapa de nuestro experimento de categorización de palabras disponemos de las 106 cues identifcadas en la etapa anterior. Cada una de ellas actúa como una marca respecto de la cual una palabra target puede ser categorizada en función de las ocurrencias de bigramas inmediatos a izquierda y a derecha. Por ejemplo, es razonable esperar que la cue '*el*' (cuarta en la lista de la *Tabla 21*) sea un muy buen indicador para relaciones bigrámicas a izquierda de las palabras target sustantivos masculinos singulares. Sintomáticamente, la ausencia o un valor sensiblemente menor de bigramas a izquierda '*el*'*-palabra_target_sustantivo_femenino_singular* también formará parte de un extenso perfil distribucional que irá agrupando a las palabras target mediante técnicas de clustering. Así pues, debido únicamente a las propiedades distribucionales de este corpus, las palabras target, en principio, serían categorizadas mediante un perfil de bigramas a izquierda y a derecha con cada cue (bigramas *cue-target* y *target-cue*, respectivamente), lo que resulta en vectores de 212 dimensiones (106 cues combinadas a cada uno de los dos lados de la palabra target).

Un espacio vectorial de 212 dimensiones representa un escollo a la hora de aplicar satisfactoriamente técnicas de clustering. Como ya vimos en los trabajos clásicos del estado de la cuestión, algunos investigadores recurrieron a técnicas de reducción de dimensionalidad como *Single Value Decomposition* (SVD) (Schütze 1993) o *Principal Component Analysis* (PCA) (Böhm *et al*. 2006). Estas técnicas de reducción de dimensionalidad han sido matemáticamente validadas y su motivación algebraica resulta irrefutable: como en un escenario de excesiva dimensionalidad es de esperar que muchas dimensiones estén "anuladas" o en cero (problema de la dispersión de datos o *data sparsity*), estas técnicas resultan en vectores "equivalentes" que se focalizan en las dimensiones más significativas en términos de la varianza (PCA) (James *et al*. 2013) o en matrices más reducidas (SVD) (Deerwester *et al*. 1990).

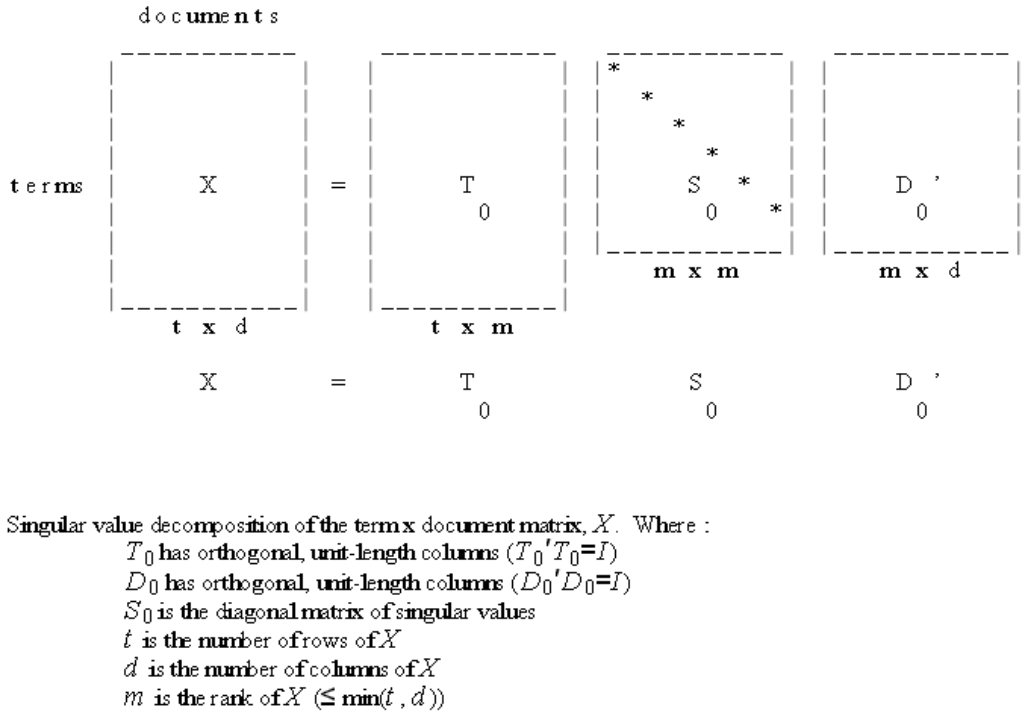

Figura 26: Esquema de reducción de la dimensionalidad de una matriz por *Single Value Decomposition* en Deerwester *et al*. (1990)

No obstante, consideramos que estas técnicas de reducción de dimensionalidad no deberían ser invocadas en nuestro experimento por dos razones:

1) Estas técnicas tienden a anular las dimensiones no significativas (en cero o con valores despreciables), cuando, en realidad, tal ausencia de relaciones bigrámicas puede ser, por contraste, un excelente indicador del comportamiento distribucional de una clase de palabras. Para ilustrar este punto recordemos el ejemplo de la cue '*el*' más arriba.

2) Estas técnicas explotan propiedades algebraicas de una matriz de vectores construidos *a posteriori* en un corpus. Así, estaríamos postulando la posibilidad de recortar dimensionalidades en función de sus respectivos valores de ocurrencia con palabras target en un corpus determinado (los valores de los vectores) y no en función de ciertas propiedades intrínsecas de dichas palabras marca (cues). Esto resulta más evidente todavía en el caso de PCA, técnica cuya aplicación requiere de la construcción vectorial del espacio dimensional bajo la forma de una matriz en forma previa a su reducción, lo cual parece contradecir la plausibilidad psicolingüística de mecanismos generales de aprendizaje disponibles para el adquiriente (Pinker 1979) durante el proceso ontogenético de adquisición de la habilidad temprana de categorización de palabras. Es decir, no parece haber ninguna motivación distribucional (en el caso de postular mecanismos generales de aprendizaje no específicos de dominio), y menos aún lingüística (en el caso de postular mecanismos de aprendizaje específicos de dominio) para apelar a dichas técnicas.

Aunque algunos otros trabajos clásicos en el campo sortearon con éxito el tratamiento de un espacio vectorial con excesiva dimensionalidad, nos propusimos la desafiante meta de investigar

la existencia de alguna propiedad intrínseca en las marcas (cues) que pudiese ser aprovechada para una reducción de la dimensionalidad más "lingüísticamente motivada". Nuestra intución apuntaba a una diferenciación en las relaciones bigrámicas a un lado y a otro en las distribuciones cue-targe y target-cue, motivada en la noción de marcación (Lorenzo y Longa 1996) y de expansión lineal de las gramáticas de los lenguajes naturales (Ćavar *et al*. 2004; Ćavar 2010). Esta idea de marcación o informatividad hacia la derecha o hacia la izquierda puede ser matemáticamente representada a partir de la *información mutua* (Shannon 1948) de una palabra respecto de un corpus o a partir de su métrica conceptualmente inversa, la *entropía* (Manning y Schütze 1999).

La primera intuición a corroborar es que las cues pueden no ser homogéneamente informativas con respecto de a qué lado se ubican de la palabra target. Esta informatividad diferenciada es una consecuencia lógica de la concepción misma de sintaxis de los lenguajes naturales en términos de una combinación estructural de consituyentes (Chomsky 1957, 1965) antes que una concatenación lineal de símbolos atómicos (véase capítulo 8 de esta tesis). Por ejemplo, si tomamos el caso de la preposición española '*de*', es esperable que encabece frases preposicionales a izquierda de las mismas, es decir, que se establezca un vínculo más estructural con la palabra siguiente antes que con la que la antecede. Por lo tanto, su papel en bigramas de la forma X-*de* debería ser menos significativo que en bigramas de la forma *de*-X, por lo que la preposición '*de*' sería más informativa hacia (marcaría hacia) la derecha. Si pudiéramos deducir para cada una de las 106 cues hacia qué lado son más informativas, el espacio vectorial quedaría reducido a la mitad: de 212 dimensiones bigrámicas a ambos lados a 106 dimensiones bigrámicas, algunas a derecha, algunas a izquierda.

Ćavar (2010) verifica nuestra intuición lingüística computando la métrica *Mutual Information* (MI) (Shannon 1948), una métrica que mide la dependencia mutua de dos variables, es decir, mide la reducción de la incertidumbre (entropía) de una variable aleatoria *x*, debido al conocimiento del valor de otra variable aleatoria *y*:

$$MI_{yx} = P(yx) * \log_2 \frac{P(yx)}{P(x) * P(y)}$$

Ecuación 17: Mutual Information entre una cue '*y*' y una palabra '*x*' siguiente (MutualInfoRight de '*y*')

Siendo *x* cualquier palabra del corpus, la MI_{yx} de la *cue y* hacia su derecha (*MutualInfoRight*) muestra la ocurrencia de *y* a la izquierda en bigramas *y-x. P(yx)* es la probabilidad de la ocurrencia del bigrama.

$$MI_{xy} = P(xy) * \log_2 \frac{P(xy)}{P(x) * P(y)}$$

Ecuación 18: Mutual Information entre una palabra '*x*' precedente y una cue '*y*' (MutualInfoLeft de '*y*')

Siendo *x* cualquier palabra del corpus, la MI_{xy} de la *cue y* hacia su izquierda (*MutualInfoLeft*) muestra la ocurrencia de *y* a la derecha en bigramas *x-y. P(xy)* es la probabilidad de la ocurrencia del bigrama.

> "In other words, local MI minima in a token sequence intuitively seem to correspond to a situation where one token does not contribute a lot about its neighboring token. This is what we expect to find when a PoS-tag to the respective side does not restrict the choice of PoS-tokens due to the lack of syntactic category or semantic selection. […] From an empirical perspective, one might expect that the MI score for a sequence of PoS-tokens *Article Noun* is much higher then for example for a sequence *Noun Verb*, given that in ca. 80% of the cases an article is followed by a noun in any common English text corpus, while the observed probability of a noun being followed by a verb is significantly smaller." [Ćavar 2010:394-395]

> "If we extend this concept of restrictive relation or selection to the lexical level, we might come to similar conclusions. The occurrence of the article *the* makes native speakers of English expect a noun to appear in the immediate local context, following the article. They would probably not have a clear intuition about some concrete noun to follow, *i.e.* they tend to have categorial intuitions associated with concrete lexical forms. On the other hand, our intuition about *the* preceding context of the seems to be rather deficient. We can test this for example in cases where the preceding word was rendered incomprehensible using noise." [Ćavar 2010:398]

Ahora que justificamos lingüística y empíricamente el criterio de reducción de dimensionalidad del espacio vectorial por validación de la informatividad de las cues hacia un lado o hacia otro de las relaciones bigrámicas, estamos en condiciones de describir la implementación de dicho criterio:

1) Para cada cue de la primera etapa (sección 7.3) calculamos todas las ocurrencias de relaciones bigrámicas en donde la cue esté a izquierda para cada type del corpus (*de-Europa* , *de-mano*, *de-la*, etc.), las sumamos y obtenemos el promedio en función de las *n* relaciones bigrámicas consideradas. Es decir la MI de la cue hacia su derecha (MutualInfoRight).

2) Repetimos 1) pero ahora con todas las ocurrencias de relaciones bigrámicas en donde la cue esté a derecha para cada type del corpus (*hombre-de*, *alegra-de*, etc.), las sumamos y obtenemos el promedio en función de las *n* relaciones bigrámicas consideradas. Es decir la MI de la cue hacia su izquierda (MutualInfoLeft).

3) Para cada cue, si MutualInfoRight>= MutualInfoLeft, entonces debemos considerar la relación bigrámica **cue-….** en los vectores que caracterizarán a las palabras target. De lo contrario, consideraremos la relación bigrámica **….-cue**.

En la Tabla 22 mostramos el cálculo de MutualInfoLeft y MutualInfoRight para cada una de las cues y la composición final de las 106 dimensiones validadas por nuestro criterio de reducción de dimensionalidad. Obsérvese, el caso paradigmático de '*y*', cuyos valores de MutualInfoLeft y MutualInfoRight son asombrosamente iguales. También es interesante hacer notar que las preposiciones y otras palabras típicamente funcionales tienden a presentar una MutualInfoRight mucho más elevada que su MutualInfoLeft, lo que se corresponde con la idea del español como una gramática linealmente expansiva hacia la derecha para la mayoría de las cues. Tal es el caso también para los adjetivos numerales cardinales (por ejemplo, '*dos*'). En cambio, los pronombres, pese a ser considerados en algunos casos palabras funcionales, tienden a tener un comportamiento que los asemeja más a las palabras de contenido (como elementos nominales que son), con MutualInfoLeft más alta que MutualInfoRight ('*ella*', '*él*', '*esto*', etc.).

	Cue	MutualInfoLeft	MutualInfoRight		Cue	MutualInfoLeft	MutualInfoRight		Cue	MutualInfoLeft	MutualInfoRight
1	de	0.000006	**0.000008**	36	él	**0.000018**	0.000005	71	vida	**0.000044**	0.000005
2	la	0.000008	**0.000021**	37	sobre	0.000004	**0.000016**	72	parte	**0.000033**	0.000010
3	que	0.000005	**0.000007**	38	cuando	0.000003	**0.000006**	73	sino	0.000002	**0.000009**
4	el	0.000006	**0.000018**	39	qué	**0.000019**	0.000006	74	puede	0.000006	**0.000009**
5	y	**0.000003**	**0.000003**	40	ha	0.000006	**0.000016**	75	bien	**0.000009**	0.000004
6	en	0.000003	**0.000016**	41	mi	0.000006	**0.000010**	76	cada	0.000004	**0.000015**
7	a	0.000005	**0.000012**	42	también	**0.000004**	**0.000004**	77	sí	**0.000014**	0.000008
8	los	0.000008	**0.000016**	43	este	0.000006	**0.000010**	78	después	0.000006	**0.000012**
9	se	0.000005	**0.000012**	44	sólo	0.000005	0.000005	79	levin	**0.000011**	0.000005
10	no	0.000004	**0.000011**	45	vez	**0.000185**	0.000005	80	hay	**0.000009**	0.000008
11	las	0.000007	**0.000018**	46	dijo	0.000010	**0.000012**	81	menos	**0.000009**	0.000004
12	un	0.000006	**0.000012**	47	hasta	0.000003	**0.000009**	82	está	0.000005	**0.000007**
13	del	0.000005	**0.000015**	48	fue	0.000004	**0.000005**	83	casa	**0.000043**	0.000005
14	por	0.000003	**0.000014**	49	mismo	**0.000051**	0.000007	84	general	**0.000013**	0.000007
15	una	0.000005	**0.000013**	50	dos	0.000007	**0.000008**	85	antes	0.000005	**0.000009**
16	con	0.000004	**0.000010**	51	entre	0.000004	**0.000012**	86	decir	**0.000031**	0.000006
17	su	0.000008	**0.000011**	52	ser	**0.000020**	0.000005	87	nada	**0.000006**	0.000005
18	lo	0.000005	**0.000010**	53	esta	0.000005	**0.000008**	88	otra	0.000006	**0.000011**
19	es	0.000004	**0.000010**	54	tiempo	**0.000029**	0.000004	89	hombre	**0.000048**	0.000005
20	para	0.000003	**0.000009**	55	todos	0.000006	**0.000020**	90	día	**0.000045**	0.000007
21	al	0.000003	**0.000011**	56	ni	0.000003	**0.000005**	91	siempre	**0.000005**	0.000004
22	más	0.000004	**0.000007**	57	valor	**0.000031**	0.000011	92	otro	**0.000009**	0.000006
23	como	0.000003	**0.000005**	58	tan	0.000004	**0.000007**	93	algo	0.000005	**0.000007**
24	pero	0.000001	**0.000006**	59	muy	0.000005	**0.000009**	94	capital	**0.000025**	0.000011
25	le	0.000007	**0.000009**	60	ella	**0.000016**	0.000004	95	uno	0.000009	0.000009
26	sus	0.000007	**0.000010**	61	son	0.000004	**0.000005**	96	producción	**0.000118**	0.000007
27	o	0.000003	**0.000004**	62	desde	0.000003	**0.000016**	97	tiene	0.000005	**0.000007**
28	había	0.000007	**0.000016**	63	años	**0.000046**	0.000007	98	modo	**0.000070**	0.000006
29	me	0.000007	**0.000010**	64	ahora	**0.000005**	0.000004	99	forma	**0.000019**	0.000009
30	trabajo	**0.000048**	0.000008	65	yo	**0.000009**	0.000008	100	te	**0.000010**	0.000009
31	sin	0.000003	**0.000011**	66	tanto	**0.000011**	0.000004	101	mundo	**0.000070**	0.000004
32	era	0.000004	**0.000007**	67	nos	0.000005	**0.000008**	102	tenía	0.000005	**0.000006**
33	si	0.000007	**0.000008**	68	estaba	0.000005	**0.000007**	103	pues	0.000003	**0.000004**
34	todo	0.000007	**0.000010**	69	porque	0.000003	**0.000005**	104	esto	**0.000008**	0.000006
35	ya	0.000003	**0.000006**	70	así	0.000005	**0.000006**	105	ana	**0.000011**	0.000006
								106	eso	**0.000013**	0.000006

Cues por izquierda: Cue - …… (dimensiones 1 a 66 de los vectores)
de la que el en a los se no las un del por una con su lo es para al más como pero le sus o había me sin era si todo ya sobre cuando ha mi este dijo hasta fue dos entre esta todos ni tan muy son desde nos estaba porque así sino puede cada después está antes otra algo uno tiene tenía pues

Cues por derecha: …… - Cue (dimensiones 67 a 106 de los vectores)
y trabajo él qué también sólo vez mismo ser tiempo valor ella años ahora yo tanto vida parte bien sí levin hay menos casa general decir nada hombre día siempre otro capital producción modo forma te mundo esto ana eso

Tabla 22: Reducción de dimensionalidad sobre las cues según nuestro criterio de Mutual Information (2º Etapa)

De todos modos, es muy llamativa la tendencia que se da entre los rankings más altos de frecuencia (cues 1 a 35) y la informatividad hacia la derecha, tendencia que disminuye en los rankings medios (cues 36 a 70), para revertirse en los rankings finales (cues 71 a 106). Esta tendencia repite un comportamiento convergente que se observa alrededor del punto de corte de la heurística de identificación de cues (véase sección anterior), donde las palabras del DFP parecen cambiar de tendencia. En otras palabras, parece ser que cuanto más arriba se está en la lista del DFP, más tiende la palabra a comportarse como una palabra funcional típica; en cambio, cuando se acerca al punto de corte la palabra parece ser más bien una de contenido. El criterio de reducción de dimensionalidad que aquí postulamos respeta una motivación lingüística y resulta inherente a la cue en sí misma, ya que no depende de los valores que arrojen el cálculo de la

relaciones bigrámicas posteriores con las palabras target (como era el caso de SVD y PCA). La preeminancia de la informatividad de una cue hacia la derecha o hacia la izquierda debería ser la misma para todo corpus masivo representativo del español.

Por otra parte, ante la evidencia que se desprende de la Tabla 22 podemos aventurar cierta repercusión en las propiedades distribucionales de las palabras, incluso en distinciones gramaticales que se creerían estancas, como la diferencia entre palabras funcionales y palabras de contenido. Para ilustrar este punto, considérese que los pronombres personales en caso nominativo tenderían a tener informatividad hacia su izquierda ('*él*', '*yo*', etc.), mientras que los pronombres personales en caso acusativo lo harían hacia la derecha ('*te*'), lo cual resulta obvio si se considera que los encíticos suelen ser usados en su gran mayoría por delante de los verbos a los que acompañan estructuralmente. Es decir, las distintas categorías morfosintácticas evidencia propiedades distribucionales diferenciadas, justificando así la discriminación de la categoría pronombres personales en al menos dos subcategorías morfosintácticas (nominativos y acusativos enclíticos). Esta necesidad de una granularidad refinada en la caracterización de las palabras morfosintácticas en función de propiedades distribucionales, más allá de los compartimentos estancos tradicionales de la gramática, será comentada en detalle en las secciones 7.9 a 7.11 de esta tesis).

7.5 Tercera etapa del algoritmo: Construcción del espacio vectorial

Una vez identificadas las palabras cues (según nuestro corpus, las 106 palabras más frecuentes), procedemos a seleccionar las 1000 palabras target siguientes en el DFP. Este recorte en los objetos del espacio vectorial nos acerca a las líneas de diseño de Redington *et al*. (1998):

> "It is not necessary (or even desirable) to record these statistics for every word in the input in order to provide useful information. From a psychological perspective, in the early stages of syntactic category acquisition, it seems unlikely that a syntactic category will be assigned to every word in the child's input, particularly given that the child's vocabulary is very limited. […] It may also be computationally appropriate to focus on a small number of target words in order to provide more reliable distributional information and to avoid unnecessarily complex computation. Moreover, it may be appropriate to be even more restrictive with respect to the set of context words (over which frequency distributions are observed). This is because each target word may occur in a relatively small number of contexts, and only the most frequent words in these contexts will provide reliable frequency information." [Redington *et al*. 1998:436]

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

En nuestro experimento, las 1000 palabras target y las 106 palabras cues tienen una cobertura total en tokens de las 2/3 partes del corpus, totalizando 1,2 millones de tokens:

gran(1318) todas(1310) fuerza(1293) ese(1290) toda(1282) entonces(1270) misma(1256)
hecho(1240) esa(1238) durante(1232) he(1228) otros(1227) cómo(1226) donde(1225) aquí(1213)
sido(1203) obreros(1195) sofía(1186) mucho(1176) habían(1174) poco(1166) aquel(1140)
vronsky(1135) poder(1121) hace(1119) han(1115) les(1108) dinero(1099) usted(1087) hacer(1085)
mayor(1076) estado(1053) nunca(1049) ellos(1034) veces(1028) aunque(1028) ejemplo(1025)
eran(1006) horas(1003) cambio(1000) contra(1000) mientras(999) tres(986) cosas(975)
medio(938) nuevo(931) hombres(930) tal(930) unos(925) casi(925) dios(924) podía(919)
mujer(915) mejor(913) sea(912) momento(912) primera(909) hacia(907) bajo(904) noche(896)
luego(894) kitty(894) madre(888) ojos(885) días(884) embargo(876) mano(876) ante(871) allí(870)
mercancías(870) proceso(868) fin(864) ver(849) obrero(842) además(841) estos(838) cual(831)
etc(824) según(820) otras(816) aquella(815) estas(807) ley(806) capitalista(806) p(800) of(794)
punto(790) país(788) siglo(787) e(785) the(779) fuera(776) libro(774) lugar(768) historia(763)
quien(740) manera(739) grandes(735) razón(733) nadie(732) todavía(730) haber(725) obra(720)
mis(719) mercancía(714) verdad(714) mí(713) año(707) éste(699) palabras(694) hoy(693)
cuenta(687) cuerpo(684) tierra(682) realidad(682) hizo(676) muchos(674) naturaleza(672)
cuanto(671) medios(669) aún(667) dice(662) número(662) niños(660) caso(659) cosa(657)
primer(656) voz(646) pronto(639) muerte(637) ello(626) tarde(625) nueva(619) preguntó(618)
posible(612) clase(610) manos(609) hora(607) debe(601) nuestro(593) dicho(591) nosotros(591)
mal(590) padre(589) producto(589) jornada(588) cabeza(587) sistema(584) habría(583)
sociedad(583) hacía(582) cierto(576) lado(575) dentro(573) iba(571) social(570) oro(569)
unas(568) alma(567) propio(565) hablar(562) tu(554) esos(553) algunos(542) saber(541)
guerra(541) demás(540) cuatro(538) idea(538) pueblo(537) va(536) parece(536) ciudad(534)
fueron(533) muchas(532) nuestra(529) tener(526) personas(525) solo(522) industria(521)
decía(517) palabra(516) tienen(515) gente(515) sentido(515) estaban(514) propia(511)

Tabla 23: Algunos ejemplos de las 1000 palabras target del experimento con su frecuencia absoluta en el DFP

Se procede entonces a realizar el cómputo de bigramas entre cada palabra *cue* (según su informatividad a derecha o a izquierda) y cada palabra target. En cuanto al cálculo de bigramas, debemos destacar que nuestro algoritmo toma en cuenta la información de límite de frases fonológicas (especialmente las marcadas en la escritura por el signo ‘,’) y límite de enunciados a través de los signos de puntuación: esto obliga a un tratamiento encapsulado de expresiones parentéticas, dislocaciones de constituyentes y algunos casos de subordinadas. Obsérvese el siguiente ejemplo:

Afortunadamente para mí , éste sería el ejemplo . El ejemplo sigue aquí .
W_1 W_2 W_3, W_4 W_5 W_6 W_7 . W_6 W_7 W_8 W_9 .

