Datos estadísticos del CORPES XXI (versión 0.9)

La visión más simple del contenido de un corpus consiste en considerarlo como un agregado de textos, de mayor o menor extensión. Sin embargo, su complejidad es considerablemente más elevada, puesto que, además del texto, es necesario tener en cuenta todos los rasgos introducidos por la codificación, tanto externa (datos de la cabecera acerca del país, tipo de texto, año de producción, autor, editorial, etc.) como interna (segmentación de oraciones, adición del lema al que corresponde cada forma, valores de las categorías gramaticales que son de aplicación, etc.). Como consecuencia de todo ello, las estadísticas pueden ser referidas a diferentes aspectos y, por tanto, mostrar discrepancias entre sí.

La estadística más sencilla es la que cuenta el número de formas ortográficas, esto es, de secuencias de caracteres situados entre dos espacios en blanco, un espacio y un signo de puntuación, dos signos de puntuación o un signo de puntuación y un espacio en blanco. Para que esa cifra dé cuenta de lo que hay realmente en un texto, es necesario retirar del recuento todas las etiquetas que han sido introducidas mediante los diferentes procesos de codificación y lo que contienen en su interior. Realizados esos procesos, obtenemos que el CORPES 0.9 contiene algo más de 277 millones de formas ortográficas procedentes de los casi 275 000 documentos que alberga en su interior.

La aplicación de procesos de análisis lingüístico supone la aparición de elementos de características diferentes y cuya suma difiere, como es lógico, de la que se obtiene de las formas ortográficas. Así, formas ortográficas como *al* o *diciéndolo* contienen dos elementos gramaticales en su interior (preposición y artículo, verbo y pronombre). En sentido contrario, a partir de un cierto grado de refinamiento, estos análisis reconocen la existencia de locuciones de diverso tipo que consideran con carácter único secuencias constituidas por dos o más formas ortográficas (*poco a poco*, *de orden de*, *a renglón seguido*, etc.) y también agrupan en un elemento único los nombres de personas o entidades de diverso tipo (*Gabriel_García_Márquez*, *Unión_europea*, etc.). Además, deben reflejar también los signos de puntuación, puesto que es forzoso atribuirles un determinado carácter. Teniendo en cuenta todos esos factores, la versión 0.9 del CORPES XXI contiene 1 615 743 elementos formales distintos (*types*)[1] que suman un total de 298 297 015 elementos (*tokens*). De ellos, 166 corresponden a signos ortográficos, que aparecen en conjunto 39 573 811 veces. Por tanto, hay en esta versión del CORPES 1 611 577 elementos formales diferentes de los signos ortográficos que suponen en conjunto un total de 258 723 204 apariciones.

Por otra parte, muchas de las formas incluidas en los textos son nombres propios de

1 Ascienden a 1 724 037 si tenemos en cuenta, además de la forma en sentido estricto, las posibles diferencias en la asignación de lema, categorías y subcategorías gramaticales.

personas, lugares geográficos, etc. o bien cifras, fechas y, en general, elementos que suman en las estadísticas generales, pero que no interesan en recuentos enmarcados en aspectos léxicos. Teniendo en cuenta todos estos factores, los datos estadísticos relevantes de la versión 0.83 del CORPES son los siguientes:

- Tamaño total:
  - 277 109 961 formas ortográficas procedentes de 274 550 documentos.
  - 1 615 743 elementos distintos (formas léxicas, locuciones, abreviaturas, cifras, signos de puntuación, etc.) que suponen en total 298 297 015 unidades.
  - 1 611 577 elementos distintos diferentes de los signos de puntuación, con un volumen total de 258 723 204 unidades.
- Elementos léxicos en sentido estricto (es decir, sin abreviaturas, cifras, fechas, signos de puntuación, nombres propios, expresiones horarias, etc.):
  - 1 259 742 elementos distintos, agrupados en 113 944 lemas, que suponen un total de 242 959 266 apariciones.
  - 7995 expresiones multipalabra distintas, agrupadas en 3502 “lemas”, y un total de 4 463 936 apariciones.