Revista de Epistemología y Ciencias Humanas

PRESENTACIÓN DE LA LINGÜÍSTICA COMPUTACIONAL

INTRODUCTION TO COMPUTATIONAL LINGUISTICS

Rodolfo Bonino

Facultad de Humanidades y Artes

Universidad Nacional de Rosario

Rosario – Argentina

rodolfobonino@yahoo.com.ar

Abstract

This paper presents a brief account of the development of the Computational Linguistics. A difference between the linguistic formalism and the computational formalism is established and the main characteristics of the linguistic formalism *5P* [1] and its compatibility with the Xerox Finite-State computational formalism (XFST) [5] are explained.

Keywords: Computational Linguistics. Linguistic Formalism. Computational Formalism. Formalism *5P*. Xerox Finite-State (XFST)

Resumen

En este trabajo se presenta una breve reseña del desarrollo de la Lingüística Computacional, se establece la diferencia entre formalismos lingüísticos y formalismos computacionales, y se explican las principales características del formalismo lingüístico *5P* [1] y su compatibilidad con el formalismo computacional Xerox Finite-State (XFST) [5].

Palabras claves: Lingüística Computacional. Formalismos lingüísticos. Formalismos computacionales. Formalismo *5P*. Xerox Finite-State (XFST).

1. INTRODUCCIÓN

La Lingüística Computacional y la Ingeniería Lingüística surgieron con gran ímpetu en Estados Unidos en la segunda mitad del Siglo XX orientadas hacia la obtención de traductores automáticos. El éxito en este campo no fue tan rápido ni eficiente como se esperaba en un primer momento y, como consecuencia de ello, se redujeron las inversiones que propiciaban su desarrollo.

A pesar de que hoy se tiene un dimensión más exacta de cuáles son sus dificultades y limitaciones, los traductores automáticos y los asistentes de traducción son una realidad en constante búsqueda de perfeccionamiento; y a partir de la década del noventa, la Unión Europea ha visto en la informática una solución a los desafíos que plantea el multilingüismo a la integración, lo que ha determinado un resurgimiento del interés, y las consiguientes subvenciones, para el desarrollo del campo.

Además de la traducción, se desarrollaron diversos modos de procesamiento automático del lenguaje natural; lo que, conjuntamente con el avance de la tecnología, produjo la popularización de la Informática y su inclusión en casi todos los ámbitos de la vida cotidiana, a tal punto que la mayoría de los usuarios olvidan o desconocen que la computadora es una máquina de cálculo y terminan considerándola una fuente de conocimiento, porque a través de Internet es posible acceder una cantidad ilimitada de documentos de todo tipo.

De modo que la Lingüística Computacional se inserta en el vasto campo de las llamadas Industrias de la Lengua, que tienen como finalidad el desarrollo de las nuevas tecnologías con los más variados fines. La utilidad práctica de este tipo de investigaciones exceden el interés de la lingüística teórica, pero abren a esta disciplina la posibilidad de utilizar a la Informática como un recurso para lograr un conocimiento más amplio y cabal de su objeto de estudio, tanto en lo que respecta a la obtención de datos, como a la formulación de hipótesis formales y procedimientos metodológicos.

Esto implica que, aunque no está exenta de problemas, la relación entre Lingüística e Informática resulta productiva para ambas disciplinas y tiene aplicaciones prácticas que la justifican. Por ello, tanto las corporaciones dedicadas a la computación como las instituciones académicas han emprendido distintos proyectos de investigación sobre el tema.

En nuestro medio se presentan algunas dificultades para llevar a cabo este tipo de estudios porque, como ni la Lingüística Informática ni la Ingeniería Lingüística se han organizado aún como especialidades universitarias, no se han desarrollado programas computacionales para el tratamiento del lenguaje natural. Este déficit no puede sortearse fácilmente ya que los programas accesibles por la red son muy limitados (se presentan como cajas negras, que

no permiten conocer los procedimientos con los que operan; en el mejor de los casos, se pueden utilizar los recursos para el análisis de palabras o secuencias) y los programas

especializados se obtienen mediante la suscripción de convenios que son costosos, difíciles de lograr e imponen ciertas restricciones al uso.