Supongamos que las cues son ‘*el*’ (a derecha) , ‘*para*’ (a derecha), ‘*aquí*’ (a izquierda)
Supongamos que las palabras target son ‘*afortunadamente*’, ‘*mí*’, ‘*sería*’, ‘*ejemplo*’, ‘*sigue*’

En este caso, los posibles bigramas a contabilizar serían W_2-W_3, W_6-W_7, W_6-W_7, W_8-W_9

El vector de 3 dimensiones de la palabra target ‘*ejemplo*’ (W_7) = (2, 0, 0)
El vector de 3 dimensiones de la palabra target ‘*afortunadamente*’ (W_1) = (0, 0, 0)
El vector de 3 dimensiones de la palabra target ‘*sigue*’ (W_8) = (0, 0, 1)
El vector de 3 dimensiones de la palabra target ‘*mí*’ (W_3) = (0, 1, 0)

Obviamente, con estas restricciones en el cómputo de bigramas significativos, un par de oraciones no aporta ninguna clarificación respecto de la información distribucional de las

palabras. No obstante, a lo largo de casi un centenar de miles de oraciones es esperable que los perfiles distribucionales (vectores de 106 dimensiones) de las palabras target sean muy significativos. Por ejemplo, la Tabla 24 muestra el vector de la palabra '*embargo*' en función de las 106 dimensiones de las cues identificadas:

[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, **1.0**, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, **875.0**, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 2.0, 0.0]

Tabla 24: Vector de 106 dimensiones de la palabra '*embargo*' (véase *Tabla 22* para identificar cada dimensión)

Obsérvese que la dimensión 29 tiene un pico claro de 875 ocurrencias bigrámicas entre tal cue (a derecha) y la palabra '*embargo*'. Tal como se indica en la Tabla 22, la dimensión 29º es la palabra '*sin*'. A su vez existen algunas ocurrencias marginales gramaticales en la dimensión 11º del vector ('*un embargo*') y otras a izquierda de las cues, cuya posible explicación sea la falta de puntuación para ejemplos como "*sin embargo, tanto*…", "*sin embargo, Ana*…" (con 2 ocurrencias en la dimensión 105º); pero en su enorme mayoría, el perfil distribucional de '*embargo*' respecto de cada una de las 106 cues es totalmente identificatorio del tipo de palabra, a tal punto que luego veremos en la salida de clusters que se trata de un objeto apartado (*outlier*) en el espacio vectorial, constituyendo un cluster de miembro único y clase indecidible (véase sección *7.7 Resultados*). Esto tiene su justificación lingüística al ser considerada la locución '*sin embargo*' como un giro lingüístico conjuntivo en español, caso único en las palabras target.

La plausibilidad de la habilidad temprana para reconocer frases fonológicas por parte de los adquirientes de un lenguaje natural ya ha sido explicada en detalle anteriormente (Mehler *et al*. 1998; Jusczyk *et al*. 1999; Christophe *et al*. 2008) y resulta un gran aporte a los experimentos de clustering tradicionales, los cuales, en el mejor de los casos, sólo consideraban el límite de las oraciones (Redington *et al*. 1998; Clark 2002). Nuestro modelo de relaciones bigrámicas resulta así más adecuado descriptiva y explicativamente, ya que la determinación de la ocurrencia del POS-tag de un token en función de su contexto inmediato no suele extenderse más allá de la frase fonológica (Balbachan y Dell'Era 2010; Ćavar 2010) -la única excepción sería el uso de la ',' en enumeraciones -véase concepto de *no-constituyente* (*dis-tituent*) en Ćavar (2010).

7.6 Cuarta etapa del algoritmo: Clustering K-means iterativo

Como explicamos en detalle en la sección *4.3 Clustering no jerárquico o partitivo*, nuestra implementación del clustering K-means incluye algunas modificaciones con el fin de analizar las distintas distribuciones posibles sobre el mismo set de los 1000 vectores de 106 dimensiones (cues). Este proceso de historización del agrupamiento en clustering, permitirá sacar algunas conclusiones acerca de la consolidación de las diversas particiones más homogéneas y

posibilitará una evaluación mucho más detallada de la salida. El algoritmo que implementamos para el clustering K-means se detalla a continuación:

1) Comenzar por ubicar 2 centroides al azar (ciclo K=2).
2) Calcular la distancia euclideana de cada uno de los objetos del espacio vectorial a dichos centroides y asignarlos a uno u otro en función de la distancia mínima.
3) Computar el error de ciclo como la sumatoria de las distancias euclideanas a sus respectivos centroides de todos los vectores de cada cluster. El error de ciclo no necesariamente se corresponde con la adecuación de la distribución en términos de las categorías inducidas, por lo que no es un criterio de finalización confiable –es decir, siempre irá disminuyendo con cada ciclo nuevo que "acomode" mejor a los vectores, dada la disponibilidad de más centroides para formar clusters.
4) Comenzar una nueva iteración con un nuevo cluster (K = n+1), inicializar los correspondientes centroides al azar y reasignar los vectores a los nuevos centroides.
5) Recalcular centroides para los nuevos clusters y el nuevo error de ciclo.
6) Iterar el algoritmo desde el paso 2) hasta que el error de ciclo de una nueva asignación sea mayor que el de la iteración actual o hasta que se alcanza el ciclo K= 106 (número de cues).

En cada ciclo adicionalmente computamos la ubicación del centroide de cada cluster como una forma de representar cuán cercanos o apartados están los clusters entre sí. Esta información será aprovechada a la hora de justificar el agrupamiento de clusters (*merging*) en *hiperclusters* (véase subsección *7.9.3*). El centroide de un cluster en un ciclo dado proporciona una interesante caracterización de los miembros del cluster respecto de la ocurrencias bigrámicas entre las cues y dichos miembros hacia un lado u otro del contexto. Por ejemplo, obsérvese el siguiente cluster del ciclo 87 con el vector que representa al centroide (cluster_centroid):

Cluster: 53 (tamaño = 4) cluster tag: (NN1) **cluster_centroid** = [4.0, 0.75, 0.0, **83.0**, 33.0, 32.25, 0.0, 0.0, 0.0, 0.0, 17.75, 7.5, 43.25, 0.0, 0.0, 10.0, 0.0, 0.5, 0.0, **197.5**, 0.0, 3.5, 0.0, 0.0, 0.0, 0.5, 0.0, 0.0, 2.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.25, 2.75, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.25, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.25, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.25, 0.0, 0.0, 21.75, 0.0, 0.25, 0.0, 0.25, 0.75, 0.0, 0.75, 0.0, 0.0, 0.0, 0.25, 0.0, 0.0, 0.0, 0.25, 0.0, 0.0, 0.0, 0.25, 0.0, 0.25, 0.25, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.25, 0.0, 0.0, 0.0, 0.0, 0.25, 0.0, 0.0, 0.0]
fin/NN1_TP principio/NN1_TP final/NN1_TP cabo/NN1_TP

Tabla 25: Ejemplo de cluster y su centroide en salida de experimento

Nótese que la granularidad en el agrupamiento nos permite inferir que estos cuatro sustantivos no sólo son singulares (Verdaderos Positivos o *True Positives* TP del tag NN1 según Tabla 26, de ahí la anotación 'NN1_TP' que acompaña a cada miembro) sino también masculinos. Pero lo que es importante de destacar es que el agrupamiento dio con una particularidad aún más refinada para este grupo, la cual los aparta de otros grupos de sustantivos singulares masculinos: constituyen giros lingüísticos adverbiales encabezados por '*al*' ('*al fin y al cabo*', '*al final*', '*al principio*'). Para hacer notar esto en el espacio vectorial, tómese en cuenta la prevalencia en la ubicación del centroide que aporta al grupo la dimensión 20º ('*al*'-), por mucho en primer lugar, y en menor medida de la dimensión 4º ('*el*'-) en segundo lugar, muy por arriba de las otras dimensiones (véase *Tabla 22* para una identificación de las cues en las 106 dimensiones del

vector). Este análisis justifica la intuición de que el centroide representa bastante bien a los miembros pertenecientes al cluster; aun cuando su ubicación en el espacio vectorial puede ser más representativa para algunos miembros más prototípicamente asignados al cluster que para otros más apartados del centroide. Por ejemplo, la dominancia de la dimensión ('*al*'-) se aplica a todos los miembros, pero éste no es el caso de la segunda componente con mayor peso ('*el*'-), cuya ocurrencia en bigramas '*el cabo*' seguramente es menos frecuente que para los otros tres miembros del cluster.

El ejemplo anterior también ilustra otra línea de diseño del experimento con importantes repercusiones a la hora de evaluar la salida, la asignación con que se diagramó el experimento es la de *hard clustering* (Manning y Schütze 1999): cada vector sólo puede ser asignado a un único cluster. Esto deja de lado la posibilidad de considerar ambigüedades a nivel del tipo de palabra morfosintáctica como sí lo hacían los experimentos de Schütze (1993) y de Clark (2002). En el caso del cluster del ejemplo, '*final*' puede ser considerado también un adjetivo. Al igual que Redington *et al*. (1998), nosotros recurrimos a un corpus de referencia de POS-tag que desambigüe formas léxicas brindando el POS-tag más frecuente (en el caso de '*final*', sustantivo). En las secciones siguientes brindaremos más detalles acerca de la construcción y tratamiento del corpus de referencia para desambiguación de POS-tag. Por ahora, también queremos hacer notar que la ambigüedad morfosintáctica en español no es un escollo tan insalvable como en el inglés (Graça *et al*. 2011) (véase en esta tesis el lineamiento de diseño 7 de la sección *7.1 Motivación de las decisiones de diseño*).

7.7 Resultados

Por motivos de espacio en las próximas páginas reproducimos sólo la salida del ciclo 87 del experimento de categorización. Dicho ciclo ha sido evaluado como el más efectivo en cuanto a la distribución de clusters (véase sección *7.10 Evaluación iterativa de los ciclos de clustering con la métrica many-to-1*) . La nomenclatura que especifica el tipo de palabra morfosintáctica inducida en función de la presencia mayoritaria de ciertos miembros será explicada en detalle en la tabla siguiente. El conjunto de etiquetas morfosintácticas (POS-tags) que usaremos se basa en el llamado *C4 tagset* (Leech *et al*. 1994; Clark 2002), un estándar que se aplicó al etiquetado del *British National Corpus*. Como vamos a etiquetar texto en español, prescindimos de algunas etiquetas del inglés y agregamos otras más apropiadas para nuestro lenguaje.

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, pero*)
12	CJS	conjunción subordinante excepto *que* (*cuando*)
13	CJT	conjunción subordinante *que* (en "*Dijo que...*")
14	CRD	adjetivo numeral cardinal (*tres*)
15	DAT	fecha (*14/02/2001*)
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 (*Londres*)
25	ORD	adjetivo numeral ordinal (*sexto, 3ro.*)
26	PND	pronombre demostrativo (*Éste***)**
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 (***se lo*** *dio*) cuasi-reflejo (él ***se*** *cayó*)
32	PRP	preposición (*sin*)
33	REL	pronombre relativo (*quien* en "*el presidente, quien avisó...*")
34	SEP	se pasivo e impersonal (***se*** *venden casas,* ***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 (*debido*)
43	VVG	gerundio de verbo léxico (*obrando*)
44	VVI	infinitivo de verbo léxico (*vivir*)
45	VVN	participio de verbo léxico (*comido*)
46	VVZ	verbo léxico conjugado (*vive*)
47	XX0	adverbio de negación (*no*)

Tabla 26: Nomenclatura adaptada del C4 para marcación morfosintáctica automática de palabras y de clusters inducidos

En las próximas 4 páginas incluimos la salida completa del ciclo 87 de clustering. Nótese que los clusters inducidos para el ciclo 87 son presentados no en un orden numérico sino justamente en función del tipo de palabra morfosintáctica predominante entre los miembros de la clase (*cluster_tag*). Finalmente, cada miembro tiene asignado automáticamente un tag /POS-TAG_TP o /POS-TAG_FP en función de que se trate de un *True Positive* TP (Verdadero

Positivo entre el POS-tag del miembro y el de la clase) o *False Positive* FP (Falso Positivo entre el POS-tag del miembro y el de la clase), respectivamente. Para conocer más sobre el criterio de asignación automática de POS-tag a un cluster (inducción de categoría) y a un miembro (evaluación de la pertenencia de dicho miembro a la clase) véase la sección siguiente. Se han resaltado en verde (resultados muy buenos), en amarillo (problemas en corpus de referencia) y en rojo (resultados malos) diversas características a ser analizadas más adelante en la sección *7.11 Discusión de los resultados y conclusiones*.

Tabla 27: Salida completa del ciclo 87 de clustering----------

```
Clustering con k = 87   cycle error= 30877.6285551
Cluster: 3  ( tamaño = 5 ) cluster tag:  (u'AJ1', 2.563296192678798)
posible/AJ1_TP  necesario/AJ1_TP  imposible/AJ1_TP  preciso/AJ1_TP  evidente/AJ1_FP
Cluster: 45  ( tamaño = 21 ) cluster tag:  (u'AJ1', 5.767416433527296) personal/AJ1_FP
material/AJ1_FP  suyo/AJ1_TP  peor/AJ1_FP  blanco/AJ1_TP  negro/AJ1_TP  público/AJ1_TP
pobre/AJ1_TP  principal/AJ1_TP  fuego/AJ1_FP  capítulo/AJ1_FP  quijote/AJ1_FP  término/AJ1_FP
pan/AJ1_FP  universal/AJ1_TP  extranjero/AJ1_TP  miedo/AJ1_FP  presente/AJ1_FP  culpable/AJ1_TP
trigo/AJ1_FP  control/AJ1_FP
Cluster: 54  ( tamaño = 4 ) cluster tag:  (u'AJ1', 1.9224721445090984)
total/AJ1_TP  inglés/AJ1_TP  silencio/AJ1_FP  francés/AJ1_TP
Cluster: 55  ( tamaño = 4 ) cluster tag:  (u'AJ1', 1.281648096339399)
mejor/AJ1_FP  único/AJ1_TP  largo/AJ1_TP  contrario/AJ1_FP
Cluster: 59  ( tamaño = 32 ) cluster tag:  (u'AJ1', 10.253184770715192)
social/AJ1_TP  unidos/AJ1_FP  arkadievich/AJ1_FP  humano/AJ1_TP  claro/AJ1_TP  natural/AJ1_TP
libre/AJ1_TP  bueno/AJ1_FP  real/AJ1_TP  constante/AJ1_FP  nacional/AJ1_TP  anterior/AJ1_TP
esterlinas/AJ1_FP  chelines/AJ1_FP  feliz/AJ1_TP  ivanovich/AJ1_FP  muerto/AJ1_TP  militar/AJ1_TP
xviii/AJ1_FP  baja/AJ1_FP  terrible/AJ1_TP  humanos/AJ1_FP  inglesa/AJ1_TP  útil/AJ1_TP
naturales/AJ1_FP  normal/AJ1_TP  sociales/AJ1_FP  alegre/AJ1_TP  sonriendo/AJ1_FP
conocido/AJ1_FP  rápidamente/AJ1_FP  espera/AJ1_FP
Cluster: 71  ( tamaño = 8 ) cluster tag:  (u'AJ1', 3.2041202408484977)
grande/AJ1_TP  cerca/AJ1_FP  lejos/AJ1_FP  fácil/AJ1_TP  alto/AJ1_TP  difícil/AJ1_TP
fuerte/AJ1_TP  arriba/AJ1_FP
Cluster: 73  ( tamaño = 4 ) cluster tag:  (u'AJ1', 1.281648096339399)
ciento/AJ1_FP  encima/AJ1_FP  completo/AJ1_TP  supuesto/AJ1_TP
----------------------------------------
Cluster: 21  ( tamaño = 9 ) cluster tag:  (u'AJ2', 2.5136508166112908)
otros/AJ2_TP  grandes/AJ2_TP  demás/AJ2_FP  estados/AJ2_FP  países/AJ2_FP  mismos/AJ2_TP
precios/AJ2_FP  primeros/AJ2_FP  cuales/AJ2_FP
Cluster: 66  ( tamaño = 21 ) cluster tag:  (u'AJ2', 7.540952449833871) casos/AJ2_FP
nuevos/AJ2_TP  oficiales/AJ2_TP  jóvenes/AJ2_FP  mayores/AJ2_FP  pocos/AJ2_TP  elementos/AJ2_FP
esclavos/AJ2_FP  pequeños/AJ2_TP  negros/AJ2_TP  objetos/AJ2_FP  diversos/AJ2_TP  ingleses/AJ2_TP
efectos/AJ2_FP  siglos/AJ2_FP  distintos/AJ2_TP  blancos/AJ2_TP  sueños/AJ2_FP
instrumentos/AJ2_FP  capitales/AJ2_FP  mejores/AJ2_FP
Cluster: 86  ( tamaño = 22 ) cluster tag:  (u'AJ2', 5.0273016332225815)
cómo/AJ2_FP  unas/AJ2_FP  algunos/AJ2_FP  muchas/AJ2_TP  alguna/AJ2_FP  ningún/AJ2_FP
ninguna/AJ2_FP  cuyo/AJ2_FP  algunas/AJ2_TP  buenos/AJ2_TP  ambos/AJ2_TP  aquellas/AJ2_FP
común/AJ2_FP  varios/AJ2_TP  tales/AJ2_TP  nuestras/AJ2_FP  cierta/AJ2_FP  moscú/AJ2_FP
santa/AJ2_FP  iquitos/AJ2_FP  francia/AJ2_FP  parís/AJ2_FP
----------------------------------------
Cluster: 49  ( tamaño = 95 ) cluster tag:  (u'AV0', 18.284305693045056)
aunque/AV0_FP  mientras/AV0_TP  bajo/AV0_FP  etc/AV0_FP  según/AV0_FP  p/AV0_FP  of/AV0_FP
e/AV0_FP  the/AV0_FP  dentro/AV0_TP  pantoja/AV0_FP  incluso/AV0_TP  demasiado/AV0_TP
precisamente/AV0_TP  aun/AV0_TP  junto/AV0_TP  alejandrovich/AV0_FP  gracias/AV0_FP
siquiera/AV0_TP  mediante/AV0_FP  quizá/AV0_TP  meses/AV0_FP  acaso/AV0_TP  siendo/AV0_FP
acerca/AV0_TP  mirando/AV0_FP  salió/AV0_FP  pensó/AV0_FP  pp/AV0_FP  empezó/AV0_FP  and/AV0_FP
simplemente/AV0_TP  especial/AV0_FP  industrial/AV0_FP  bastante/AV0_TP  completamente/AV0_TP
constantemente/AV0_TP  d/AV0_FP  escribió/AV0_FP  finalmente/AV0_TP  directamente/AV0_TP
seguro/AV0_FP  humana/AV0_FP  latina/AV0_FP  comenzó/AV0_FP  entró/AV0_FP  vamos/AV0_FP
solamente/AV0_TP  mundial/AV0_FP  mas/AV0_TP  palacios/AV0_FP  vivo/AV0_FP  haciendo/AV0_FP
v/AV0_FP  alguno/AV0_FP  quizás/AV0_TP  in/AV0_FP  detrás/AV0_TP  capaz/AV0_FP  querido/AV0_FP
minutos/AV0_FP  siguió/AV0_FP  fabril/AV0_FP  aires/AV0_FP  individual/AV0_FP  excedente/AV0_FP
exclamó/AV0_FP  on/AV0_FP  penal/AV0_FP  añadió/AV0_FP  productiva/AV0_FP  exactamente/AV0_TP
totalmente/AV0_TP  alejandrovna/AV0_FP  ed/AV0_FP  delante/AV0_TP  naturalmente/AV0_TP
variable/AV0_FP  distinto/AV0_FP  obrera/AV0_FP  pensando/AV0_FP  cualquiera/AV0_FP
alrededor/AV0_TP  relativa/AV0_FP  sentado/AV0_FP  entero/AV0_FP  mamá/AV0_FP  realmente/AV0_TP
dicen/AV0_FP  t/AV0_FP  ah/AV0_FP  panta/AV0_FP  l/AV0_FP  pública/AV0_FP
Cluster: 51  ( tamaño = 4 ) cluster tag:  (u'AV0', 1.3543930142996337)
```

aún/AV0_TP importante/AV0_FP allá/AV0_TP importantes/AV0_FP
Cluster: 81 (tamaño = 3) cluster tag: (u'AV0', 2.0315895214494506)
entonces/AV0_TP luego/AV0_TP allí/AV0_TP

Cluster: 67 (tamaño = 11) cluster tag: (u'CRD', 7.784332528288198) tres/CRD_TP muchos/CRD_FP
cuatro/CRD_TP cinco/CRD_TP diez/CRD_TP seis/CRD_TP veinte/CRD_TP ocho/CRD_TP siete/CRD_TP
treinta/CRD_TP doce/CRD_TP

Cluster: 6 (tamaño = 5) cluster tag: (u'DT1', 3.779763149684619)
todas/DT1_FP ese/DT1_TP toda/DT1_TP esa/DT1_TP ellos/DT1_FP

