Propuesta de una clasificación general y dinámica para la definición de errores

Itziar Aldabe, Bertol Arrieta, Arantza Díaz de Ilarraza, Montse Maritxalar, Maite Oronoz, Larraitz Uria

En este trabajo presentamos una clasificación dinámica que ha sido definida con el objetivo de almacenar y clasificar errores. En concreto, los datos recopilados nos sirven como punto de partida para estudiar el proceso de aprendizaje del euskera y para llevar a cabo investigaciones en diversos campos de estudio como el Análisis de Errores (AE) y el Procesamiento del Lenguaje Natural (PLN). La clasificación que aquí presentamos se encuentra integrada en un conjunto de herramientas de PLN desarrolladas en el grupo IXA (EHU-UPV): por un lado, en la base de datos DESBIDERATZEAK, diseñada para recopilar información sobre el proceso de aprendizaje del euskera dentro del área del Aprendizaje y Enseñanza de Lenguas Asistido por Ordenador; y por otro, en la base de datos ERROREAK, creada para el estudio del tratamiento automático de errores con la idea de desarrollar un corrector gramatical y de estilo para la lengua vasca.

Palabras clave: Clasificación de errores relacionados con el euskera; corpus de errores; aprendizaje y enseñanza de lenguas asistido por ordenador; tratamiento automático de errores.

In this work we present a dynamic classification defined to store and classify errors. The collected information will be the starting point for the study of Basque language learning process and for research on different fields such as Error Analysis (EA) and Natural Language Processing (NLP). This error classification is integrated in some NLP tools developed in the IXA research group at the University of the Basque Country. On the one hand, the classification is the basis of the DESBIDERATZEAK database, designed to gather information concerning the Basque learning process within the Computer Assisted Language Learning field. On the other hand, it is also the basis of the ERROREAK database, created for the study of automatic error treatment for the development of a grammar and style checker for Basque.

Key words: Error classification for Basque; error corpora; computer assisted language learning and teaching; automatic error treatment.

INTRODUCCIÓN

Las investigaciones llevadas a cabo en torno al Análisis de Errores (AE) y el tratamiento automático de errores han adquirido una gran relevancia debido a los resultados obtenidos en los últimos años dentro del Aprendizaje y Enseñanza de Lenguas (AEL). Nosotros pensamos contribuir con aportaciones interesantes, ya que estamos avanzando en la creación de un entorno global, compuesto por varias herramientas lingüístico-computacionales, para el aprendizaje y la enseñanza de lenguas asistido por ordenador; y paralelamente, continuamos trabajando también en la creación de un corrector gramatical y de estilo para el euskera. Partiendo de estos dos objetivos, hemos tipificado una clasificación de errores lingüísticos que va a ser el punto de partida de nuestro trabajo.

Esta clasificación de errores se concreta básicamente en dos bases de datos:

- DESBIDERATZEAK es un producto diseñado principalmente para profesores que imparten la enseñanza de la lengua vasca, con la pretensión de recopilar las desviaciones y la información psicolingüística de sus alumnos, y al mismo tiempo realizar una diagnosis de la(s) interlengua(s) de los mismos;

- ERROREAK, por su parte, es la base de conocimiento necesario para desarrollar un corrector gramatical y de estilo para el euskera; recopila información técnica y lingüística de los errores.

Aunque las pretensiones de ambas bases de datos no sean coincidentes, las dos tienen un objetivo común, que es la recopilación y la clasificación de los errores que se cometen en la producción escrita de la lengua vasca. Por lo tanto, la información recogida en DESBIDERATZEAK y en ERROREAK nos resultará imprescindible para poder llevar a cabo investigaciones sistemáticas en el campo del Aprendizaje y Enseñanza de Lenguas Asistido por Ordenador, así como en el tratamiento automático de errores. En este estudio nos centraremos fundamentalmente en el área del aprendizaje y la enseñanza de la lengua vasca asistido por ordenador, dejando para otra ocasión el tratamiento automático de los errores y el corrector gramatical de textos.

ANÁLISIS DE ERRORES

El Análisis de Errores fue una teoría muy criticada hasta la publicación, en 1967, de "*The Significance of Learners' Errors*" de S. Pit Corder, quien defendía que los estudios realizados en torno al análisis de errores mostraban resultados muy positivos y aportaciones interesantes para el estudio del aprendizaje y la enseñanza de lenguas. Por su parte, S. Fernández (1997) también afirma que "*existe una nueva concepción de los errores... los errores se valoran ahora, además de como paso obligado para llegar a apropiarse de la lengua, como índices del proceso que sigue el aprendiz en ese camino... las producciones 'incorrectas' del aprendiz de una L2 serían marcas, también, de los diferentes estadios del proceso de apropiación de la lengua*". Y de hecho, es cierto que clasificando errores e infiriendo las estrategias adoptadas por los aprendices de las lenguas, podemos profundizar en la problemática sobre los procesos que conllevan tanto el aprendizaje como la enseñanza de un idioma.

