Proyecto COWS-L2H

(Corpus of Written Spanish of L2 and Heritage Speakers)

1. Presentación del proyecto

El equipo de investigación ha recopilado un corpus de textos escritos por estudiantes de español como segunda lengua de diferentes niveles. En este momento, el corpus contiene textos de más de 2000 estudiantes, que suman casi 900 000 palabras.

Entre los objetivos que persigue la elaboración de este corpus están los siguientes:

1. Desarrollar un programa de corrección automática de errores, para lo que se necesitan versiones corregidas de cada texto del corpus que sirvan de entrenamiento al programa.
2. Determinar patrones en la evolución de determinados errores presentes en las composiciones escritas de los estudiantes. Para ello es necesario etiquetar en cada texto dichos errores.

Estos objetivos han motivado la participación en este proyecto de investigación. Así, la labor del equipo de trabajo de la USAL consistirá en (1) proporcionar versiones corregidas de cada texto y (2) etiquetar los errores seleccionados que aparecen en cada texto.

2. Plan de trabajo

A cada miembro del equipo se le asignará una serie de textos en formato txt. Los textos que recibiréis estarán nombrados según el esquema siguiente:

> Clave del informante_Tema del texto.txt

Por ejemplo:

> 157048.S17_Vacation.txt

A partir de cada texto el revisor tendrá que realizar tres tareas:

1. crear un nuevo documento en Word con control de cambios con el texto corregido. Este texto deberá guardarse añadiendo las palabras "control cambios":
   157048.S17_Vacation.control cambios.docx
2. crear un nuevo documento txt con una versión corregida del texto. Este texto deberá guardarse añadiendo la palabra "corrected":
   157048.S17_Vacation.corrected.txt
3. crear un nuevo documento txt con el texto original etiquetado. Este texto deberá guardarse añadiendo la palabra "annotated":
   157048.S17_Vacation.annotated.txt

Se creará un calendario para la asignación de textos y la entrega de las nuevas versiones.

Cada miembro del equipo tendrá acceso a una carpeta personal en Dropbox en la que encontrará tres subcarpetas:

1. Textos originales. Aquí estarán disponibles los textos originales que se vayan asignando.

2. Textos corregidos. Aquí deberán colgarse, por cada texto corregido, las dos versiones señaladas arriba: una en docx con el texto corregido con control de cambios y otra en txt solo con el texto corregido.
3. Textos anotados. Aquí deberán colgarse los textos etiquetados.

2.1. Instrucciones para la corrección de textos (tarea 1)

Los textos se intervendrán lo menos posible. La corrección se limitará a errores ortográficos, gramaticales y léxicos evidentes, y se evitará cambiar cuestiones de estilo, aun cuando la construcción no resulte totalmente natural para un hablante nativo. Véanse los ejemplos de textos corregidos.

2.2. Instrucciones para el etiquetado de errores (tarea 2)

Con independencia de los errores corregidos en la tarea 1, en esta tarea 2 solo se van a etiquetar **5 tipos de errores**:

1. Errores de concordancia de género y/o de número (*Agreement errors*)
2. Atribución indebida del género o del número (*Attribution errors*)
3. Presencia o ausencia indebida de pronombres de sujeto o de artículos (*Presence / Absence errors*)
4. Confusión de preposiciones o de los verbos *ser* y *estar* (*Exchange errors*)
5. Errores en la colocación de los adjetivos (*Placement errors*)

Se recomienda tener abierto el documento con el texto corregido previamente para mantener la coherencia en el proceso de etiquetado, pero identificando y etiquetando solo aquellos errores que entran en uno de estos 5 tipos. El resto de errores que se hayan corregido en la tarea 1 pero no pertenezcan a estos 5 tipos se dejarán como estaban en el texto original, sin etiquetar de ninguna manera.