Cluster: 19 (tamaño = 13) cluster tag: (u'DT2', 6.729563462959896)
dios/DT2_FP estos/DT2_TP aquella/DT2_FP estas/DT2_TP mis/DT2_FP nuestro/DT2_FP esos/DT2_TP
nuestra/DT2_FP aquellos/DT2_TP cualquier/DT2_FP américa/DT2_FP ellas/DT2_FP san/DT2_FP

Cluster: 1 (tamaño = 33) cluster tag: (u'NN1', 12.54936650018316)
puesto/NN1_TP objeto/NN1_TP interés/NN1_TP recuerdo/NN1_TP resultado/NN1_TP grado/NN1_TP
destino/NN1_TP jefe/NN1_TP asunto/NN1_TP favor/NN1_TP juego/NN1_TP nivel/NN1_TP
estilo/NN1_TP consumo/NN1_TP pensamiento/NN1_TP interior/NN1_TP aspecto/NN1_TP
volumen/NN1_TP período/NN1_TP doble/NN1_FP deseo/NN1_TP dolor/NN1_TP secreto/NN1_TP
tono/NN1_TP cuento/NN1_TP aumento/NN1_TP profesor/NN1_TP empleo/NN1_TP doctor/NN1_TP
lector/NN1_TP método/NN1_TP origen/NN1_TP motivo/NN1_TP
Cluster: 4 (tamaño = 5) cluster tag: (u'NN1', 1.568670812522895) padre/NN1_TP alma/NN1_TP
propio/NN1_FP nombre/NN1_TP rostro/NN1_TP
Cluster: 8 (tamaño = 5) cluster tag: (u'NN1', 1.568670812522895)
plusvalía/NN1_TP vista/NN1_FP plata/NN1_TP filosofía/NN1_TP circulación/NN1_TP
Cluster: 9 (tamaño = 4) cluster tag: (u'NN1', 1.568670812522895) relación/NN1_TP
función/NN1_TP proporción/NN1_TP marcha/NN1_TP
Cluster: 11 (tamaño = 6) cluster tag: (u'NN1', 1.568670812522895)
cosa/NN1_TP sola/NN1_FP persona/NN1_TP serie/NN1_TP especie/NN1_TP determinada/NN1_FP
Cluster: 12 (tamaño = 38) cluster tag: (u'NN1', 14.510205015836778) propiedad/NN1_TP
causa/NN1_TP mirada/NN1_TP conversación/NN1_TP conciencia/NN1_TP necesidad/NN1_TP
novela/NN1_TP diferencia/NN1_TP libertad/NN1_TP magnitud/NN1_TP base/NN1_TP fe/NN1_TP
suma/NN1_TP materia/NN1_TP atención/NN1_TP memoria/NN1_FP cara/NN1_TP fábrica/NN1_TP
explotación/NN1_TP posibilidad/NN1_TP impresión/NN1_TP cuestión/NN1_TP paz/NN1_TP
revolución/NN1_TP práctica/NN1_TP sala/NN1_TP comida/NN1_TP empresa/NN1_TP boca/NN1_TP
capacidad/NN1_TP región/NN1_TP ciencia/NN1_TP condesa/NN1_TP comunicación/NN1_TP
noticia/NN1_TP policía/NN1_TP plaza/NN1_TP demanda/NN1_TP
Cluster: 13 (tamaño = 18) cluster tag: (u'NN1', 4.706012437568685)
buena/NN1_FP simple/NN1_FP sonrisa/NN1_TP respuesta/NN1_TP nota/NN1_TP suerte/NN1_TP
pregunta/NN1_TP pequeña/NN1_FP frase/NN1_TP unidad/NN1_TP larga/NN1_FP fiesta/NN1_TP
escala/NN1_TP niña/NN1_TP prueba/NN1_TP tercera/NN1_FP imagen/NN1_TP verdadera/NN1_FP
Cluster: 16 (tamaño = 14) cluster tag: (u'NN1', 5.490347843830133)
poder/NN1_TP dinero/NN1_TP estado/NN1_TP momento/NN1_TP proceso/NN1_TP obrero/NN1_TP
país/NN1_TP siglo/NN1_TP libro/NN1_TP año/NN1_TP cuerpo/NN1_TP producto/NN1_TP
sistema/NN1_TP pueblo/NN1_TP
Cluster: 20 (tamaño = 4) cluster tag: (u'NN1', 1.1765031093921712)
siguiente/NN1_FP orden/NN1_TP joven/NN1_TP menor/NN1_TP
Cluster: 22 (tamaño = 13) cluster tag: (u'NN1', 4.706012437568685)
servicio/NN1_TP campo/NN1_TP río/NN1_TP café/NN1_TP castigo/NN1_FP ejército/NN1_TP
suelo/NN1_TP crimen/NN1_TP enunciado/NN1_TP mar/NN1_TP sur/NN1_TP universo/NN1_TP
norte/NN1_TP
Cluster: 24 (tamaño = 5) cluster tag: (u'NN1', 1.9608385156536188) amigo/NN1_TP hijo/NN1_TP
esposa/NN1_TP hija/NN1_TP hermana/NN1_TP
Cluster: 25 (tamaño = 6) cluster tag: (u'NN1', 1.1765031093921712) nuevo/NN1_FP haber/NN1_FP
pronto/NN1_FP oro/NN1_TP uso/NN1_TP acuerdo/NN1_TP
Cluster: 26 (tamaño = 12) cluster tag: (u'NN1', 3.9216770313072375)
capitalista/NN1_TP punto/NN1_TP número/NN1_TP caso/NN1_TP primer/NN1_FP precio/NN1_TP
mercado/NN1_TP gobierno/NN1_TP camino/NN1_TP último/NN1_FP fondo/NN1_TP capitán/NN1_TP
Cluster: 27 (tamaño = 4) cluster tag: (u'NN1', 1.1765031093921712)
madre/NN1_TP propia/NN1_FP hermano/NN1_TP marido/NN1_TP
Cluster: 34 (tamaño = 12) cluster tag: (u'NN1', 3.9216770313072375)
fuerza/NN1_TP misma/NN1_FP primera/NN1_FP noche/NN1_TP mano/NN1_TP ley/NN1_TP
historia/NN1_TP tierra/NN1_TP naturaleza/NN1_TP jornada/NN1_TP cabeza/NN1_TP puerta/NN1_TP
Cluster: 36 (tamaño = 38) cluster tag: (u'NN1', 11.372863390790988)
mal/NN1_FP sentido/NN1_TP señor/NN1_TP amor/NN1_TP agua/NN1_TP derecho/NN1_FP
primero/NN1_FP carácter/NN1_TP paso/NN1_TP régimen/NN1_TP corazón/NN1_TP presidente/NN1_TP
desarrollo/NN1_TP aire/NN1_TP viaje/NN1_TP papel/NN1_TP segundo/NN1_FP viejo/NN1_FP
autor/NN1_TP cuarto/NN1_FP movimiento/NN1_TP salario/NN1_TP problema/NN1_TP espíritu/NN1_TP
médico/NN1_FP arte/NN1_TP sol/NN1_FP comercio/NN1_TP cielo/NN1_TP tema/NN1_TP coche/NN1_TP
centro/NN1_TP espacio/NN1_TP coronel/NN1_TP príncipe/NN1_TP brazo/NN1_TP resto/NN1_TP
curso/NN1_FP
Cluster: 38 (tamaño = 36) cluster tag: (u'NN1', 12.157198797052436) política/NN1_TP
frente/NN1_TP falta/NN1_TP importancia/NN1_TP mala/NN1_FP moral/NN1_TP felicidad/NN1_TP

alegría/NN1_TP alta/NN1_FP figura/NN1_TP ayuda/NN1_TP voluntad/NN1_TP sangre/NN1_TP
costumbre/NN1_TP salud/NN1_TP opinión/NN1_TP decisión/NN1_TP actividad/NN1_TP ocasión/NN1_TP
condición/NN1_TP suya/NN1_FP técnica/NN1_FP experiencia/NN1_TP vuelta/NN1_TP teoría/NN1_TP
carne/NN1_TP acción/NN1_TP venta/NN1_TP educación/NN1_TP compra/NN1_TP levita/NN1_FP
música/NN1_TP disciplina/NN1_TP visita/NN1_TP culpa/NN1_TP sombra/NN1_TP
Cluster: 41 (tamaño = 12) cluster tag: (u'NN1', 2.7451739219150664)
tipo/NN1_TP oficial/NN1_FP vestido/NN1_FP escritor/NN1_TP sitio/NN1_TP verdadero/NN1_FP
extraño/NN1_FP filósofo/NN1_TP encuentro/NN1_TP mes/NN1_TP determinado/NN1_FP
esfuerzo/NN1_TP
Cluster: 50 (tamaño = 4) cluster tag: (u'NN1', 1.1765031093921712)
medio/NN1_TP cuenta/NN1_TP efecto/NN1_TP seguida/NN1_FP
Cluster: 53 (tamaño = 4) cluster tag: (u'NN1', 1.568670812522895)
fin/NN1_TP principio/NN1_TP final/NN1_TP cabo/NN1_TP
Cluster: 57 (tamaño = 10) cluster tag: (u'NN1', 3.5295093281765135)
manera/NN1_TP obra/NN1_TP nueva/NN1_FP hora/NN1_TP palabra/NN1_TP carta/NN1_TP
cantidad/NN1_TP máquina/NN1_TP expresión/NN1_TP semana/NN1_TP
Cluster: 58 (tamaño = 10) cluster tag: (u'NN1', 3.13734162504579)
lugar/NN1_TP lado/NN1_TP buen/NN1_FP niño/NN1_TP sueño/NN1_TP instante/NN1_TP
sentimiento/NN1_TP pequeño/NN1_FP par/NN1_TP grupo/NN1_TP
Cluster: 61 (tamaño = 30) cluster tag: (u'NN1', 10.588527984529541)
tarde/NN1_TP clase/NN1_TP época/NN1_TP situación/NN1_TP edad/NN1_TP maquinaria/NN1_TP
masa/NN1_TP última/NN1_FP luz/NN1_TP pena/NN1_TP existencia/NN1_TP división/NN1_TP
señora/NN1_TP prisión/NN1_TP mitad/NN1_TP calle/NN1_TP acumulación/NN1_TP justicia/NN1_TP
segunda/NN1_FP mayoría/NN1_TP iglesia/NN1_TP escuela/NN1_TP riqueza/NN1_TP literatura/NN1_TP
única/NN1_FP habitación/NN1_TP cama/NN1_TP manufactura/NN1_TP lengua/NN1_TP ventana/NN1_TP
Cluster: 62 (tamaño = 16) cluster tag: (u'NN1', 6.27468325009158)
razón/NN1_TP mercancía/NN1_TP verdad/NN1_TP muerte/NN1_TP sociedad/NN1_TP guerra/NN1_TP
idea/NN1_TP ciudad/NN1_TP industria/NN1_TP gente/NN1_TP mañana/NN1_TP familia/NN1_TP
economía/NN1_TP mesa/NN1_TP población/NN1_TP princesa/NN1_TP
Cluster: 76 (tamaño = 2) cluster tag: (u'NN1', 0.7843354062614475) realidad/NN1_TP
voz/NN1_TP
Cluster: 79 (tamaño = 12) cluster tag: (u'NN1', 4.313844734437962)
caballo/NN1_TP maestro/NN1_TP discurso/NN1_TP individuo/NN1_TP jardín/NN1_TP poeta/NN1_TP
acto/NN1_TP género/NN1_TP antiguo/NN1_FP banco/NN1_TP partido/NN1_TP periódico/NN1_TP

Cluster: 29 (tamaño = 23) cluster tag: (u'NNP', 6.511939230802893)
tu/NNP_FP saber/NNP_FP tener/NNP_FP quién/NNP_FP ésta/NNP_FP esteban/NNP_TP dolly/NNP_TP
sergio/NNP_TP alberto/NNP_TP alexey/NNP_TP vivir/NNP_FP trabajar/NNP_FP karenin/NNP_TP
maría/NNP_TP alguien/NNP_FP escribir/NNP_FP oblonsky/NNP_TP bush/NNP_TP comer/NNP_FP
leer/NNP_FP dante/NNP_TP éstos/NNP_FP platón/NNP_TP
Cluster: 30 (tamaño = 6) cluster tag: (u'NNP', 1.775983426582607)
sofía/NNP_TP vronsky/NNP_TP usted/NNP_FP kitty/NNP_TP nadie/NNP_FP éste/NNP_FP
Cluster: 33 (tamaño = 23) cluster tag: (u'NNP', 2.9599723776376785)
aquí/NNP_FP unos/NNP_FP esas/NNP_FP estar/NNP_FP londres/NNP_TP algún/NNP_FP hilde/NNP_TP
nuestros/NNP_FP dónde/NNP_FP don/NNP_TP pie/NNP_FP algodón/NNP_FP brasil/NNP_TP ahí/NNP_FP
europa/NNP_TP lienzo/NNP_FP visitadoras/NNP_FP hierro/NNP_FP hambre/NNP_FP dólares/NNP_FP
azúcar/NNP_FP vapor/NNP_FP repente/NNP_FP

Cluster: 23 (tamaño = 13) cluster tag: (u'NN2', 5.797656098571021) personas/NN2_TP
ideas/NN2_TP relaciones/NN2_TP formas/NN2_TP tierras/NN2_TP clases/NN2_TP máquinas/NN2_TP
casas/NN2_TP diversas/NN2_FP empresas/NN2_TP mismas/NN2_TP minas/NN2_FP ciudades/NN2_TP
Cluster: 44 (tamaño = 5) cluster tag: (u'NN2', 2.108238581298553)
obreros/NN2_TP hombres/NN2_TP ojos/NN2_FP medios/NN2_TP niños/NN2_TP
Cluster: 68 (tamaño = 22) cluster tag: (u'NN2', 10.541192906492764)
días/NN2_TP libros/NN2_TP productos/NN2_TP valores/NN2_TP tiempos/NN2_TP caballos/NN2_TP
trabajos/NN2_TP brazos/NN2_TP pobres/NN2_FP últimos/NN2_FP hechos/NN2_TP salarios/NN2_TP
pies/NN2_TP seres/NN2_TP soldados/NN2_TP animales/NN2_TP cambios/NN2_TP individuos/NN2_TP
filósofos/NN2_TP derechos/NN2_TP géneros/NN2_TP campesinos/NN2_TP
Cluster: 69 (tamaño = 8) cluster tag: (u'NN2', 4.216477162597106)
cosas/NN2_TP palabras/NN2_TP manos/NN2_TP mujeres/NN2_TP condiciones/NN2_TP fuerzas/NN2_TP
leyes/NN2_TP fábricas/NN2_TP
Cluster: 82 (tamaño = 7) cluster tag: (u'NN2', 3.1623578719478296)
hijos/NN2_TP amigos/NN2_TP padres/NN2_TP propios/NN2_FP pensamientos/NN2_TP
sentimientos/NN2_TP labios/NN2_TP
Cluster: 83 (tamaño = 18) cluster tag: (u'NN2', 7.905894679869573)
horas/NN2_TP mil/NN2_FP partes/NN2_TP nuevas/NN2_FP distintas/NN2_FP páginas/NN2_TP
piezas/NN2_TP noches/NN2_TP obras/NN2_TP necesidades/NN2_TP preguntas/NN2_TP penas/NN2_TP
lágrimas/NN2_TP circunstancias/NN2_TP cartas/NN2_TP razones/NN2_TP escuelas/NN2_TP
materias/NN2_TP

Cluster: 80 (tamaño = 5) cluster tag: (u'PRP', 1.4925071260935914)
contra/PRP_TP hacia/PRP_TP ante/PRP_TP tras/PRP_TP repuso/PRP_FP

Cluster: 18 (tamaño = 5) cluster tag: (u'VMZ', 4.827446923028148)

```
podía/VMZ_TP  sé/VMZ_FP  puedo/VMZ_TP  pudo/VMZ_TP  obstante/VMZ_FP
-----------------------------------
Cluster: 0  ( tamaño = 12 ) cluster tag:  (u'VVI', 4.102769762025776)
fines/VVI_FP  creer/VVI_TP  propósito/VVI_FP  diario/VVI_FP  perder/VVI_TP  buscar/VVI_TP
cargo/VVI_FP  daria/VVI_FP  principios/VVI_FP  entender/VVI_TP  mirar/VVI_TP  gusto/VVI_FP
Cluster: 40  ( tamaño = 29 ) cluster tag:  (u'VVI', 4.923323714430932)
donde/VVI_FP  libras/VVI_FP  media/VVI_FP  igual/VVI_FP  millones/VVI_FP  josé/VVI_FP  c/VVI_FP
cuya/VVI_FP  respecto/VVI_FP  varias/VVI_FP  papá/VVI_FP  encontrar/VVI_TP  tantas/VVI_FP
tantos/VVI_FP  marx/VVI_FP  juan/VVI_FP  diferentes/VVI_FP  comprender/VVI_TP  aquello/VVI_FP
cambiar/VVI_TP  b/VVI_FP  aristóteles/VVI_FP  tanta/VVI_FP  nicolás/VVI_FP  recordar/VVI_TP
parecer/VVI_TP  cincuenta/VVI_FP  llevar/VVI_TP  luis/VVI_FP
Cluster: 56  ( tamaño = 5 ) cluster tag:  (u'VVI', 3.282215809620621) hacer/VVI_TP  ver/VVI_TP
nosotros/VVI_FP  hablar/VVI_TP  pensar/VVI_TP
Cluster: 70  ( tamaño = 18 ) cluster tag:  (u'VVI', 13.94941719088764)
ir/VVI_TP  llegar/VVI_TP  dar/VVI_TP  pasar/VVI_TP  salir/VVI_TP  volver/VVI_TP  dejar/VVI_TP
entrar/VVI_TP  seguir/VVI_TP  tomar/VVI_TP  hacerlo/VVI_TP  poner/VVI_TP  ti/VVI_FP
producir/VVI_TP  dormir/VVI_TP  morir/VVI_TP  conocer/VVI_TP  comprar/VVI_TP
-------------------------------------------
Cluster: 32  ( tamaño = 12 ) cluster tag:  (u'VVN', 7.540952449833871)
dicho/VVN_TP  visto/VVN_TP  pasado/VVN_FP  dado/VVN_TP  llegado/VVN_TP  tenido/VVN_TP
escrito/VVN_TP  ido/VVN_TP  oído/VVN_TP  podido/VVN_FP  perdido/VVN_FP  dejado/VVN_TP
-------------------------------------------
Cluster: 5  ( tamaño = 17 ) cluster tag:  (u'VVZ', 5.563304791242049)
has/VVZ_TP  creo/VVZ_TP  será/VVZ_FP  hubo/VVZ_FP  resulta/VVZ_TP  conocía/VVZ_TP  podían/VVZ_FP
puedes/VVZ_FP  basta/VVZ_TP  pocas/VVZ_FP  sabes/VVZ_TP  tuve/VVZ_TP  habrá/VVZ_FP
sabemos/VVZ_TP  creía/VVZ_TP  quiso/VVZ_TP  digo/VVZ_TP
Cluster: 10  ( tamaño = 17 ) cluster tag:  (u'VVZ', 8.092079696352071) hizo/VVZ_TP  dice/VVZ_TP
debe/VVZ_FP  hacía/VVZ_TP  decía/VVZ_TP  volvió/VVZ_TP  da/VVZ_TP  sabe/VVZ_TP  quedó/VVZ_TP
veía/VVZ_TP  sintió/VVZ_TP  llama/VVZ_TP  trataba/VVZ_TP  encuentra/VVZ_TP  encontraba/VVZ_TP
llamaba/VVZ_TP  pone/VVZ_TP
Cluster: 15  ( tamaño = 15 ) cluster tag:  (u'VVZ', 7.586324715330067)
miró/VVZ_TP  contestó/VVZ_TP  vio/VVZ_TP  dejó/VVZ_TP  daba/VVZ_TP  voy/VVZ_TP  encontró/VVZ_TP
queda/VVZ_TP  habla/VVZ_TP  abrió/VVZ_TP  llevó/VVZ_TP  llamó/VVZ_TP  miraba/VVZ_TP  deja/VVZ_TP
siento/VVZ_TP
Cluster: 47  ( tamaño = 18 ) cluster tag:  (u'VVZ', 5.563304791242049)
he/VVZ_TP  eran/VVZ_FP  habría/VVZ_FP  tienen/VVZ_TP  sabía/VVZ_TP  pueden/VVZ_FP  sería/VVZ_FP
hubiera/VVZ_TP  quería/VVZ_TP  tengo/VVZ_TP  quiere/VVZ_TP  podría/VVZ_FP  tuvo/VVZ_TP
podemos/VVZ_FP  quiero/VVZ_TP  soy/VVZ_FP  tenían/VVZ_TP  existe/VVZ_TP
Cluster: 60  ( tamaño = 4 ) cluster tag:  (u'VVZ', 1.011509962044009)
habían/VVZ_FP  hace/VVZ_TP  han/VVZ_TP  les/VVZ_FP
Cluster: 64  ( tamaño = 6 ) cluster tag:  (u'VVZ', 2.023019924088018) preguntó/VVZ_TP
parece/VVZ_TP  dio/VVZ_TP  parecía/VVZ_FP  pareció/VVZ_FP  dije/VVZ_TP
Cluster: 75  ( tamaño = 41 ) cluster tag:  (u'VVZ', 13.149629506572115)
durante/VVZ_FP  casi/VVZ_FP  además/VVZ_FP  todavía/VVZ_FP  fueron/VVZ_FP  apenas/VVZ_FP
estoy/VVZ_TP  pensaba/VVZ_TP  tampoco/VVZ_FP  jamás/VVZ_FP  tenemos/VVZ_TP  llegó/VVZ_TP
sigue/VVZ_TP  pasó/VVZ_TP  pasa/VVZ_TP  van/VVZ_TP  ocurre/VVZ_TP  viene/VVZ_TP  somos/VVZ_FP
hablaba/VVZ_TP  pueda/VVZ_FP  seguía/VVZ_FP  significa/VVZ_TP  quieres/VVZ_TP  representa/VVZ_TP
pudiera/VVZ_FP  estamos/VVZ_TP  esperaba/VVZ_TP  sean/VVZ_FP  lleva/VVZ_TP  permite/VVZ_TP
iban/VVZ_TP  estuvo/VVZ_TP  vemos/VVZ_TP  produce/VVZ_TP  estás/VVZ_TP  deben/VVZ_FP
tienes/VVZ_TP  llega/VVZ_TP  eres/VVZ_TP  fuese/VVZ_FP
Cluster: 77  ( tamaño = 10 ) cluster tag:  (u'VVZ', 5.057549810220045) sentía/VVZ_TP
trata/VVZ_TP  ve/VVZ_TP  puso/VVZ_TP  levantó/VVZ_TP  convierte/VVZ_TP  refiere/VVZ_TP
acercó/VVZ_TP  dirigió/VVZ_TP  sentó/VVZ_TP
Cluster: 78  ( tamaño = 13 ) cluster tag:  (u'VVZ', 4.046039848176036)
fuera/VVZ_FP  hoy/VVZ_FP  iba/VVZ_TP  va/VVZ_TP  estaban/VVZ_TP  están/VVZ_TP  hemos/VVZ_TP
tú/VVZ_FP  debía/VVZ_FP  haya/VVZ_FP  llevaba/VVZ_TP  hacen/VVZ_TP  trabajan/VVZ_TP
-----------------------------------------------------
Cluster: 2  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
modos/INDECIDIBLE_FP
Cluster: 7  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
cierto/INDECIDIBLE_FP
Cluster: 14  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
duda/INDECIDIBLE_FP
Cluster: 17  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
hecho/INDECIDIBLE_FP
Cluster: 28  ( tamaño = 2 ) cluster tag:  (u'INDECIDIBLE', 0)
otras/INDECIDIBLE_FP  inglaterra/INDECIDIBLE_FP
Cluster: 31  ( tamaño = 3 ) cluster tag:  (u'INDECIDIBLE', 0)
aquel/INDECIDIBLE_FP  cambio/INDECIDIBLE_FP  cuanto/INDECIDIBLE_FP
Cluster: 35  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
embargo/INDECIDIBLE_FP
Cluster: 37  ( tamaño = 2 ) cluster tag:  (u'INDECIDIBLE', 0)
nunca/INDECIDIBLE_FP  sea/INDECIDIBLE_FP
Cluster: 42  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
```

```
ejemplo/INDECIDIBLE_FP
Cluster: 43  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
ello/INDECIDIBLE_FP
Cluster: 46  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
mercancías/INDECIDIBLE_FP
Cluster: 48  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
mucho/INDECIDIBLE_FP
Cluster: 52  ( tamaño = 2 ) cluster tag:  (u'INDECIDIBLE', 0)
mayor/INDECIDIBLE_FP  cual/INDECIDIBLE_FP
Cluster: 63  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
sido/INDECIDIBLE_FP
Cluster: 65  ( tamaño = 3 ) cluster tag:  (u'INDECIDIBLE', 0)
veces/INDECIDIBLE_FP  pesar/INDECIDIBLE_FP  través/INDECIDIBLE_FP
Cluster: 72  ( tamaño = 1 ) cluster tag:  (u'INDECIDIBLE', 0)
tal/INDECIDIBLE_FP
Cluster: 74  ( tamaño = 2 ) cluster tag:  (u'INDECIDIBLE', 0)
términos/INDECIDIBLE_FP  absoluto/INDECIDIBLE_FP
Cluster: 84  ( tamaño = 2 ) cluster tag:  (u'INDECIDIBLE', 0)
gran/INDECIDIBLE_FP  mujer/INDECIDIBLE_FP
Cluster: 85  ( tamaño = 2 ) cluster tag:  (u'INDECIDIBLE', 0)
poco/INDECIDIBLE_FP  solo/INDECIDIBLE_FP
```