De todos modos, y a consecuencia de los importantes avances que han aportado las nuevas tecnologías en los últimos años, el Análisis de Errores Automatizado (AEA) ha superado con creces muchos de los límites del Análisis de Errores tradicional, y nos permite una investigación mucho más rápida, eficaz, controlada y sistemática. Por ejemplo, utilizando adecuadamente las posibilidades que nos ofrece el AEA, resulta posible efectuar diferentes estudios sobre las desviaciones en las que incurren normalmente los alumnos de lenguas y conocer, en la medida de lo posible:

- hasta que punto el alumno domina el idioma que está estudiando;
- el proceso de aprendizaje de una segunda lengua;
- cuáles son las principales dificultades de los alumnos para aprender un nuevo idioma;
- etc.

Conociendo de cerca estos factores, podremos adaptar nuestras herramientas lingüístico-computacionales a las necesidades didácticas de los alumnos y profesores; y ello nos permite también indagar, entre otros temas, cuáles pueden ser las herramientas de ayuda más idóneas para cada nivel de estudio de la lengua. Para todo esto, nos va a resultar imprescindible la recopilación y clasificación de las desviaciones cometidas por los alumnos. Tal y como afirmaba Cowan (2003), el análisis del corpus lingüístico puede ser válido para profundizar en el desarrollo de los programas de aprendizaje de lenguas asistido por ordenador, de forma que las evidencias negativas detectadas actúen positivamente para erradicar definitivamente los errores gramaticales persistentes. Dagneaux *et al.* (1998) y Granger (2002), por su parte, defienden que el Análisis de Errores Automatizado de los textos escritos por alumnos proporciona información inestimable a los investigadores de la Adquisición de Segundas Lenguas (ASL) y Enseñanza de Lenguas Extranjeras (ELE), ya que las nuevas perspectivas aportadas por el AEA se pueden incorporar perfectamente a los sistemas pedagógicos, en general, y a los programas de Aprendizaje de Lenguas Asistido por Ordenador, en particular.

CONCEPTO DE ERROR

En lo que se refiere al concepto de 'error', se emplean varios términos y definiciones en el ámbito objeto de esta investigación. En este estudio no pretendemos clarificar el alcance de este término. Pero teniendo en cuenta que la clasificación de errores que aquí presentamos se emplea en diferentes campos de investigación, consideramos importante definir y diferenciar estos tres conceptos básicos: "error", "desviación" y "estilo".

Según los criterios que nosotros hemos definido, un "error" es cualquier estructura u output no gramatical. Por otra parte, diremos que una estructura lingüística es una "desviación" cuando se cumpla alguna de estas tres condiciones (Maritxalar *et al.*, 1996):

- la estructura lingüística producida por el alumno es "incorrecta" (lo que habitualmente se ha llamado error);
- la estructura se utiliza con demasiada frecuencia, sustituyendo a otras estructuras más adecuadas que el alumno pretende evitar (ej: usar la conjunción '*eta' = 'y'* repetidamente en lugar de alternar con '*ordea' = 'sin embargo')*;
- el alumno evita una determinada estructura (ej: '*nor da?' = '¿quién es?'* en lugar de '*nor ote da?' = '¿quién será?')*.

Las desviaciones son parte de la interlengua de los alumnos y se producen bien sea porque los alumnos todavía no conocen esas estructuras o bien porque no las dominan. Por último, nos referimos al "estilo" para clasificar aquellas estructuras, producidas tanto por alumnos como por nativos de una lengua, que son gramaticales pero no siempre adecuadas (ej: la constante repetición del adverbio *'entonces'*, sin ir alternando sinónimos).

Así, no consideramos errores, sino desviaciones, los ejemplos clasificados en la base de datos DESBIDERATZEAK; por el contrario, en ERROREAK todos los ejemplos son considerados errores, ya que esta base de datos no representa el punto de vista psicolingüístico. En lo que se refiere al estilo, aunque estos fenómenos no se pueden considerar "errores", nos ha parecido oportuno incluir también este apartado en la clasificación de errores, ya que esos datos nos serán muy útiles en el futuro para abordar el desarrollo de un corrector de estilo para el euskera.

Así pues, consideramos muy importante esta especificación ya que en nuestras bases de datos guardamos, clasificamos, analizamos y tratamos las desviaciones en las que incurren los alumnos, los errores en general y la estilística de los textos.

