MODELIZACIÓN DE LA COMPETENCIA GRAMATICAL EN LA INTERLENGUA BASADA EN EL ANÁLISIS DE CORPUS

Maritxalar M., Díaz de Ilarraza A., Alegría, I., Ezeiza N.

e-mail: jipmaanm@si.ehu.es
Lengoaia eta Sistema Informatikoak saila
Euskal Herriko Unibertsitatea (UPV/EHU)

649 postakutxa
20080 Donostia

RESUMEN

En este artículo presentamos un sistema semi-automático para la modelización de la competencia gramatical de la Interlengua por niveles en estudiantes de Segundas Lenguas, basada en un análisis de corpus. Así mismo, proponemos una metodología Top-down para dicha modelización. La Interlengua, compuesta por un conjunto de estructuras lingüísticas contextualizadas, incluye las *desviaciones* del idioma. Estas las representamos de igual forma que las estructuras *estándares.* En consecuencia, la filosofía de modelización que proponemos contrasta con la seguida en la mayoría de los correctores ortográficos y sintácticos desarrollados en lingüística computacional.

ABSTRACT

In this article we present a semiautomatic system for modelling Interlanguage Level Models of second language learners, based on the study of corpora. Also, we propose a Top-down methodology for such a modelling. The interlanguage is composed by a set of contextualized linguistic structures, and the deviations of the language are part of the interlanguage as well. Such deviations are represented in the same way as the *standards* structures. Thus, we can say than the modelling approach is different from most of the spelling and syntactic checkers which have been developed in computational linguistics.

1. Introducción

En este artículo presentamos un sistema semi-automático para la modelización de la competencia gramatical de la Interlengua por niveles de estudiantes de Segundas Lenguas, a partir de un análisis de corpus. Así mismo, proponemos una metodología Top-down para dicha modelización empezando por estudiantes con un nivel alto del idioma y continuando progresivamente hacia niveles más bajos. Los ejemplos de implementación que se plantean están basados en estudiantes castellano-parlantes cuya segunda lengua de estudio es el Euskara.

En el segundo apartado, definiremos el concepto de Interlengua. Por una parte, especificaremos qué entendemos por desviación *versus* el concepto error que tradicionalmente se ha venido aplicando en la mayoría de los correctores ortográficos y sintácticos (Agirre et. al. 1992). Por otra parte, comentaremos diferentes perspectivas en la formulación de un mismo fenómeno lingüístico a la hora de definir la competencia gramatical del alumno. Además, mencionaremos brevemente el proceso de construcción que, a través de los diferentes niveles de la Interlengua, siguen tanto el conocimiento compilado como el conocimiento variable. Por último, no olvidaremos la contextualización de la Interlengua inseparable, según Selinker et. al. (1994), de la definición de la competencia gramatical del alumno.

En el tercer punto, presentaremos una propuesta de metodología Top-down para la modelización de la Interlengua a diferentes niveles. Dicha metodología se basa en el estudio de corpus de textos *escritos* por alumnos, resultado de diversos tipos de ejercicios planteados (ej: redacción de un cuento, resumen de un artículo...).

En el siguiente apartado explicaremos diferentes aspectos de la implementación que estamos desarrollando para la modelización, centrándonos fundamentalmente en el estudio de la competencia léxica, morfológica y morfosintáctica del alumno.

Finalmente, expondremos las conclusiones basadas en los resultados obtenidos hasta el momento y el trabajo a realizar en un futuro próximo.

2. La Competencia Gramatical en la Interlengua

Los conceptos *dialectos transicionales (Corder 1971)* y *sistemas aproximados (Nemser 1971)* corresponden a ideas previas a la *Interlengua (Selinker 1972, 1992)* cuyo objetivo era definir la competencia comunicativa y gramatical sobre la segunda lengua. No obstante, todos ellos recogen las siguientes características comunes: a) el discurso de un alumno es **independiente de la lengua materna (L1) y la lengua de estudio (L2)** y es producto de un **sistema lingüístico estructurado**; b) el sistema lingüístico es **variable** durante el proceso de aprendizaje y para estudiantes del **mismo nivel es muy parecido**, a excepción de algunas diferencias,

consecuencia de la experiencia personal de aprendizaje del sujeto. Las citadas características serán la base para la modelización de la Interlengua que presentamos en el artículo.

