Generación de un diccionario y herramientas de análisis morfológico para el español

Aiala Rosá
Grupo de Procesamiento de Lenguaje Natural (PLN),
Instituto de Computación (InCo),
Facultad de Ingeniería, Universidad de la República,
Montevideo, Uruguay
aialar@fing.edu.uy

Marzo de 2005

Resumen
En este documento se describe el trabajo de elaboración de un diccionario de formas flexivas para el español. Se define un modelo morfológico para nombres y adjetivos y se implementan procedimientos para la generación de formas flexivas. Estos procedimientos se aplican a un diccionario de lemas. El proyecto incluye la incorporación del diccionario al etiquetador que se utiliza en el grupo de Procesamiento de Lenguaje Natural, a partir de lo cual se comprueba una notoria mejora en los resultados del etiquetado de textos.

Palabras clave
procesamiento de lenguaje natural, diccionario, análisis morfológico, etiquetador

1. Introducción

El grupo de Procesamiento de Lenguaje Natural (PLN) dispone de varias herramientas de base para el tratamiento informático de textos en español. Entre estas herramientas se cuenta con un etiquetador de distribución libre, FreeLing [9], cuyo funcionamiento se basa, entre otras cosas, en la realización de consultas a un diccionario. El diccionario contenido en la versión libre del etiquetador es bastante limitado e introduce una cantidad importante de errores al procesar textos.

El trabajo que se presenta en este documento surge por la necesidad de mejorar el diccionario del etiquetador mencionado, el cual es considerado un recurso lingüístico fundamental para el grupo PLN. En particular, es necesario incluir palabras nuevas para las categorías nombre, adjetivo y verbo.

El grupo PLN cuenta, entre sus recursos lingüísticos, con listas de nombres, adjetivos y verbos de tamaño considerable que sirven como punto de partida para la creación del diccionario. Estas listas contienen solamente lemas: nombres en singular, adjetivos masculinos en singular y verbos en infinitivo. Este material no resulta suficiente para alcanzar el objetivo planteado, ya que el diccionario del etiquetador debe contener todas las formas flexivas correspondientes a cada lema: singulares y plurales, masculinos y femeninos, conjugaciones verbales.

Este requerimiento llevó a proponer un trabajo más general de definición de un modelo morfológico para nombres, adjetivos y verbos del español, incluyendo la implementación de procedimientos para generar todas las formas flexivas correspondientes a cada lema.

Más allá de su aplicación directa al etiquetador, el diccionario junto con el modelo morfológico asociado a él y los procedimientos de generación de formas flexivas, constituyen un recurso muy valioso para el grupo PLN. Generalmente, este tipo de recurso no se encuentra disponible en forma gratuita, y este proyecto permite construirlos a muy corto plazo.

El alcance del proyecto se restringe al trabajo con nombres y adjetivos, la generación de las formas flexivas de los verbos se aborda en el marco de otro proyecto del grupo. Queda pendiente un trabajo de depuración del diccionario, ya que es necesaria una revisión minuciosa de los elementos que lo componen.

2. Descripción general del proyecto

El trabajo involucra diferentes tareas que se describen a continuación.

Confección de un diccionario de lemas

Se recopilaron listas de nombres, adjetivos y verbos, conformándose un diccionario bastante completo para el español. Estas listas están compuestas por lemas, es decir, palabras sin

flexión morfológica. Por lo tanto, se incluyen: sustantivos en singular, adjetivos de género masculino en singular y verbos en infinitivo.

Definición del modelo morfológico

Para definir el modelo morfológico para nombres y adjetivos del español, se tomó como base un modelo ya existente, disponible en el InCo, que, si bien sirvió de punto de partida para esta tarea, no contempla en su totalidad los rasgos necesarios para este trabajo y además contiene numerosos errores. El modelo definido en este proyecto permite describir todas las palabras de las categorías mencionadas, nombre y adjetivo, en función de su comportamiento respecto al número y el género.

Implementación del programa para generación de formas flexivas

La información que aporta el modelo morfológico permite generar todas las formas flexivas de las palabras del diccionario. Dada una palabra cualquiera, según su categoría y la información morfológica que lleva asociada, se generan las formas flexivas correspondientes. Por ejemplo, para el nombre *archivo* se genera el plural *archivos*, para el adjetivo *abatido* se generan el femenino *abatida* y los dos plurales *abatidos* y *abatidas*.