CLASIFICACIÓN DE ERRORES

La clasificación que presentamos parte de una estructura dinámica y jerárquica y está basada, prácticamente, en *Euskal Gramatika Osoa* (Zubiri *et al.*, 1995). No obstante, hemos consultado una bibliografía muy exhaustiva para comparar diferentes clasificaciones gramaticales realizadas tanto para el euskera como para otras lenguas, y conocer, así, diversos modelos de categorización. Además, nos han sido de ayuda algunas clasificaciones realizadas previamente por el grupo IXA. Y también hemos tomado en cuenta tanto los contenidos gramaticales que se imparten en los diferentes niveles de aprendizaje de los Euskaltegis como los datos obtenidos en un estudio llevado a cabo por profesores de euskera, que analizaron y detectaron las desviaciones de los alumnos en textos reales.

Existen varias formas para constituir una clasificación de errores. Los errores pueden ser categorizados en función de la estructura superficial de las palabras/frases: omisión, adición, substitución y ordenamiento de las letras/palabras. Se pueden clasificar también por su apariencia: errores abiertos (que son obvios fuera de contexto) y errores encubiertos (evidentes solamente en un determinado contexto). Existen, además, clasificaciones basadas en niveles lingüísticos, tales como errores fonológicos, léxicos, sintácticos, etcétera.

Nosotros, a la hora de definir la taxonomía principal de la clasificación, consideramos muy positivo y apropiado la fusión de diferentes enfoques, es decir, la combinación de una clasificación lingüística y una clasificación en base a las características superficiales de los errores. Granger (2003), por ejemplo, también se mostró a favor de combinar interpretaciones diferentes en una misma clasificación: "*there are two major descriptive error taxonomies:*

- *one based on linguistic categories (general ones such as morphology, lexis, grammar, and more specific ones such as auxiliaries, passives, and prepositions)*.

- *the other focusing on the way surface structures have been altered by learners (e.g., omission, addition, misformation, and misordering).*
  There is a great benefit to combining them".

Y tomando de referencia estas afirmaciones, nosotros hemos aunado dos perspectivas distintas: la clasificación correspondiente a las características superficiales de los errores y la correspondiente al nivel lingüístico. De este modo, categorizamos los errores ortográficos en base a las características superficiales (omitir, añadir, sustituir, cambiar el orden de las letras...); los errores no-ortográficos, por su parte, los clasificamos en categorías lingüísticas generales (léxico, morfosintaxis, semántica...) y en subcategorías más específicas (verbo, pronombre, declinación, etc.). En la actualidad, la clasificación consta de 7 categorías principales y 155 subcategorías, de las cuales 129 corresponden a los nodos 'hoja' o extremidades del árbol.

Cada categoría/subcategoría consta de un código, la definición del código y, al menos, un ejemplo de ese tipo de error. De este modo, hemos elaborado una especie de árbol jerárquico: las categorías y subcategorías formarían los nodos intermedios del árbol; las hojas contendrían la información técnica y psicolingüística correspondiente a los errores o las desviaciones; y los errores y desviaciones serían ejemplos de las hojas (cuadro 1):

Cuadros 1 y 2: ejemplos de la estructura jerárquica de la clasificación.

A continuación, resumimos algunos criterios principales que hemos tenido que definir a la hora de determinar la jerarquía de la clasificación; criterios que consideramos importantes explicar para entender mejor la estructura de esta categorización:

- Primeramente, hemos diferenciado los errores ortográficos de los no ortográficos: consideramos errores ortográficos aquellas palabras que como unidad léxica son incorrectas y consideramos errores no ortográficos las estructuras que ortográficamente están bien escritas pero que son sintáctica y/o semánticamente incorrectas en un contexto determinado.

- Dentro de los errores ortográficos establecemos subcategorías como omitir, añadir, sustituir y cambiar el orden de las letras. Entre los errores no ortográficos distinguimos errores léxicos; errores morfológicos, sintácticos y morfosintácticos; errores de noción; errores semánticos; signos de puntuación; y errores de estilo.

- Los errores que computacionalmente tratamos a nivel léxico (errores como **haundi* → √*handi*, **laister* → √*laster*, **eritzi* → √*iritzi*) podrían ser clasificados como *Errores ortográficos* o *Errores léxicos*. De momento, hemos considerado más oportuno clasificar este tipo de errores dentro de los *Errores ortográficos* ya que los usuarios no disponen de un punto de vista computacional; y por consiguiente, les sería imposible conocer cuáles son las palabras que tratamos automáticamente a nivel léxico.

- Como ya disponemos de herramientas que detectan errores ortográficos, analizaremos los errores ortográficos recopilados para crear nuevas aplicaciones pedagógicas que sirvan a los alumnos para dominar la ortografía correctamente. La recopilación de errores morfosintácticos, por su parte, será una fuente de información esencial tanto para el desarrollo del corrector gramatical y de estilo, así como para la creación de nuevas aplicaciones en el campo del aprendizaje y enseñanza de lenguas asistido por ordenador.