Los errores de los 5 tipos señalados deberán marcarse de acuerdo con el siguiente **formato**:

[] = error escrito por el estudiante

{} = forma correcta, proporcionada por el revisor

<> = etiquetas de anotación de errores

Así, todos los errores identificados en el texto se dispondrán de la siguiente manera: [error del estudiante]{forma corregida}<etiquetas de error>. No deben dejarse espacios entre el corchete de cierre y la llave de apertura (]{) ni entre la llave de cierre y el ángulo de apertura (}<). Si tras las etiquetas del error aparece uno nuevo, entonces sí debe dejarse un espacio:

[error1]{corrección}<etiquetas> [error 2]{corrección}<etiquetas>

- Solo la palabra que contiene el error debe ser etiquetada (no todo el sintagma).

- Si el error es una omisión, los corchetes deben estar vacíos (sin ningún símbolo o espacio entre ellos): []

- Si el error es una palabra sobrante, las llaves deben estar vacías (sin ningún símbolo o espacio entre ellas): {}

2.2.1. Etiquetas de los 5 tipos de errores

1. Errores de concordancia de género y/o de número (*Agreement errors*)

- ga = gender agreement error that affect determiners (*unos niñas pequeñas*), adjectives (*una ventana pequeño*) and, in some cases, nouns (*El chico es una buena jugadora*)
- na = number agreement error (*ventanas grande*)
- ga:na = gender agreement error + number agreement error

Estas etiquetas deben ir seguidas, tras dos puntos sin espacio, de la etiqueta correspondiente a la clase de palabra afectada por el error:

- det = error associated with determiners
- adj = error associated with adjective
- noun = error associated with noun
- pron = error associated with pronoun
- adv = error associated with adverb

Por ejemplo:

El chico es [una]{un}<ga:det> [buena]{buen}<ga:adj> [jugadora]{jugador}<ga:noun>.

- En estas anotaciones, las palabras *una buena jugadora* están etiquetadas con <ga> porque hay un problema de concordancia entre *una buena jugadora* y el sujeto *El chico*.

Se consideran errores de concordancia y, por tanto, deben marcarse con las etiquetas <ga> y <na>, los errores de género y número en determinantes y adjetivos en los que sea evidente que el género o el número del sustantivo está bien atribuido, como, por ejemplo, en *una ventana pequeño*. Este error se etiquetaría, por tanto, así:

una ventana [pequeño]{pequeña}<ga:adj>

2. Errores de atribución indebida de género o número (*Attribution errors*)

- gat = gender attribution error (*una problema, un problemo, un problemo pequeño, el persono, el gran casa rojo*…)
- in = number invention (*la vacación*)

Estas etiquetas también deben ir seguidas, tras dos puntos sin espacio, de la etiqueta correspondiente a la clase de palabra afectada por el error:

- det = error associated with determiners
- adj = error associated with adjective
- noun = error associated with noun
- pron = error associated with pronoun
- adv = error associated with adverb

Aparte de algunos adverbios (*Me gusta mucha*), los errores de atribución de género siempre afectan al sustantivo, aunque esto tenga consecuencias en los determinantes o adjetivos con los que concuerde. Por ejemplo, si alguien escribe *Vivo en el gran casa rojo*, el error solo se produce en el sustantivo *casa*, al que se le asigna un género incorrecto (masculino, cuando debería ser femenino), porque la concordancia del determinante (*el*) y el adjetivo (*rojo*) es, en realidad, "correcta", teniendo en cuenta la atribución de género incorrecta del sustantivo.

Básicamente, si hay evidencia (por lo menos 2 determinantes o adjetivos que concuerdan entre sí pero no con el sustantivo) de que el aprendiz piensa que el sustantivo es del género incorrecto, hay que aplicar la etiqueta <gat>. Por ejemplo:

Yo vivo en [el]{la}<gat:det> gran casa<gat:noun> [rojo]{roja}<gat:adj>

El artículo y el adjetivo se etiquetan como <gat> porque muestran el error de atribución de género del sustantivo, pero la concordancia es correcta. La etiqueta <ga> no debe aplicarse cuando las palabras erróneas concuerdan entre sí. En este sintagma el error está en la atribución de género del sustantivo (*casa*), pero la concordancia de determinantes y adjetivos es coherente.

