EUSLEM: Un lematizador/etiquetador de textos en euskara

Aldezabal I., Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K. (*)
Aduriz I., Urkia M. (**)

(*) Informatika Fakultatea. 649 P.K. 20080 DONOSTIA (Euskal Herria).
acpalloi@si.ehu.es
(**) UZEI. Aldapeta, 20. 20009 DONOSTIA (Euskal Herria)

Resumen

Un lematizador/etiquetador automático es una herramienta básica para aplicaciones tales como indexación automática, bases de datos documentales, análisis sintáctico y semántico, análisis de corpus, etc.

Las características con las que se ha diseñado el lematizador/etiquetador EUSLEM han sido las siguientes: extenso, robusto, de propósito general, modular y eficiente.

Los módulos básicos en los que se divide son: preprocesador para la detección y etiquetado de números, signos de puntuación, etc.; analizador morfológico general basado en la morfología de dos niveles con tratamiento de formas dialectales, errores debidos al desconocimiento del idioma estándar (tratamiento de variantes lingüísticas) y gestión de léxicos específicos; lematización sin léxico para que el sistema sea robusto; tratamiento de unidades léxicas complejas (términos de más de una palabra) y desambiguación basada en estadísticas. En una fase posterior se complementará con la desambiguación basada en conocimiento lingüístico.

A la hora de abordar este proyecto un inconveniente que hemos tenido que superar es la falta de un sistema de etiquetas para el euskara. Esto combinado con que el euskara es una lengua aglutinante que puede acumular gran cantidad de información en una palabra (varios casos, elipsis, etc.) hace que la definición del sistema de etiquetas no sea evidente.

Temas: Análisis de corpus, análisis morfológico.

estricta, fuerte dependencia (marcada por la congruencia) y no admite ser interrumpida físicamente por intercalación de material, excepto en casos marcados. Este tipo de relación es el más "rígido" de cuantos expondremos. Remitiéndonos a algunos experimentos psicolingüísticos de los años sesenta, podemos afirmar que en este tipo de relaciones las pausas son inexistentes. La existencia de éstas implicaría romper o relajar este tipo de dependencia entre dos elementos.

En cuanto a la relación de especificación a nivel intrasintagmático, ésta es más difícil de ilustrar, puesto que aparentemente es un tipo de dependencia que sólo se da entre sintagmas (recordemos lo dicho en la nota 4 que para algunos autores hay que abandonar este tipo de relación y concentrarse en a otras dos). A nivel extrasintagmático, esta relación se da entre el NP sujeto de una frase y su predicado, en el marco de la IP. Si bien a este nivel se mantiene todavía la congruencia, se admite la interrupción física de la relación con intercalación de material, como frases parentéticas. Otros elementos que están en esta posición son los interrogativos, sean ya partículas o expresiones más complejas. La relación de especificación no es tan rígida como la de complemento. Por ello postularemos aquí una pausa de mediana duración y carácter optativo.

Para acabar con las relaciones intrasintagmáticas, definiremos las dependencias de los adjuntos como las más "libres" de todas. Como es sabido, estos elementos son optativos por definición, y su presencia no es regulada léxicamente y estructuralmente son los más independientes dentro de un sintagma. Dicho de otro modo, su desaparición comporta una pérdida de información pero no de gramaticalidad. Estos elementos se consideran a menudo como "sintagmas dentro de un sintagma", y no modifican gramaticalmente al elemento que acompañan. Este tipo de dependencia es ejemplificado por las PPs que pueden acompañar a un nombre. En estos casos, en el interior de un sintagma, podrán existir pausas más prolongadas (obligatorias en casos marcados como p. ej. en enumeraciones).

Vemos pues que el principio de dependencia sintáctica - pausas prosódicas se cumple en las relaciones intrasintagmáticas.

Fijémonos ahora si nuestras predicciones se ajustan también a las relaciones extrasintagmáticas, es decir, a las relaciones que se establecen entre sintagmas. También aquí distinguimos entre dependencias de complementariedad, especificidad y adjunción.

Como en el caso de las dependencias complementarias intrasintagmáticas, las dependencias complementarias extrasintagmáticas son las más rígidas. Establecidas por selección léxica (lo que también se conoce por "subcategorización", estas relaciones son las más estrechas que puedan darse entre sintagmas. El ejemplo clásico es la relación entre un verbo y su objeto directo, en muchas lenguas regulada configuracionalmente.