No obstante, en la Facultad de Humanidades y Artes de la UNR, durante los años 2004 y 2005, la Maestría en Teoría Lingüística y Adquisición del Lenguaje ha organizado seminarios de Lingüística Computacional dictados por el Dr. Gabriel Bès, de la Universidad Blaise Pascal (Clermont-Ferrant, Francia); y la Dra. Zulema Solana, con asesoramiento científico del Dr. Bès, ha elaborado el proyecto "INFOSUR. Investigación y desarrollo" (PID de la SECYT UNR), que actualmente trabaja en distintos proyecto vinculados al tema.

Estos fueron, en nuestro medio, los primeros pasos en una disciplina que, como se dijo, tiene una trayectoria de varias décadas en los países centrales. A pesar de la demora en el abordaje de la temática, los resultados obtenidos en algunos aspectos la descripción formal del español, como el sistema verbal, pueden medirse en relación con los más sofisticados que se hallan en Internet.

Una de las desventaja en la que ocasionan las limitaciones para acceso a las licencias de programas computacionales es la imposibilidad de mostrar los logros del trabajo con la presentación atractiva que tienen los analizadores desarrollados por las universidades y centros de estudios más importantes; pero esto no es un impedimento para llevar a cabo una investigación seria acerca de las propiedades formales del lenguaje.

A pesar de ello, queda abierta la posibilidad de elaborar un proyecto interdisciplinario con instituciones educativas dedicadas a la informática, lo cual acarrearía, además de ventajas económicas, otras más importantes, como la de ser colaboradores activos en la producción de programas y no meros consumidores.

Independientemente de la herramienta informática que se utilice, la meta es obtener formalismos lingüísticos compatibles con los formalismos computacionales. En el próximo apartado se presenta una breve reseña de los avatares que ha tenido la relación entre ambos.

2. FORMALISMOS LINGÜÍSTICOS Y COMPUTACIONALES

Desde sus primero trabajos, Chomky [2] plantea que el lenguaje natural, a pesar de su enorme complejidad, tiene una estructura regular y lógica, cuyas propiedades deben ser especificadas de manera rigurosa y explícita por la lingüística, mediante la formulación de las reglas que lo generan (de ahí la denominación de "gramática generativa").

Hasta la actualidad, esta premisa domina gran parte los estudios lingüísticos, chomskianos o no, e implica la necesidad de apelar a un metalenguaje similar al de la lógica o las matemáticas, a partir del cual se constituye la gramática; de este modo, la gramática se identifica con un formalismo capaz de generar construcciones del lenguaje natural.

Según el tipo de reglas que son capaces de generar, [2] distingue distintos tipos de gramáticas, con diferentes posibilidades de ser traducidas a formalismos computacionales. Una regla permite reescribir un símbolo: por ejemplo, la regla *O --\>SN SV* expresa simplemente que oración se reescribe como sintagma nominal y sintagma verbal.

a) Gramáticas de tipo 0 (irrestrictas): no imponen ninguna restricción a la formulación de reglas.

b) Gramáticas de tipo 1 (dependientes del contexto): la parte derecha de la regla tiene, cuanto menos, la misma longitud que la parte izquierda. Por ejemplo: la regla *SN VA SN --\> SN VP SP*. La parte derecha de la regla expresa que, en el contexto SN y SP, verbo activo se reescribe como verbo pasivo. Como se advierte, una gramática de este tipo no permitiría generar segundas de pasiva, porque en este caso a la parte derecha de la regla le faltaría un elemento, el SP.

