# Aproximación de la teoría de la inferencia gramatical a los estudios de adquisición del lenguaje

Leonor Becerra Bonache[1]

Yale University, Department of Computer Science & Universitat Rovira i Virgili, Grupo de Investigación en Lingüística Matemática.
51 Prospect Street, New Haven, CT, USA & Pl. Imperial Tarraco, 1, Tarragona, España
leonor.becerra-bonache@yale.edu

**Resumen**

En este artículo proponemos utilizar la teoría de la inferencia gramatical para entender el proceso de adquisición del lenguaje natural. En busca de este objetivo, intentamos responder a dos preguntas: ¿en qué clases de lenguajes formales deberían centrarse los estudios de inferencia gramatical?, ¿qué tipo de información deberíamos tener en cuenta durante el proceso de aprendizaje?. Proponemos también nuevas direcciones en el campo de la inferencia gramatical motivadas por estudios de adquisición del lenguaje natural.

**Palabras clave:** Inferencia gramatical, adquisición del lenguaje natural, suavemente dependiente del contexto, preguntas de corrección

**Abstract**

In this paper, we propose to use the theory of the grammatical inference in order to understand the process of natural language acquisition. In pursuit of this goal, we try to answer two questions: Which classes of formal languages should grammatical inference studies be focused on? What kind of information should be taken into account during the learning process? We also propose new research directions in the field of grammatical inference motivated by studies of natural language acquisition.

**Keywords:** Grammatical Inference, natural language acquisition, mildly context-sensitive, correction queries

**Resum**

En aquest article es proposa la utilització de la teoria d'inferència gramatical per entendre el procés d'adquisició del llenguatge natural. Intentem respondre a dues preguntes: Quina classe de llenguatges formals haurien de considerar els estudis d'inferència gramatical?; Quin tipus d'informació hauríem de tenir en compte durant el procés d'aprenentatge? Es proposen, així mateix, noves direccions en l'àmbit de la inferència gramatical motivades pels estudis d'adquisició del llenguatge natural.

**Paraules clau:** Inferència gramatical, adquisició del llenguatge natural, suaument dependent del context, preguntes de correcció.

**Tabla de contenidos**



[1] Este trabajo se ha podido llevar a cabo gracias a una *Marie Curie International Fellowship* (OIF) dentro del *6th European Community Framework Programme*.

## 1. Introducción

¿Cómo adquieren los niños el lenguaje? Independientemente de la cultura y la lengua a la que estén expuestos, los niños son capaces de adquirir el lenguaje de manera máximamente eficiente y con una tasa de éxito que no alcanzan ni de lejos en otras tareas cognitivas de mucho menor complejidad. No obstante, a pesar de la facilidad con la que los niños adquieren el lenguaje, existe gran dificultad para poder explicar cómo se lleva a cabo esta tarea. Por lo tanto, dar respuesta a esta pregunta supone un gran reto para investigadores de distintas áreas, incluyendo la lingüística, la ciencia cognitiva y la informática.

A pesar de las investigaciones realizadas hasta el momento, no se ha logrado entender las reglas, estrategias y otros procesos que subyacen a esta capacidad humana. El deseo de poder entender mejor el proceso de adquisición del lenguaje natural ha motivado la investigación de *modelos formales de adquisición del lenguaje*. El interés por estudiar estos modelos radica no sólo en poder ahondar así en la comprensión de cómo los niños adquieren su primera lengua, sino también en las numerosas aplicaciones prácticas que pueden tener estos modelos (Parekh y Honavar 2002).

El campo de investigación conocido como Aprendizaje Automático tiene como objetivo estudiar y modelar informáticamente procesos de aprendizaje. Dentro del Aprendizaje Automático, encontramos un subcampo especializado en el aprendizaje de lenguajes formales, conocido como *Inferencia Gramatical* (IG). A grandes rasgos, un problema de IG podría definirse como un juego en el que hay dos participantes: un profesor y un aprendiz. El profesor proporciona datos al aprendiz, y a partir de ellos, el aprendiz tiene que ser capaz de identificar el lenguaje subyacente (Clark 2004).

La IG como campo teórico de estudio empezó con el intento por parte de E.M. Gold (Gold 1967) de formalizar la adquisición del lenguaje natural. Tras este trabajo, se han realizado gran cantidad de investigaciones centradas especialmente en la obtención de resultados formales (por ejemplo, caracterización formal de los lenguajes que se intentan aprender, demostraciones formales sobre si ciertos algoritmos pueden aprender de acuerdo con definiciones concretas, etc.). No obstante, además de esa vertiente teórica, los algoritmos de IG también se han intentado aplicar, con cierto éxito, a problemas prácticos (por ejemplo, procesamiento del lenguaje natural, biología computacional). Para una introducción general a la IG el lector puede consultar (Sakakibara 1997, Yokomori 2004, de la Higuera 2004).