- Tratamos los errores ortográficos a nivel de palabras y los no ortográficos, al contrario, a nivel de sintagma u oración, ya que en este caso el contexto resulta imprescindible. Por ejemplo, muchas veces ocurre que una palabra correctamente escrita sea un 'error' en un contexto determinado:

neskek etxea ikusi du

neskek, como unidad léxica, es una palabra correcta. Pero en esta frase, es sintácticamente incorrecta ya que el sujeto plural *neskek* necesitaría el auxiliar plural *dute* o, de otro modo, el auxiliar singular *du* requeriría el sujeto singular *neskak*.

- Tal y como lo preveíamos, se dan un tipo de errores que se pueden clasificar en más de una categoría, lo cual dificulta la tarea de categorización.

Guk geu

Éste es un ejemplo que podría ser clasificado en dos categorías: *Pronombres* o *Falta de concordancia dentro del Sintagma Nominal*. Hemos acordado que en estos casos sea el usuario quien decida que categoría asignar al error.

- También tuvimos algunas dudas en cuanto a la corrección de errores; tales como:

• Si se produce más de un error en una misma frase, ¿debemos corregir la frase en su conjunto o solamente el error que estamos clasificando?:

Tsakurra etorri du

1. error ortográfico: ***tsakurra***
¿Cómo corregir? Como los errores ortográficos se clasifican y se corrigen a nivel léxico: ***txakurra***

2. error sintáctico: mezcla de paradigma de los verbos: ***etorri du***
¿Cómo corregir?
- ***tsakurra etorri da*** (→ corregir solo el error sintáctico, dejando el error ortográfico tal cual)
- ***txakurra etorri da*** (→ corregir la frase completa, recalcando el error sintáctico)

• Si un error puede tener más de una corrección posible, ¿debemos presentar todas las posibilidades o con una sola sería suficiente?:

Katuak ikusi dute

1. tipo de error: falta de concordancia entre sujeto y verbo → Posibles correcciones:
- *Katu**ak** ikusi **du***
- *Katu**ek** ikusi **dute***

2. tipo de error: falta de concordancia entre objeto y verbo → Posibles correcciones:
- *Katu**ak** ikusi **dituzte***
- *Katu**ak** ikusi **ditu***

• Al corregir un error, ¿debemos reescribir toda la oración o es suficiente con corregir solamente la parte errónea?

Tsoria etorri du → ¿Cómo corregir?
corrección 1: **txoria etorri da**
corrección 2: **tsoria etorri da**
corrección 3: **etorri da**
corrección 4: **da**
corrección 5: **tsoria da**

La cuestión es que las dos bases de datos (DESBIDERATZEAK y ERROREAK) han sido creadas con fines diferentes: aprendizaje y enseñanza de la lengua vasca y corrector gramatical, respectivamente. Por tanto, los lingüistas que introduzcan errores en ERROREAK, y los profesores que introduzcan las desviaciones en DESBIDERATZEAK, no tendrán las mismas consideraciones y puntos de vista a la hora de corregir los errores. Además, la corrección suele ser algo muy subjetivo que depende también de otros factores (como el nivel de cada alumno, el tipo de ejercicio asignado, la importancia del error, etc.). Por lo tanto, resulta muy difícil adoptar criterios específicos en torno a la corrección de errores y, consecuentemente, consideramos que lo más acertado es ofrecer la oportunidad de facilitar todas las correcciones posibles.

De todos modos, y como ya hemos afirmado, la clasificación que presentamos es dinámica, lo cual nos permite la posibilidad de realizar cambios dependiendo de los ejemplos y los datos recogidos. Es, por tanto, un modelo general que se puede modificar, actualizar o adaptar a los objetivos y a las necesidades de cada usuario. Sin duda alguna, el dinamismo que ofrece la clasificación es una característica muy significativa.

Tal y como hemos mencionado anteriormente, esta clasificación de errores se alimenta de dos bases de datos, ERROREAK y DESBIDERATZEAK, que se encuentran integradas y cuya integración permite el uso compartido de la clasificación en dos aplicaciones web: *Erreus* (http://ixa.si.ehu.es/Erreus) e *Irakazi* (http://ixa.si.ehu.es/ikasleDB/menua).

Cuadro 3: Integración y unión de las bases de datos ERROREAK y DESBIDERATZEAK.

Los usuarios de *Erreus* serán principalmente lingüistas computacionales, mientras que *Irakazi* está enfocado más a profesores de euskera. Ambas aplicaciones están disponibles para poder realizar consultas e introducir información en las dos bases de datos presentadas: ERROREAK y DESBIDERATZEAK. Los ejemplos recogidos en estas bases de datos son fuente de información relevante para a) analizar la interlengua de los alumnos de euskera basándonos en textos reales y b) seguir investigando en el tratamiento automático de errores con el fin de desarrollar el corrector gramatical y de estilo.