c) Gramáticas de tipo 2 (independientes del contexto): la parte izquierda solo puede tener un símbolo. Por ejemplo, la regla ya citada: *O --\>SN SV*. Son recursivas, porque permiten utilizar el mismo símbolo de la izquierda a la derecha de la regla; también admiten elementos opcionales y alternativos (α o β).

d) Gramáticas de tipo 3 (regulares o de estados finitos): la parte derecha de la regla contiene un símbolo terminal (un elemento del vocabulario) o bien un símbolo terminal y un símbolo no terminal. Por ejemplo: permiten reglas del tipo *PRED --\> v* o *PRED --\> v SN*, que expresan:

"predicado se reescribe como verbo" y "predicado se reescribe como verbo y sintagma nominal". La limitación de este tipo de gramáticas es que no pueden dar cuenta de la recursividad del lenguaje natural, ya que no admiten reglas del tipo O --> SN SV, porque a la derecha no hay ningún símbolo terminal.

Los lenguajes generados por estas gramáticas son diferentes, pero no se ha podido demostrar que ninguna de ellas sea completamente adecuada para explicar el lenguaje natural: las gramáticas de tipo 0 son las más expresivas, porque pueden generar cualquier construcción; pero no imponen restricciones para evitar construcciones agramaticales; por el contrario, las gramáticas de tipo 3 son las más restringidas y, por esa razón, no permiten generar muchas construcciones gramaticales. Como consecuencia de esto, los formalismos lingüísticos son objeto de constantes reformulaciones.

Cualquier intento formalización, choca con las ambigüedades categoriales y estructurales que son inherentes todas las lenguas naturales. Por ejemplo: *la amenaza de muerte* puede interpretarse como un sintagma nominal o como un sintagma verbal; en las construcciones *le hizo abrir la boca* y *le hizo abrir el vientre* las secuencia clítico – verbo – SN tienen

interpretaciones diferentes: mientras las primera tiende a interpretarse como *hizo que él abriera la boca*, la segunda se interpreta como *hizo que le abrieran el vientre*.

Los lenguajes informáticos, en cambio, son por definición estrictamente formales; es decir, la relación significante – significado es uno a uno: a cada símbolo de estos lenguajes le corresponde uno y solo un significado, y viceversa. De modo que los formalismos computacionales resultan más controlables que el lenguaje natural, por eso se los ha visto como un modelo de formalización y un recurso para validar las gramáticas propuestas por las teorías lingüísticas.

Durante mucho tiempo los lingüistas computacionales intentaron trabajar a partir del modelo transformacional de Chomsky; pero las dificultades de adaptación que presentaba este modelo al tratamiento computacional llevaron proponer la reducción del componente transformacional de la teoría.

Las propuestas teóricas de Chomky también evolucionaron en este sentido, sin embargo, se focalizaron más en el aspecto psicolingüístico. No he hallado referencia a intentos de utilizar las últimas teorías chomkyanas [3] y [4] en Lingüística Computacional, en cambio, han cobrado auge modelos teóricos de motivación computacional como la Gramática Léxico Funcional (LFG), la

Gramática de Estructura Sintagmática Generalizada (GPSG) y la Gramática de Estructura Sintagmática Nuclear (HPSG), entre otros.

A pesar de estar elaboradas con miras al tratamiento informático; estos modelos tienen en común con los chomskyanos los presupuestos de que una teoría lingüística debe dar cuenta principios universales del lenguaje (Gramática Universal) y postular un modelo realista del lenguaje, o sea, explicar cómo se genera "realmente" el lenguaje.

Paralelamente a estas teorías, otro aspecto de la relación entre la Informática y el lenguaje natural es el interés por obtener resultados prácticos, lo que hace abandonar la búsqueda de la Gramática Universal y conocimiento de la lengua natural; y centra los esfuerzos en resolver problemas específicos que, en muchos casos, no pretenden cubrir todas las oraciones de unas lengua y privilegian el método estadístico.