Lo que proponemos aquí es aproximar la teoría de la IG a los estudios de adquisición del lenguaje (áreas relacionadas pero que responden a tradiciones científicas diferentes), en pos de los siguientes objetivos: por una parte, mejorar técnicas y modelos en IG utilizando ideas procedentes de estudios sobre adquisición del lenguaje; por otra parte, proporcionar un modelo para explicar la adquisición del lenguaje natural utilizando la teoría de la IG. En este artículo nos vamos a centrar especialmente en el primer punto.

Proporcionar un modelo computacional de adquisición del lenguaje supone uno de los retos más interesantes para la IG, pero para poder alcanzar tal propósito se deben tener en cuenta los siguientes aspectos. En primer lugar, la clase de lenguajes que deberíamos aprender debería tener al menos el mismo poder expresivo que los lenguajes naturales. En segundo lugar, los datos proporcionados a nuestro algoritmo deberían ser del mismo tipo

que los que dispone el niño durante el proceso de aprendizaje. Dando un repaso a las investigaciones realizadas en el campo de IG, nos damos cuenta de que la mayoría de estudios no sólo se basan en clases de lenguajes con limitado poder expresivo, además, los algoritmos propuestos se basan en la disponibilidad de distintos tipos de información que son discutibles desde el punto de vista de la adquisición del lenguaje. Así pues, en este artículo intentaremos responder a las siguientes preguntas: ¿qué clase de lenguajes deberíamos aprender?, ¿qué información deberíamos proporcionar a nuestro algoritmo? Además, propondremos nuevas direcciones en el campo de la IG motivadas por estudios de adquisición del lenguaje natural.

El artículo queda organizado de la siguiente manera. En el apartado 2, hacemos un repaso de las clases de lenguajes en las que se han centrado los estudios de IG, y discutimos su adecuación para el estudio de la adquisición del lenguaje. En el subapartado 2.1 proponemos una clase de lenguajes relevante desde un punto de vista lingüístico. Posteriormente, en el apartado 3, repasamos distintas consideraciones sobre el tipo de datos que podrían estar disponibles para el niño. En el subapartado 3.1 proponemos por una parte, un nuevo tipo de datos que podría tenerse en cuenta en los estudios de IG y, por otra parte, un modelo alternativo que intenta reflejar de manera más fiel el proceso de adquisición del lenguaje. Finalmente, en el apartado 4, presentamos conclusiones y trabajo futuro.

## 2. ¿Qué clase de lenguajes deberíamos aprender?

El campo de la IG ha centrado su investigación en el aprendizaje de lenguajes regulares (REG) y lenguajes independientes del contexto (CF). Encontramos muy pocos estudios que intenten identificar clases de lenguajes con más poder expresivo que CF utilizando técnicas de IG. Un repaso general sobre los resultados más importantes en este ámbito se puede encontrar en (Sakakibara 1997, de la Higuera 2004).

REG y CF constituyen los dos primeros niveles en la jerarquía de Chomsky[2]. Cabe destacar que N. Chomsky introdujo sus gramáticas formales como herramientas para formalizar la sintaxis del lenguaje natural. No obstante, la jerarquía de Chomsky tiene ciertas limitaciones que deben tenerse en cuenta en el estudio del lenguaje natural. Una de sus principales limitaciones aparece cuando intentamos situar los lenguajes naturales en esta jerarquía.

Poco después de la publicación de (Chomsky 1956), empezó un largo debate sobre el lugar que ocupan los lenguajes naturales dentro de la jerarquía de Chomsky. Este debate se centró en intentar determinar si los lenguajes naturales son CF o no. Durante la década de los 60 y los 70, encontramos tanto argumentos de apoyo como de rechazo de esta idea. Sin embargo, a finales de los 80, se descubrieron en varias lenguas ejemplos convincentes de estructuras no-CF, tales como la estructura de concordancia múltiple, concordancia cruzada y duplicación (el lector interesado en estos ejemplos puede consultar (Bresnan et al. 1987, Culy 1987, Shieber 1987)). Gracias a estos hallazgos, los lingüistas logran ponerse de acuerdo y aceptan que los lenguajes no son CF.

[2] La jerarquía de Chomsky consta de 4 niveles (de menos poder generativo a más): REG (regulares), CF (independientes del contexto), CS (dependientes del contexto) y RE (recursivamente enumerable). Para más información consultar (Rozenberg y Salomaa 1997).