Ambas aplicaciones, *Erreus* e *Irakazi*, tienen una parte pública, una registrada y una privada. La parte pública se utiliza para realizar consultas; la registrada para introducir datos y visualizar, cambiar o actualizar los ejemplos introducidos por cada usuario (para acceder a la parte registrada, el usuario necesitará una contraseña); por último, en la parte privada se controlarán el dinamismo de la clasificación y los datos introducidos por los usuarios, y se completará la información necesaria para el tratamiento automático de los errores y las desviaciones.

La clasificación de errores que aquí presentamos ha sido evaluada y utilizada por los lingüistas del grupo IXA y los alumnos del postgrado Hiztek. Y como ya se encuentra disponible en la red, confiamos en que progresivamente ira incrementando el número de usuarios, lo que nos permitirá enriquecer en el tiempo nuestras bases de datos. Para ello, ya hemos contactado con algunos Euskaltegis y en un plazo breve comenzarán los profesores de euskera a utilizar nuestras herramientas computacionales.

EVALUACIÓN DE LA CLASIFICACIÓN DE ERRORES

Para evaluar la categorización de errores, preparamos un ejercicio que constaba de 28 frases, con uno o más errores en cada una de ellas, para que once lingüistas procedieran a su clasificación.

La pretensión de este ejercicio de evaluación consistía en observar qué tipo de errores/desviaciones detectaban los lingüistas, cómo los clasificaban y cómo los corregían:

- si detectaban y clasificaban todos los errores de cada frase, o solo se fijaban en los errores más significativos.
- si clasificaban cada error en todas las posibles categorías/subcategorías, o se limitaban a clasificarlos bajo un único código.
- si primero corregían la frase y luego clasificaban el error, o si la corrección la hacían una vez el error había sido clasificado.
- si corregían solamente los errores a clasificar (sin corregir el resto de la frase), o si corregían la frase en su totalidad.
- si entendían que se trata de una clasificación completa y sencilla, o si por el contrario aportaban alguna propuesta de mejora.

De este ejercicio de evaluación, nos gustaría subrayar las siguientes conclusiones principales:

• Todos los lingüistas no siempre coincidieron en la clasificación y corrección de los errores. Pero en general, podemos afirmar que los resultados obtenidos son ciertamente positivos.

• Los lingüistas comentaron que no tuvieron ningún problema especial a la hora de clasificar muchos de los errores, pero que dudaron principalmente con las incorrecciones que hacen referencia a los errores morfosintácticos.

• Hay errores que se pueden clasificar en más de una categoría o subcategoría. Esta posibilidad quedó representada en once ejemplos. A pesar de esto, la tendencia general entre los lingüistas ha sido clasificar cada error en una sola categoría. En un par de frases, dos lingüistas han clasificado el mismo error en dos categorías diferentes, y en una sola frase seis de ellos han coincidido en asignar dos categorías al mismo error.

• Uno de los lingüistas no ha especificado con demasiada concreción la subcategoría de los errores, es decir, ha adjudicado las subcategorías principales sin profundizar en definir las sub-subcategorías, con lo que su clasificación ha resultado más superficial.

• En las frases en las que nosotros habíamos previsto un gran número de errores, los lingüistas no han coincidido plenamente en la determinación de los errores y su clasificación. En el caso de las frases excesivamente largas y complejas, la tendencia ha sido la de corregir la frase en su totalidad (reescribiendo el párrafo), ya que resulta bastante complicado asignar una sola posible categoría/subcategoría a cada uno de los errores detectados.

• Cinco lingüistas nos han comentado que les ha resultado más fácil corregir primero las frases y luego clasificar los errores. Por el contrario, tres lingüistas que ya estaban familiarizados con la clasificación, han categorizado los errores antes de proceder a su corrección. Y otros tres, primero clasifican y luego corrigen los errores, y viceversa.

• Tal y como habíamos previsto, las actitudes y los puntos de vista adoptados por los lingüistas en lo que se refiere a la corrección de errores tuvieron un signo diverso: dos de ellos reescribieron las frases enteras corrigiendo y subrayando únicamente el error a clasificar; otros dos lingüistas a veces reescribieron las frases en su totalidad y otras veces solo la parte errónea; y el resto optó por reescribir y corregir solamente el error a clasificar, proponiendo una única corrección posible. Precisamente porque ya intuíamos que las actitudes adoptadas para la corrección de errores serían un tanto subjetivas y diversificadas, hemos considerado oportuno definir unos criterios principales sobre el proceso de categorización y uso de las aplicaciones web mencionadas para que todos actuemos de la forma más coherente posible.