Incorporación de palabras al diccionario del etiquetador

Las formas flexivas generadas se incorporaron al diccionario del etiquetador para lo cual se hicieron algunos ajustes de modo de adaptarlo a ciertas restricciones de formato. Como última etapa del trabajo se evaluaron los resultados de la aplicación del etiquetador usando el diccionario completo, comparándolos con los resultados obtenidos utilizando el diccionario original y la demo disponible on-line [11].

3. Definición del modelo morfológico

La definición del modelo consiste en identificar, dentro de cada categoría, diferentes grupos de palabras con igual comportamiento en lo que respecta a los rasgos morfológicos que intervienen en la flexión de la categoría.

Para esta tarea se tomó como punto de partida un modelo que, si bien no es del todo apropiado para los fines de este trabajo, contiene información que permitió resolver este punto en un tiempo razonable dentro del cronograma global del proyecto. Además, este modelo está asociado a un conjunto considerable de nombres y adjetivos, lo cual simplifica la tarea de asociar a cada palabra su grupo morfológico.

En el modelo de partida, se agrupan palabras de igual comportamiento morfológico bajo un identificador de grupo. Así por ejemplo se tiene el grupo de los nombres que se comportan como *escuela,* que son de género femenino y forman su plural por adición de *s*. Por otro lado, se tiene el grupo de los nombres que se comportan como *sombrero*, que son de género masculino y forman su plural por adición de *s*. Este tipo de clasificación no se consideró apropiado para este proyecto, ya que se pretende generar las formas de plural de los nombres

de manera automática. En este sentido, resulta más natural que *escuela* y *sombrero* lleven la misma descripción en lo que respecta a la formación del plural y se diferencien en su género.

Otra característica negativa del modelo de partida es que no todos los grupos de palabras tienen información de género. Por ejemplo, para el grupo de las palabras terminadas en vocal+*z*, no es posible deducir a qué género pertenecen. Se encuentran dentro de este grupo, por un lado, *antifaz, matraz, ajedrez* de género masculino y, por otro lado, *paz, acidez, rapidez* de género femenino[1].

El modelo que se propone en este trabajo tiene en cuenta los dos rasgos morfológicos relevantes para las categorías nombre y adjetivo: género y número. Cada rasgo se describe por separado; se define el comportamiento en cuanto al número y el comportamiento en cuanto al género de cada grupo morfológico en forma independiente.

Cada grupo se identifica mediante una etiqueta de la forma *categoría_número_género*; los tres componentes de la etiqueta se describen a continuación.

- *categoría*: Indica la categoría a la que pertenece la palabra: nombre o adjetivo.
- *número*: Indica el comportamiento de la palabra respecto al número, es decir, describe la formación del plural.
- *género*: Indica el comportamiento de la palabra respecto al género, es decir, describe la formación del femenino.

Una vez definidos los grupos de una categoría particular, se le asigna a cada palabra de esa categoría el identificador del grupo al cual pertenece. De este modo, se genera un diccionario morfológico compuesto por elementos de tipo *lema:grupo*.

El siguiente ejemplo muestra un extracto del diccionario morfológico:

```
...
abonador:a_sp21_mf2
abonador:n_sp2_m
abonadora:n_sp1_f
abonaré:n_sp1_m
abono:n_sp1_m
abordable:a_sp1_c
abordador:a_sp21_mf2
abordaje:n_sp1_m
abordo:n_sp1_m
aborigen:a_sp2_c
aborigen:n_sp2_c
...
```

Como se observa en los ejemplos, algunas palabras están duplicadas en el diccionario por pertenecer a las dos categorías (*abonador* y *aborigen* son a la vez nombres y adjetivos). Además, se puede ver que cada nombre tiene un género fijo (el nombre *abonador* es masculino, el nombre *abonadora* es femenino), en cambio, los adjetivos contienen en el elemento *género* de la etiqueta, el modelo de flexión para generar la forma de femenino (el adjetivo *abonador* es masculino y a partir de él se genera el adjetivo femenino *abonadora* que no se incluye en el diccionario morfológico).

[1] En el apéndice A se presenta este punto con más detalle.

3.1 Nombres

Para generar las formas flexivas de un nombre solamente se considera el rasgo número. El género se considera un rasgo fijo de la categoría, es decir, cada lema de la categoría nombre es femenino, masculino o común[2].

Cada grupo morfológico de nombres es identificado con una etiqueta de tipo n_*número_género*, tal como se describió anteriormente.

Los valores posibles para los elementos *número* y *género* se detallan en las siguientes tablas.