Ahora, la cuestión es poder determinar cuanto poder más allá de CF es necesario para describir estas construcciones no-CF que aparecen en el lenguaje natural. Debemos tener en cuenta que: por una parte, la familia CF no es suficientemente expresiva para describir fragmentos importantes del lenguaje natural, pero tiene buenas propiedades computacionales; por otra parte, la familia CS contiene todas las construcciones importantes que aparecen en el lenguaje natural, pero computacionalmente es muy compleja. Por lo tanto, lo ideal sería encontrar algún mecanismo capaz de generar construcciones CF y no-CF, pero teniendo bajo control el poder generativo. Esta idea ha dado lugar a la noción de mecanismos *suavemente dependientes del contexto* (MCS), introducida originalmente por A.K. Joshi (Joshi 1985).

En este artículo, entendemos que una familia de lenguajes es MCS si reúne las siguientes condiciones (para más detalles ver (Kudlek et al. 2002)):

- cada lenguaje en esa familia es semilineal
- para cada lenguaje en esa familia el problema de pertinencia se puede resolver en tiempo polinómico
- esa familia contiene los siguientes lenguajes no-CF:
    - concordancias múltiples: $L_1 = \{a^n b^n c^n \mid n \geq 0\}$
    - concordancias cruzadas: $L_2 = \{a^n b^m c^n d^m \mid n, m \geq 0\}$
    - duplicación: $L_3 = \{w\, w \mid w \in \{a, b\}^*\}$

En la literatura sobre el tema podemos encontrar otras variantes de esta definición. Por ejemplo, en ocasiones se asume que esta familia contiene todos los lenguajes CF (por ejemplo, ver (Joshi 1987, Roach 1987, Steedman 1985)); esto significa que la familia MCS ocuparía una posición concéntrica en la jerarquía de Chomsky, entre CF y CS. Sin embargo, teniendo en cuenta la motivación lingüística de MCS, se nos plantea la siguiente pregunta: ¿es necesario generar todos los lenguajes CF? Tal y como otros autores han señalado (Manaster-Ramer 1999, Kudlek et al. 2002), los lenguajes naturales podrían ocupar una posición ortogonal en la jerarquía de Chomsky, es decir, sólo contendrían algunos lenguajes regulares, algunos CF, pero estarían incluidos en CS (la Figura 1 muestra una familia que ocupa esta posición ortogonal); de hecho, algunas estructuras REG y CF no aparecen de manera natural en frases, y podemos encontrar ejemplos de construcciones en el lenguaje natural que no son ni REG ni CF.

Por lo tanto, teniendo en cuenta todas estas ideas, consideramos que sería de gran interés investigar mecanismos que generen lenguajes MCS (cumpliendo las 3 condiciones especificadas anteriormente), pero que a su vez ocupen una posición ortogonal en la jerarquía de Chomsky.

### 2.1 Simple p-dimensional External Contextual

Un mecanismo que reúne tales características es la clase de gramáticas *Simple p-dimensional External Contextual* ($SEC_p$).[3] A diferencia de las gramáticas de Chomsky, este mecanismo no utiliza símbolos no terminales y no tiene reglas de derivación excepto

[3] Estas gramáticas tienen sus raíces en las llamadas gramáticas *contextuales*, que fueron introducidas con motivaciones lingüísticas por S. Marcus (Marcus 1969). Las gramáticas contextuales se basan en la idea de modelar algunos aspectos naturales, como por ejemplo, la aceptación de una palabra (construcción) sólo en ciertos contextos.

una regla general: añadir contextos. Además, a diferencia de otras gramáticas contextuales, estas gramáticas añaden los contextos en los extremos finales de la palabra[4] (de ahí el nombre de *external*), funcionan con vectores *p*-dimensionales de palabras y vectores *p*-dimensionales de contextos (de ahí el término de *p-dimensional*) y su base se reduce a una única palabra (de ahí que se llame *simple*). A continuación pasamos a detallar la definición formal de $SEC_p$.

Sea $p \geq 1$ un entero fijado y sea $\Sigma$ un alfabeto. Una p-palabra *x* sobre $\Sigma$ es un vector p-dimensional cuyos componentes son palabras sobre $\Sigma$ (por ejemplo, $x = (x_1, x_2, ..., x_p)$, donde $x \in \Sigma^*$, $1 \leq i \leq p$). Un p-contexto *c* sobre $\Sigma$ es un vector p-dimensional cuyos componentes son contextos sobre $\Sigma$ (por ejemplo, $c = [c_1, c_2, ..., c_p]$, donde $c_i = (u_i, v_i)$, $u_i, v_i \in \Sigma^*$, $1 \leq i \leq p$).