• En lo que se refiere a la estructura de la clasificación, la mayoría de los lingüistas comentaron que resulta demasiado detallada. Otros, en cambio, se mostraron a favor de una clasificación lo más exhaustiva posible porque consideran que así se podrán obtener categorías/subcategorías suficientes para poder clasificar los errores detectados en un corpus real donde la diversidad es muy amplia. Desde el punto de vista computacional también creemos que una clasificación detallada asegura mayor consistencia, y pensamos que será una fuente de información más completa que ayudará en el tratamiento automático de errores.

• Tal como ya hemos mencionado anteriormente, el tratamiento, la categorización y la corrección de errores tienen unas connotaciones bastante subjetivas, ya que cada lingüista/profesor cuenta con sus criterios propios y puntos de vista. Pero de todos modos, conviene insistir en el hecho de que esta evaluación nos ha permitido asegurar que la clasificación presentada responde sin problemas a diferentes perspectivas y puntos de vista.

CONCLUSIONES

Por lo tanto, esta clasificación de errores ha sido creada con objetivos diferentes. La pretensión principal, como ya hemos comentado, es clasificar errores y desviaciones de los alumnos de euskera en dos bases de datos: ERROREAK y DESBIDERATZEAK. Ello conlleva unos trabajos previos y nos abre diferentes líneas de investigación para ir avanzando en campos como el Análisis de Errores Automatizado, Aprendizaje y la Enseñanza de Lenguas Asistida por Ordenador y el tratamiento automático de errores.

Por el momento, y para conseguir textos de alumnos en proceso de aprendizaje de la lengua vasca, estamos en contacto con varios Euskaltegis que se han mostrado muy interesados en nuestra línea de investigación y nos han facilitado material escrito suficiente. Con los textos obtenidos, hemos recopilado, hasta ahora, un corpus de errores compuesto de 258.231 palabras, que nos es imprescindible como fuente de información para nuestras investigaciones. Vamos categorizando los errores y las desviaciones detectadas en ese corpus en la clasificación de errores global, jerárquica y dinámica aquí presentada. Dicha clasificación, además, ya se encuentra accesible en la red, por lo que en breve esperamos comenzar a enriquecer la información de las dos bases de datos mencionadas.