7.8 Corpus de referencia para etiquetamiento automático de POS-tag

Al momento de correr el experimento nos encontramos con algunas típicas dificultades para organizar la salida:

1) La falta de un gold standard o benchmark contra el cual comparar la distribución adecuada o no de clusters en cada ciclo (Redington *et al*. 1998). Si bien las categorías morfosintácticas de la gramática del español podrían parecer a simple vista como candidatas idóneas para tal función, los compartimentos estancos tradicionales del tipo de palabra morfosintáctica suelen estar motivados más por necesidades teóricas de la gramática antes que por los usos concretos de las palabras (Nath *et al*. 2008). Considérese, por ejemplo, la evidencia del cluster 53 de la Tabla 25. La gramática tradicional clasificaría a los miembros de ese cluster como sustantivos comunes o, en el mejor de los casos, como sustantivos masculinos singulares. Sin embargo, un refinamiento más granular de las categorías podría demostrar la especificidad del cluster en cuanto a locuciones adverbiales de la forma '*al*'+sustantivos_masculinos_singulares.

2) Otro gran problema era cómo explotar la discriminación de una categoría sintáctica mayor (como verbos o sustantivos) en sus diferentes propiedades morfosintácticas derivadas (¿rasgos morfológicos?, ¿rasgos de subcategorización para los verbos?).

3) Por último, ¿cómo lidiar con las formas léxicas ambiguas en español? Aun cuando por decisiones de diseño nuestro algoritmo asignaba cada vector a una única clase, restaba considerar bajo qué criterio se realizaría tal asignación. En todo caso, ante una forma ambigua en sus etiquetas POS-tags, cabría esperar que la interpretación más frecuente fuese la que prevaleciera ('*reparo*' sustantivo por sobre '*reparo*' verbo, '*como*' adverbio por sobre '*como*' verbo por sobre '*como*' conjunción, etc.).

La solución a todos estos problemas radica en encontrar un corpus anotado que pueda desambiguar formas léxicas según su POS-tag. Con ello podríamos asignarle a cada cluster inducido un POS-tag de clase provisorio que represente la prevalencia de los POS-tag de sus respectivos miembros. Entre los recursos de acceso gratuito sólo con fines académicos

analizamos los dos 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 marcación adicional de funciones sintácticas (Civit 2003)	Automático: *chunking* y *POS-tagging* Validación manual por muestreo aleatorio (Moreno Sandoval *et al.* 1999)

Tabla 28: Comparación entre corpora CAST-3LB y Spanish Treebank

También analizamos CRATER, un corpus masivo multilingüe 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. Todos estos corpora de referencia partían de definiciones de POS-tag demasiado granulares para nuestro escenario (200, 350 y hasta 500 etiquetas POS-tag).

La cantidad de etiquetas POS-tag de referencia no es un dato menor al momento de diseñar el experimento. Clark (2002) trabaja con 77 etiquetas del estándar CLAWS-4 y un corpus de 12 millones de tokens del BNC para su experimento. Si extrapolamos la relación entre etiquetas y tokens de corpus, un escenario de 300 etiquetas resulta inviable. Además, existe también una limitación algebraica en el algoritmo de clustering: la relación entre cues y palabras target. En nuestro experimento, como en el de Redington *et al.* (1998), trabajamos con 1000 vectores a clusterizar. Sería muy improbable que esos 1000 objetos deban repartirse entre 300 categorías a razón de 3 miembros por categoría. Aunque obviamente la cantidad de vectores entre las categorías del benchmark no resulte uniformemente distribuida, la proporción de etiquetas POS-tag respecto de los vectores resultará igualmente elevada.

Reconocemos que el español tiene una complejidad morfológica superior a la del inglés. Sin embargo, tal riqueza morfológica no siempre se ve mapeada en fenómenos sintácticos diferenciados. Por ejemplo, la comparación entre la primera persona y la segunda persona de los paradigmas verbales (por ejemplo , '*digo*' vs. '*dices*') no parece ser conducente para la selección diferenciada de marcos sintácticos del contexto inmediato, ya que el único evento

morfosintácticamente discriminante en tales caso (el pronombre '*yo*' vs. el pronombre '*tú*') está mayormente ausente (sujeto tácito). Análogamente, puede decirse algo similar de la distinción morfológica entre algunos tiempos verbales (pretérito perfecto simple vs. pretérito imperfecto, etc.).

Por todo lo expuesto arriba, nos vimos obligados a encarar la esforzada tarea de generar un *corpus de referencia* propio en español, etiquetado según nuestros lineamientos morfosintácticos adaptados del C4 (Leech *et al.* 1994), los cuales están expresados en la Tabla 26. Entre los lineamientos principales a la hora de adpatar los criterios de anotación del estándar C4 priorizamos aquellos rasgos morfosintácticos que muy problamente determinen contextos de ocurrencia inmediatos diferenciados (sustantivos singulares NN1 vs. sustantivos plurales NN2 vs. nombres propios NNP, verbos léxicos finitos VVZ vs. verbos léxicos en infinitivo VVI, etc.).

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. El corpus de referencia resultante y el instructivo describiendo la metodología utilizada se encuentran disponibles para la comunidad científica bajo los alcances de una licencia *Creative Commons*. Nuestro corpus de referencia alcanzó una dimensión comparable a la de los corpora de referencia estándares en español disponibles con anotación morfosintáctica (compárense *Tabla 28* y *Tabla 29*).

	Nuestro corpus de referencia
Tamaño en palabras	49.925
Tamaño en oraciones	2.108
Extensión promedio de oraciones	≈24 palabras
Anotación morfosintáctica	Manual: 48 etiquetas adaptadas del estándar C4 (Tabla 26 + 1 signo de fin de oración)
Criterio de anotación	Propio, adaptado del C4

Tabla 29: Corpus de referencia para etiquetamiento automático de POS-tag

A continuación presentamos un ejemplo de la anotación manual de texto en el corpus de referencia. Recuérdese consultar la Tabla 26 para la nomenclatura de las etiquetas asignadas a cada palabra:

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/NNP ./$$$

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/DAT , 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 ./$$$

Tabla 30: Ejemplo de texto etiquetado morfosintácticamente en el corpus de referencia

Este corpus de referencia nos permite desambiguar una forma léxica como '*ama*' y saber que su uso más frecuente es como verbo conjugado (VVZ) y no como sustantivo singular (NN1). Además, un corpus de referencia morfosintácticamente anotado cumple otra función importante: ofrece un perfil de frecuencias de ocurrencias de las etiquetas POS-tag.

POS-tag	n	%	*POS-tag*	n	%	*POS-tag*	n	%
PRP	9613	19,25	PPE	621	1,24	AT0	64	0,13
NN1	8280	16,58	VBZ	571	1,14	EX0	58	0,12
AT1	6484	12,99	DPS	424	0,85	NN0	54	0,11
VVZ	3857	7,73	SEP	347	0,7	VBI	47	0,09
NN2	3409	6,83	CJS	250	0,5	AJC	46	0,09
NNP	2405	4,82	DT1	250	0,5	PNQ	11	0,02
AJ1	1897	3,8	XX0	226	0,45	AJS	8	0,02
AV0	1610	3,22	ORD	153	0,31	VBG	7	0,01
CJC	1574	3,15	PNP	135	0,27	AVQ	5	0,01
AT2	1501	3,01	AJ0	127	0,25	VMN	1	0
CRD	1056	2,12	VMZ	118	0,24	ITJ	0	0
VVI	902	1,81	PNI	118	0,24	POS	0	0
AJ2	851	1,7	DAT	113	0,23	VBN	0	0
VVN	847	1,7	VVG	112	0,22	VM0	0	0
REL	835	1,67	DT2	104	0,21	VMG	0	0
CJT	730	1,46	PND	103	0,21	TOTAL	49929	100%

Tabla 31: Distribución de POS-tag en el corpus de referencia

Tomando como criterio la distribución de frecuencias de los POS-tag en el corpus de referencia, estamos en condiciones de ponderar la incidencia de los POS-tag de cada miembro de un cluster al etiquetamiento del cluster (*cluster_tag*), de modo de disponer de un criterio más adecuado que el mero conteo de la mayoría de los POS-tag presentes en la clase. En efecto, a lo largo de todas las distribuciones de los ciclos iterativos, debemos asignar al cluster una etiqueta en función de qué tipos de miembros contiene a los fines de la evaluación. Pero en algunos ciclos -sobre todo, en los ciclos iniciales- las purezas de los clusters no están bien consolidadas y es probable que coexistan en los clusters diversos miembros pertenecientes a diversas POS-tags, con poco

margen de diferencia entre ellos. En tales casos, sería inadecuado guiarse por un criterio de simple mayoría, ya que como observamos a partir de la Tabla 31, la ocurrencia de NN1 (sustantivos singulares) sería mucho más problable que, por ejemplo, la de CRD (adjetivos numerales cardinales). En el siguiente cluster del ciclo 46 podemos observar la problemática:

Cluster: 6 (tamaño = 33)
aquí/AV0 unos/AT2 muchos/AJ2 ello/PNP nuestro/DPS esos/DT2 nuestra/DPS alguna/DT1 aquellos/DT2 cualquier/AJ1 esas/DT2 estar/VVI londres/NNP algún/AJ1 ellas/PNP hilde/NNP nuestros/DPS san/NNP dónde/AVQ buenos/AJ2 ambos/AJ2 pie/NN1 algodón/NN1 brasil/NNP ahí/AV0 europa/NNP visitadoras/NN2 hierro/NN1 hambre/NN1 dólares/NN2 moscú/NNP vapor/NN1 repente/AV0

Tabla 32: Ejemplo de cluster de baja pureza

Entre los 33 miembros del cluster observamos: 6 miembros NNP , 5 miembros NN1 , 3 miembros AV0 , 3 miembros DT2 , 3 miembros AJ2 , 3 miembros DPS , 2 miembros NN2 , 2 miembros PNP , 2 miembros AJ1 , 1 miembro VVI , 1 miembro AT2 , 1 miembro AVQ y 1 miembro DT1.

La primera impresión es que el cluster está dominado por los nombres propios (NNP). Sin embargo, según el corpus de referencia, la frecuencia relativa de los NNP (4,82%) es casi 25 veces la de, por ejemplo, los determinantes plurales DT2 (0,21%). Estas distribuciones no uniformes nos obligaban a cuantificar la incidencia del POS-tag de cada miembro en el POS-tag del cluster (*cluster_tag*) en función de una métrica conocida en recuperación de la información (*information retrieval*): frecuencia de término–frecuencia inversa de documento (tf-idf) (Manning y Schütze 1999). Esta métrica es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. El valor tf-idf aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia inversa de la palabra en la colección de documentos, lo que permite manejar el hecho de que algunas palabras son mucho más comunes que otras. Si consideramos el POS-tag de cada miembro de un cluster como la "palabra" del documento que sería cada cluster en una colección que sería el corpus de referencia, entonces el POS-tag del cluster puede calcularse como:

tf-idf (de cada POS-tag en cluster)= frec.absol POS-tag en cluster*log(1/frec.relativa de POS-tag en corpus referencia)

$$tf - idf = f_{abs\_en\_cluster} * \log \frac{1}{f_{relativa\_en\_corpus\_referencia}}$$

Ecuación 19: tf-idf adaptado al cálculo ponderado del tag de un cluster en función de los POS-tags de sus miembros

De este modo, para el ejemplo de la Tabla 32 los posibles POS-tag del cluster (*cluster_tag*) se presentan en la Tabla 33. Notemos que la importancia ponderada del tag DT2 prevalece como el tag elegido por el criterio para todo el cluster (*cluster_tag*). Para completar el criterio de la

asignación automática del *cluster_tag* con vistas a la evaluación total del ciclo, debemos asegurarnos de que este criterio se aplique sólo para los POS-tag presentes en al menos dos miembros de un cluster. De este modo, algunos clusters quedarán INDECIDIBLES en cuanto a su *cluster_tag*. Afortunadamente, debemos mencionar que en la mayoría de los casos la elevada pureza de los clusters hace que un simple conteo por POS-tag mayoritario en cada clase sea suficiente para asignarle dicho POS-tag a la clase. No obstante, la métrica tf-idf resulta imprescindible como criterio auxiliar en el caso de clusters de baja pureza. Con la métrica tf-idf para la asignación automática del tag del cluster podemos entender más en detalle la salida presentada en la Tabla 27.

POS-tag	frec absoluta	%	frec relativa	1/frec relativa	log (1/frec rel)	**tf-idf**
NNP	6	4,82	0,0482	20,746888	1,316952962	**7,90171777**
NN1	5	16,58	0,1658	6,03136309	0,780415474	**3,90207737**
AV0	3	3,22	0,0322	31,0559006	1,492144128	**4,47643238**
DT2	3	0,21	0,0021	476,190476	2,677780705	**8,03334212**
AJ2	3	1,7	0,017	58,8235294	1,769551079	**5,30865324**
DPS	3	0,85	0,0085	117,647059	2,070581074	**6,21174322**
NN2	2	6,83	0,0683	14,6412884	1,165579296	**2,33115859**
PNP	2	0,27	0,0027	370,37037	2,568636236	**5,13727247**
AJ1	2	3,8	0,038	26,3157895	1,420216403	**2,84043281**
VVI	1	1,81	0,0181	55,2486188	1,742321425	**1,74232143**
AT2	1	3,01	0,0301	33,2225914	1,521433504	**1,5214335**
AVQ	1	0,01	0,0001	10000	4	**4**
DT1	1	0,5	0,005	200	2,301029996	**2,30103**

Tabla 33: Cálculo del *cluster_tag* del cluster ejemplo de la Tabla 32 en función de tf-idf

Una vez asignado el cluster_tag, fácilmente logramos evaluar automáticamente una salida como la presentada en la Tabla 27, en donde el cluster de nuestro ejemplo ahora se presentará evaluado en forma interna en cuanto a la pertenencia (*True Positives* TP) o no (*False Positives* FP) de sus miembros con respecto al POS-tag de la clase (DT2):

Cluster: 6 (tamaño = 33) **cluster tag: (u'DT2', 8.03334212)**
aquí/DT2_FP unos/DT2_FP muchos/DT2_FP ello/DT2_FP nuestro/DT2_FP **esos/DT2_TP** nuestra/DT2_FP alguna/DT2_FP **aquellos/DT2_TP** cualquier/DT2_FP **esas/DT2_TP** estar/DT2_FP londres/DT2_FP algún/DT2_FP ellas/DT2_FP hilde/DT2_FP nuestros/DT2_FP san/DT2_FP dónde/DT2_FP buenos/DT2_FP ambos/DT2_FP pie/DT2_FP algodón/DT2_FP brasil/DT2_FP ahí/DT2_FP europa/DT2_FP visitadoras/DT2_FP hierro/DT2_FP hambre/DT2_FP dólares/DT2_FP moscú/DT2_FP vapor/DT2_FP repente/DT2_FP

Tabla 34: Evaluación automática de la pertenencia de los miembros de un cluster a la clase

Pese al enorme esfuerzo de supervisión de la anotación manual, el corpus de referencia no es infalible, tal como se aprecia en los casos resaltados en amarillo de la salida de la Tabla 27. No obstante, consideramos que haber implementado está metodología de evaluación interna del cluster es sumamente útil no sólo a los fines de una evaluación general del ciclo de clustering -

como explicaremos en las secciones siguientes-, sino principalmente como una muestra de la elevada pureza de los clusters obtenidos en cada uno de esos ciclos.

7.9 Métricas de evaluación de un ciclo de clustering

7.9.1 ¿Métricas propias de la distribución o propias de un modelo HMM a partir de la distribución?

Es sabido que el problema de hallar el número de clusters que optmiza una distribución de objetos en el espacio vectorial mediante técnicas de clustering no es sencillo. Hasta ahora nos centramos en describir un criterio de asignación de una etiqueta morfosintátcia (*Part-Of-Speech tag* POS-tag) a cada cluster de un ciclo (*cluster_tag*), en función de la pertenencia de cada uno de sus miembros a un cierto POS-tag y considerando las distribuciones probabilísticas no uniformes de POS-tag en corpora masivos (corpus de referencia). No obstante, esta evaluación inicial de la salida de un ciclo de clustering no es una métrica representativa de la efectividad del clustering, sino más bien, de la composición interna de cada cluster.

Para las métricas generales de evaluación de la salida de un experimento de clustering, tal como observamos en el capítulo 5 de esta tesis, algunos trabajos se enfocaron en la evaluación del mapeo de dicha distribución de clusters respecto de un set inicial de POS-tags (gold standard) (Redington *et al*. 1998; Wang 2012) y algunos otros evaluaron las clases inducidas a partir de la menor perplejidad posible de un modelo HMM entrenado en los bigramas o trigramas resultantes de tales POS-tags sobre las palabras del corpus (Brown *et al*. 1992; Martin *et al*. 1998; Clark 2000, 2002, 2003). El primer caso presenta el problema de la disponibilidad de un gold standard suficientemente adecuado y granular sobre el cual mapear la salida (véanse subsecciones 7.9.2 y 7.9.3). El segundo caso no necesita de un gold standard, pero adolece de otro problema: la transformación de un criterio de adecuación de la distribución de clusters en un criterio de perplejidad markoviana asociada. Consideramos que este segundo caso de evaluación general, en última instancia, no está midiendo la adecuación de la inducción de las categorías de POS-tags sino en forma vicaria, tal como el mismo Clark (2002) lo reconoce:

> "Evaluation is in general difficult with unsupervised learning algorithms. Previous authors have relied on both informal evaluations of the plausibility of the classes produced, and more formal statistical methods. Comparison against existing tag-sets is not meaningful –one set of tags chosen by linguists would score very badly against another without this implying any fault as there is no 'gold standard'. I therefore chose to use an objective statistical measure, the perplexity of a very simple finite state model, to compare the tags generated with this clustering technique against the BNC tags, which uses the CLAWS-4 tag set […] which had 76 tags. **This is by no means an ideal measure, since the perplexity does not directly relate to what I am trying to achieve here**." [Clark 2002:66-69] (*las negritas y el subrayado son nuestros*)

Es decir, las métricas de evaluación de clusters de categorías morfosintácticas basadas en la menor perplejidad de modelos HMM resultarían validadas únicamente bajo la premisa de plausibilidad psicolingüística de que en todo lenguaje natural la aparición de categorías gramaticales morfosintácticas responde a un criterio óptimo de cobertura en bigramas o

trigramas de las combinaciones posibles de las palabras en oraciones, lo cual puede no ser necesariamente el caso. Esto sería equivalente a sostener que todo adquiriente de un lenguaje puede identificar en un corpus de PLD un número determinado de POS-tags, tal que el reemplazo de las palabras por dichos POS-tags redunde en un modelo óptimo de determinación bigrámica o trigrámica de las cadenas. Esto implicaría una premisa general adicional en nuestro enfoque tendiente a demostrar que las categorías morfosintácticas ontogenéticamente tempranas de un lenguaje natural pueden ser inducidas únicamente a partir de un corpus PLD, aun cuando estas distribuciones de categoría no sean necesariamente las óptimas en términos de modelos markovianos de dicho lenguaje natural. Así pues, en el diseño de nuestro experimento decidimos seguir un criterio de evaluación general de las categorías inducidas a partir de un criterio más apegado a la disponibilidad ontogenética de un criterio de adecuación (un enfoque comparable al de Redington *et al*. 1998), a diferencia de los criterios basados en el cómputo de la perplejidad de modelos markovianos entrenados en las categorías inducidas (Brown *et al*. 1992; Martin *et al*. 1998; Clark 2002). Tales criterios serán explicados en detalle en las subsecciones siguientes.