Sea $p \geq 1$ un entero. Una gramática *Simple p-dimensional External Contextual* se define como $G = (\Sigma, B, C)$, donde $\Sigma$ es el alfabeto de G, B es una p-palabra sobre $\Sigma$ llamada la *base* de G, y C es un conjunto finito de p-contextos sobre $\Sigma$. El conjunto de contextos de G es llamado C.

Sean $x = (x_1, x_2, ..., x_p)$ y $z = (z_1, z_2, ..., z_p)$ dos p-palabras sobre $\Sigma$. Por definición, $x \Rightarrow_G z$ sii $z = (u_1x_1v_1, u_2x_2v_2, ..., u_px_pv_p)$ para algún p-contexto $c = [(u_1v_1), (u_2v_2), ..., (u_pv_p)] \in C$.

El lenguaje generado por G, denotado como L(G), se define como: $L(G) = \{z \in \Sigma \mid$ existe $(x_1, x_2, ..., x_p) \in B$ tal que $(x_1, x_2, ..., x_p) \Rightarrow_G^* (z_1, z_2, ..., z_p)$ y $z = z_1z_2 ... z_p\}$

Veamos a continuación un ejemplo concreto de $SEC_p$. Imaginemos que tenemos la siguiente gramática $SEC_2$: $\Sigma = \{a, b, c\}$, $B = \{(\lambda, \lambda)\}$, $C = [(a, b), (c, \lambda)]$.[5] Si aplicamos una vez el contexto a la base, obtendremos: (aλb, cλλ). La concatenación de los elementos internos da lugar a la cadena *abc*. Si aplicamos dos veces el mismo contexto, ahora obtendremos (aaλbb, ccλλλ), cuya cadena correspondiente es *aabbcc*. Y así sucesivamente. Por lo tanto, mediante esta gramática tan simple somos capaces de generar el siguiente lenguaje no-CF: $\{a^nb^nc^n \mid n \geq 0\}$.

Tal y como se demuestra en (Becerra-Bonache 2006), $SEC_P$ es un mecanismo para generar familias de lenguajes MCS. Esto implica que $SEC_P$ tiene varias propiedades relevantes desde un punto de vista lingüístico; además de que los lenguajes que puede generar son semilineales y se pueden analizar en tiempo polinómico, $SEC_P$ es capaz de generar las tres básicas construcciones no-CF que aparecen en el lenguaje natural, es decir, concordancia múltiple, concordancia cruzada y duplicación.

Además, $SEC_P$ tiene otra propiedad relevante: es incomparable con las familias REG y CF, pero está incluida en CS (por lo tanto, ocupa una posición ortogonal en la jerarquía de

[4] Utilizaremos el término *palabra* (en ocasiones, *cadena*) como sinónimo del término inglés *word/string*.
[5] λ es utilizado para designar la cadena vacía.

Chomsky, tal y como parece que pueden tener los lenguajes naturales). Por lo tanto, $SEC_P$ podría ser un candidato apropiado para modelar la sintaxis del lenguaje natural.[6]

La siguiente figura muestra la localización de $SEC_P$ en la jerarquía de Chomsky.

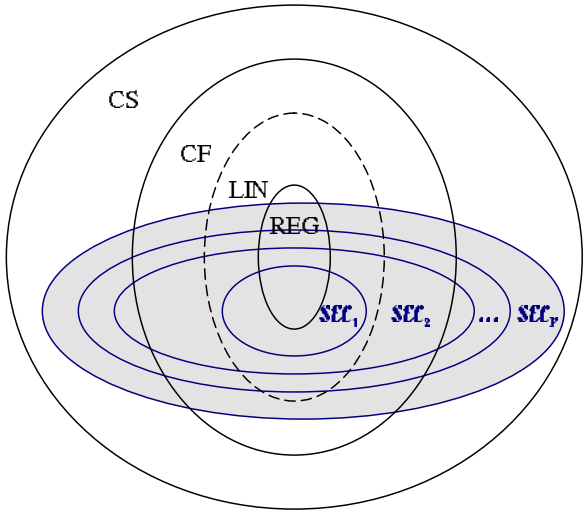


Figura 1: $SEC_p$ ocupa una posición ortogonal en la jerarquía de Chomsky

Teniendo en cuenta que la investigación en IG se ha centrado en mecanismos que tienen una capacidad expresiva limitada para describir algunos aspectos de la sintaxis del lenguaje natural, proponemos que los estudios de IG se centren en clases de lenguajes más relevantes desde un punto de vista lingüístico, como por ejemplo, $SEC_P$.