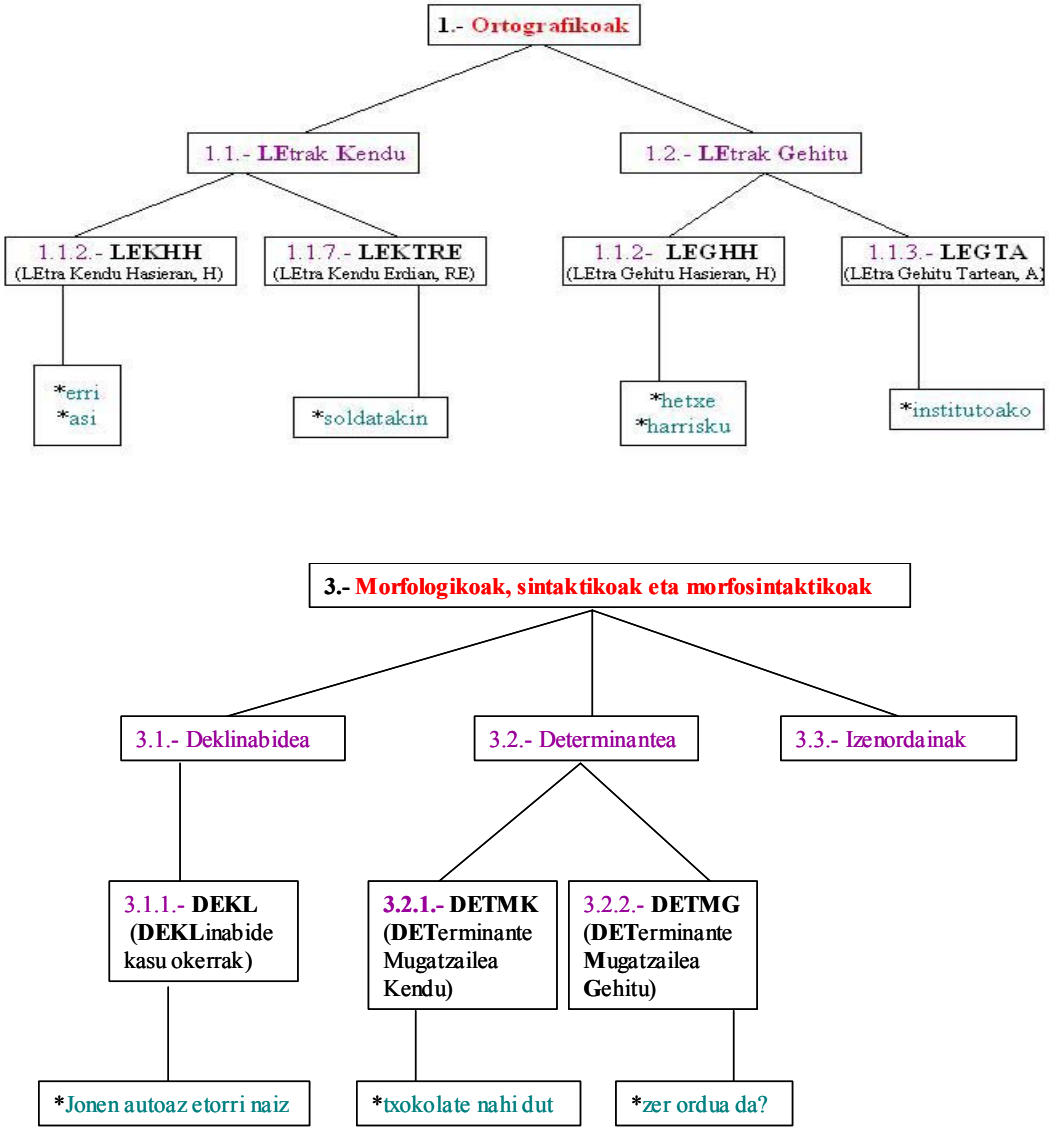
De ahí en adelante, y considerando la información psicolingüística recogida en la base de datos DESBIDERATZEAK, podremos diagnosticar el proceso y las características propias del aprendizaje del euskera, tanto de un alumno concreto como de un grupo de alumnos; podremos conocer también las características principales que se presentan en cada nivel, las formas verbales más y/o menos utilizadas en cada nivel, las estructuras más comunes entre los alumnos, etc.; podremos adaptar las estrategias y los materiales de aprendizaje a las necesidades reales de los alumnos y/o profesores. En efecto, es muy importante conocer sus necesidades; de otro modo, plantearíamos soluciones para problemas inexistentes y fallaríamos en proporcionar soluciones a los problemas que realmente tienen (Borin, 2002). Esta información nos será de utilidad también para estudiar los resultados que proporcionan las metodologías que se siguen en los Euskaltegis y prever el tipo de sistemas o herramientas lingüístico-computacionales de ayuda que podemos ofrecer a los alumnos/profesores en base a las dificultades y necesidades que presentan.

REFERENCIAS

- Aduriz I. (1994). *Errore ortografikoen azterketa eta zuzenketa bi mailatako morfologiaren arabera*. UPV/EHU Euskal Filologia. Barne-txostena-doktoregoa.
- Aduriz I., Aldezabal I., Aranzabe M., Arrieta B., Arriola J., Atutxa A., Díaz de Ilarraza A., Gojenola K., Oronoz M., Sarasola K., Urizar R. (2002). *The design of a digital resource to store the knowledge of linguistic errors*. DRH2002 (Digital Resources for the Humanities). Edimburgo.
- Aldabe I., Amoros L., Arrieta B., Díaz de Ilarraza A., Maritxalar M., Oronoz M., Uria L. (2005). *Learner and Error Corpora Based Computational Systems*. In Proceedings of the PALC 2005 Conference. Poland.
- Arrieta B., Díaz de Ilarraza A., Gojenola K., Maritxalar M., Oronoz M. (2003). *A database system for storing second language learner corpora*. Learner corpora workshop. Corpus linguistics 2003. Lancaster, UK.
- Becker M., Bredenkamp A., Crysmann B., Klein J. (1999). *Annotation of Error Types for German News Corpus*. In Proceedings of the ATALA workshop on Treebanks, Paris.
- Borin L. (2002). *What have you done for me lately? The fickle alignment of NLP and CALL*. Reports from Uppsala Learning Lab.
- Cowan R. (2003), http://www.public.iastate.edu/~apling/TLLB.html
- Dagneaux E., Denness S. & Granger S. (1998). *Computer-Aided Error Analysis*. System, Vol. 26, 163-174.
- Díaz, A.M.; Tipología de errores gramaticales para un corrector automático; en Proceedings del XXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 409-416.
- Díaz de Ilarraza A., Maritxalar A., Maritxalar M., Oronoz M. (1999). *IDAZKIDE: an intelligent CALL environment for second language acquisition*. Proceedings of a one-day conference "Natural Language Processing in Computer-Assisted Language Learning" organised by the Centre for Computational Linguistics, UMIST, in association with EUROCALL, a special ReCALL publication, 12-19. UK.
- Euskaltzaindia (1993). *Euskal Gramatika Laburra*. Bilbo.
- Fernández, S. (1997). Interl*engua y Análisis de Errores en el aprendizaje del español como lengua extranje*ra. Ed. Edelsa.
- Gojenola, K. (2000). *EUSKARAREN SINTAXI KONPUTAZIONALERANTZ. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan*. Informatika Fakultatea, UPV-EHU, 2000ko uztailaren 20a, Kepa Sarasola Donostiako Informatika Fakultateko (UPV/EHU) irakaslearen zuzendaritzapean eginiko tesia.
- Gojenola, K. eta Oronoz, M. (2000). Co*rpus-Ba*s*ed Syntactic Error Detection Using Syntactic Pattern*s. NAACL-ANLP00, Student Research Workshop. Seattle.
- Granger S. (2002). *A Bird's-eye view of learner corpus research*. Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, p. 3-33, Benjamins, Amsterdam and Philadelphia.
- HABE (1999). *Helduen Euskalduntzearen Oinarrizko Kurrikulua*. 120-131. or., Donostia.
- James, C. (1998). *Errors in Language Learning and Use: Exploring Error Analysis*. Longman.