Aunque aquí sí es posible intercalar p. ej. una PP entre el verbo y su objeto, ésta será generalmente pronunciada entre pausas y con cambio entonativo. En cualquier caso, estudios experimentales nos confirman que la tendencia es no separar demasiado

mejor, estas "tendencias", seguro que avanzaremos en nuestro propósito.

Antes de entrar a presentar nuestro estudio, destacaremos que, paralelamente a la investigación llevada a cabo por los autores que basan sus datos en la lectura, también la investigación de pausas por parte de quienes basan sus datos en el habla espontáneo. Estos autores, como Butcher (1980), Caldignetto (1992), o Misono (1990) son muy partidarios de establecer una relación entre las pausas y la estructura sintáctica del habla. Aunque, como hemos dicho más arriba, cuestiones de metodología nos alejan de estos autores, sí es cierto que tenerlos en cuenta puede ayudarnos de algún modo a mejorar nuestro trabajo.[2]

Teniendo en cuenta estas ideas, vamos a ver si es posible establecer un cierto tipo de relación entre la sintaxis y las pausas dentro de una frase. Teorías sintácticas actuales nos dicen que, dentro de un sintagma, esto es, dentro del ámbito que forma uno de los elementos léxicos (N, V, A, P) o funcionales (C, D, I) y los elementos que los acompañan, existen tres tipos de dependencia, de mayor a menor. Estas son:

-dependencia de complemento (o léxica)
-dependencia de especificador (o estructural)
-dependencia de adjunto (o "libre")[3]

Estas dependencias imponen una serie de restricciones al posible orden de las palabras y al material que se podrá intercalar entre elementos. Veremos como estas dependencias se dan tanto en el interior de los sintagmas (aquí las llamaremos "relaciones intrasintagmáticas" como entre los sintagmas (a lo que llamaremos "relaciones extrasintagmáticas")[4]

Empezaremos estudiando las dependencias en las relaciones intrasintagmáticas. Una de ellas es la relación de complemento en el DP, lo que implica una contigüidad

---

[2] Formulamos esta idea con una cierta reserva, ya que hay que tener claro que de hecho se está hablando de dos tipos distintos de producción de pausas, en lo que a la competencia del hablante se refiere: en el habla espontánea, tenemos que considerar la competencia propia de un determinado hablante, mientras que en la lectura el hablante intenta "reproducir" la competencia de otra persona.

[3] Hay que remarcar que sin embargo algunas teorías no distinguen entre especificadores y adjuntos. Nosotros la haremos para obtener una tipología más detallada de las pausas.

[4] Probablemente alguien pueda argumentar aquí que la única relación de carácter realmente intrasintagmático es la primera que expondremos, mientras que el resto son relaciones entre sintagmas. Ello es debido a la fuerte recursividad de las lenguas naturales, que hace que diferenciar los dos niveles que acabamos de postular sea hasta cierto punto superfluo. A pesar de que una afirmación de este tipo puede tener su certeza, nosotros hacemos tal distinción ya que creemos que queda justificada por la búsqueda de una correcta tipificación las pausas en un texto.

...gía respiratoria implicada en el proceso del habla. En este nivel podremos definir ...ptos como el de "grupo respiratorio", "grupo fónico", "grupo acentual" u otros. ...podemos definir también la importancia de los niveles segmental o ...asegmental o bien su interacción para postular una teoría de las pausas. ...lmente, aquí podremos hacer afirmaciones más generales sobre la prosodia, p. ...interacción de las pausas con otros factores prosódicos como la entonación o ...ergía.

Factores (morfo)sintácticos

...primera vista, el nivel sintáctico puede parecer uno de los más alejados de la ...osodia. No obstante, creemos que de este nivel se pueden extraer algunos ...ncipios importantes de cara al tema que nos interesa, ya que la sintaxis es un área ...la lingüística que ha alcanzado uno de los mayores niveles de formalización.

...conseguimos establecer pues una correspondencia más o menos estrecha entre ...estructura sintáctica y las pausas de un enunciado, podremos establecer en ...consecuencia una mejor sistematización de las pausas y a la vez podremos postular ...una (estrecha) relación entre dos niveles lingüísticos bastante diversos. Ello ...significaría una evidencia suplementaria para el paradigma de la economía cognitiva, ...esto es, para el principio teórico que argumenta en pro de la reducción de las ...manifestaciones psicológicas externamente observables a un número mucho más bajo ...de facultades mentales que parecerían estar implicadas en los procesos cognitivos ...cuando éstos son observados externamente. Dicho de otro modo, la pluralidad de ...fenómenos externos cuantificables y teorizables de varios de varios modos sería la ...manifestación de unos pocos procesos cognitivos - incluso en nuestro caso de uno ...solo: la facultad humana del lenguaje - que sin embargo mostraría diversos niveles ...de exteriorización.