Resultados iniciales en cuanto a aprendizaje de $SEC_P$ se pueden encontrar en (Becerra-Bonache y Yokomori 2004, Oates et al. 2006).

## 3. ¿Qué información deberíamos proporcionar a nuestro algoritmo?

Para simular correctamente el aprendizaje del lenguaje natural, los ejemplos que deberíamos proporcionar a nuestro algoritmo deberían ser del mismo tipo de los que puede disponer un niño. Pero ¿de qué tipo de datos dispone el niño?

Existe una tendencia general a reducir los datos disponibles para el niño a dos tipos: positivos o negativos. Por datos positivos se entiende construcciones que son gramaticalmente correctas y todo el resto es considerado como datos negativos. Teniendo en cuenta que los niños están expuestos a un gran número de frases gramaticalmente correctas producidas por los adultos, la disponibilidad de datos positivos es trivialmente

[6] No estamos afirmando que sea el mejor modelo para la sintaxis del lenguaje natural o que podamos describir todas las construcciones de los lenguajes naturales mediante gramáticas SEC. Lo único que intentamos remarcar aquí es la relevancia lingüística de esta clase y el interés que puede tener su estudio.

aceptada. No obstante, la disponibilidad de datos negativos sigue siendo un tema de gran controversia.

Cabe destacar que esta distinción entre datos positivos y negativos fue utilizada por E.M. Gold en su trabajo *Language identification in the limit* (Gold 1969). Dentro del marco de los lenguajes formales, está distinción queda clara, ya que datos positivos hacen referencia a cadenas que pertenecen al lenguaje y los datos negativos a cadenas que no pertenecen. Sin embargo, en el marco de los lenguajes naturales, este tipo de distinción no parece tan acertada, ya que no todos los datos que un niño puede recibir se pueden clasificar claramente como positivos o negativos; en ocasiones, podemos encontrar frases que son gramaticalmente correctas pero que pueden contener a su vez información negativa (más adelante veremos un ejemplo de ello), de manera que, su clasificación se hace complicada.

Por lo tanto, en el caso del lenguaje natural, las definiciones sobre el tipo de datos disponibles para el niño deberían ser más refinadas. Especialmente, encontramos problemas con la definición de datos negativos; ésta es tan general que cada autor hace su propia interpretación (por ejemplo, ver (Marcus 1993, Saxton 1997)). Así pues, es importante definir qué son exactamente datos negativos. Además, consideraciones sobre si los niños reciben o no datos negativos pueden depender de cómo uno defina este concepto.

Por ejemplo, si consideramos que datos negativos son frases completamente incorrectas, o respuestas (a una frase incorrecta producida por el niño) del tipo "Esto está mal", "No, no deberías hablar así", o "Voy a decir algo incorrecto, así que presta atención", podemos afirmar que es raro encontrar este tipo de datos. Sin embargo, cada vez hay mayor evidencia de que los niños también disponen de input correctivo (Farrar 1992, Morgan et al. 1995). En ese caso, ¿las correcciones deberían considerarse datos positivos o datos negativos?

Un tipo de estas correcciones son las llamadas *expansiones*.[7] Durante las primeras etapas de adquisición del lenguaje,[8] tenemos constancia de que los adultos, ante frases incompletas producidas por el niño, intentan repetir la misma idea pero utilizando una frase gramaticalmente correcta (con respecto a la gramática adulta). Por ejemplo, reproducimos a continuación un diálogo entre Rafael y su madre, cuando él tenía 22 meses (Hernández-Pina 1984: 284):

Madre: ¿Quieres ir de paseo?
Rafael: calle
Madre: ¿quieres ir de paseo a la calle?
Rafael: carrito
Madre: ¿en el carrito?
Madre: corre, tráelo que está en la cocina
Rafael: carrito cocina
Madre: sí, el carrito está en la cocina. Tráelo.

[7] Cabe destacar que aquí entendemos por corrección la repetición de una frase agramatical producida por el niño, pero corregida con respecto a la gramática adulta. No entendemos por correcciones respuestas del tipo *Esto esta mal* (tal y como hemos señalado anteriormente, los adultos no suelen dirigirse así a los niños).
[8] Etapa holofrástica, etapa de las 2 palabras y etapa telegráfica (Brown 1976).