Valores posibles para el número de los nombres			
Valor	**Descripción**		**Ejemplo**
spX	Existen dos formas diferenciadas para singular y plural. X indica cómo se genera el plural.		
	X=1	formación de plural por adición de *s*	*casa/casas*
	X=2	formación de plural por adición de *es*	*doctor/doctores*
	X=3	formación de plural por sustitución de *z* por *ces*	*actriz/actrices*
s	Sólo existe la forma singular.		*sed, cenit*[3]
p	Sólo existe la forma plural.		*víveres, afueras*
i	Número invariable. La forma de singular y la forma de plural son iguales.		*la crisis/las crisis*

Valores posibles para el género de los nombres		
Valor	**Descripción**	**Ejemplo**
m	Nombre de género masculino.	*servicio, doctor*
f	Nombre de género femenino.	*casa, doctora*
c	Género común. El nombre tiene valor femenino y masculino.	*el dentista/la dentista*

Para nombres con formación de plural por adición de *es*, grupo sp2, se deben tener en cuenta posibles modificaciones para las palabras que llevan tilde. Se tienen los casos siguientes:

camión → *camiones*
joven → *jóvenes*
régimen → *regímenes*

Se resolvió no considerar estas variaciones dentro del modelo, por no formar parte de la descripción clásica de formación del plural de los nombres [1,2,6]. Este punto se resuelve en la implementación del procedimiento para la generación de las formas flexivas (sección 4.1).

[2] Según la bibliografía consultada, es común considerar al nombre como perteneciente a un género determinado [1,6], existen sin embargo algunos textos donde se incluye el género dentro del paradigma flexivo de los nombres [2].

[3] Las formas plurales de nombres de este tipo pueden usarse con sentido figurado [2]. Si solamente se pretende reconocer palabras en un texto dado, podrían incluirse estos plurales sin que esto introdujera errores de procesamiento.

En la siguiente tabla se muestran diferentes nombres con sus grupos morfológicos correspondientes.

Grupo morfológico	Ejemplo
n_sp1_m	*perro/perros*
n_sp1_f	*perra/perras*
n_sp1_c	*dentista/dentistas*
n_sp2_m	*emperador/emperadores*
n_sp2_f	*misión/misiones*
n_sp3_f	*emperatriz/emperatrices*
n_sp3_m	*aprendiz/aprendices*
n_i_m	*el paracaídas/los paracaídas*
n_i_f	*la crisis/las crisis*
n_p_m	*alrededores*
n_p_f	*catacumbas*

3.2 Adjetivos

Para definir el modelo flexivo para los adjetivos se consideran los rasgos morfológicos género y número. Cada grupo morfológico de adjetivos es identificado con una etiqueta de tipo a*_número_género*, tal como se describió anteriormente.

Los valores posibles para *número* y *género* se detallan en las siguientes tablas.

Valores posibles para el número de los adjetivos			
Valor	**Descripción**		**Ejemplo**
spX	Existen dos formas diferenciadas para singular y plural. X indica cómo se genera el plural.		
	X=1	formación de plural por adición de *s*	*grande/grandes*
	X=2	formación de plural por adición de *es*	*acogedor/acogedores*
	X=3	formación de plural por sustitución de *z* por *ces*	*voraz/voraces*
	$X=X_1X_2$	para el masculino, formación del plural ídem a spX_1; para el femenino, formación del plural ídem a spX_2	(sp31) *andaluz/andaluces* *andaluza/andaluzas*
s	Sólo existe la forma singular.		*uno*
p	Sólo existe la forma plural.		*veinte*
i	Número invariable. La forma de singular y la forma de plural son iguales.		*demás*

Valores posibles para el género de los adjetivos			
Valor	**Descripción**		**Ejemplo**
mfX	Existen dos formas diferenciadas para masculino y femenino. X indica cómo se genera el femenino.		
	X=1	formación del femenino por sustitución de *o* por *a*	*bueno/buena*
	X=2	formación del femenino por adición de *a*	*ganador/ganadora*
	X=3	formación de femenino por sustitución de *os* por *as* (adjetivos sin forma singular)	*doscientos/doscientas*
m	No existe la forma femenina.		*párroco*
f	No existe la forma masculina.		*parturienta*
c	Género común. El adjetivo tiene valor femenino y masculino.		*inteligente*

Para los adjetivos que forman el plural por adición de *es,* grupo sp2, se hace la misma aclaración que en el caso de los nombres (sección 3.1): posibles variaciones en los tildes son resueltas por el procedimiento que genera las formas flexivas (sección 4.1).