Una forma de aproximarse a lo dicho en el párrafo anterior es la de hacer coincidir los emplazamientos de las pausas con puntos muy determinados de la estructura sintáctica. Seguimos aquí una línea que se remonta por lo menos a Goldman-Eisler (1972) y que ha hallado más o menos aceptación, pero que se considera la más practicable. Aunque parece difícil que esta teoría pueda dar cuenta de todas las pausas en un texto, sí que nos parece una teoría lógica, sencilla y de fácil implementación. De hecho, es la teoría que más o menos de manera implícita se ha impuesto (cf. Cruttenden, 1990, págs. 34-37).

No obstante, los trabajos experimentales mencionados al principio ven dificultades en querer establecer una relación entre las pausas y la sintaxis. Se constata que las pausas no ortográficas en la lectura son de menor duración y de carácter más opcional, aunque por otra parte se afirma que "en un texto poco puntuado todo lector aprovecha un límite sintáctico para hacer una pausa" (Puigví/Fernández, 1993, pág. 39) con o cual se da pie a establecer una correlación como la que nosotros proponemos. También en este sentido se expresan Martí/Gudayol (1993, pág. 6), quienes, tras afirmas en su estudio que las pausas marcadas son casi siempre respetadas por los lectores, afirman que "cuando el fragmento es excesivamente largo hay una tendencia a añadir pausas en puntos [no marcados] muy específicos." Si podemos saber más exactamente cuáles son esos "puntos específicos" o, todavía

de pausas sea menos importante. Parece ser que las pausas no ortográficas son, en la lectura, las más cortas y las de más "libre" colocación. Por tanto, serán también las más difíciles de sistematizar (cf. los trabajos citados más arriba).

Esta separación entre pausas ortográficas y no ortográficas, cuya relación parece ser de intersección parcial, hace que nos hallemos delante de dos sistemas distintos que definirán la recurrencia y características de las pausas en una determinada lengua: un "sistema normativo", enfocado a establecer normas de puntuación, fijado por los académicos, y un "sistema espontáneo", que ayudará en la lectura a vocalizar mejor el texto.

En general podemos afirmar que las pausas resultan de la interacción de los siguientes factores, que comentaremos en las siguientes secciones y que parecen ser los responsables últimos de la localización y la duración de las pausas en el habla humana:

-factores semántico-pragmáticos
-factores fonético-prosódicos
-factores (morfo)sintácticos

3. Factores semántico-pragmáticos

Estos factores son los más idiosincrásicos de todos, y a este nivel es muy difícil formular restricciones concretas o formalizaciones. En la lectura, y en mayor medida en el habla espontánea, el hablante podrá cambiar el emplazamiento de las pausas y su duración de manera considerable, dependiendo de las necesidades enfáticas del discurso. Por ejemplo, los silencios podrán ser alargados voluntariamente para captar la atención del oyente y con el mismo fin las pausas podrán ser alargadas, acortadas, suprimidas o colocadas en otro punto del discurso. La entonación es el parámetro sobre el que más se opera en este contexto, aunque la manipulación de pausas tampoco es despreciable. En el resto del presente artículo, no nos ocuparemos más de estos factores, de difícil teorización y codificación.

Tampoco nos ocuparemos de las pausas dubitativas, que en psicolingüística se ven como claves de organización del enunciado a la hora de producirlo, o de las pausas estudiables desde un punto de vista más "sociolingüístico", cuyo número según algunos investigadores sería decreciente a mayor edad y formación del individuo en cuestión.

Dejamos este tipo de pausas de lado no tan sólo porque en la sección 1 dijimos que estábamos interesados en las áreas más formalizables, sino también por una obvia cuestión de fondo: la producción de pausas (marcadas o no) en lectura obedece a principios psicolingüísticos muy distintos de los que pueden valer en esta sección. Mientras que en la producción de pausas vemos en acción a la competencia lingüística propia de un hablante, en la lectura se intenta reproducir la competencia de otro hablante. Por ello, los mecanismos que subyacen a los factores semántico-pragmáticos son de índole distinta al sistema que nos interesa estudiar a nosotros.

4. Factores fonético-prosódicos

Aquí las pausas obedecen fundamentalmente a las restricciones impuestas por la