Así pues, el adulto intenta moldear las emisiones producidas por el niño y construir frases completas, ya sea de una manera consciente o inconsciente. Además, tal y como podemos ver, una expansión es una frase gramaticalmente correcta producida por el adulto, por lo tanto, podríamos decir que es un dato positivo. No obstante, una expansión también contiene información negativa, ya que al ser recibida indica que la frase producida por el niño no era gramaticalmente correcta. Por lo tanto, ¿las expansiones deberían ser consideradas como datos positivos o negativos?.

Cabe destacar que las expansiones también juegan un papel importante en la comunicación entre niño y adulto; si la construcción del niño fuera totalmente correcta, esta expansión no sería necesaria, ya que supondría repetir de nuevo lo mismo y la comunicación se vería cortada (en una conversación entre adultos no vamos repitiendo lo que el otro dice ya que eso obstaculizaría la comunicación, pero en el caso de niño-adulto, la expansión también ayuda a que haya una comunicación más fluida entre ambos, ya que el adulto se asegura de que ha entendido bien lo que quiere decir el niño, y a su vez, el niño también se asegura de que se le ha entendido perfectamente).

Por lo tanto, las expansiones constituyen un tipo de las posibles correcciones que los niños pueden recibir y, tal y como hemos visto, son difíciles de clasificar como datos positivos o negativos (contienen información positiva y negativa al mismo tiempo). Curiosamente, en la mayor parte de las investigaciones, todo tipo de correcciones se consideran datos negativos y no se tienen en cuenta para el proceso de aprendizaje.

Aunque los datos positivos son esenciales en el proceso de aprendizaje y juegan un papel principal en este proceso, consideramos que las correcciones pueden jugar un papel complementario, proporcionando información adicional que puede ser de gran ayuda durante el proceso de aprendizaje del niño. Además, la información inherente en una corrección podría también mejorar el aprendizaje, haciendo que algunos aspectos del lenguaje se aprendan con más rapidez.[9] Por lo tanto, las correcciones podrían también jugar un papel relevante en términos de eficiencia. Así que, ¿por qué no tener en cuenta las correcciones en los estudios de aprendizaje del lenguaje?

### 3.1 Correcciones en Inferencia gramatical

La mayor parte de investigaciones en IG se han centrado especialmente en el aprendizaje de lenguajes formales a partir sólo de datos positivos (cadenas que pertenecen al lenguaje). En los estudios en los que se tienen en cuenta también los datos negativos, éstos hacen referencia únicamente a cadenas que no pertenecen al lenguaje (por lo tanto, bastante alejado de la idea de *expansión*). Teniendo en cuenta todo lo expuesto, proponemos aplicar la idea de correcciones a los estudios de IG.

Por una parte, consideramos que los modelos de IG pueden beneficiarse de las correcciones, por ejemplo, el modelo de *aprendizaje a partir de preguntas* introducido por D. Angluin (Angluin 1987). En este modelo, tenemos un profesor que conoce el lenguaje y tiene que ser capaz de responder correctamente a las preguntas formuladas por el aprendiz. Existen distintos tipos de preguntas, pero la combinación estándar utilizada es preguntas de pertinencia (MQs) y preguntas de equivalencia (EQs). En el caso de una MQ, el aprendiz pregunta si una cadena concreta está en el lenguaje y el profesor

[9] Existen estudios en los que los niños que reciben expansiones hacen progresos más rápidos en el lenguaje que aquellos que no las reciben (Hernández-Pina 1984).

responde *sí* o *no* (únicamente). En una EQ, el aprendiz hace una conjetura y pregunta si es correcta; el profesor responde *sí* si esa conjetura genera el mismo lenguaje, y en caso contrario, devuelve un *contraejemplo*. Un profesor capaz de responder a ambos tipos de preguntas es conocido como *Minimally Adequate Teacher*. Teniendo en cuenta que este tipo de preguntas no es muy natural en un contexto de aprendizaje real, nuestra idea es modelar una forma más natural de responder. De esta manera, hemos introducido una extensión de las MQs llamada *preguntas de corrección* (CQs); en lugar de responder *sí* o *no*, si la cadena no pertenece al lenguaje el profesor devuelve una corrección. Resultados iniciales y prometedores utilizando esta idea se pueden encontrar en (Becerra-Bonache et al. 2006, Becerra-Bonache et al. 2007).

Por otra parte, consideramos que sería de gran interés desarrollar un modelo de aprendizaje que refleje mejor la interacción real entre niño y adulto. Tal y como hemos visto, hay evidencia de que los niños reciben tanto datos positivos como correcciones. Entonces, ¿por qué no combinarlos en un único modelo?. Teniendo en cuenta que ninguno de los modelos en IG ha utilizado tal combinación, intentamos abrir aquí una nueva línea de investigación. Proponemos así un modelo interactivo entre aprendiz y profesor basado en la disponibilidad de datos positivos y correcciones. Consideramos que tal modelo puede incluso reflejar mejor la interacción real entre el niño y el adulto. Los objetivos de este modelo serían: a) utilizar información relevante en el proceso de adquisición del lenguaje; b) tener en cuenta más aspectos de procesos de aprendizaje real; c) utilizar herramientas más naturales (por ejemplo, preguntas empíricamente motivadas).