En la siguiente tabla se muestran diferentes adjetivos con sus grupos morfológicos correspondientes.

Grupo morfológico	Ejemplo
a_sp1_mf1	*lindo/linda/lindos/lindas*
a_sp21_mf2	*colaborador/colaboradora/colaboradores/colaboradoras*
a_sp3_c	*feliz/felices*
a_p_mf3	*doscientos/doscientas*
a_sp1_m	*acetilsalicílico/acetilsalicílicos*
a_sp1_f	*nodriza/nodrizas*

4. Generación de formas flexivas

A partir del diccionario morfológico, se desarrolló un procedimiento que permite generar todas las formas flexivas correspondientes a cada palabra. Se obtuvo, de este modo, un diccionario de formas flexivas, cada una con su lema y una etiqueta que describe sus características morfológicas.

Para el siguiente fragmento del diccionario morfológico:

```
...
abominación:n_sp2_f
abonable:a_sp1_c
abonado:a_sp1_mf1
abonado:n_sp1_m
abonador:a_sp21_mf2
abonador:n_sp2_m
abonadora:n_sp1_f
...
```

Se generan las siguientes formas flexivas:

```
...
abominación abominación NCFS
abominaciones abominación NCFP
abonable abonable AQCS
abonables abonable AQCP
abonado abonado AQMS
abonados abonado AQMP
abonada abonado AQFS
abonadas abonado AQFP
abonado abonado NCMS
abonados abonado NCMP
abonador abonador AQMS
abonadores abonador AQMP
abonadora abonador AQFS
abonadoras abonador AQFP
abonador abonador NCMS
abonadores abonador NCMP
abonadora abonadora NCFS
abonadoras abonadora NCFP
...
```

Las etiquetas del diccionario de formas flexivas siguen el formato del etiquetador FreeLing [10], el cual resultó apropiado para este trabajo. Este formato es fácilmente sustituible según el uso que se pretenda darle al diccionario.

La etiqueta para nombres está formada por cuatro letras mayúsculas:

N -> categoría nombre
C -> común (en contraposición a propio)
M|F|C -> género (masculino, femenino o común)
S|P|N -> número (singular, plural o invariable)

La etiqueta para adjetivos está formada por cuatro letras mayúsculas:

A -> categoría adjetivo
Q -> calificativo (no se trabajó con otro tipo de adjetivo)
M|F|C -> género (masculino, femenino o común)
S|P|N -> número (singular, plural o invariable)

El programa se implementó en Prolog y procede del siguiente modo:

- Se lee una entrada de tipo **palabra:categoría_número_género** del diccionario morfológico (archivo de entrada).

- Para los nombres, se genera el plural según la información contenida en **número** (si es de la forma spX). Tanto para la forma singular como para la forma plural, se mantiene el género representado por el elemento **género**.
  Por ejemplo, para **camión:n_sp2_m** se genera *camiones* de género masculino; para **orden:n_sp2_m** se genera *órdenes* de género masculino (*el orden*); para **orden:n_sp2_f** se genera *órdenes* de género femenino (*la orden*). Para **catacumbas:n_p_f** no se genera ninguna forma nueva.

- Para los adjetivos, en un primer paso se genera el femenino según la información contenida en **género** (si es de la forma mfX). En un segundo paso se generan los dos plurales según la información contenida en **número**.
  Por ejemplo, para **feliz:a_sp3_c** se genera sólo el plural *felices*; para **conservador:a_sp21_mf1** se generan el femenino *conservadora* y los dos plurales *conservadores* y *conservadoras*.

- Se escriben todas las formas flexivas generadas para la entrada leída, junto con el lema y la etiqueta asociados a cada una, en el diccionario de formas flexivas (archivo de salida).

- Se vuelve al primer paso.

4.1 Consideraciones respecto del plural formado por adición de "es"

Para las palabras que forman el plural por adición de *es*, grupo sp2, hay que tener en cuenta la presencia de tildes, ya sea en la forma singular, en la forma plural o en ambas. Para resolver este punto se tuvieron en cuenta las siguientes consideraciones.

- Las palabras que forman el plural en forma similar a *camión*, es decir, por adición de *es* al final y eliminación de tilde sobre la última vocal, son identificadas por terminar en vocal con tilde + letra *n* o letra *s*. La forma plural es generada sustituyendo la vocal con tilde y la *n* o la *s* por la vocal sin tilde + *nes* o *ses* (*tobog-án/tobog-anes, and-én/and-enes, piol-ín/piol-ines, avi-ón/avi-ones, at-ún/at-unes, comp-ás/comp-ases, rev-és/rev-eses*).