Antes de entrar en las diferentes perspectivas a la hora de explicar un fenómeno lingüístico, consideramos importante precisar el concepto de **desviación** como alternativa a la noción de *error*. Diremos que una estructura lingüística es una desviación (Maritxalar et. al. 1996) cuando se cumpla una de las tres condiciones siguientes: a) la estructura lingüística es "incorrecta" – lo que habitualmente se ha llamado error –; b) la estructura se utiliza demasiado sustituyendo a otras estructuras más adecuadas que se pretenden evitar – ej: usar la conjunción *eta (= y)* en lugar de *ordea (= sin embargo) –;* c) se evita una determinada estructura (ej: *nor da? (= ¿quién es?)* en lugar de *nor ote da? (=¿quién será?)*). Las desviaciones son parte de la Interlengua y estarán representadas de la misma forma que las estructuras estándares que no se consideren desviaciones en el nivel de lengua de estudio. Debemos tener en cuenta que una desviación tipo C en el nivel 10 como: *nor da?* no se considera desviación a niveles bajos de la lengua.

2.1. Perspectivas superficial y lingüística para un mismo fenómeno

Una vez fijada la noción de estructura lingüística en la competencia gramatical de la Interlengua pasaremos a explicar las diferentes perspectivas que se pueden dar para un mismo fenómeno lingüístico. Tomemos como ejemplo las siguientes desviaciones tipo A:

a) *oihanan* *(=> oihan**e**an)* *(= en el bosque)*

b) *dakilako* *(=> daki**e**lako)* *(= porque lo sabe)*

Tanto en el caso de *oihanan* como en *dakilako* desde una perspectiva **superficial pura** podríamos codificar la estructura como LEBME (Borrar la LEtra E en Mitad de la palabra). Sin embargo, desde una perspectiva **superficial lingüística** diríamos que en el caso a) el alumno ha borrado la *-e* epentética para la unión de morfemas, mientras que en b) simplemente ha sustituido el morfema causal *-elako* por *-lako.*

Ambas perspectivas, la superficial (pura) y la lingüística, estarán representadas para todo fenómeno lingüístico en que se puedan identificar. Debemos tener en cuenta que para ciertos fenómenos sólo existirá una de las perspectivas y para otros, en cambio, incluso encontraremos más de una interpretación posible dentro de la misma perspectiva. Veamos dos desviaciones del tipo A:

DESVIACION 1

institutoako *(=> institutoko)*
(= del instituto) (locativo)

Dos interpretaciones superficiales:

a) LEAFA (Añadir la LEtra A al Final del lema de la palabra)
b) MOA-A (Añadir el MOrfema A)

Dos interpretaciones lingüísticas:

a) Lema *institutoa* + morfema *-ko*

b) Lema *instituto* + morfema *-a* + morfema *-ko*

DESVIACION 2

haurra ekarri du *(=> haurr**ak**)*

(= el niño lo ha traido) / (él ha traido el niño)

Dos interpretaciones lingüísticas:

a) *haurra* =>

@SUBJ => morfema *-a* (ergativo) (@SUBJ = función sintáctica (sujeto))

@OBJ => morfema *-a* (absolutivo) (@OBJ = función sintáctica (objeto))

Teniendo en cuenta la variabilidad del sistema lingüístico de la Interlengua es perfectamente viable que un alumno pueda tener en su base de conocimientos el lema *institutoa* e *instituto* al mismo tiempo, o utilice el morfema *-a* tanto para el absolutivo como para el ergativo.