7.9.2 Mapeo 1-to-1: El problema del gold standard

Idealmente, podríamos esperar que todos los miembros de las categorías morfosintácticas mayores se agrupen en un solo cluster (por ejemplo, todos los sustantivos, todos los verbos conjugados, etc.). Si éste fuera el caso, entonces bastaría el procedimiento descripto en la sección anterior para dar con una métrica que evalúe la distribución general de dichas categorías en la inducción de clusters, mapeando una categoría esperada al cluster más representativo de dicha categoría (considerando la composición de sus miembros). Una vez mapeadas todas las categorías del *tag set esperado (gold standard)*, el resto de los clusters recibiría cero crédito en la distribución. Este criterio de evaluación es conocido como *1-to-1 mapping* (Haghighi y Klein 2006; Christodoulopoulos *et al*. 2010), aunque no es muy efectivo.

> "One difficulty in evaluating POS induction systems is that there is no straightforward way to map the clusters found by the algorithm onto the gold standard tags; moreover, some systems are designed to induce the number of clusters as well as their contents, so the number of found clusters may not match either the gold standard or that of another system." [Christodoulopoulos *et al.* 2010:575]

El gran problema a la hora de la evaluación de clusters es que las categorías morfosintácticas mayores no son tán monolíticas como la gramática las suele postular. Las técnicas de clustering son tan exhaustivas en el rastreo de particularidades del contexto inmediato de los usos generalizados de las palabras que terminan "distribuyendo" cada categoría morfosintáctica mayor en numerosos clusters (Redington *et al*. 1998). Uno podría pensar que esta proliferación de clusters de una misma categoría morfosintáctica puede deberse, en última instancia, a la discriminación de las categorías en subcategorías, pero no siempre es el caso. Por ejemplo, a primera vista parecería que toda la categoría sustantivos bien podría ser distribuida entre varios clusters: masculinos vs. femeninos, concretos vs. abstractos, singulares vs. plurales, etc. El gran problema es que estas distinciones gramaticales no son necesariamente ortogonales a los

contextos de uso. Consideremos nuevamente la evidencia del cluster 53 de la Tabla 25. Como ya explicamos en la sección 7.6, la información distribucional de estos cuatro sustantivos singulares masculinos privilegió su apartamiento de los clusters con el grueso de los sustantivos singulares masculinos reflejando la prevalencia de sus usos como giros lingüísticos adverbiales, algo que no sucedió, por ejemplo, con el sustantivo masculino singular '*lado*', pese a que también está muy marcado por la contracción '*al*' (aunque ciertamente no al punto de una lexicalización de giro lingüístico). Este sorprendente grado de refinamiento en la salida del clustering no se vería apropiadamente reflejado en un mapeo *1-to-1* sin más, según el cual se aplicaría la etiqueta s*ustantivos singulares* NN1 (eventualmente, Masculino) al cluster más homogéneamente numeroso para tal categoría, dejando sin categorizar otros clusters muy representativos de dicha categoría (pero menos numerosos) u otros clusters representativos de un refinamiento más granular de dicha categoría (como el caso del cluster 53 de la Tabla 25).

Entonces, el criterio de mapeo *1-to-1* no es muy viable porque tiende a disminuir mucho con el sucesivo aumento del número de clusters por sobre el de categorías del *tag set* y porque tiende a desechar refinamientos granulares importantes, ante la ausencia de ciertos fenómenos en el tag set del gold standard, el cual, en nuestro caso, estaría representado por los criterios de anotación morfosintáctica del corpus de referencia (véase *Tabla 26*).

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

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

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

En nuestro experimento, adoptamos esta decisión de diseño, tal como se había anticipado en el lineamiento 9 de la sección *7.1 Motivación de las decisiones de diseño*. Más allá de la justificación metodológica, existe una inituición gramatical en adoptar este criterio de evaluación general de la distribución de un ciclo de clustering. Es de esperar que la ubicación de los clusters en el espacio vectorial refleje en alguna medida el criterio de agrupamiento de clusters en función de la similitud de los miembros preeminentes en cada uno de ellos. Así, pues a dos o más clusters del mismo tipo (indicado por el valor del *cluster_tag* en nuestra salida de la Tabla 27) corresponde un mismo *hipercluster*. Si regresamos por un momento a la salida completa del ciclo 87 de nuestro experimento (véase *Tabla 27*) notaremos que los clusters se presentan agrupados con este criterio de hiperclusters. La siguiente Tabla 35 muestra la ubicación de los

respectivos centroides en las 106 dimensiones del espacio vectorial. Para la nomenclatura de POS-tags véase *Tabla 26*. Para las referencias de las 106 dimensiones véase *Tabla 22*.

Si un centroide representa prototípicamente la ubicación espacial de un cluster, al menos en cuanto a la concentración mayoritaria de sus miembros, entonces al computar la distancia euclideana de los centroides entre sí podemos darnos una idea de qué clusters están más cercanos o más alejados entres sí. Nuestra intución metodológica de los hiperclusters podría verse justificada empíricamente si, por ejemplo, los clusters *sustantivos singulares* NN1 que conforman el hipercluster NN1 aparecen de algún modo más cercanos entre sí, en comparación con, por ejemplo los clusters que conforman el hipercluster *verbos en infinitivo* VVI. La siguiente Figura 27 ilustra este punto.

El concepto de hipercluster, tal como denominamos en este trabajo al agrupamiento de clusters, resulta muy significativo. Desde un punto de vista metodológico permite una evaluación que resuelve el problema del mapeo de un número creciente de clusters inducidos en las categorías del gold standard. Desde un punto de vista algebraico el hipercluster se ve justificado en gran medida por la ubicación en el espacio vectorial de los centroides de los clusters que lo conforman, lo cual, a su vez, refleja particularidades morfosintácticas propias del dominio lingüístico al que pertenecen los datos.

Por ejemplo, si comparamos la información de la Figura 27 con la salida completa de la Tabla 27, notaremos que los clusters INDECIDIBLES (especialmente aquellos de un único miembro) son los objetos más apartados (*outliers*) del espacio vectorial. En algunos trabajos de clustering, se observó un especial esfuerzo en evitar los clusters de miembros únicos, penalizando incluso su inducción para la evaluación general de los resultados (Böhm *et al*. 2006). Sin embargo, a la luz de la naturaleza lingüística de los objetos a clusterizar, la aparición de *outliers* conlleva otro significado. De hecho, en la Figura 27 el objeto más apartado de todos los otros centroides es el cluster 35 de un único miembro '*embargo*' (véase *Tabla 27*), y la explicación de por qué este hallazgo de este cluster de un único miembro debería ser premiado antes que castigado fue esbozada en la sección 7.5 (véase *Tabla 24*):

> "Using classification data for the purpose of evaluating clustering results, however, encounters several problems since the class labels do not necessarily correspond to natural clusters. A typical example includes the clustering specific identification of outliers, i.e., of objects that do not belong to any cluster. In classification data, however, usually each object has assigned a certain class label. **Thus, a clustering algorithm that detects outliers is actually punished in an evaluation based on these class labels even though it should be rewarded for identifying outliers as not belonging to a common cluster albeit outliers represent a genuine class of objects**. Similar dificulties occur if the labeled classes split up in difierent sub-clusters or if several classes cannot be distinguished leading to one larger cluster. Consequently, in general, classes cannot be expected to exactly correspond to clusters." [Färber et al. 2010:1-2] (*las negritas y el subrayado son nuestros*)

Para nuestra métrica de evaluación general iterativa de los ciclos de clustering, optaremos por una posición salomónica en cuanto a que los clusters INDECIDIBLES no serán considerados como motivo de premios ni de castigos (véase sección 7.10).

Tabla 35: Ubicación en 106 dimensiones de los centroides de los clusters del ciclo 87. Nomenclatura de POS-tags en Tabla 26. Referencias de las 106 dimensiones en Tabla 22.

Tabla 35: Ubicación en 106 dimensiones de los centroides de los clusters del ciclo 87. Nomenclatura de POS-tags en Tabla 26. Referencias de las 106 dimensiones en Tabla 22.

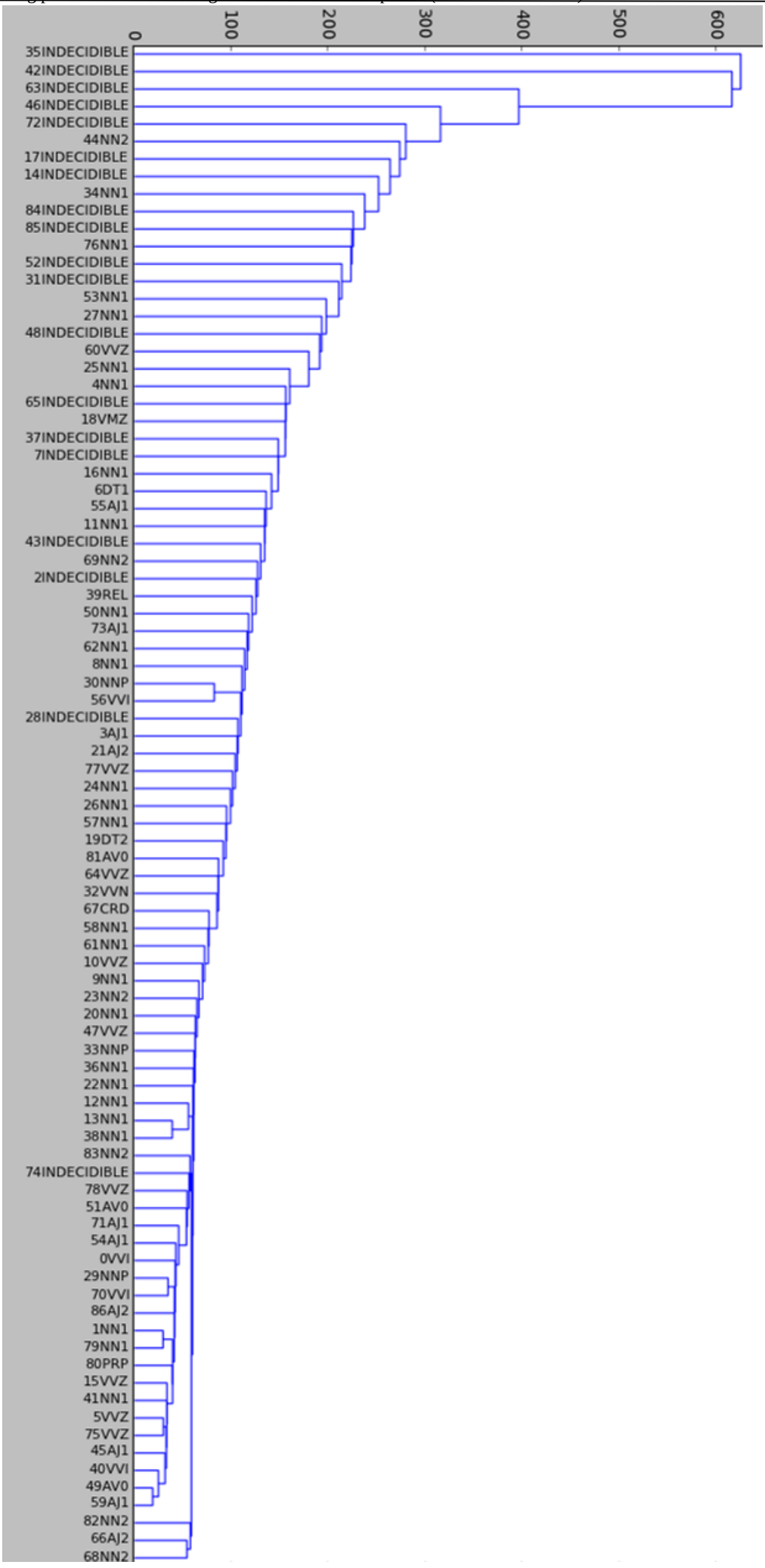

Figura 27: Distancias euclideanas entre los centroides de de los clusters del ciclo 87, según Tabla 35

En el eje x se incluye el número de cluster según Tabla 27 y el POS-tag según Tabla 26.

Por otro lado, es notable cómo la ubicación de los centroides en el espacio vectorial revela interesantes relaciones entre los clusters que conforman un hipercluster. Por ejemplo, los centroides de los cluster 1 y 79 están muy cercanos entre sí y ambos, a su vez, lo están respecto del centroide del cluster 41. Notablemente todos estos clusters de alta pureza agrupan sustantivos singulares (NN1) masculinos. En contraste, los clusters 38 y 13 también están casi pegados entre sí, agrupando sustantivos singulares (NN1) pero en este caso, femeninos. Si continuamos la exploración de las distancias euclideanas en la Figura 27, observamos que el par de clusters 38 y 13 se unen luego al cluster 12. Aunque el cluster 12 está un poco más lejano en el espacio, esta unión no es llamativa ya que el cluster también agrupa NN1 femeninos. En este sentido, la distribución de clusters nos está indicando que en nuestro *gold standard* habría sido necesaria una separación de los NN1 en función del género del sustantivo.

Habiendo justificado el concepto de hipercluster, resta entonces definir las métricas de evaluación general que utilizaremos bajo este criterio de agrupamiento de clusters. Siguiendo a Redington *et al*. (1998), apelaremos a las dos métricas clásicas para medir la efectividad de un sistema: *Precisión* ***P***, que toma en cuenta falsos positivos FP (*Precision*, véase *Ecuación 7*) y *Cobertura* ***C*** o *Exhaustividad*, que toma en cuenta falsos negativos FN (*Recall o Completeness*, véase *Ecuación 8*). Como estas dos métricas deben actuar armónicamente para que la efectividad del sistema sea alta -de poco sirve un sistema que sea extremadamente preciso en sus juicios (alta ***P***), pero que actúe muy raramente (baja ***C***), y, a la inversa, lo mismo sucedería con un sistema que emite juicios de pertenencia siempre (alta ***C***) pero se erquivoca mucho (baja ***P***)-, adoptaremos el promedio armónico entre ambas métricas, conocido como *medida F* (*F-score*) (Manning y Schütze 1999) para la representación final de la efectividad del agrupamiento en hiperclusters.

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

Ecuación 20: Medida F (Con β = 1 para asignar igual peso a Precisión o *Precision P* y a Cobertura o *Recall C*)

De este modo, en cada ciclo de clustering evaluaremos la medida F para cada uno de los 16 POS-tags que se inducen en el experimento –únicamente dejamos de lado la evaluación de algunos POS-tag marginales como REL o AJC (véase *Tabla 26*), que sólo fueron inducidos muy intermitentemente en uno o dos ciclos y en algún cluster muy poco denso (de pocos miembros). A diferencia de otros experimentos que se centraron exclusivamente en las palabras de contenido (Vlachos *et al*. 2009) o en las palabras funcionales (Wang 2012), en nuestro experimento veremos cómo son inducidos exitosamente tanto POS-tags típicamente de contenido como POS-tags típicamente funcionales en estas 16 categorías de hiperclusters.

7.9.4 Otras métricas: Variación de la información

Durante la última década surgieron algunas propuestas alternativas para las métricas clásicas basadas en el mapeo de categorías entre las inducidas en los clusters y aquellas del gold standard (véanse subsecciones 7.9.2 y 7.9.3). Meilă (2003) propuso la *Variación de Información* (*Variation of Information* VI) como una forma de considerar no sólo la prevalencia mayoritaria de los miembros de un cluster, sino también la homogeneidad de los miembros remanentes. La métrica está basada en los conceptos fundamentales de la teoría de la información (Shannon 1948). Dadas las distribuciones de clusters en la salida *C* y la distribución de cluster del gold standard *T*, VI es la cantidad de información perdida (*Mutual Information-loss* MI-loss, véase *Ecuación 5*) más la cantidad de información ganada en ir de *C* a *T*; es decir, la suma de la entropía condicional de cada cluster de *C* en función de cada cluster de *T*, $H(C|T)$ y la de cada cluster *T* en función de *C*, $H(T|C)$:

$$VI(C, T) = H(T|C) + H(C|T) = H(C) + H(T) - 2\,MI(C,T)$$

Ecuación 21: Variación de la Información como métrica de evaluación general de distribuciones de clusters en Meilă (2003)

Por tratarse de una métrica basada en la entropía cuanto menor sea el número final de VI más validada resultará la distribución *C* en función del gold standard *T*. Si se calcula la VI para distintas distribuciones de *C* (agrupando clusters o dividiéndolos), podremos hallar aquella distribución que presente la VI más baja. Esta métrica evolucionó luego como una VI orientada a cubrir falsos positivos y falsos negativos -véase concepto de *Validity-measure* o *V-measure* en Rosenberg y Hirschberg (2007) y de *V-beta* en Vlachos *et al*. (2009). No obstante, estas métricas basadas en la teoría de la información presentan dos falencias graves:

1) Christodoulopoulos *et al*. (2010) señalan el principal problema: el resultado final es una escala de bits que no tiene un correlato conceptual claro para ser interpretado en los escenarios de clustering, una ventaja competitiva que sí tienen las métricas basadas en el mapeo de categorías, donde el resultado final es un porcentaje de efectividad.

2) Graça *et al*. (2011) critican esta métrica por su rango de valores dependientes de la cantidad de POS-tags y, consecuentemente, por la dificultad que entraña el intento de comparación de diferentes escenarios de clustering basándose en esta métrica:

> "VI has desirable geometric properties -it is a metric and is convexly […]. However, the range of VI values is dataset-dependent (VI lies in [0 ; 2log*N*] where *N* is the number of POS tags) which does not allow a comparison across datasets with different *N*. The validity-measure (V) is also an entropy-based measure and always lies in the range [0 ; 1], but does not satisfy the same geometric properties as VI." [Graça *et al*. 2011:548]

7.9.5 Otras métricas: Medida F de sustitución

Todas las métricas anteriores parten de la premisa de la disponibilidad de un gold standard. Recientemente Frank *et al*. (2009) propusieron la *Medida F de sustitución* (*sustitutable F-score* SF) como una forma de evaluar escenarios de clustering sin una distribución inicial de categorías de referencia (gold standard). Para ello, los autores comparan los clusters inducidos con un set *S*

de clusters de referencia que surgen de marcos de ocurrencia *sustituibles* (Harris 1954) en el corpus de datos:

> "Ideally a substitutable frame would be created by sentences differing in only one word (*e.g.* "I want the blue ball." and "I want the red ball.") and the resulting cluster would contain the words that change (*e.g.* [blue, red]). However since it is almost impossible to find these types of sentences in real-world corpora, the authors use frames created by two words appearing in the corpus with exactly one word between (e.g. the —- ball). Once the substitutable clusters have been created, they can be used to calculate the [Sustitutable] Precision (*SP*), [Sustitutable] Recall (*SR*) and [Sustitutable] F-score (*SF*) of the system's clustering." [Christodoulopoulos *et al*. 2010:641]

$$SP = \frac{\sum_{s \in S} \sum_{c \in C} |s \cap c|(|s \cap c| - 1)}{\sum_{c \in C} |c|(|c| - 1)}$$

$$SR = \frac{\sum_{s \in S} \sum_{c \in C} |s \cap c|(|s \cap c| - 1)}{\sum_{s \in S} |s|(|s| - 1)}$$

$$SF = \frac{2 \cdot SP \cdot SR}{SP + SR}$$

Ecuación 22: Medida F de sustitución en Frank *et al*. (2009)

Sin embargo, el problema insalvable de esta métrica es que la sustituibilidad de los marcos de ocurrencia no depende únicamente de las propiedades sintácticas puras que se desean elucidar con estos experimentos, sino que también están involucradas propiedades semánticas y hasta pragmáticas de conocimiento de mundo. Por ejemplo, a pesar de formar una clase muy diferenciada en varios experimentos (incluso en el nuestro), las distintas instancias de adjetivos numerales cardinales (CRD) podrían no presentar los mismos marcos de ocurrencias sustituibles en toda la extensión del corpus: (?) "*lo vi con mis tres ojos*. Consecuentemente, esta métrica no estaría capturando adecuadamente el concepto de clase morfosintáctica que justamente se pretende evaluar.

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

Ahora que explicamos en detalle en qué consistió nuestro experimento de clustering para inducción de categorías sintácticas en español, su plausibilidad de modelización, sus lineamientos de diseño y sus métricas de evaluación, llegó el momento de analizar la salida completa de los 106 ciclos. Recordemos que el experimento corre iterativamente en ciclos que van desde K=2 clusters hasta K = 106 clusters. La siguiente tabla muestra la composición de los objetos (palabras target) a ser clusterizados. Si bien el corte inicial era de 1000 palabras target, 89 de esas palabras correspondía a categorías morfosintácticas marginales: categorías funcionales de poquísimos miembros y de prevalencia intermitente (en muy asialdas ocasiones) en los clusters (REL, AJC, CJC, CJS, etc.). Las restantes 911 palabras target, entonces, se distribuyeron entre 16 categorías de inducción casi permanente a lo largo de todo el experimento, con elevados valores de pureza consolidados a partir de los ciclos medios (véase *Figura 28*).

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

Tabla 36: Palabras target a ser clusterizadas según POS-tag de corpus de referencia y baseline de cada POS-tag

En cada ciclo calculamos Precisión (véase *Ecuación 7*), Cobertura (véase *Ecuación 8*) y medida F (véase *Ecuación 20*) para cada uno de los 16 POS-tags, prevalezcan o no como el *cluster_tag*, en cada uno de los hiperclusters inducidos. Sobre estas 16 medidas F calculamos el promedio común y el promedio ponderado (según el peso de cada POS-tag en la distribución de 911 palabras target). Para un ejemplo de estas métricas calculadas para un ciclo, véase la Tabla 38.

A continuación, la Tabla 37 y la Figura 28 muestran las métricas completas para todos los ciclos, destacándose en amarillo el ciclo 87 como el ciclo con promedios de medida F máximos. Los valores resaltados en celeste representan los valores máximos aislados para cada hipercluster.

Tabla 37: Evaluación general iterativa de todos los ciclos de clustering según medida F bajo criterio de mapeo *many-to-1*

Figura 28: Evaluación general iterativa de todos los ciclos de clustering según medida F bajo criterio de mapeo *many-to-1*

Tabla 37: Evaluación general iterativa de todos los ciclos de clustering según medida F bajo criterio de mapeo many-to-1

Figura 28: Evaluación general iterativa de todos los ciclos de clustering según medida F bajo criterio de mapeo many-to-1

En la Tabla 38 presentamos el detalle de la evaluación para el ciclo 87, cuya distribución en clusters puede consultarse en la Tabla 27. De todos modos, es de destacar que a partir de los ciclos medios (ciclo 52 en adelante), las medidas F de la mitad de los POS-tag se presentan consolidadas en valores relativamente estables, especialmente para las categorías mayores de sustantivos y verbos (NN1, NN2, VVZ, VMZ, VVI, VVN). Esto se aprecia en los gráficos de la Figura 28 como líneas "planchadas" que no fluctúan demasiado, lo cual significa que a partir de cierto momento de la "historización" de la inducción, las clases están mayormente consolidadas en cuanto a la pertenencia de sus miembros (con mínimas fluctuaciones).