[1] propone el modelo *5P*, que cuestiona tanto el método estadístico como los presupuestos de las teorías realistas mencionadas anteriormente. *5P* está constituido por módulos que organizan la investigación con miras a la formalización lingüística, según su propio autor son los siguientes:

> *À partir de la numérotation des P de 5P en P1 (P2, P3, P4) P5) on décline les composants du cadre ici proposé de la manière suivante.*
>
> - *P1 ou P de Protocoles, un Protocole étant la représentation d'une donnée à laquelle aboutit un Observateur dûment modélisé, cette donnée ayant ou non été observée dans un corpus effectif.*

> - *P2 ou P de Propriétés, une Propriété pouvant s'identifier formellement à un axiome. Un ensemble fini de Propriétés spécifie en intention un ensemble (fini ou infini) de suites d'une langue donnée. On appellera modèles ces suites. Un* modèle *es ainsi une suite qui satisfait un ensemble de Propriétés. Les modèles vont être associés aux énoncés de la langue décrite.*
> - *P3 ou P de Projections, une Projection étant une abstraction d'un ensemble de Propriétés d'une langue spécifiant les caractéristiques communes à toutes ces Propriétés.*
> - *P4 ou P de Principes, un Principe –tout comme une Propriété – pouvant formellement s'identifier à un axiome, les principes étant plus abstraits et généraux que les Propriétés. Les Principes vont introduire des contraintes générales, valables pour toutes les langues ou pour des groupes de langues sur le type des Propriétés –et donc des modèles – qui décrivent les langues particulières.*
>
> - *P5 ou P de Processus, un Processus étant un procédure effective implantable ou implantée en machine, permettant d'analyser et/ou produire un énoncé d'une langue particulière, cet énoncé étant explicitement associé à un modèle qui satisfait les Propriétés (P2) décrivant cette langue particulière.*
>
> *Les Propriétés (P2) sont testées en les confrontant aux Protocoles (P1). Le résultat de la confrontation exprime l'adéquation externe de ce système d'hypothèses que son les P2.*
>
> (pág. 280 -281).

Como puede inferirse de la cita precedente, *5P* no toma como punto de partida hipótesis generales acerca de la gramática universal ni siquiera de una lengua en su totalidad, sino un subconjunto de cadenas del francés, analiza la frase verbal núcleo del francés, con lo que establece un objeto lingüístico mayor a la palabra con propiedades formales estables.

La descripción de las propiedades (P2) de ese subconjunto no se propone en sí mismo como una explicación ni como una teoría. La explicación y la teoría consisten la formulación de hipótesis por medio de formalismos calculables:

> *Ici* expliquer *est assimilé à* prévoir: *si l'on connaît les P2 nécessaires à la description d'un corpus C fini [...] ces P2 doivent, conjointement avec les P3 et le P4, permettre de calculer –i.e. de prévoir– d'autres P2 –les P2'– qui soit sont également nécessaires à la description de C, soit son nécessaires à la description d'autres corpus de la même langue* (pág. 281).

La novedad de esta propuesta consiste, entonces, en que la estructura formal del lenguaje no se postula como un sistema de reglas sino como un sistema de hipótesis que, en su conjunto, permiten aproximarse al objeto, sin necesidad de que individualmente se asocie cada hipótesis a un aspecto particular del objeto.

Los principales postulados de *5P* pueden sintetizarse en las siguientes afirmaciones:

1) Es necesario proponer hipótesis descriptivas y explicativas, calcular sus consecuencias y validarlas en relación con lo observable, tal como lo hacen las ciencias empíricas: *il y a interaction entre hipothèses et observations, mais pas d'exigence de précédence des unes sur les autres*. (pág. 282).

2) El modelo no tiene una pretensión realista, en cuanto considera que el objeto se construye a partir de la observación y que cada punto de vista introduce su propia perspectiva que recorta necesariamente el objeto.

