

# Breve aproximación al análisis informático de la lengua

Geoffrey Sampson<sup>1</sup>

Por lo que concierne al inglés, e imagino que también al español, no hay duda de que Internet ha originado en la escritura una suerte de libertinaje. A través de la Red se intercambian textos con una cantidad de errores ortográficos y una falta de estilo que resultarían intolerables en un documento impreso en papel. En cualquier caso, dado que los buenos materiales quedan sepultados en Internet por una montaña de escoria, cabe argumentar que un bajo nivel de escritura sirve al menos como señal de ayuda a los usuarios, útil para filtrar las necesidades. No obstante, considero más probable que todo ello destruya la autoridad de quienes velan por el trabajo literario y la ortografía, de modo que *todo* escrito resulte cada vez más tosco, una circunstancia que, a mi modo de ver, será muy desafortunada para el público lector.

A decir verdad, más que con la enseñanza, durante los últimos años he estado comprometido de forma constante con la investigación. Me ocupo del procesamiento de lenguajes naturales, un tema acerca del cual doy clase de cuando en cuando. Mis dos últimos libros son *English for the computer* (Oxford University Press, 1995), donde se publica por primera vez un registro, comentado y exhaustivo, con muestras de construcciones gramaticales de inglés hablado y escrito, recopiladas informáticamente e intercambiables entre distintos centros de investigación<sup>2</sup>, y *Educating Eve: The language instinct debate* (Cassell, 1997), refutación de la teoría (defendida por Steven Pinker y Noam Chomsky) de que la estructura del lenguaje

<sup>1</sup> Lingüista británico, especialista en informática. Docente en la Escuela de Ciencias Cognitivas e Informática de la Universidad de Sussex. Fundador de UCREL (Unit for Computer Research on English Language) en la Universidad de Lancaster. Director y fundador de CCALAS (Centre for Computer Analysis of Language and Speech) en la Universidad de Leeds.

<sup>2</sup> English for the computer deriva del «SUSANNE Corpus», un proyecto informático mediante el cual Sampson quiso establecer una taxonomía del inglés. Este proyecto surge de una constatación: las bases de datos con muestras lingüísticas anotadas y analizadas servirán como fuente de información y matriz estadística para el desarrollo de los algoritmos de procesamiento del lenguaje humano. Con todo ello será posible el reconocimiento y entendimiento del habla humana por parte de la máquina, aproximando la posibilidad de una comunicación oral entre hombres y computadoras. El «SUSANNE Corpus» fue producido en Inglaterra con el patrocinio del Economic & Social Research Council.

humano está codificada genéticamente en la mente humana desde el nacimiento.

En este momento, mi actividad investigadora se centra fundamentalmente en la producción de muestras de inglés hablado y escrito, analizadas en su estructura y legibles mediante máquinas, con el fin de obtener estadísticas para el uso de *software* que analice nuestro idioma de modo automático.

Si bien trabajo en un departamento de informática, la mayor parte del tiempo utilizo el ordenador sólo para escribir y usar el correo electrónico. Disfruto al idear fragmentos sueltos de un programa, pero la mayor parte de las tareas en lo que se refiere al *software* que necesito para mi proyecto son realizadas por un investigador empleado con este propósito. Desafortunadamente, tampoco dispongo de tiempo para esa actividad.

Dado que las investigaciones de mi equipo dan lugar a un gran número de archivos electrónicos, en lugar de libros u otros textos impresos, necesitamos utilizar la Red como un escaparate a través del cual dar a conocer esta labor. Los libros presentan habitualmente una reseña en la sobrecubierta donde se ofrece alguna información acerca del autor, incluyendo detalles de interés humano. Por idéntica razón, me pareció buena idea intercalar dos o tres páginas de ese tipo entre todo el material académico que hay en mi página electrónica, creada en febrero de 1996. Según mi experiencia, resulta más fructífera la lectura de un libro si conozco algo de la faceta personal del autor, e imagino que lo mismo sucede a los usuarios de nuestros productos.

*Traducción de Guzmán Urrero*