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

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

Tabla 38: Detalle de evaluación de ciclo 87 (véase *Tabla 27* para la distribución de miembros en cluster para dicho ciclo)

7.11 Discusión de los resultados y conclusiones

7.11.1 Consideraciones cuantitativas y cualitativas

A partir del análisis detallado de las Tablas 27, 28 y 38, podemos elaborar las siguientes observaciones:

1) Todas las categorías sintácticas mayores fueron inducidas con un alto grado de pureza. Se observan refinamientos granulares en rasgos de género y número (para sustantivos) y

en otras caracterizaciones morfosintácticas (verbos modales VMZ vs. verbos léxicos VVZ).

2) Al igual que en Redington *et al*. (1998), las categorías sintácticas mayores, coincidentes con palabras de contenido (verbos y sustantivos), reportan medidas F altísimas, del orden del 80% y hasta 90%.

3) En el otro extremo, uno de los hiperclusters con menor medida F (40,7%) son los adverbios (AV0). Este grupo quedó confinado a un cluster único y masivo de 95 miembros muy heterogéneos, con objetos claramente marginales (caracteres únicos como '*d*', '*p*', '*v*', etc.). Como reporta Nath *et al*. (2008), es normal que en el clustering partitivo quede en cada ciclo uno o dos clusters masivos que actúan como receptáculo indiferenciado de objetos del espacio vectorial. Posiblemente éste sea el caso.

4) Si bien los adjetivos presentan medidas F bajas, en muchos casos el refinamiento por cluster es sumamente interesante. Obsérvese por ejemplo el cluster 3, cuyos miembros resultan ser adjetivos que en general son usados con una proposición ("*es preciso que…*", "*es necesario que…*", etc.).

5) En todos los casos, es notable la consolidación de los agrupamientos a partir de los ciclos medios (ciclo 52 en adelante).

7.11.2 Comparación con el baseline

Una forma usual en lingüística computacional de poder apreciar los resultados de un experimento es calcular la línea de base (*baseline* o *lower bound*) (Manning y Schütze 1999) contra el cual los resultados son comparados. El baseline varía con la copmplejidad de la tarea que el experimento se propone modelizar, pero en tareas de clasificación -dada una de las 1000 palabras target, asignarle el correcto POS-tag puede ser entendido como una tarea de clasificación- se suele recurrir al cómputo de las probabilidades de cada una de las etiquetas a asignar. Esta medida refleja estadísticamente el comportamiento de un sistema que actuara eventualmente por azar, sin ninguna inferencia previa o inducida.

En nuestro caso la Tabla 39 muestra el baseline para cada uno de los 16 POS-tags inducidos en hiperclusters y la efectividad de predicción del ciclo óptimo 87. La medida F también es una probabilidad de acertar correctamente la asignación del POS-tag, por lo que es directamente comparable con el baseline. Observamos que nuestro sistema predice el POS-tag de una palabra para las categorías mayores (verbos, sustantivos y adjetivos) entre 3 y 14 veces mejor que el azar, y para las categorías menos numerosas (cardinales, preposiciones, determinantes, etc.) la efectividad de nuestro sistema trepa a entre 50 y 70 veces más que el mero azar.

baseline	Fciclo 87	Fciclo87/base	POS-tag
0,106	0,4457	4 veces	AJ1
0,038	0,4000	**11 veces**	AJ2
0,055	0,4076	7 veces	AV0
0,014	0,8000	**57 veces**	CRD
0,007	0,0000	0 veces	DPS
0,007	0,5000	**71 veces**	DT1
0,007	0,4000	**57 veces**	DT2
0,342	0,8748	3 veces	NN1
0,092	0,7758	8 veces	NN2
0,043	0,4000	9 veces	NNP
0,005	0,0000	0 veces	PND
0,008	0,6154	**77 veces**	PRP
0,014	0,3158	**23 veces**	VMZ
0,042	0,6038	**14 veces**	VVI
0,014	0,6923	**49 veces**	VVN
0,117	0,7984	7 veces	VVZ
0,0569		**12 veces**	PONDERADO

Tabla 39: Comparación de la efectividad máxima del experimento de inducción (ciclo 87) sobre el baseline, por POS-tag y promedio ponderado

El promedio ponderado de todo el sistema (0,69% de medida F) es de una efectividad 12 veces superior al baseline ponderado (tomando en cuenta la distribución de POS-tags iniciales en las 1000 palabras).

7.11.3 Comparación con los trabajos clásicos y con el estado del arte

Como ya se explicó anteriormente, evaluar distintos experimentos de clustering no resulta una tarea sencilla. La variación en los distintos escenarios de objetos a ser clusterizados, para distintos idiomas e, incluso, con diferentes métricas nos obliga a ser cuidadosos con cualquier afirmación surgida de la mera comparación de resultados.

En principio estamos muy interesados en compararnos con el experimento de Redington *et al*. (1998), cuyo trabajo es el que más se asemeja al nuestro por las metodologías aplicadas en el input a tratar, en el algoritmo de clustering y en la evaluación de los resultados (véase sección 5.4).

		Observed	
Class	n	Accuracy	Completeness
noun	407	.90	.53
adjective	81	.38	.45
numeral	10	.09	.82
verb	239	.72	.24
article	3	.10	1.00
pronoun	52	.25	.24
adverb	60	.17	.18
preposition	21	.33	.53
conjunction	9	.06	.33
interjection	16	.18	.67
complex contraction	58	.55	.47
Overall	956	.72	.47

Tabla 40: Precisión y Cobertura para cada POS-tag en el experimento 3 en Redington *et al*. (1998)

Recordemos que en lugar de la medida F, Redington *et al*. (1998) calculan la informatividad (*informativeness*, véase *Ecuación 9*) como una forma de sopesar a la vez los falsos positivos y los falsos negativos. No obstante, a partir de la Tabla 40, en donde recordamos los valores de *Precisión P* (*Precision*, o como se denomina erróneamente en ese trabajo, *Accuracy*) y *Cobertura C* (*Completeness*) que habían surgido del experimento 3 de Redington *et al*. (1998) (véase sección 5.4.3), estamos en condiciones de calcular las medidas F (véase *Ecuación 20*) para el total (*Overall*) y así compararla con nuestra efectividad de 0,69. **El valor de medida F total para Redington *et al*. (1998) es de 0,57 es decir, 12 puntos porcentuales por debajo de la efectividad de nuestro sistema.**

$$\text{Medida } F = \frac{2 * P * C}{P + C} = (2*0.72*0.47) / (0.72+0.47) = 0.57$$

Ecuación 23: Medida F total (*Overall*) en el experimento 3 de Redington *et al.* (1998)

En comparación con otros trabajos del estado del arte, nuestro experimento reporta una efectividad menor, aunque con valores muy cercanos, como en el caso de Clark (2003) con medida F de 0,724 para el inglés (véase sección 5.6) o el de Berg-Kirkpatrick *et al*. (2010) con medida F de 0,755 para el inglés –este último trabajo no contempla una modelización psicolingüísticamente plausible de los PLD, como se explicó en la sección 5.7.

No resulta un detalle menor que nuestro experimento esté enfocado al español. En efecto, un lenguaje con libre orden de constituyentes sintácticos podría atentar contra una buena perfomance de las técnicas distribucionales de inducción (Clark 2002). Por otro lado, si bien esta desventaja podría ser mitigada con una disponibilidad mayor de información morfológica previa para la tarea de categorización que para el inglés, nuestro experimento no contemplaba el tratamiento de dicha infromación morfológica plausiblemente disponible. En ese sentido, seguíamos los lineamientos clásicos de Clark (2002), aunque es de esperar que en nuestro trabajo a futuro podamos investigar la viabilidad de un enfoque que incluya morfología previa (Clark 2003).

Existen muy pocos trabajos de inducción de POS-tag en español. Para el caso, Graça *et al*. (2011) adaptaron los algoritmos clásicos (véase capítulo 5) (Brown *et al*. 1992 BROWN y Clark 2002 CLARK5 y Clark 2003 CLARK10 en Tabla 41) a nuestro idioma, reportando medidas F sensiblemente menores que para el inglés en el criterio *many-to-1*. Así pues, si la comparación se hace en base al español, nuestro experimento reportaría una efectividad equiparable a la de los mejores trabajos clásicos del campo.

		1-Many					
		En45	En17	PT	BG	DK	ES
1	BROWN	68.7	68.7	69.6	63.2	69.6	69.7
2	CLARK5	72.4	63.5	66.0	62.3	57.3	67.6
3	CLARK10	72.5	63.2	67.1	57.0	58.2	70.1

Tabla 41: Medidas F porcentuales para Brown *et al*. (1992), Clark (2002) y Clark (2003), adaptados a distintos idiomas (inglés EN45 y EN 17, portugués PT, búlgaro BG, danés DK y español ES), según Graça *et al*. (2011)

7.11.4 Plausibilidad psicolingüística de la modelización

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

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

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

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

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

7.12 Trabajo a futuro para el experimento de categorización

Existe una gran área de mejora del experimento en relación con su escalabilidad. Para ello, sería fundamental ampliar el corpus de PLD a decenas de millones de tokens, en función de las cantidades que se manejan en trabajos más abarcativos (Clark 2002). A su vez, sería importante ampliar el corpus de referencia para la desambiguación POS-tag y mejorar las anotaciones manuales del mismo. Por ejemplo, más allá de los errores e inconsistencias esperables en la tarea de anotación manual, resulta evidente que la colección de etiquetas morfosintácticas elegidas para el criterio de evaluación podría refinarse más (sustantivos singulares masculinos NN1M y femeninos NN1F, etc.). Se trata de una movida sutil porque no es deseable incurrir en una eclosión de categorías morfosintácticas que atentaría contra las posibilidades de inducción en función de la dispersión de datos -la explicación para este refinamiento problemático del *tag se*t de categorías ya ha sido dada en secciones anteriores.

Por el lado algebraico de la modelización del espacio vectorial se podría experimentar con escenarios de decisiones matemáticas diversas, como por ejemplo la normalización de los valores de frecuencia absoluta de las relaciones bigrámicas que componen los vectores (Manning y Schütze 1999) o el cambio del criterio de similitud entre objetos (distancia Manhattan por distancia euclideana). También sería importante estudiar la viabilidad de extender el mecanismo propuesto para la categorización de todas las palabras de un vocabulario, más allá de las limitaciones impuestas por el desarrollo ontogenético en nuestra modelización, en pos de corroborar una mayor plausibilidad cognitiva del modelo.

Posiblemente la línea de investigación más desafiante tenga que ver con una reconsideración de los indicios facilitadores del clustering, en cuanto a considerar información morfológica previa a la categorización (Clark 2003). En el los trabajos actuales no hay un claro consenso acerca de si la información morfológica es plausiblemente un posible input para la categorización, un producto de la misma o incluso parte del mismo proceso de categorización (Clark 2002, 2003). Por ejemplo, en el caso típico del morfema amalgama '*-o*' en español, la marcación morfológica debería afectar únicamente a verbos. Es probable que la modelización de la morfología a través de la ortografía de la palabra (procedimiento al que recurren los trabajos del estado del arte) en forma previa a la categorización de la misma, si es que no se combina apropiadamente con un procedmiento de desambiguación morfológica, pueda resultar en concepciones erróneas que nunca serían atestiguadas en el desarrollo ontogenético. No obstante, algunos trabajos proponen modelizar la información morfológica como un facilitador más para la categorización (Clark 2003) con resultados potenciados respecto de enfoques distribucionales más asépticos (Clark 2002):

> "Clearly, a learning algorithm would have to learn morphology at the same time as, or prior to, learning the set of syntactic categories. […] Languages with comparatively rich morphology tend to have rather free word order, which might cause problems with distributional induction techniques. However these languages tend to signal the part of speech in the surface form of words, so it would be possible to use that information to learn." [Clark 2002:75-76]

En todo caso, como mencionamos en las decisiones de diseño de nuestro experimento, todas estas faciliatciones adicionales son una demostración *a fortiori* del enfoque.

Si bien nuestro experimento ha suministrado evidencia contundente de la viabilidad de modelos que recurren a 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 categorías morfosintácticas en sí mismo, sino la posibilidad de inducir una gramática completa a partir del mismo (Clark 2002; Balbachan y Dell'Era 2010). Aunque reconocemos que esta meta de largo plazo es por demás ambiciosa para el estado de arte actual de la disciplina, en el próximo capítulo de esta tesis describimos sucintamente cómo podría aprovecharse la salida del experimento de categorización en un eventual *pipeline* de inducción de sintaxis rudimentaria.

Capítulo 8. Continuación del experimento de categorización hacia una sintaxis rudimentaria: inducción de constituyentes sintácticos

8.1 El estado actual de la cuestión en inducción de gramáticas formales (grammar inference)

Retomando lo que habíamos adelantado en la sección 2.1, los experimentos de inducción de categorías sintácticas son el punto de partida para enfoques integrales sobre la inducción no supervisada de gramáticas formales a partir de corpora no anotado. **Con este tipo de experimentos se podría postular la plausibilidad algorítmica de aprovechar la información de salida del experimento del capítulo 7 como punto de partida para la construcción de una sintaxis rudimentaria, mediante la inducción de constituyentes sintácticos a partir de la etiquetación morfosintáctica de palabaras (categorización).**

Clark (2002) aplicó técnicas de clustering sobre un corpus anotado con información de clase morfosintáctica de palabra -surgida de su propio experimento de categorización, véase sección 5.6- para agrupar secuencias y hallar constituyentes mediante un criterio de información mutua contextual. Los constituyentes así agrupados se consideran reescrituras de símbolos. A partir de una gramática sin sesgos, donde hay una regla para cada símbolo a reescribir, Clark (2002) procede a reagrupar las reescrituras halladas para formar reglas que minimicen la longitud de descripción de la gramática (criterio *Minimum Description Length* MDL). En cada iteración agrega los símbolos no terminales que hagan falta y realiza un análisis parcial del corpus con la gramática obtenida hasta el momento, continuando hasta que un criterio de medida determine que la gramática independiente de contexto hallada deja de producir constituyentes plausibles.

Usando una estrategia similar a la de Clark (2002), basada en árboles de estructura de frase y métodos de clustering distribucional, Klein y Manning (2002) obtuvieron buenos resultados en la tarea de inducir estructuras de constituyentes de manera no supervisada para el aprendizaje de clases de palabras sobre cálculos de similitud basados en información contextual. Luego extendieron su trabajo con modelos de dependencia (Mel'čuk 1988; Klein y Manning 2004) que usan la co-ocurrencia de palabras como forma de facilitar los cálculos probabilísticos en estructuras de núcleo y símbolos dependientes. Combinando su método de constituyentes lineales y su método de dependencia, Klein y Manning (2004) obtienen el mejor resultado hasta la fecha en la tarea de inducción de gramática y logran capturar algunos rasgos elementales de la estructura del lenguaje. Evaluados sobre corpora translingüísticos, todos estos modelos explotan las regularidades de la información de cada lenguaje y proveen análisis sintácticos de alta calidad a partir de ningún conocimiento previo específico de dominio.

Spitkovsky, Alshawi y Jurafsky (2009) parten del método de Klein y Manning (2004) para inducir gramáticas, pero en lugar de ajustar a mano ciertos parámetros apriorísticos requeridos por el enfoque bayesiano, optan por considerar el tamaño del corpus como una variable

significativa. Su algoritmo aprende a partir de corpora de tamaño y complejidad gradualmente crecientes.

Entre los enfoques aplicados al español, uno de los primeros trabajos de inducción de gramáticas es el algoritmo de inducción de Juárez Gambino y Calvo (2007). Basándose en la noción de *sustituibilidad* de Harris (1954) para hallar regularidades estructurales, los investigadores desarrollaron un algoritmo no supervisado para entrenar el sistema de inducción de gramática ABL (*Alignment-Based Learning*) (van Zaanen 2000) con un corpus de español bastante reducido (CAST-3LB).

8.2 Diseño de corpus propio para inducción de constituyentes

Para su tesis y, en particular, para el experimento de inducción de constituyentes sintácticos Clark (2002) recurre al *British National Corpus* (BNC) en su primera edición del año 1994, un *corpus* sincrónico de inglés británico que contiene 100 millones de tokens de registro variado (periódicos, obras literarias, etc.), etiquetados automáticamente según el estándar C4 (CLAWS-4) –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 en este capítulo es el estudio de factibilidad del experimento de Clark (2002) acerca de la inducción de constituyentes sintácticos para el español, nuestro trabajo se propuso inicialmente 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, recurriendo a un corpus anotado morfosintácticamente comparable al de Clark (2002). Los lineamientos que se siguieron en la creación de este corpus de referencia en español han sido desciptos en la sección 7.8 de esta tesis.

Dadas la masividad y complejidad de las consultas sobre la distribución de las etiquetas morfosintácticas y la considerable proliferación de las mismas en virtud de criterios demasiado granulares, lo cual podría generar un consecuente problema de dispersión de datos, 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.

	Clark (2002)	**Nuestro corpus**
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 C4	Manual: 48 etiquetas adaptadas del estándar C4
Criterio de anotación	BNC (Leech *et al.* 1994)	Propio, adaptado del BNC

Tabla 42: Comparación entre corpora de entrada para ambos experimentos de inducción de constituyentes

8.3 Algoritmo de inducción de constituyentes sintácticos en Clark (2002)

8.3.1 Descripción general

El algoritmo de Clark (2002) que nos ocupa en este caso 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).

8.3.2 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 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 (2002) se intenta hallar los constituyentes más básicos para construir reglas gramaticales.

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); 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 (2002). 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 29.

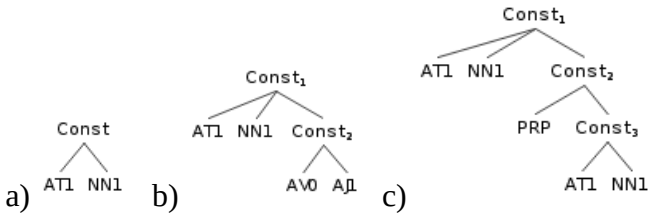

Figura 29: Constituyentes de a) 1 nivel, b) 2 niveles y c) 3 niveles de imbricación

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

Para el primer paso de su algoritmo, Clark (2002) 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 $77^2 \approx 5500$ parámetros, 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*:

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

Ecuación 24: Cálculo de umbral mínimo de ocurrencias según distribución de etiquetas morfosintácticas

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 (2002) al nuestro. Con todo, consideramos que estos lineamientos en cuanto al umbral de ocurrencias son comparativamente significativos: Clark (2002) obtiene 753 secuencias candidatas y en nuestro caso obtenemos 198 para el escenario más efectivo de 110 ocurrencias (véase *Tabla 44*).

Una idea importante que opera en el experimento de Clark (2002) 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 o no un constituyente.

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 de tales pares.

8.5 Paso 2: Clustering de secuencias candidatas a constituyentes

Una vez obtenida la Tabla 43 de información distribucional, el siguiente paso en el algoritmo de Clark (2002) consiste en agrupar 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 2002:132]

Clark (2002) sugiere como algoritmo de clustering el método 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 (2002) 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 (2002) 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.). No obstante, en nuestro caso, sólo estamos interesados en el sub-proceso de inducción de constituyentes (véase sección *8.7 Modificaciones al experimento original de inducción de constituyentes*).

	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 43: Muestra de tabla de información distribucional (secuencias y contextos) para inducción de constituyentes

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

Una vez concluido el paso 2, Clark (2002) obtiene 100 clusters. En nuestro caso, como se ve en el *Anexo I*, obtuvimos 25 clusters. Sin embargo, como Clark (2002) 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 2002:133]

Para determinar cuáles son los grupos de secuencias válidas como constituyentes, Clark (2002) 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 MI*):

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

Ecuación 25: 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; $p_2(y)$ 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 30 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:

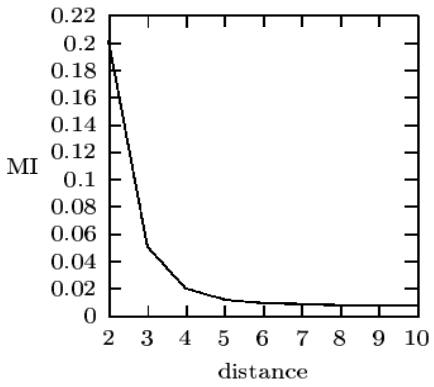

Figura 30: La información mutua entre el contexto previo y el contexto posterior desciende conforme crece la distancia que los separa (medida en símbolos), según Li (1990)

Por esta razón, Clark (2002) 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 (2002) 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 nuestra réplica del experimento hasta este punto obtuvimos resultados similares a los reportados por Clark (2002). Mediante una evaluación manual de los clusters determinamos una medida F (promedio armónico entre precisión y cobertura) del 65% (véase Tabla 44).

8.7 Modificaciones al experimento original de inducción de constituyentes

En su experimento original, Clark (2002) 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 nosotros. 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:

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

Ecuación 26: *Pointwise mutual information* (información mutua punto a punto)

En la Figura 31 se puede apreciar un diagrama que compara el experimento original de Clark(2002) y nuestra adaptación al español. Clark (2002) 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 la *MI punto a punto* 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.

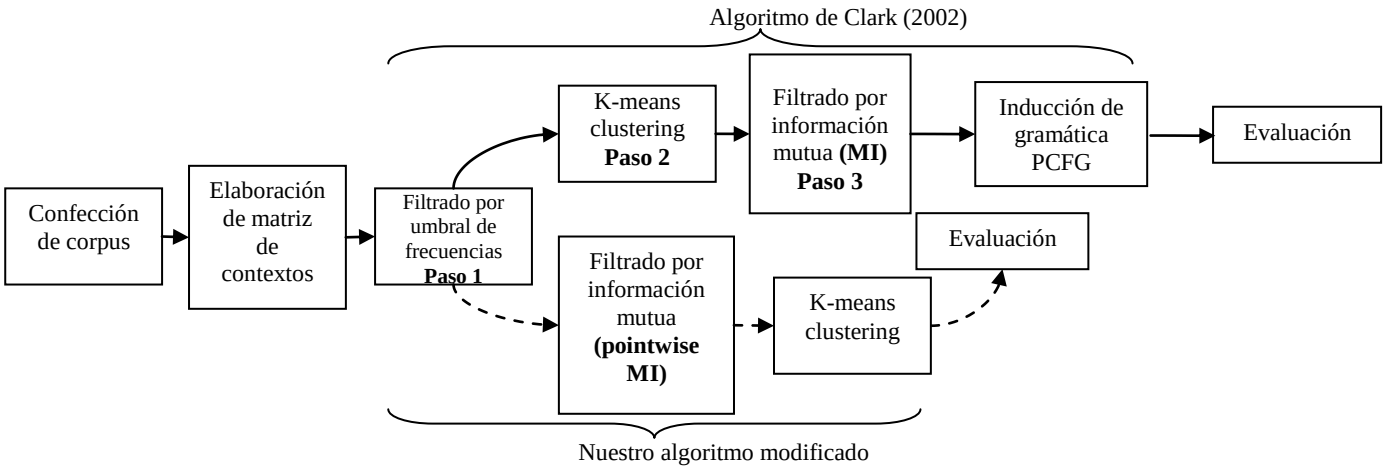

Figura 31: Experimento original de Clark (2002) de inducción de sintaxis y nuestra adaptación al español para la inducción de constituyentes

8.8 Evaluación de los resultados de inducción de constituyentes

A continuación resumimos la descripción de nuestro experimento de inducción de constituyentes sintácticos:

i. Dividimos las 2.108 oraciones en dos grupos: 2000 oraciones para entrenamiento y 108 para la evaluación de los constituyentes inducidos (véase *Anexo III*).
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. Agrupamos las 107 secuencias con clustering basado en K-means: obtuvimos 25 clusters de alta pureza (véase *Anexo I*).
v. Repetimos el experimento con distintos umbrales en (i.), obteniendo los resultados de la Tabla 44.
vi. Evaluamos manualmente la salida de (iii.) con nuestros propios juicios de gramaticalidad (véase *Anexo II*), de modo de calcular la medida F para cada escenario de (v.): obtuvimos distintos valores para la columna de constituyentes válidos (*nº Positivos* en la Tabla 44).