- Las palabras que forman el plural en forma similar a *joven*, es decir, por adición de *es* al final y transformación a esdrújula, son identificadas por terminar en *en* (sin tilde sobre la e) y tener al menos dos sílabas. La última característica se evalúa en forma aproximada, exigiendo que tengan un largo mayor a 4 letras (este número se fijó a partir de los ejemplos observados). De este modo se incluyen dentro de este grupo palabras como *aborigen/aborígenes*, *orden/órdenes*, pero se excluyen palabras monosilábicas como *tren/trenes*, *sien/sienes*.

- Las palabras que forman el plural en forma similar a *régimen*, es decir, por adición de *es* al final y corrimiento del tilde (*regímenes*), son identificadas por terminar en *en* y tener tilde sobre la antepenúltima vocal.

- Las palabras restantes que llevan el modelo sp2, forman el plural agregando *es* al final, sin otras consideraciones: *ombú/ombúes*, *director/directores*.

5. Aplicaciones

Las herramientas generadas dentro del marco de este proyecto tienen aplicación en la gran mayoría de los tratamientos que pueden realizarse sobre textos.

El análisis morfológico de las palabras de un texto suele ser uno de los pasos básicos de las aplicaciones de procesamiento de lenguaje natural. En particular, el diccionario, incorporado al etiquetador FreeLing, será utilizado en varios proyectos del grupo PLN, entre otros, la plataforma generada como parte del proyecto Clatex de segmentación de textos en proposiciones.

Además, el etiquetador integra el conjunto de herramientas que se ofrece a los estudiantes de los cursos que dicta el grupo, para desarrollar pequeños proyectos.

En el apartado que sigue, se describe el proceso de incorporación del nuevo diccionario al etiquetador FreeLing.

5.1 Enriquecimiento del diccionario del etiquetador FreeLing

5.1.1 Ajuste del diccionario de formas flexivas

El diccionario del etiquetador, además de contener nombres, adjetivos y verbos, contiene las restantes categorías gramaticales. Por lo tanto, este diccionario no puede ser sustituido; es necesario incorporar los elementos nuevos al diccionario original. Para esto se ajustó el formato del diccionario generado en este proyecto al formato requerido por el etiquetador. Esto implicó agrupar todas las etiquetas posibles para una misma palabra en una sola línea, ya que el diccionario del etiquetador no puede contener entradas duplicadas.

Para el siguiente extracto del diccionario de formas flexivas:

```
...
abonado abonado AQMS
abonados abonado AQMP
abonada abonado AQFS
abonadas abonado AQFP
abonado abonado NCMS
abonados abonado NCMP
abonador abonador AQMS
abonadores abonador AQMP
abonadora abonador AQFS
abonadoras abonador AQFP
abonador abonador NCMS
abonadores abonador NCMP
abonadora abonadora NCFS
abonadoras abonadora NCFP
...
```

Se agruparon algunas líneas del siguiente modo:

```
...
abonado abonado AQMS abonado NCMS
abonados abonado AQMP abonado NCMP
abonada abonado AQFS
abonadas abonado AQFP
abonador abonador AQMS abonador NCMS
abonadores abonador AQMP abonador NCMP
abonadora abonador AQFS abonadora NCFS
abonadoras abonador AQFP abonadora NCFP
...
```

El ajuste se llevó a cabo luego de unir los dos diccionarios ya que algunas de las palabras nuevas están incluidas en el diccionario original, generándose así entradas duplicadas. Por ejemplo, la forma verbal *canto*, incluida en el diccionario del etiquetador, no puede incorporarse como nombre en una línea diferente; el lema y la etiqueta correspondientes a esa segunda acepción deben ser agregados en la misma línea donde se encuentra la forma verbal.

Para realizar este ajuste de formato se unieron y ordenaron los dos diccionarios con comandos Unix (cat y sort –u) y luego se procesó el diccionario completo con un programa en Prolog que efectuó la agrupación de entradas iguales en una misma línea.

5.1.2 Evaluación de la aplicación del etiquetador usando el nuevo diccionario

Se evaluó el efecto de la incorporación de nombres y adjetivos al nuevo diccionario en base a los resultados de su aplicación sobre 5 textos[4] de un tamaño promedio de 1400 palabras.