## 4. Conclusión

La investigación en modelos formales de adquisición del lenguaje puede ser de gran relevancia para ahondar en la comprensión del modo en que los niños adquieren su primera lengua. Basándonos en esta idea, hemos propuesto aplicar la teoría de la IG al estudio de adquisición del lenguaje.

Teniendo en cuenta que las investigaciones realizadas en IG se han centrado en los aspectos matemáticos de los modelos y no han explotado su relevancia lingüística, con este artículo hemos intentado aportar ideas lingüísticas que pueden ser de utilidad para los estudios de IG. No sólo se pueden obtener nuevas perspectivas y nuevas soluciones a problemas de IG, sino que se puede lograr que los modelos de IG pueden sean más realistas. Además, hemos propuesto nuevas direcciones en IG, las cuales pasamos a especificar a continuación.

Por un lado, hemos visto que los estudios de IG se centran en el aprendizaje de REG o CF, pero ninguna de estas clases tiene suficiente poder expresivo para describir algunas construcciones que aparecen en los lenguajes naturales. Por lo tanto, lo interesante sería estudiar una clase de lenguajes capaz de captar las construcciones no-CF básicas que aparecen en el lenguaje natural y que, a su vez, tuviera buenas propiedades computacionales. Justamente estas son las características de la clase de lenguajes conocida como *suavemente dependiente del contexto* (MCS). Además, nos interesaría que esa clase fuera incomparable con REG y CF, pero estuviera incluida en CS (es decir, que ocupara un posición ortogonal en la jerarquía de Chomsky, tal y como parece que tienen los lenguajes naturales).

Teniendo en cuenta todas estas consideraciones, hemos propuesto que los estudios de IG se centren en clases de lenguajes más relevantes desde un punto de vista lingüístico, es decir, clases de lenguajes que reúnan todas las interesantes propiedades lingüísticas descritas anteriormente (que sean MCS y que ocupen una posición ortogonal en la jerarquía de Chomsky). Además, hemos presentado una clase de lenguajes que podría ser un posible candidato: *Simple External Contextual* ($SEC_P$). Consideramos que sería de gran interés explotar la relevancia lingüística de esta clase; su estudio podría contribuir a comprender mejor algunos aspectos de la adquisición del lenguaje natural.

Por otro lado, hemos visto que hay una tendencia a reducir los datos disponibles para el niño a dos tipos: positivos o negativos. Además de que no existe una definición clara de qué se entiende exactamente por datos negativos, podemos encontrar ciertos tipos de datos que ofrecen tanto información positiva como negativa y que, por lo tanto, son difíciles de clasificar. Un ejemplo de estos datos son las *correcciones* (concretamente hemos hecho referencia a un tipo concreto de corrección conocida como *expansión*). Teniendo en cuenta que las correcciones podrían jugar un papel relevante en el proceso de aprendizaje, hemos aplicado esta idea a los estudios de IG, concretamente al modelo de aprendizaje a partir de preguntas, introduciendo así un nuevo paradigma de aprendizaje: *aprendizaje a partir de preguntas de corrección* (CQs). Además, hemos propuesto un modelo alternativo de aprendizaje basado en la combinación de datos positivos y correcciones.

En el futuro, nos gustaría poder explorar la relevancia lingüística de las CQs dentro del campo de la IG. Una dirección de investigación prometedora sería formalizar el modelo propuesto y desarrollar algoritmos para implementarlo. Además, nos gustaría poder estudiar clases de lenguajes utilizando este modelo, centrando especialmente nuestra atención en clases de lenguajes lingüísticamente relevantes, como por ejemplo, $SEC_P$. También queremos explorar las aplicaciones de nuestros resultados a otras áreas, tales como la robótica, la traducción automática y el procesamiento del lenguaje natural.

## 5. Referencias bibliográficas


Angluin, D. (1987). "Learning regular sets from queries and counterexamples", *Information and Computation* 75: 87-106.

Becerra-Bonache, L. (2006). "On the Learnability of Mildly Context-Sensitive Languages using Positive Data and Correction Queries". Tesis doctoral, Universidad Rovira i Virgili.