3) Los postulados anteriores excluyen el punto de vista psicolingüístico: no se supone que el sistema de hipótesis propuesto por el modelo tenga ningún correlato con los procesos mentales de producción del lenguaje, que no pueden validarse empíricamente.

4) Para constituir un conocimiento interesante, las proposiciones deben formar un conjunto estructurado por relaciones lógicas. Esto implica que, aunque las propiedades del objeto hayan sido descriptas con anterioridad, el trabajo científico consiste en expresar esas propiedades en un formalismo calculable.

5) Las fuentes declarativas para la generación no son las propiedades sino la **función** entre las categorías y las propiedades:

> *Les Propriétés sont exprimées en utilisant des catégories et des prédicats qui expriment des relations entre catégories. Les catégories sont des ensembles de traits, chaque trait dans une catégorie étant une étiquette associée a une seule valeur. On distingue trois grands types de Propriétés : les Propriétés d'existence, de linéarité et de fléchage* (pág. 290)

6) Para el formalismo lingüístico las categorías se asocian con símbolos operacionales[1], con lo que tampoco hay ninguna pretensión realista con respecto a las categorías utilizadas.

7) Se distingue formalismo lingüístico de formalismo computacional y se busca la interacción entre ambos:

> *...la description des langues, exprimée para les Propriétés (P2), les projections (P3) et les Principes (P4) n'est ni négligé ni subordonnée au traitement automatique. Mais cette description est exprimée dans un formalisme calculable, de telle manière qu'un lien formel puisse être établi entre ces descriptions et les Processus (P5) de traitement automatique* (pág. 275).

Por otra parte, el formalismo lingüístico no queda sujeto al formalismo computacional, ya que las propiedades pueden ser calculables independientemente su implantación en un programa concreto.

[1] [1] toma de Bochensky la noción de sentido *eidético*. Según menciona el sentido *eidético* es un sentido puramente operacional: se sabe únicamente cómo se debe emplear, pero no qué significa.

7) Consecuentemente, no es prioritaria la aplicación práctica del formalismo lingüístico:

> *Mais dans 5P* outil de calcul *n'es pas identifié avec* outil d'analyse ou de génération automatique d'énoncés ou de textes *et encore moins, avec logiciels des industries de la langue susceptibles de devenir merchandise.* (pág 348).

[6] analiza las posibilidades de traducir las Propiedades (P2) propuestas en el formalismo lingüístico *5P* a expresiones regulares de un formalismo computacional como Xerox Finite-State (XFST) y establece una comparación de P2 con las gramáticas clásicas de estructura de frase.

Las propiedades que utiliza Bès para el estudio de la frase verbal núcleo del francés son:

a) Propiedades de existencia, que incluyen:

a.1. *amod*: es el alfabeto del modelo, es decir, las categorías utilizadas en el lenguaje L.

a.2. *oblig*: son las categorías que deben aparecer obligatoriamente en el lenguaje L solo una vez.

a.3. *uniq*: son las categorías que no pueden aparecer más de una vez en el lenguaje L.

a.4. *exig*: expresa que una categoría de L exige la presencia de otra categoría de L.

a.5. *exigac*: es una subclase de *exig* que expresa que dos categorías de L deben concordar entre sí.

a.6. *exclu*: expresa que una categoría de L excluye la presencia de otra categoría de L.

b) Propiedades de linealidad: especifican el orden de las categorías.

c) Propiedades de flechaje: especifican relaciones de dependencia entre los categorías del lenguaje L.

Troullieux observa que XFST tiene amplias posibilidades de expresar las (P2), con excepción de dos aspectos:

1) Según Bès, La Propiedad *amod* también especifica que para cada categoría existe una cadena que contiene, al menos, una palabra de esa categoría (es decir todas las categorías se usan por lo menos una vez). Esta es una condición general asignada a las categorías, no una propiedad de las cadenas en sí, y no puede expresarse por medio de expresiones regulares. Esto implica que, a diferencia de *5P*, XFST admitiría categorías inútiles.