Las evaluaciones del algoritmo original de Clark y de nuestra implementación revelaron que ambos métodos convergen en resultados similares. Obtuvimos una medida F de alrededor del 65% para el escenario más efectivo de nuestro experimento (véase *Tabla 44*).

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 numeral cardinal)- bien integradas a sintagmas nominales. En otros casos, la extensión de los constituyentes llega a 5 y 6 etiquetas (véase *Anexo II*), lo que demuestra la viabilidad del enfoque para inducción de complejas estructuras de constituyentes.

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 44: Evaluación de la medida F para distintos escenarios en experimento de inducción de constituyentes, según umbral de ocurrencias

$$\text{Precisión } P = \frac{nºConstituyentesPositivos}{nºConstituyentesPositivos + FalsosPositivos}$$

Ecuación 27: Precisión para inducción de constituyentes

$$\text{Cobertura } C = \frac{nº\,Constituyentes\,Positivos}{nº\,Constituyentes\,Positivos + Falsos\,Negativos}$$

Ecuación 28: Cobertura para inducción de constituyentes

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

Ecuación 29: Medida F para inducción de constituyentes (Con $\beta = 1$ para asignar igual peso a *P* y a *C*)

8.9 Discusión de los resultados del experimento de inducción de constituyentes

Este experimento verifica la factibilidad de encarar la tarea de inducción de constituyentes sintácticos en español con un enfoque basado en la información distribucional, únicamente a partir de categorías morfosintácticas como input, de modo de conectar la salida del experimento central de nuestra tesis a un *pipeline* en cascada de inducción integral de sintaxis (véase *Figura 1*). 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 definirlo. 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 este experimento a escala 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 44*).

Por otro lado, como el mismo Clark (2002) 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 II*). 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 de 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.

Capítulo 9. Conclusiones generales

9.1 Una nueva visita al APS: Mecanismos cognitivos de aprendizaje por inducción

A lo largo de toda esta tesis hemos abordado el problema de la adquisición de lenguaje y, en particular, de sintaxis a partir del isomorfismo entre lenguajes naturales y formales. Para ello hemos planteado una posible modelización de dicho proceso, tomando la categorización de palabras como punto de partida y explorando la plausibilidad empírica de nuestro modelo, en función de la disponibilidad de indicios posibles en una etapa tan temprana del desarrollo ontogenético del lenguaje. Finalmente, hemos demostrado la factibilidad algorítmica de dicho modelo, obteniendo categorías morfosintácticas únicamente a partir de las propiedades distribucionales del corpus, sin recurrir a premisas o mecanismos de aprendizaje de dominio específico.

Hasta aquí nuestra tesis podría ser un modelo viable para demostrar la invalidez del APS, en el sentido de que no es necesario postular conocimiento lingüístico innato para echar a rodar el algoritmo de inducción general de sintaxis (Clark 2002; Lappin y Shieber 2007; Clark y Lappin 2013). No obstante, el que un modelo sea viable no significa que coincida necesariamente con los mecanismos concretos de dicho proceso de adquisición del lenguaje:

> "The APS rests on the premise that there are no general purpose learning algorithms that could learn a plausible grammar for any natural language based on the small sample of data available to the infant child. It can therefore be refuted by demonstrating that there are "generalised learning mechanisms" that can do just that. The most convincing way of doing this would be to exhibit a fully implemented computer program that can perform this task, and it is this that I attempt to do later on. Note that we are not concerned here with whether the human child actually uses these mechanisms or not. There are a number of requirements for a learning algorithm to constitute a refutation of the APS. I shall mention the main criteria here, and then discuss them further below. The data that the algorithm learns from should be as close as possible to the data that is available to the child, both in quantity and type. The algorithm should not have access to any linguistic or domain specific information. It must produce a "plausible" grammar. It must be a general purpose learning algorithm. It should work on all natural languages, and not make languagespecific assumptions. However, there are some criteria that it need not satisfy: in particular it need not be a cognitive model for language. This is an important point: we are not trying to find a cognitive model of language learning. All we are trying to do is demonstrate the existence of algorithms that can learn from the data available to the child." [Clark 2002:11]

En efecto, sostenemos que los mecanismos generales de aprendizaje pueden ser modelizados a través de las técnicas de clustering para el agrupamiento de objetos en múltiples dimensiones. Pero, ¿acaso estos mecanismos de aprendizaje están efectivamente disponibles para los adquirientes de un lenguaje natural? Todos los enfoques modernos sobre el problema con pretenciones de adecuación descriptiva y adecuación explicativa adscriben a una postura mentalista del proceso de adquisición del lenguaje (Chomsky 1965, 1975; Clark 2002; Eguren y Soriano 2004). En el fondo, la gran polémica se centra en torno al estado inicial del sistema mente/cerebro: ya sea dotado de estructuras ricas y principios de dominio específico, cuyos parámetros son deducidos a partir de los PLD –postura del innatismo que propone un *sesgo fuerte* (*strong bias*)-, ya sea dotado de mecanismos generales de aprendizaje no específicos de dominio que actúan por inducción sobre los PLD –postura del empirismo que postula un *sesgo débil* (*weak bias*). Se suele considerar que si adscribimos al innatismo, podemos "barrer bajo la

alfombra" el problema de los mecanismos cognitivos específicos, ya que postular en forma innata principios ricamente estructurados constreñiría notablemente el espacio posible de generación de gramáticas particulares parametrizadas para PLD dados. Sin embargo, se ha demostrado que esto no es así:

> "Estas observaciones llevarían a C a la conclusión natural de que la organización intelectual de un ser humano maduro es un sistema complejo e integrado que incluye ciertas estructuras adquiridas sobre la base de adaptaciones iniciales bastante específicas." [Chomsky 1975:200]

> "Results derived from complexity considerations offer a more promising source of insight into the boundary conditions of learning. The positive results that we have surveyed here indicate that we should be modeling the target of language acquisition as an objective representation -a grammar whose primitive symbols can be directly identified from the measurably observable properties of the language itself. Such representations are efficiently learnable. Therefore, these results suggest that the types of grammars that have been posited within different variants of the Principles and Parameters program are not plausible candidates for learnable representations. This is because the values of the proposed parameters are remote from the data of natural language, and so they cannot be efficiently estimated or learned. In positive terms, although the most elementary formalisms initially considered on the objective representation approach are too weak, more sophisticated grammars seem to achieve the right level of expressive power for capturing the properties of natural language syntax, while remaining efficiently learnable." [Clark y Lappin 2013:105]

En cambio, si adscribimos a la hipótesis empirista, el gran problema que subyace es cuáles son los mecanismos cognitivos concretos involucrados en la inducción de gramáticas a partir de los PLD. En última instancia, de un modo u otro, nuestra investigación acerca de una modelización plausible de inducción de sintaxis nos conduce inexorablemente a interrogarnos acerca del funcionamiento cognitivo de los adquirientes de una lengua ante los PLD:

> "¿Qué clase de estructuras cognitivas desarrollan los seres humanos sobre la base de su experiencia, específicamente en el caso de la adquisición del lenguaje? ¿Cuál es la base para la adquisición de tales estructuras y cómo se desarrollan?" [Chomsky 1975:173]

Chomsky y Miller (1963) y Chomsky (1975) argumentan que un mecanismo de aprendizaje inductivo para la adquisición del lenguaje basado en procesos markovianos de primer orden no converge hacia sistemas adecuados para representar la sintaxis de los lenguajes naturales. Este argumento, sería formalmente demostrado luego mediante el *Teorema de Gold* (Gold 1967), tal como explicamos en la sección *1.4 El Teorema de Gold revisitado*:

> "Supongamos , por ejemplo que puede demostrarse que una teoría del aprendizaje particular tienen la siguiente propiedad: un sistema, que de otra manera carecería de estructura y que puede ser modificado de acuerdo con los mecanismos de esta teoría, puede aproximarse en su límite a cualquier mecanismo de estados finitos que produzca cadenas de izquierda a derecha a medida que pasa de estado a estado, pero nada más que este mecanismo. Puesto que es bien conocido que éste no puede representar ni siquiera la sintaxis de sistemas extremadamente simples (por ejemplo, el cálculo proposcicional) y con seguridad no los de la sintaxis de la lengua, podemos llegar a la conclusión inmediata de que la teoría es inadecuada para explicar el aprendizaje lingüístico." [Chomsky 1975:198-199]

La gran falacia que subyace a esta descalificación de las teorías inductivas de aprendizaje del lenguaje consiste en suponer que todo mecanismo inductivo debe necesariamente basarse en cadenas markovianas, aun de órdenes superiores (Cohen 1970; Manning y Schütze 1999).

Cohen (1970) polemiza explícitamente con Chomsky y Miller (1963) al proponer un mecanismo de aprendizaje inductivo no markoviano para la adquisición del lenguaje. Cohen (1970) comienza por identificar el germen de la generalización de la descalificación chomskyana hacia las teorías de aprendizaje inductivo en lo que Bacon denominó *inducción enumeradora*

(*enumerative induction*) (Cohen 1970; Chomsky 1975) como correlato de la lógica de primer orden, la cual no alcanza para describir adecuadamente el lenguaje natural (Quine 1970). En cambio, Cohen (1970) sugiere una estrategia inductiva diferente: la *inducción variacional* (*variational induction*).

> "He seems preoccupied, and Chomsky and Miller likewise, with the kind of casual influencing of the mind that is analogous to what Bacon called induction by simple enumeration, and ignores altogether the kind of casual influencing that is analogous to induction by variation of circumstance. But there is no plausible reason why the title «empirical» should be withheld from the latter method of learning, if it is granted to the former. Indeed, if science in fact prefers variational induction to Markovian learning, there is a strong presumption that the former is a better method than the latter for accomplishing complex learning tasks. […] But I shall try to show that variational induction -on an appropiate construction- can in principle produce results in a way quite different from Markovian learning, and that therefore the Chomsky-Miller argument against empirical language-learning is invalid." [Cohen 1970:300]

Cohen (1970) considera que tal estrategia de aprendizaje inductivo no actuaría como un compilador de probabilidades markovianas en relación con una lista predeterminada de categorías, sino como un formulador y testeador de hipótesis. Así pues, este tipo de aprendizaje se basaría en sobregeneralizaciones, que luego serían acotadas. En este sentido, nuestra hipótesis de la abstracción de categorías sintácticas a partir de los datos del propio lenguaje podría ser compatible con este mecanismo cognitivo inductivo. Hemos demostrado en la evaluación iterativa de la historización de nuestro experimento (véase *capítulo 7*) que las categorías sintácticas son inducidas por refinamiento de clases sobregeneralizadas (al principio, todos las palabras tienden a ser sustantivos, luego van apareciendo verbos y más tarde las demás categorías granulares). A su vez, esto se condice con la evidencia empírica en producción del desarrollo ontogenético de las categorías de palabras (Brown 1973).

Aun aceptando lo anterior, resta sortear el escollo de la necesidad de un mecanismo cognitivo poderoso para inducir toda la sintaxis de un lenguaje natural -no sólo la inducción de clases del proceso de categorización de palabras sino principalmente las reglas sintácticas que se aplican sobre ellas. Recientes investigaciones en ciencias cognitivas aportan una renovada evidencia sobre el problema. Dewar y Xu (2010) demuestran que a partir de una edad tan temprana como los 9 meses los niños ya disponen de la capacidad para adquirir conocimiento generalizable a partir de evidencia muy escasa (sobregeneralizaciones como evidencia de lógica de segundo orden, la cual sí es suficiente para describir formalmente el lenguaje natural):

> "This is critical for formation of an overhypothesis –a second-order inductive generalization that involves postulating two variables (e.g. bags and colors of marbles) and some correspondence rule to link the specific values of the two variables. The present study provides evidence that as early as 9 months of age, infants are able to make second-order inductive generalizations. […] Future research is needed to undertstand the nature of the underlying inductive learning mechanism, which may be responsible for the acquisition of several presumed-innate inductive biases in a number of knowledge domains." [Dewar y Xu 2010:6]

Incluso mecanismos más complejos como la recursividad, que otrora constituyeran una fortaleza inexpugnable del domino lingüístico específico, el germen mismo de la dependencia estructural de la sintaxis (Chomsky 1957, 1975; Chomsky y Miller 1963), hoy en día son considerados como mecanismos generales de aprendizaje verificables en otros dominios (Cohen

1970), una habilidad innata propia de la especie que se puede rastrear filogenéticamente hasta la talla de huesos y la práctica de nudos del *Homo Neanderthalensis* (Balari *et al*. 2008), lo cual invalida la férrea renuencia original de Chomsky (1975) hacia las teorías de aprendizaje de sesgo débil:

> "Finalmente Cohen arguye que otros enfoques más simples son suficientes para dar cuenta de la adquisición de la lengua y señala algunas posibilidades. Lamentablemente éstas no logran siquiera empezar a tratar las más elementales propiedades lingüísticas que han sido discutidas en la literatura sobre el tema, por ejemplo, la de dependencia estructural. En consecuencia, estas propuestas tal como aparecen no pueden ser tomadas en serio. En cuanto a la posibilidad de que la capacidad para usar transformaciones sea un caso especial de «alguna habilidad genérica», la propuesta es completamente vacua hasta que se especifique de qué habilidad se trata y, por la razones ya mencionadas, no resulta particularmente plausible." [Chomsky 1975:260]

Resulta llamativa esta férrea oposición. Desde la lingüística chomskyana actual se ha abandonado el concepto transformacional hace décadas por las razones expuestas en el capítulo 1 de esta tesis, pero se sigue hablando de capacidades del Lenguaje en Sentido Amplio (LSA) y capacidades del Lenguaje en Sentido Estricto (LSE) (Hauser *et al*. 2002), postulando la recursividad como una de estas últimas (Watumull *et al*. 2014).

9.2 Una reflexión final

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

> "This problem does not entail that formal learning theory has nothing to offer the study of language acquisition. On the contrary, it is highly relevant. However, we argue that the crucial problems are not information theoretic, as suggested in the Gold results. Instead, they are complexity theoretic. By modeling the computational complexity of the learning process, we can, under standard assumptions, derive interesting result concerning the types of representations (or grammars) that are efficiently learnable. It is uncontroversial that the human capacity to learn is bounded by the same computational limitations that restrict human abilities in other cognitive domains. The interaction of this condition with the complexity of inducing certain types of representations from available data constitutes a fruitful object of study." [Clark y Lappin 2013:90-91]

El progreso de las técnicas estadísticas y el avance de las investigaciones sobre corpora abarcativos revelan que incluso los más simples mecanismos estadísticos pueden contribuir al esclarecimiento del proceso de adquisición del lenguaje. En particular, el conjunto de marcas e indicios provistos por la información distribucional constituye una herramienta válida para la inducción de juicios acerca de la pertenencia de palabras a categorías morfosintácticas. Hemos demostrado empíricamente la estrecha correlación entre palabras cue vs. palabras target, distinción operativamente homologable a las nociones lingüísticas de palabras funcionales vs. palabras de contenido, y hemos señalado el importante papel que podrían desempeñar dichas palabras funcionales en la adquisición del lenguaje, aunando las respectivas agendas de investigación de la lingüística computacional y de la psicolingüística. Justamente, una deuda pendiente en el campo de la psicolingüística es la necesidad de compatibilizar evidencia contradictoria acerca del momento ontogenético de la adquisición de las palabras funcionales en

producción y en comprensión, lo cual contribuirá a la mayor adecuación explicativa de los enfoques computacionales, en función de los diversos pre-requisitos de modelización (el pre-requisito son las cues, no la categorización de las cue).

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

Finalmente, resulta imperioso situar este tipo de investigaciones en el marco más general de un proyecto de inducción integral de sintaxis (Clark 2002; Klein y Manning 2004). El aprendizaje no supervisado de sintaxis o, en otras palabras, el problema de la inducción de una gramática a partir de un corpus sin anotaciones, todavía presenta interesantes desafíos desde el punto de vista de la lingüística teórica y de sus aplicaciones prácticas. Algunos algoritmos tempranos (Lari y Young 1990) resultaron técnicamente muy eficientes, pero sus resultados no necesariamente convergían a una gramática óptima. Otros algoritmos bayesianos más recientes para inducir gramáticas probabilísticas independientes de contexto PCFG (Johnson 2008) cuentan con más garantías sobre los resultados, pero resultan ineficientes. Los aportes de la teoría de lenguajes formales indican que el problema de aprendizaje de PCFGs no es tratable directamente, a menos que se cambie la representación de la gramática o que el método utilizado incluya conocimientos lingüísticos capaces de guiar una heurística correcta y eficiente.

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

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

Consideramos entonces que el mérito de la presente tesis es abarcar modelos de inducción de fenómenos sintácticos que puedan aportar renovada evidencia al debate acerca de la adquisición del lenguaje; en especial, si consideramos que 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, las ciencias cognitivas o incluso la biolingüística, pero la plausibilidad de dicha evidencia mediante una modelización efectiva es claramente un asunto para la agenda actual de la lingüística computacional.

Referencias bibliográficas

1. Abney, Steven. 1996. Statistical methods and linguistics. En Judith Klavans y Philip Resnik (eds.). *The balancing act.* Cambridge, Massachusetts. MIT Press.

2. Abney, Steven. 2008. *Semisupervised learning for computational linguistics*. Boca Raton. Chapman & Hall.

3. Balari, Sergio, Antonio Benítez Burraco, Marta Camps, Víctor M. Longa, Guillermo Lorenzo y Juan Uriagereka. 2008. Homo loquens neanderthalensis? En torno a las capacidades simbólicas y lingüísticas del Neandertal. En *Munibe Antropologia-Arkeologia* (59):3-24.

4. Balbachan, Fernando. 2006. Killing time: metaphors and their implications in lexicon and grammar. En *metaphorik.de* (10):6-30.

5. 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.

6. Balbachan, Fernando y Diego Dell'Era. 2010. Inducción de constituyentes sintácticos en español con técnicas de clustering y filtrado por información mutua. En *Linguamática* 2(2):39-57.

7. Barry, Anita. 2002. *Linguistic perspective on language and education*. Westport. Greenwod

8. Bates, Elizabeth, Virginia Marchman, Donna Thal, Larry Fenson, Philip Dale, J. Steven Reznick, Judy Reilly y Jeff Hartung. 1994. Developmental and stylistic variation in the composition of early vocabulary. En *Journal of Child Language* (21):85-123.

9. Baum, Leonard. 1972. An inequality and associated maximization techniques in statistical estimation of probabilistic functions of Markov processes. En *Inequalities* (3):1-8.

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

11. Biemann. 2006a. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. En *Proceedings of HLTNAACL'06 workshop on TextGraphs*, pp.73-80. Nueva York.

12. Biemann, Chris. 2006b. Unsupervised part-of-speech tagging employing efficient graph clustering. En *Proceedings of COLING ACL 2006*, pp.7-12. Morristown.

13. Bloom, Lois. 1973. *One word at a time: The use of single word utterances before syntax*. La Haya. Mouton.

14. Bloomfield, Leonard. 1933. *Language*. Nueva York. Holt.

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

16. Bowerman, Melissa. 1973. Structural relationships in children's utterances: syntactic or semantic? En Moore, Timothy (ed.). *Cognitive development and the acquisition of language.* Nueva York. Academic Press.

17. Braine, Martin. 1987. What is learned in acquiring word classes? A step toward an acquisition theory. En MacWhinney, Brian (ed.). *Mechanisms of language acquisition*. Hillsdale. Erlbaum.

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

19. Brown, Roger. 1973. *A first language: The early stages*. Cambridge, Massachusetts. Harvard University Press.

20. Cabré, María Teresa y Mercè Lorente. 2003. Panorama de los paradigmas en lingüística. En Estany, Anna (coord.). *Enciclopedia iberoamericana de filosofía*. Madrid. Consejo Superior de Investigaciones Científicas.

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

22. Cattell, Raymond. 1943. The description of personality: Basic traits resolved into clusters. En *Journal of Abnormal and Social Psychology* (38):476-506.

23. Ćavar Damir, Paul Rodrigues y Giancarlo Schrementi. 2004. Syntactic parsing using mutual information and relative entropy. En *Proceedings of the Midwest Computational Linguistics colloquium (MCLC)*. Bloomington.

24. Ćavar Damir. 2010. On statistical metrics for selection and phrasality. En Thomas Hanneforth y Gisbert Fanselow (eds.). *Language and logos*. Berlin. Akademie Verlag.

25. Chater, Nick y Peter Conkey. 1993. Sequence processing with recurrent neural networks. En Mike Oaksford y Gordon Brown (eds.), *Neurodynamics and Psychology*, pp.269-294. Londres. Academic Press.

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

27. Chemla, Emmanuel, Toben Mintz, Savita Bernal y Anne Christophe. 2009. Categorizing words using ´frequent frames´: what cross-linguistic analyses reveal about distributional acquisition strategies. En *Developmental Science* 12(3):396-406.

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

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

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

31. Chomsky, Noam y George Miller. 1963. Finitary models of language users. En R. Duncan Luce, Robert Bush y Eugene Galanter (eds.). *Handbook of mathematical psychology* (Vol. 2), pp.419-491. Nueva York. John Wiley and Sons.

32. Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, Massachusetts. MIT Press.

33. Chomsky, Noam. 1966. *Topics in the theory of generative grammar*. París. Mouton.

34. Chomsky, Noam. 1975. *Reflexiones sobre el lenguaje*. Buenos Aires. Sudamericana.

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

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

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

38. Chistodoulopoulos, Christos, Sharon Goldwater y Mark Steedman. 2011. A Bayesian mixture model for Part-of-Speech induction using multiple features. En *Proceedings of the 2011 conference on empirical methods in Natural Language Processing*, pp.638-647. Edimburgo.

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

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

41. Clark, Alexander. 2002. *Unsupervised language acquisition: theory and practice*. Tesis de doctorado. University of Sussex.