Becerra-Bonache, L. y T. Yokomori (2004). "Learning mild context-sensitiveness: Toward understanding children's language learning", *ICGI'04*: 53-64.

Becerra-Bonache, L., A.H. Dediu y C. Tîrnauca (2006). "Learning DFA from correction and equivalence queries", *ICGI'06* : 281-292.

Becerra-Bonache, C. de la Higuera, J.C. Janodet y F. Tantini (2007). "Learning balls of strings with correction queries", *ECML'07*: 18-29.

Bresnan, J., R. Kaplan, S. Peters y A. Zaenen (1987). "Cross-serial dependencies in dutch". En W. Savitch, E. Bach, W. Marsh y G. Safran-Naveh, eds., *The Formal Complexity of Natural Language*. Dordrecht: D. Reidel, pp. 286-319.

Brown, R. (1976). *A first language*. Cambridge: Harvard University Press.

Chomsky, N. (1956). “Three models for the description of language”, *IRE Transactions on Information Theory* 3: 113-124.

Clark, A. (2004). “Grammatical Inference and First Language Acquisition”, *Workshop on Psycho-Computational Models of Human Language Acquisition*: 25-32.

Culy, C. y D. Reidel (1987). “The complexity of the vocabulary of bambara”. En W. Savitch, E. Bach, W. Marsh y G. Safran-Naveh, eds., *The Formal Complexity of Natural Language*. Dordrecht: D. Reidel, pp. 349-357.

de la Higuera, C. (2004). “A bibliographical study of grammatical inference”, *Pattern Recognition* 38: 1332-1348.

Farrar, M. (1992). “Negative evidence and grammatical morpheme acquisition”, *Developmental Psychology* 28: 90-98.

Gold, E.M. (1967). “Language identification in the limit”, *Information and Control* 10: 447-474.

Hernández-Pina, F. (1984). *Teorías Psico-sociolingüísticas y su aplicación a la adquisición del español como lengua materna*. Madrid: Siglo XXI.

Joshi, A.K. (1985). “How much context-sensitivity is required to provide reasonable structural descriptions: Tree Adjoining grammars”. En D. Dowty, L. Karttunen y A. Zwicky, eds., *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*. New York: Cambridge University Press, pp. 206-250.

Kudlek, M., C. Martín-Vide, A. Mateescu y V. Mitrana (2002). “Contexts and the concept of mild context-sensitivity”, *Linguistics and Philosophy* 26(6): 703-725.

Manaster-Ramer, A. (1999). “Some uses and abuses of mathematics in linguistics”. En C. Martín-Vide, ed., *Issues in Mathematical Linguistics*. Amsterdam: John Benjamins, pp. 73-130.

Marcus, G. (1993). “Negative evidence in language acquisition”, *Cognition* 46: 53-95.

Marcus, S. (1969). “Contextual Grammars”, *Revue Roumaine des Mathématiques Pures et Appliquées* 14: 1525-1534.

Morgan, J., K. Bonamo y L. Travis (1995). “Negative evidence on negative evidence”, *Developmental Psychology* 31: 180-197.

Oates, T., T. Armstrong, L. Becerra-Bonache, M. Atamas (2006). “Inferring grammars for mildly context-sensitive languages in polynomial time”, *ICGI’06*: 137-147.

Parekh, R., y V. Honavar (2000). “Grammar inference, automata induction and language acquisition”. En R. Dale, H. Moisl y A. Somers, eds., *Handbook of Natural Language Processing*. New York: Marcel Dekker, pp. 727-774.

Roach, K. (1987). “Formal properties of head grammars”. En A. Manaster-Ramer, ed., *Mathematics of Language*. Amsterdam: John Benjamins, pp. 293-348.

Rozenberg, G. y A. Salomaa, eds. (1997). *Handbook of formal languages (volume 1-3)*. Berlin: Springer.

Sakakibara, Y. (1997). “Recent advances of grammatical inference”, *Pattern Recognition* 38: 15-45.

Saxton, M. (1997). “The contrast theory of negative input”, *Journal of Child Language* 24: 139-161.

Shieber, S. (1987). “Evidence against the context-freeness of natural languages”. En W. Savitch, E. Bach, W. Marsh y G. Safran-Naveh, eds., *The Formal Complexity of Natural Language*. Dordrecht: D. Reidel, pp. 320-334.

Steedman, M. (1985). “Dependency and coordination in the grammar of Dutch and English”, *Language* 61: 523-568.

Yokomori, T. (2004). “Grammatical Inference and Learnning”. En C. Martín-Vide y G. Păun, eds., *Formal Languages and Applications*. Berlin: Springer, pp. 507-528.