Todas las palabras inventadas con nuevos "morfemas de género" (*persono, artisto*...) se consideran errores de atribución de género. Si los determinantes o adjetivos del nombre concuerdan con él en ese género incorrecto, se utiliza la etiqueta de atribución (gat). Si no, la de concordancia (ga). Ejemplos:

[el]{la}<gat:det> [persono]{persona}<gat:noun>

[el]{la}<gat:det> [persone]{persona}<gat:noun>

[la]{la}<ga:det> [persono]{persona}<gat:noun>

[Las]{Los}<gat:det> [estereotipicas]{estereotipos}<gat:noun>

[Los]{Los}<ga:det> [estereotipicas]{estereotipos}<gat:noun>

La etiqueta <in> identifica errores en la interpretación de singulares o plurales especiales. Los determinantes y adjetivos referidos al nombre con un error en la atribución del número llevan esta etiqueta siempre que no haya un problema de concordancia. Ejemplos:

[La]{Las}<in:det> [vacación]{vacaciones}<in:noun> [perfecta]{perfectas}<in:adj>

[La]{Las}<in:det> [vacación]{vacaciones}<in:noun> [perfecto]{perfectas}<ga:in:adj>

[El]{Las}<gat:in:det> [vacación]{vacaciones}<gat:in:noun>
[perfecta]{perfectas}<ga:in:adj>

[El]{Las}<gat:in:det> [vacación]{vacaciones}<gat:in:noun> [perfecto]{perfectas}<gat:in:adj>

Se presentan a continuación otros ejemplos con errores de género y de número, tanto de concordancia como de atribución:

Ella es [simpático]{simpática}<ga:adj>

[Lo]{La}<ga:pron> vi tres veces.

Juan es [un]{una}<ga:det> persona fantástica.

Me gusta [mucha]{mucho}<gat:adv> la playa.

Hay [muchas]{muchos}<gat:det> programas[programas]{programas}<gat:noun>.

McAdams es muy simpática con [su]{sus}<na:det> amigos.

3. Presencia o ausencia indebida de pronombres de sujeto o de artículos (*Presence / Absence errors*)

La primera etiqueta indica el **tipo de error:**

pr = *presence*: presencia de un elemento que no debería aparecer.

ab = *absence*: ausencia de un elemento necesario.

A continuación se coloca la etiqueta del **elemento** que sobra o que falta:

su:pron = pronombre sujeto

det:art = determinante artículo

Tras la etiqueta <det:art>, separada por dos puntos sin espacio, debe aparecer una de las dos etiquetas siguientes:

def = para *el, la, los, las*

indef = para *un, una, unos, unas*

Ejemplos:

Johnny Depp es un actor famoso. [Él]{}<pr:su:pron> nació en 1963 en Kentucky.

[]{El}<ab:det:art:def> Señor Sandler es muy feliz.

trabajaba como [un]{}<pr:det:art:indef> ingeniero

4. Confusión de preposiciones o de los verbos *ser* y *estar* (*Exchange errors*)

La primera etiqueta indica el **tipo de error:**

conf = confusión

A continuación se coloca la etiqueta de los **elementos** que se han intercambiado:

prep:popa = preposiciones *por* y *para*

v:seta = verbos *ser* y *estar*

Ejemplos:

Te doy gracias [para]{por}<conf:prep:popa> todo.

[Soy]{Estoy}<conf:v:seta> cansado de estudiar.

[Estaba]{Era}<conf:v:seta> un hombre feliz.

5. Errores en la colocación de los adjetivos (*Placement errors*)

La primera etiqueta indica el **tipo de error:**

col = colocación

A continuación se coloca la etiqueta del **elemento** mal colocado:

adj = adjetivo

Ejemplos:

Es mi [favorito papel]{papel favorito}<col:adj>.

2.2.2. Resumen del etiquetado de errores