

Nicoletta Calzolari

Istituto di Linguistica Computazionale
di Pisa

Lexicografía Computacional

El procesamiento del lenguaje aplicado a corpus textuales a gran escala requiere la construcción de léxicos computacionales que contengan cientos de miles de entradas. Cada entrada debe contener de manera explícita información de todo tipo, incluida la semántica.

En este marco, se considera fundamental la reusabilidad de material lexicográfico ya existente. El concepto de reusabilidad apareció por vez primera en el Workshop realizado en Grosseto (1986) y financiado por la CE, donde se planteó la necesidad de diseñar herramientas lingüísticas reusables, multifuncionales y multilingües.

El concepto de reusabilidad debe entenderse en dos sentidos: a) para la extracción y uso de información léxica presente de manera explícita o implícita en fuentes de información léxica diversa (MRDs, Bases de Datos terminológicas, corpus textuales etc.) como ayuda para construir léxicos computacionales a gran escala y reusables en el sentido de b, y b) - para construir léxicos computacionales destinados a diversos tipos de usuarios (diferentes sistemas de PLN, lexicógrafos, lingüistas, usuarios no especializados).

Los léxicos computacionales actuales responden a uno de estos dos tipos. La necesidad de disponer de recursos léxicos que responda a estas características se pone de manifiesto en el impulso dado por la CE a este tema, que se ha traducido en la puesta en marcha de diversos proyectos como son Aquilex (Esprit BRA), Eurotra 7 Multilex (Esprit) y Genelex (Eureka) y la subvención de otro como TEI (Text Encoding Initiatives) y Survey on Linguistic Resources.

El proyecto Esprit Aquilex puede considerarse como un prototipo del primer grupo, mientras que otros proyectos, como Eurotra-7 como un prototipo del segundo.

El proyecto Aquilex

El objetivo de la investigación en el marco de Aquilex se centra en el desarrollo de técnicas y metodologías para extraer información, tanto explícita como implícita, de tipo sintáctico y semántico, a partir de diccionarios en soporte magnético (MRDs) para construir el componente léxico de sistemas de PLN.

Esta metodología se basa en la posibilidad de extraer de manera

automática la información que los MRDs contienen de manera implícita. Así, el diccionario se considera como una fuente de conocimiento básico y los objetivos fundamentales son la formalización de este conocimiento básico general en forma de conceptos y relaciones semánticas.

Entre los temas de investigación conectados con el objetivo último de adquisición de conocimiento cabe destacar: a- el diseño de programas para la extracción de los superordinados a partir de las definiciones, su desambiguación y la construcción de taxonomías; b- el diseño de analizadores para el análisis de las definiciones con el objetivo de extraer la información semántica implícita; c- el estudio de como representar la información semántica extraída (i.e.: en forma de conceptos, atributos, relaciones entre conceptos); d- el estudio de como relacionar las taxonomías y la representación conceptual; e- el diseño e implementación de software básico para la creación, acceso y procesamiento de bases de datos léxicas y una base de conocimiento léxico.

Las taxonomías y las plantillas (estructuras de rasgos en que se representa la información semántica) desarrolladas en el marco de Aquilex constituyen un primer grado de normalización y estandarización en la representación de la semántica y del conocimiento general a partir de diccionarios (unos diez en total) y de lenguas diversas (cuatro).

En el proyecto Aquilex colaboran las universidades de Cambridge, Dublín, Amsterdam, la Universidad Politécnica de Cataluña y el Instituto de Lingüística Computacional de Pisa. Se trabaja sobre diccionarios monolingües del inglés, holandés, italiano, castellano y sobre diccionarios bilingües inglés-holandés, inglés-italiano.

Como complemento indispensable de los trabajos lexicográficos debemos citar las bases de datos textuales. En este marco cabe destacar el programa TEI (Text Encoding Initiative) consistente en una iniciativa para definir criterios generales de codificación e intercambio de textos en soporte magnético. Se trata de un proyecto internacional de cuatro años de duración que tiene como objetivos: a)- especificar un formato uniforme para los textos en soporte magnético; b)- difundir las líneas generales de codificación acordadas; y c)- documentar los esquemas de documentación más importantes y desarrollar un metalenguaje en el que describirlos.