Se trabajó con los siguientes conjuntos de archivos:
C1: archivos etiquetados usando el diccionario original (diccionario D1).
C2: archivos etiquetados usando el diccionario completo (diccionario D2).
C3: archivos con las palabras de cada texto que no se encuentran en el diccionario D1 (sólo adjetivos y nombres).
C4: archivos con las palabras de C3 que sí se encuentran en el diccionario D2.
C5: archivos con las diferencias entre los archivos de C1 y los archivos de C2.

De los conjuntos C3 y C4 se evaluó el porcentaje de palabras inicialmente desconocidas que quedaron incluidas en el diccionario completo:

Texto	Cantidad de palabras	N y A no incluidos en D1	N y A no incluidos en D1 pero sí en D2
P01	1382	103	77
P02	511	28	25
P03	2370	108	89
P04	1338	47	38
P05	1506	80	65
Total	7107	366	294

[4] Los textos pertenecen al corpus CorIn [8], compuesto por artículos de prensa uruguayos.

El 80% de las palabras que inicialmente eran desconocidas son reconocidas usando la nueva versión del diccionario.

Los archivos de C5 muestran las diferencias entre los dos etiquetados. Estas diferencias responden a varios motivos:

- Palabras que no están en el diccionario D1 pero sí están en el diccionario D2, para las cuales el etiquetado con D1 asigna una etiqueta probable que resulta incorrecta y el etiquetado con D2 asigna una etiqueta correcta.

Etiquetado usando D1	Etiquetado usando D2
¿ ¿ Fia quiénes quién PT0CP000 habitaban habitar VMII3P0 entonces entonces RG el el DA0MS0 **litoral litoral RG**[5] ? ? Fit	¿ ¿ Fia quiénes quién PT0CP000 habitaban habitar VMII3P0 entonces entonces RG el el DA0MS0 **litoral litoral NCMS000** ? ? Fit

- Palabras que no están en el diccionario D1 pero sí están en el diccionario D2, para las cuales el etiquetado con D1 asigna una etiqueta probable que resulta correcta pero incompleta (se pierde, por ejemplo, información sobre género y número) y el etiquetado con D2 asigna la etiqueta completa.

Etiquetado usando D1	Etiquetado usando D2
en en SPS00 un uno DI0MS0 área área NCFS000 de de SPS00 cinco cinco DN0CP0 **quilómetros quilómetros NC00000**	En en SPS00 Un uno DI0MS0 área área NCFS000 de de SPS00 cinco cinco DN0CP0 **quilómetros quilómetro NCMP000**

- Palabras que están en el diccionario D1 (y, por lo tanto, en D2), para las cuales el etiquetado con D1 asigna mal la categoría; se realiza mal la desambiguación por tener un conjunto incompleto de categorías posibles. En el ejemplo que sigue, *cubierta* está en D1 pero falta la etiqueta correspondiente a la categoría adjetivo.

Etiquetado usando D1	Etiquetado usando D2
estaba estar VMII1S0 totalmente totalmente RG **cubierta cubierta NCFS000** por por SPS00 diseños diseño NCMP000	Estaba estar VMII1S0 totalmente totalmente RG **cubierta cubierto AQ0FS0** por por SPS00 diseños diseño NCMP000

- Palabras que están en el diccionario D1 (y, por lo tanto, en D2), para las cuales el etiquetado con D1 asigna mal la categoría; se realiza mal la desambiguación dentro de la lista de categorías posibles por errores en categorías cercanas (palabras que no están en D1). En el ejemplo que sigue, la palabra *quilómetros* no está en D1 por lo que se le asigna mal la categoría (adverbio), este error repercute en el etiquetado de *que* para el cual se asigna la categoría conjunción (CS) en vez de pronombre (PR0CN000).

[5] La etiqueta RG significa adverbio, ver [10].

