

Procesamiento estadístico del léxico

Esta conferencia presenta la fundamentación teórica y metodológica y los resultados principales de nuestro proyecto *Léxico Básico del Español de Chile*, en el que, para determinar el léxico más usado en el país, se operó sobre un corpus de referencia propio de quinientas mil palabras en contexto, extraídas de cerca de mil libros y de más de tres mil ejemplares de diarios chilenos publicados por vez primera entre 1981 y 2006. Se trabajó con distintos géneros que constituyen “mundos” diferentes —Narrativa, Ensayo, Drama, Técnico-Científico y Prensa— y se establecieron comparaciones entre el léxico de cada uno de ellos, usando herramientas de estadística inferencial. El procesamiento del corpus ha utilizado herramientas de léxico-estadística y lingüística computacional y ha considerado aspectos semánticos y gramaticales. Los principales resultados de esta investigación son: a) La creación de un corpus de referencia etiquetado de 501211 palabras en contexto. b) La determinación del léxico básico del español de Chile. c) La creación de un diccionario no definitorio de 4832 entradas.