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

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

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

45. Clark, Eve. 2009. *First Language Acquisition*. Cambridge, Reino Unido. Cambridge University Press.

46. Cohen, Jonathan. 1970. Some applications of inductive logic to the theory of language. En *American Philosophical Quaterly* (7):299-310.

47. Cowie, Fiona. 1999. *What's within? Nativism reconsidered*. Oxford. Oxford University Press.

48. Cramer, Bart. 2007. Limitations of current grammar induction algorithms. En *Proceedings of the ACL 2007 student research workshop*:43-48.

49. Cutting, Douglas, Jan Pedersen, David Karger y John Tukey. 1992. Scatter/Gather: a cluster-based approach to browsing large document collections. En *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval SIGIR '92*, pp.318-329. Copenhague.

50. Deacon, Terrence. 1997. *The symbolic species*. Nueva York. W.W. Norton.

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

52. Dewar, Kathryn y Fei Xu. 2010. Induction, overhypothesis, and the origin of abstract knowledge. En *Psychological Science* 21(12):1871-1877.

53. Diaz, Michele y Gregory McCarthy. 2009. A comparison of brain activity evoked by single content and function words: An fMRI investigation of implicit word processing. En *Brain Research* (1282):38-49.

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

55. Eguren, Luis y Olga Fernández Soriano. 2004. *Introducción a una sintaxis minimista*. Madrid. Gredos.

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

57. Elman, Jeffrey. 1991. Distributed representations, simple recurrent networks, and grammatical structure. En *Machine Learning* 7(2/3):195-225.

58. Färber, Ines, Stephan Günnemann, Hans-Peter Kriegel, Peer Kröger, Emmanuel Müller, Erich Schubert, Thomas Seidl y Arthur Zimek. 2010. On using class-labels in evaluation of clusterings. En *Proceedings of the 16th ACM SIGKDD International Conference knowledge discovery and data mining*. Washington.

59. Fenson, Larry, Virginia Marchman, Philip Dale, Steven Reznick, Donna Thal y Elizabeth Bates. 1993. *The MacArthur communicative development inventories: User's guide and technical manual*. Baltimore. Paul H. Brokes Publishing Co.

60. Fillmore, Charles. 1992. 'Corpus linguistics' vs. 'computer-aided armchair linguistics'. Directions in Corpus Linguistics. En *Proceedings from a 1992 Nobel symposium on corpus Linguistics*, pp.35-60. Estocolmo.

61. Finch, Steve y Nick Chater. 1992. Bootstrapping syntactic categories. En *Proceedings of the 14th annual conference of the Cognitive Science Society of America*, pp.820-825. Bloomington, Indiana University.

62. Finch, Steve. 1993. Finding structure in language. Tesis de doctorado. University of Edinburgh.

63. Finch, Steve, Nick Chater y Martin Redington. 1995. Acquiring syntactic information from distributional statistics. En Levy, Joseph, Dimitrios Bairaktaris, John Bullinaria y Paul Cairns (eds.). *Connectionist models of memory and language*. Londres. UCL Press.

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

65. 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*, pp.539-544. Austin.

66. Frank, Stella, Sharon Goldwater y Frank Keller. 2009. Evaluating models of syntactic

category acquisition without using a gold standard. En *Proceedings of CogSci09*, pp.2576-2581. Amsterdam.

67. Frigo, Lenore y Janet McDonald. 1998. Properties of phonological markers that affect the acquisition of gender-like subclasses. En *Journal of Memory & Language* (39):218-245.

68. Galicia Haro, Sofía y Alexander Gelbukh. 2007. *Investigaciones en Análisis Sintáctico para el Español*. México. Instituto Politécnico Nacional.

69. Gambino, Omar J. y Hiram Calvo. 2007. On the usage of morphological tags for grammar induction. En Alexander Gelbukh y Ángel Kuri Morales (eds.). *MICAI 2007, LNAI 4827*, pp. 912-921. Berlín. Springer-Verlag.

70. Gao, Jianfeng y Mark Johnson. 2008. A comparison of bayesian estimators for unsupervised hidden markov model pos taggers. En *Proceedings of EMNLP 2008*, pp.344-352. Morristown.

71. Gareth, James, Daniela Witten, Trevor Hastie, Robert Tobshirani. 2013. *An introduction to statistical learning with applications in R*. Nueva York. Springer.

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

73. Goldwater, Sharon y Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. En *Proceedings of ACL 2007*, pp.744-751. Praga.

74. Gómez, Rebbeca. 2002. Variability and detection of invariant structure. En *Psychological Science* 13(5):431-436.

75. Gómez, Rebbeca y Jessica Maye. 2005. The developmental trajectory of nonadjacent dependency learning. En *Infancy* 7(2):183-206.

76. Goodluck, Helen. 1991. *Language acquisition: a linguistic introduction*. Oxford. Blackwell.

77. Goodwin, Anthony. 2013. *Nonadjacent dependency learning in typical develoment and autism spectrum disorders*. Tesis de doctorado. University of Connecticut.

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

79. Grishman, Ralph. 1986. *Computational linguistics: an introduction*. Cambridge. Reino Unido. Cambridge University Press.

80. Haghighi, Aria y Dan Klein. 2006. Prototype-driven learning for sequence models. En *Proceedings of HLT/NAACL 2006*, pp.320-327. Nueva York.

81. Harris, Zellig. 1954. Distributional structure. En *Word IO*:140-162.

82. Hauser, Marc, Noam Chomsky y Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? En *Science* (198):1569-1579.

83. Headen, William, David McClosky y Eugene Charniak. 2008. Evaluating unsupervised part-of-speech tagging for grammar induction. En Proceedings of *COLING 2008*, pp.329-336. Stroudsburg.

84. Hochmann, Jean-Rémy, Ansgar Endress y Jacques Mehler. 2010. Word frequency as a cue for identifying function words in infancy. En *Cognition* 115(3):444-457.

85. Horning, James. 1969. *A study of grammatical inference*. Tesis de doctorado. Stanford University.

86. Infante-López, Gabriel y Maarten de Rijke. 2006. A note on the expressive power of Probabilistic Context Free Grammars. En *Journal of Logic, Language and Information* 15(3):219-231.

87. Jackendoff, Ray. 1977. *X'-Syntax: a study of phrase structure*. Cambridge, Massachusetts. MIT Press.

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

89. Johnson, Mark. 2008. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. En *Proceedings of the 46th Annual Meeting of the ACL*, pp.398-406.

90. Jusczyk, Peter y Richard Aslin. 1995. Infants' detection of sound patterns in fluent speech. En *Cognitive Psychology* (29):1-23.

91. Jusczyk, Peter. 1997. *The discovery of spoken language*. Cambridge, Massachusetts. MIT Press.

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

93. Kedar, Yarden, Marianella Casasola y Barbara Lust. 2006. Getting there faster: 18-and 24-month-old Infants' use of function words to determine reference. En *Child Development* (77): 325-338.

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

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

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

97. Kupiec, Julien. 1992. Robust part-of-speech tagging using a hidden Markov model. En *Computer Speech & Language* (6):225-242.

98. Lafferty, John y Robert Mercer. 1993. Automatic classification using features of spelling. En *Proceedings of the 9th annual conference of the University of Waterloo Centre for the new OED and Text Research*, pp.89-103. Oxford.

99. Lakoff, George. 1987. *Women, fire and dangerous things*. Chicago. University of Chicago Press.

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

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

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

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

104. Levy, Yonata. 1983. It's frogs all the way down. En *Cognition* (15):75-93.

105. Li, Wentian. 1989. Mutual information functions of natural language texts. Reporte técnico. Santa Fe, NM. Santa Fe Institute.

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

107. Lorenzo, Guillermo y Víctor Longa. 1996. *Introducción a la sintaxis generativa*. Madrid. Alianza.

108. Manning, Christopher. y Nick Chater. 2006. Probabilistic models of language processing and acquisition. En *Trends in Cognitive Sciences* 10(7):335-344.

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

110. Maratsos, Michael y Anne Chalkley. 1981. The internal language of children's syntax: the ontogenesis and representation of syntactic categories. En Keith Nelson (ed.). *Children's language* (Vol. 2), pp.127-214. Nueva York. Gardner Press.

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

112. Marx, Zvika, Igo Danan y Eli Shamir. 2002. Cross-component clustering for template induction. En *Workshop on text learning (TextML-2002)*, pp.66-75.

113. Mc. Murray, Bob y George Hollich. 2009. Core computational principles of language acquisition: can statistical learning do the job?. En *Developmental Science* 12(3):365-368.

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

115. Meilă, Marina. 2003. Comparing clusterings by the variation of information. En *Proceedings of the 6th annual conference on computational learning theory and 7th kernel workshop*, pp.173-187. Washington.

116. Meyer, Charles. 2002. *English Corpus Linguistics*. Cambridge, Reino Unido. Cambridge University Press.

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

118. Mintz, Toben. 2002. Category induction from distributional cues in an artificial language. En *Memory & Cognition* 30(5):678-686.

119. Mintz, Toben, Elissa Newport y Thomas Bever. 2002. The distributional structure of grammatical categories in speech to young children. En *Cognitive Science 26*(4):393-424.

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

121. Mintz, Toben. 2006. Finding the verbs: distributional cues to categories available to young learners. En Hirsh-Pasek, Kathy y Roberta Michnick Golinkoff (eds.). *Action meets word: How children learn verbs*. Nueva York. Oxford University Press.

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

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

124. Moreno Sandoval, Antonio. 2001. *Gramáticas de unificación y rasgos*. Madrid. Antonio Machado.

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

126. Ney, Hermann, Ute Essen y Reinhard Kneser. 1995. On the estimation of small probabilities by leaving one out. En *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17(12):1202-1212.

127. Norvig, Peter. 2011. On Chomsky and the two cultures of statistical learning. http://norvig.com/chomsky.html (sitio web accedido el 02/12/2013).

128. Nowak, Martin, Komarova, Natalia y Partha Niyogi . 2001. Evolution of universal grammar. En *Science* (291):114-118.

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

130. Paskin, Mark. 2002. Grammatical bigrams. En Dietterich, Thomas, Sue Becker y Zoubin Ghahramani (eds.). *Advances in neural information processing systems 14*. Cambridge, Massachusetts. MIT Press.

131. Peters, P. Stanley y Robert Ritchie. 1973. On the Generative Power of Transformational Grammars. En *Information Sciences* (6):49-83.

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

133. Pine, Julian y Helen Martindale. 1996. Syntactic categories in the speech of young children: The case of the determiner. En *Journal of Child Language* (23):369-395.

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

135. Pinker, Steven. 1984. *Language learnability and language development*. Cambridge, Massachusetts. Harvard University Press.

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

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

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

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

140. Quine, Willard.1970. *Philosophy of Logic*. Englewood Cliffs. Prentice-Hall

141. 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*, pp.970-975. Mahwah, New Jersey. Erlbaum.

142. 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.

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

144. Roark, Brian y Richard Sproat. 2007. *Computational approaches to morphology and syntax*. Oxford University Press. Oxford.

145. Rosenberg, Andrew y Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. En *Proceedings of EMNLP-CoNLL 2007*, pp.410-420. Praga.

146. Russell, Stuart y Peter Norvig. 1995. Inteligencia Artifical: un enfoque moderno. Englewood Cliffs. Prentice Hall.

147. Saffran, Jenny, Richard Aslin y Elissa Newport. 1996. Statistical learning by 8-months infants. En *Science* (274):1926-1928.

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

149. Santorini, Beatrice. 1991. Part-of-speech tagging guidelines for the Penn treebank project. Reporte técnico. Philadelphia. University of Pennsylvania.

150. Shieber, Stuart. 1985. Evidence against the context-freeness of natural language. En *Linguistics and Philosophy* (8):333-343.

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

152. Schütze, Hinrich. 1995. Distributional Part-of-Speech tagging. En *Proceedings of the 7th conference of the EACL*, pp.141-148. Dublín.

153. Shannon, Claude. 1948. A mathematical theory of communication. En *Bell System Technical Journal* (27):379-423.

154. Shi, Rushen. 1995. *Perceptual correlates of content words and function words in early language input.* Tesis de doctorado. Brown University.

155. Shi, Rushen, James Morgan y Paul Allopenna. 1998. Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. En *Journal of Child Language* 25(1):169-201.

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

157. Shi, Rushen y Andréane Melançon. 2010. Syntactic Categorization in French-Learning Infants. En *Infancy* 15(5):517-533.

158. Smith, Noah y Jason Eisner. 2004. Annealing techniques for unsupervised statistical language learning. En *Proceedings of ACL 2004*, pp.487-494. Barcelona.

159. Smith, Kirk. 1966. Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? En *Journal of Experimental Psychology* (72):580-588.

160. Solan, Zach, David Horn, Eytan Ruppin y Shimon Edelman. 2005. Unsupervised learning of natural languages. En *Proceedings of the National Academy of Sciences of the United States PNAS* 102(33):11629-11634.

161. Sparck Jones, Karen y Julia Galliers. 1996. *Evaluating natural language processing systems*. Berlín. Springer-Verlag.

162. Spitkovsky, Valentin, Hiyan Alshawi, y Dan Jurafsky. 2009. Baby Steps: How 'Less is More' in unsupervised dependency parsing. En *NIPS: Grammar Induction, Representation of Language and Language Learning*. Whistler.

163. Suppes, Patrick. 1973. Semantics of natural languages. En Jaakko Hintikka, Julius Moravcsik y Patrick Suppes (eds.). *Approaches to natural language*, pp.370-394. Stanford. Stanford University Press.

164. Tager-Flusberg, Helen. 1997. Language acquisition and theory of mind: contributions from the study of autism. En Lauren Adamson y Mary Ann. Romski (eds.). *Research on communication and language disorders: Contributions to theories of language development*. Baltimore. Paul Brookes Publishing.

165. Tomasello, Michael. 2000a. Do young children have adult syntactic competence? En *Cognition* 74(3):209-253.

166. Tomasello, Michael. 2000b. The item-based nature of children's early syntactic development. En *Trends in Cognitive Sciences* 4(4):156-163.

167. Valian, Virginia. 1986. Syntactic categories in the speech of young children. En *Developmental Psychology* (22):562-579.

168. Valian, Virginia, Stephanie Solt y John Stewart. 2009. Abstract categories or limited-scope formulae? The case of children's determiners. En *Journal of Child Language* 36(4):743-778.

169. Van Zaanen, Menno. 2000. ABL: Alignment-based learning. En *Proceedings of the 18th*

international conference on Computational Linguistics COLING, pp.961-967. Montreal.

170. Vlachos, Andreas, Anna Korhonen y Zoubin Ghahramani. 2009. Unsupervised and constraint dirichlet process mixture models for verb clustering. En *Proceedings of GEMS 2009*, pp.74-82. Morristown.

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

172. Watumull, Jeffrey, Marc Hauser, Ian Roberts y Norbert Horstein. 2014. On recursion. En *Frontiers in Pschology* (Vol.4) Art. 1017.

173. Wexler, Kenneth. 1999. *The MIT encyclopedia of the cognitive sciences*. Cambrigde, Massachusetts. MIT Press.

174. Wolff, J. Gerard. 1988. Learning syntax and meanings through optimization and distributional analysis. En Yonata Levy, Izchak Schlesinger y Martin Braine (eds.). *Categories and Processes in Language Acquisition*. Hillsdale. Erlbaum.

175. Yang, Charles. 2004. Universal Grammar, statistics or both? En *Trends in Cognitive Sciences* (10):451-456.

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

177. Zipf, George. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, Massachusetts. Addison-Wesley.

Listado de abreviaturas y siglas

Sigla	***Descripción***
ABL	**Alignment-Based Learning***: tipo de algoritmo*
ACUAH	**Análisis de la Conversación de la Universidad de Alcalá de Henares***: corpus de referencia*
ADIOS	**Automatic Distillation Of Structure:** *tipo de algoritmo*
ALFAL	**Corpus de Asociación de Lingüistica y Filología de América Latina**: *corpus de referencia*
APS	**Argument from Poverty of Stimulus**: *concepto lingüístico*
BNC	**British National Corpus**: *corpus de referencia*
CDC	**Context Distribution Clustering***: tipo de algoritmo*
CDD	**Cumulative Degree Distribution**: *concepto de teoría de redes*
CEAP	**Corpus de Encuestas de Asunción de Paraguay**: *corpus de referencia*
CFG	**Context-Free Grammar**: *tipo de gramática formal*
CHILDES	**Child Language Data Exchange System**: *corpus de referencia*
COVJA	**Corpus Oral de la Variedad Juvenil en Alicante**: *corpus de referencia*
CRATER	**Corpus Resources for Terminology Extraction**: *corpus de referencia*
CREA	**Corpus de Referencia del Español Actual**: *corpus de referencia*
CSC	**Corpus de Santiago de Compostela**: *corpus de referencia*
CSMV	**Corpus Sociolingüístico de Mérida, Venezuela**: *corpus de referencia*
C4	**CLAWS-4**: *nomenclatura*
DFP	**Decreasing Frequency Profile**: *técnica estadística*
DLG	**Distributional Lattice Grammars**: *tipo de algoritmo*
DMV	**Dependency Model with Valence**: *tipo de algoritmo*
GU	**Gramática Universal**: *concepto lingüístico*
HMM	**Hidden Markovian Model**: *tipo de gramática formal*
IIL	**Identification In the Limit**: *concepto de teoría de lenguajes formales*
IULA CT	**Corpus Técnico del IULA** (Instituto Universitario de Lingüístca Aplicada): *corpus de referencia*
KLD	**Kullback-Leibler Divergence**: *métrica*
MCSG	**Mildly Context-Sensitive Grammar**: *tipo de gramática formal*
MDL	**Maximum Description Length**: *técnica estadística*
MI	**Mutual Information**: *métrica*
MLU	**Mean Length Utterance**: *métrica*
MLUm	**Mean Length Utterance in morphemes**: *métrica*
NLP	**Natural Language Processing**
NP	**Noun Phrase**: *concepto lingüístico*
PCA	**Principal Component Analysis**: *técnica algebraica*
PCFG	**Probabilistic Context-Free Grammar**: *tipo de gramática formal*
PHMM	**Probabilistic Hidden Markovian Model**: *tipo de gramática formal*
PLD	**Primary Linguistic Data**: *concepto lingüístico*
POS-tag	**Part-Of-Speech tag**: *concepto lingüístico*
SCFG	**Stochastic Context Free Grammar**: *tipo de gramática formal*
SRN	**Simple Recurrent Network**: *concepto de redes neuronales*
SVD	**Single Value Decomposition**: *técnica algebraica*
SVO	**Sujeto-Verbo-Objeto**: *concepto lingüístico*
TF-IDF	**Term Frequency – Inverse Document Frequency**: *métrica*
TP FP TN FN	**True Positives, False Positives, True Negatives, False Negatives**: *métricas*
VI	**Variation of Information**: *métrica*
VP	**Verb Phrase**: *concepto lingüístico*
WSJ	**Wall Street Journal**: *corpus de referencia*

Índice alfabético de conceptos

A

B

C

Ch

C

D

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

V

Anexo I Clustering de secuencias candidatas a constituyentes (capítulo 8)

0	1	2	3	4	5	6	7	8
PRP DT1	AV0 AT1NN1	NNP NNP	PRP AT1NN1PRP AT1NN1	XX0	NN2 PRP AT1	REL VVZ	NN1 AT1NN1	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	NN1 AV0	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	NN1 PRP AT1NN1AJ1	VVZ NN2
AV0 AT1	CJT		PRP AT1NN1PRP NNP		DT1		NN1 AJ1	VVZ PRP VVI
PRP AT1			PRP CRD NN2		DPS		NN1 PRP AT1NN1	PPE VVZ
PRP AT1NN1PRP AT1			CJC NN2		NNP PRP AT1		NN1 PRP CRD NN2	VVZ AT1NN1
PRP AT1AJ1			PRP AT1NN1		NNP AT1		DAT	VVZ VVN
			CJC AT1NN1		AT2 NN2 PRP AT1		NN1 PRP NNP	
			AV0 PRP AT1NN1		VVI AT1		NN1 PRP NN1	
			PRP AT1NNP				NN1 NNP	
			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	NN1 PRP AT2 NN2	NN1 PRP CRD
AT1NN1 PRP	VBZ VVN PRP	AT2		NN2 REL	AJ2 NN2	CJC AT1	NN1 PRP NN2	NN1 REL VVZ
AT2 NN2 PRP	VVZ PRP				NN2 PRP AT1NN1	VVZ AT1NN1 PRP AT1	NN1 VVN	NN1 VBZ
NNP PRP	AT1NN1VVZ PRP				NN2	AJ1 PRP AT1		NN1 CJC
CRD NN2 PRP	AJ2 PRP				NN2 PRP NN2	VVN PRP AT1		NN1 PRP AT2
NN2 PRP AT1NN1PRP	VVZ PRP AT1NN1PRP				NN2 PRP NN1	VVZ PRP AT1		NN1 VVZ
NN2 AJ2 PRP	NNP VVZ PRP				NN2 AV0			
AT1NN1PRP NN1PRP	PRP NN1PRP				NN2 VVZ			
VVI PRP	PRP NNP PRP							
AT1NN1AJ1 PRP	CRD PRP							
AT1NN1PRP AT1NN1PRP	PRP AT2 NN2 PRP							
	PRP NN2 PRP							
	PRP AT1NN1 PRP							
	VVN PRP							
	PRP VVI PRP							
	PRP							
	CJC PRP							
	SEP VVZ PRP							

AV0 PRP
PRP AT1NN1AJ1PRP
VVZ AT1NN1PRP
PPE VVZ PRP
AJ1PRP

Anexo II Muestra de salida final del experimento con constituyentes: filtrado por MI (capítulo 8)

	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	$$$--AJ1	0.032
error	AV0 PRP	2	9.209	VVZ--DPS	0.081
	AV0 PRP AT1 NN1	4	6.107	PRP--AJ1	0.032
	AV0 VVZ	2	7.798	PRP--DPS	0.049
error	CJC	1	9.969	NN1--PNI	0.270
error	CJC AT1 NN1	3	8.002	NN2--ORD	0.036
	CJC VVZ	2	8.537	NN1--DPS	0.046
error	CJT AT1 NN1	3	6.951	VVZ--PRP	0.022
	CJT VVZ	2	7.798	PRP--DPS	0.045
	CRD NN2	2	8.979	PRP--CJC	0.061
	NN1	1	12.421	AT1--REL	0.344
	NN1 AJ1	2	11.061	AT1--CJC	0.071
error	NN1 AJ1 PRP AT1	4	8.852	AT1--NN1	0.023
error	NN1 PRP	2	11.943	AT1--NN0	0.129
error	NN1 PRP AT1	3	11.137	AT1--ORD	0.049
	NN1 PRP AT1 NN1	4	10.639	AT1--AJ1	0.082
	NN1 PRP AT1 NN1 AJ1	5	6.965	AT1--$$$	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 III Muestra de constituyentes inducidos sobre algunas oraciones de prueba (capítulo 8)