Etiquetado usando D1	Etiquetado usando D2
Los el DA0MP0 arqueólogos arqueólogos RG recorren recorrer VMIP3P0 todos todo DI0MP0 los el DA0MP0 días día NCMP000 los el DA0MP0 50 50 Z **quilómetros quilómetros RG** **que que CS** separan separar VMIP3P0 el el DA0MS0 sitio sitio NCMS000 de de SPS00 Salto Salto NP00000	Los el DA0MP0 arqueólogos arqueólogo NCMP000 recorren recorrer VMIP3P0 todos todo DI0MP0 los el DA0MP0 días día NCMP000 los el DA0MP0 50 50 Z **quilómetros quilómetro NCMP000** **que que PR0CN000** separan separar VMIP3P0 el el DA0MS0 sitio sitio NCMS000 de de SPS00 Salto Salto NP00000

Si bien la utilización del diccionario D2 permitió reducir la cantidad de errores, se introdujeron algunos errores nuevos, es decir, a algunas palabras que son etiquetadas en forma correcta usando D1 se les asigna una categoría equivocada usando D2. En algunos casos, se agregaron categorías nuevas para palabras ya existentes en D1, y estas nuevas categorías produjeron problemas de desambiguación. Errores de este tipo fueron procesados por la demo on-line de la página de FreeLing [11] a fin de verificar si tienen relación con el hecho de haber agregado elementos al diccionario. A continuación se muestra un ejemplo.

Etiquetado usando D1	Etiquetado usando D2
Cerámicas cerámica NCFP000 , , Fc puntas punta NCFP000 de de SPS00 flecha flecha VMN0000 , , Fc mortero mortero RG , , Fc	**Cerámicas cerámico AQ0FP0** , , Fc puntas punta NCFP000 de de SPS00 flecha flecha NCFS000 , , Fc mortero mortero NCMS000 , , Fc

Aplicación de la Demo:

Análisis morfológico generado por la Demo:

Analysis Results

Cerámicas	*cerámica* NCFP000 0.458711	*cerámico* AQ0FP0 0.521418	*Cerámicas* NP00000 0.0199823
,	, Fc 1		
puntas	*punta* NCFP000 0.578193	*puntar* VMIP2S0 0.00809969	*punto* AQ0FP0 0.413707
de	*de* NCFS000 9.91473e-05	*de* SPS00 0.999901	
flecha	*flecha* NCFS000 0.666667	*flechar* VMIP3S0 0.166667	*flechar* VMM02S0 0.166667
,	, Fc 1		
mortero	*mortero* NCMS000 1		
,	, Fc 1		

En este texto, la palabra *Cerámicas* es un nombre, por lo que el etiquetado usando D1 es correcto, mientras que los resultados de la utilización de D2 y la demo son incorrectos.

El análisis morfológico realizado por la Demo permite ver el conjunto completo de etiquetas disponibles para la palabra *Cerámicas*, sin aplicar el procedimiento de desambiguación. Este conjunto es igual al que contiene el diccionario D2. Podemos concluir que el error en el etiquetado se debe al método de desambiguación que se aplica, es decir, depende de la forma en que se elige la etiqueta apropiada para cada palabra según el contexto de ocurrencia.

6. Conclusiones

Se construyó el diccionario morfológico tal como se había planteado, el cual constituye un recurso de utilización inmediata en diferentes trabajos del grupo PLN.

Este diccionario, en conjunto con el procedimiento que genera las formas flexivas de todas las palabras incluidas en él, permitió la generación de un diccionario de formas flexivas que ya fue incorporado a un software de uso frecuente del grupo, mejorando notoriamente los resultados de su aplicación.

Se prevén diversas mejoras y extensiones a este trabajo, que fueron tenidas en cuenta entre las tareas planificadas por el grupo para el año próximo.

Como primer paso, es necesario incorporar al diccionario morfológico todas las formas flexivas correspondientes a los infinitivos del diccionario de lemas con el que cuenta el grupo.

En cuanto al modelo morfológico definido, se debería contemplar la coexistencia de diferentes formas de plural para una misma palabra. Por ejemplo, el adjetivo *maravedí* permite tres variantes para el plural: *maravedís/maravedíes/marvedises* [2]. Muchas palabras provenientes de otros idiomas presentan esta peculiaridad: suelen admitir diferentes formas de plural.

Este proyecto se propuso obtener un resultado en un corto plazo, por esta razón se trabajó con las listas de palabras originales en bruto. Será necesario trabajar en la depuración del diccionario haciendo una revisión de las palabras incluidas en él y de la asignación de los grupos morfológicos.

Bibliografía

[1] Alarcos Llorach, Emilio. *Gramática de la lengua española.* Madrid: Espasa Calpe, 1994.

[2] Ambadiang, Théophile. *La flexión nominal. Género y número.* En: Bosque, Ignacio y V. Demonte, *Gramática Descriptiva de la Lengua Española.* Madrid: Espasa Calpe, 1999. Cap. 74, págs. 4844-4913.