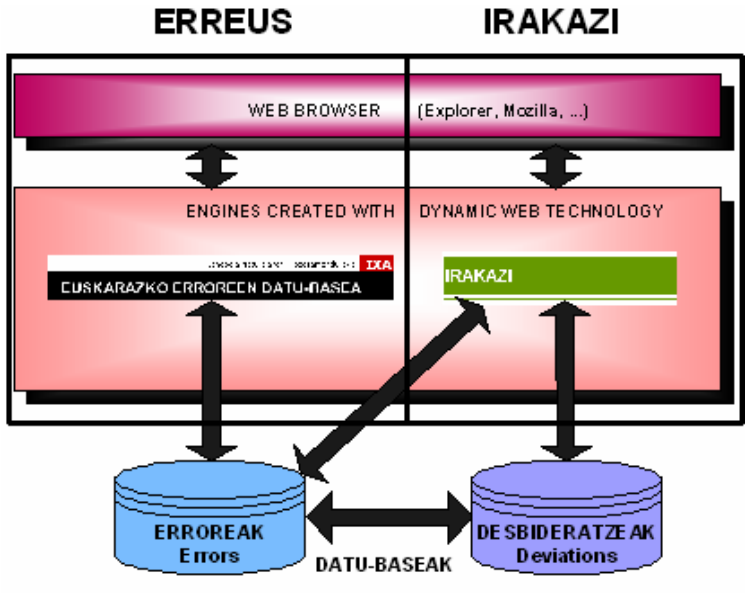
- Laka, I. *A Brief Grammar of Euskera*. Internet: http://www.ehu.es/grammar/
- Larsen-Freeman D. & H. Long M. (1994). *Estudio de la adquisición de segundas lenguas*. Gredos, p. 61.
- Maritxalar M., Díaz de Ilarraza A., Alegria I., Ezeiza N. (1996). *Modelización de la competencia gramatical en la interlingua basada en el análisis de corpus.* Procesamiento del Lenguaje Natural (SEPLN), Revista no. 19, 166-178. Sevilla.
- Maritxalar M. (1999). *MUGARRI: A multisystem environment to acquire the linguistic knowledge of second language learners*. PhD, EHU, Donostia.
- Norrish J. (1983). *Language learners and their errors*. London: The Macmillan Press Ltd.
- Ortiz de Urbina, J. (2001). *A Grammar of Basque*. University of Deusto.
- Richards, J.C. (1974). Error Analysis: Perspectives on Second Language Acquisition. Longman, London.
- Richards, J.C. (1974). A non-contrastive approach to error analysis. Longman, London.
- Sarrionandia B. (1999). *Perfiles lingüísticos: modelo alternativo de evaluación de la competencia del euskera.* Tesis doctoral; director: Jon Franco; fecha de lectura: 26-11-99.
- Selinker, L. (1972). *Interlanguage; International Review of Applied Linguistics*. Vol. 10 No. 3: 209-231.
- Selinker, L. (1992). *Rediscovering Interlanguage*. Longman, London and New York.
- Txillardegi (1978). *Euskal Gramatika*. Ediciones Vascas, Bilbo.
- Vilius Juozulynas (1994). *Errors in the compositions of 2nd year german students: an empirical study for parser-based ICALI*. CALICO Journal, Vol. 12, No. 1.
- Zubiri, I. eta Zubiri, E. (1995). *Euskal Gramatika Osoa*. Didaktiker SA, Bilbo.

Cuadros 1 y 2



Cuadros 1 y 2: ejemplos de la estructura jerárquica de la clasificación.

Cuadro 2



Integración y unión de las bases de datos ERROREAK y DESBIDERATZEAK**.**