2) Las propiedades de *flechaje* no pueden traducirse a expresiones regulares, porque operan por encima de las cadenas definidas por las otras Propiedades y, por lo tanto, no pueden expresarse por suma o sustracción.

Con respecto a la comparación entre P2 y las gramáticas de estructuras de frases, señala que la novedad es que P2 hace uso sistemático de la intersección y no usa explícitamente la concatenación; por el contrario, las gramáticas de estructura de frase privilegian la unión y la concatenación. Es decir, en lugar de reglas, P2 utiliza la intersección de todas las propiedades, es decir, una descomposición de la información expresada por los sistemas de reglas tradicionales, lo que lo hace más adecuado para el tratamiento de los lenguajes naturales.

La traducción de P2 a expresiones regulares de XFST limitan el poder expresivo de P2 a la especificación de lenguajes regulares (gramáticas de tipo 3), porque los símbolos usados como argumentos de las propiedades son variables que se definen no recursivamente, es decir, las propiedades se pueden aplicar a cadenas de símbolos terminales; pero P2 también puede expresar lenguajes independientes del contexto (gramáticas del tipo 2) si los símbolos que se usan como argumentos de las propiedades son indistintamente símbolos terminales o símbolos no terminales; o sea, si las propiedades no se aplican directamente a las cadenas del lenguaje, sino a las cadenas de sus constituyentes inmediatos.

En otros aspectos, el poder expresivo de P2 es menor que el de las expresiones regulares de XFST, por ejemplo, es imposible definir con Propiedades el lenguaje denotado por la expresión regular [*a* (*a*)], es decir, el conjunto {*a, aa*} porque no es posible controlar con precisión mediante P2 cuántas palabras de la misma categoría se permiten en una cadena (por ejemplo incluya "una o dos *a*s"); solo se puede declarar que tales palabras pueden o deben aparecer (con las propiedades amod y oblig, y que puede haber solo una determinada palabra (con la propiedad *uniq*).

Como resultado de este análisis Troullieux, concluye que P2 también pertenece al paradigma de las gramáticas de estructura de frase porque cualquier conjunto de P2 puede traducirse, en un equivalente regular o en una gramática independiente del contexto.

3. CONCLUSIONES

La distinción entre formalismos lingüísticos y formalismos computacionales permite establecer la especificidad de la Lingüística Computacional, que es una disciplina complementaría, pero diferente de la Ingeniería Lingüística.

En este marco, el formalismo *5P* se perfila como una metodología adecuada para formular propiedades del lenguaje humano accesibles al tratamiento informático. Estas propiedades aportan un conocimiento teórico acerca del lenguaje que excede la utilidad práctica que se pueda hacer de él.

XFST resulta una herramienta útil para comprobar la adecuación de gran parte de las formalizaciones producidas por *5P*.

4. REFERENCIAS

[1] G. Bès. "La phrase verbale noyau en français" en *Recherches sur le français parlé*, Nº 15, págs. 273 – 358. Publications de l'Université de Provence, 1999.

[2] N. Chomky. *Syntactic Structures*. La Haya, Mouton. Versión en español: *Estructuras sintácticas*, México, Siglo XXI, 1980.

[3] N. Chomsky. *Lectures on Government and Binding*. Dordrecht. Foris.

[4] N. Chomsky. *The Minimalist Program*. Cambridge, Mass., MIT Press.

[5] L. Karttunen, T. Gaál, and A. Kempe. *Xerox Finite-State Tool*. The Document Company - Xerox, 1997.

[6] F. Trouilleux. "Specifying Properties of a Language with Regular Expressions". En *TAL*, 2007.

Este trabajo ha sido publicado como primer capítulo de *La interlengua de los aprendientes del español como L2,* Solana,Z. (ed) 2009, Centro de Estudios de Adquisición del Lenguaje, Facultad de Humanidades y Artes, UNR