[3] Lacuesta, Ramón S. y Eugenio Bustos. *La derivación nominal.* En: Bosque, Ignacio y V. Demonte, *Gramática Descriptiva de la Lengua Española.* Madrid: Espasa Calpe, 1999. Cap. 69, págs. 4505-4594.

[4] Pena, Jesús. *Partes de la morfología. Las unidades del análisis morfológico.* En: Bosque, Ignacio y V. Demonte, *Gramática Descriptiva de la Lengua Española.* Madrid: Espasa Calpe, 1999. Cap. 66, págs. 4305-4366.

[5] Rainer, Franz. *La derivación adjetival.* En: Bosque, Ignacio y V. Demonte, *Gramática Descriptiva de la Lengua Española.* Madrid: Espasa Calpe, 1999. Cap. 70, págs. 4595-4643.

[6] Real Academia Española. *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa Calpe, 1973.

Corpus utilizados

[7] Corpus de Referencia del Español Actual de la Real Academia Española (CREA) [en línea]
http://www.rae.es [consulta: febrero de 2005].

[8] Couto, Javier, M. Grassi, M. Malcuori, J. J. Prada y D. Wonsever. *Corpus informatizado: textos del español del Uruguay (CORIN)*. En *SLPLT-2 - Second International Workshop on Spanish Language Processing and Language Technologies.* Jaén, 2001.

Referencias correspondientes al etiquetador FreeLing

[9] Página principal de FreeLing [en línea]
http://garraf.epsevg.upc.es/freeling/ [consulta: febrero de 2005].

[10] Documentación de FreeLing, descripción de etiquetas [en línea]
http://www.lsi.upc.es/~nlp/freeling/parole-es.html [consulta: febrero de 2005].

[11] Demo on-line de FreeLing [en línea]
http://www.lsi.upc.es/~nlp/freeling/demo.php [consulta: febrero de 2005].

Apéndice: Género de los nombres y adjetivos terminados en letra vocal + z

Nombres

Los nombres terminados en *-az, -ez, -iz, -oz, -uz* pertenecen a un mismo grupo morfológico, según el modelo del cual se partió para este trabajo. Si bien todas estas palabras tienen el mismo comportamiento en cuanto al número, es decir, forman el plural del mismo modo (sustitución de *z* por *ces*), no todas tienen el mismo género. La información de género no es brindada por el modelo y resulta fundamental para este trabajo.

A continuación se hace un breve análisis del género de estos nombres según las diferentes terminaciones.

Nombres en *-az*

El diccionario de lemas contiene 27 nombres terminados en *-az*. La mayoría son de género masculino: *antifaz, capataz, haz, matraz*; algunos pocos son de género femenino: *interfaz, paz*.

Algunos son adjetivos usados como nombres, en estos casos tienen género común: un (hombre) *rapaz*, una (mujer) *rapaz*.

Nombres en *-ez*

El diccionario contiene 27 nombres terminados en *-ez*. La gran mayoría son nombres derivados, formados por adición del sufijo *-ez* a un adjetivo base (*acidez, aridez, candidez*). Este sufijo genera nombres de género femenino, a menos de *doblez* que tiene género ambiguo [3].

Algunos nombres terminados en *-ez* no son derivados, por lo tanto, no puede anticiparse su género: *tez* es femenino, *ajedrez* es masculino.

Algunos nombres en *-ez* de género masculino son: *ajedrez, alférez, diez, jaez, jerez, juez, pez.*

Nombres en *-iz*

El diccionario contiene 49 nombres terminados en *-iz*. Muchos de ellos son de género femenino, como *actriz, adoratriz, institutriz*, pero muchos son de género masculino, como *aprendiz, desliz, tapiz*. Algunos son adjetivos usados como nombres y tienen género común como *infeliz*.

Nombres en *-oz* y *-uz*

El diccionario contiene 13 nombres terminados en *-oz* y 20 terminados en *-uz*. Tienen género variado: *hoz, voz, luz* femeninos y *arroz, portavoz, trasluz* masculino.

Para los nombres en *-ez* y en *-iz*, se consideró conveniente asignar el género femenino a todos y luego cambiarlo manualmente para los (pocos) nombres masculinos. En los restantes casos mencionados, es necesario hacer una recorrida "manual" para asignar el género a cada nombre.

Adjetivos

Los adjetivos terminados en vocal+z como *feliz, asaz, atroz*, tienen género común, por lo que la asignación del modelo para generación del femenino no introduce ninguna dificultad. De los adjetivos que se encuentran en el diccionario, el único que varía según el género es *andaluz/andaluza*.