2.2. Conocimiento Variable versus Conocimiento Compilado

Como hemos comentado previamente, la competencia gramatical en los alumnos de L2 es variable, sin embargo, en el estudio realizado se observa que existen estructuras que se estabilizan durante el proceso de aprendizaje. Dichas estructuras componen lo que llamaremos el **conocimiento compilado** mientras que las estructuras cambiantes que se encuentran en proceso de aprendizaje serán lo que llamaremos **conocimiento variable**. Estas últimas pueden evolucionar a nuevas estructuras variables o pasar a formar parte del conocimiento compilado. Al analizar corpus nos encontramos con cuatro tipos de estructuras en función de la utilización de las mismas: intentos, sistemáticas, cambiantes y lapsus (Maritxalar et. al. 1996). Las estructuras sistemáticas forman parte del conocimiento compilado y las estructuras cambiantes pertenecen al conocimiento variable. Sin embargo, Los intentos, outputs incomprensibles del alumno, y los lapsus, desviaciones momentáneas causadas por la falta de concentración, no se considerarán parte de la competencia gramatical del alumno.

En consecuencia, vemos necesaria la introducción de un valor de estabilización en las estructuras de la Interlengua, en función del cual determinaremos cuándo una estructura pertenece al conocimiento compilado o al variable.

Por otro lado, la variabilidad de la Interlengua cambia en función del contexto en el que se active la estructura lingüística (Selinker et. al. 1992). Por lo tanto, la especificación de una estructura lingüística dentro de un nivel de Interlengua vendrá caracterizada por: características lingüísticas, valor de estabilización y contexto. Este último podrá ser de tres tipos: lingüístico (ej: aparición de subordinadas en la frase), textual (ej: longitud de la frase o de la palabra; tipo de texto: redacción, carta...) y temático (ej: tema del texto, ejercicio: creación, resumen...). Esta forma de

contextualización de las estructuras lingüísticas se podría aplicar igualmente a la modelización de idiomas como primera lengua. Las estructuras lingüísticas que activamos cuando redactamos un cuento son diferentes a las utilizadas cuando escribimos textos científico-técnicos.

3. Metodología Top-down para la Modelización de Interlenguas por Niveles

En este apartado explicaremos una metodología Top-down para la modelización de Interlenguas por niveles, basada en el análisis de corpus, que hemos aplicado en (Maritxalar et. al. 1993, Andueza et. al. 1996, Maritxalar et. al. 1996). La base de partida de dicha metodología es: a) el sistema lingüístico de estudiantes del mismo nivel es muy parecido (Corder 1971, Nemser 1971, Selinker 1972,1992); b) el conocimiento compilado a un nivel de Interlengua incluye los conocimientos compilados de Interlenguas a niveles inferiores (Maritxalar et. al. 1994). La metodología seguida parte de niveles altos para ir progresivamente estudiando niveles más bajos, lo cual nos permite analizar el proceso evolutivo de los modelos de Interlengua. Las razones principales por las que el estudio parte de niveles altos hacia niveles más bajos se basan en: por un lado las herramientas lingüístico-computacionales que existen para el euskara (corrector ortográfico, etc.) son más adaptables a niveles superiores; por otro lado, las herramientas para el análisis automático de textos escritos por alumnos de niveles altos son en general más robustas que las correspondientes a niveles bajos y, por último, el material escrito disponible en cursos superiores es mayor que en los cursos de iniciación.

El proceso de modelización se divide en tres fases diferentes:

Fase 1ª => Determinación del corpus de estudio.

Fase 2ª => Construcción del modelo de la competencia gramatical en la Interlengua para el nivel n más alto de estudio.

Fase 3ª => Construcción de los modelos de Interlengua para niveles inferiores

3.1. Determinación del corpus de estudio

Criterios en la recopilación del corpus:

* Textos escritos por alumnos de diferentes escuelas de idiomas.

* Agrupación de textos en base al tipo de ejercicio propuesto por el profesor (ej: creación, resumen, tema ...).

* Elección de alumnos de asistencia constante durante el curso académico.

* Selección de alumnos en base al perfil del alumno de forma que queden representados estudiantes con: ritmos de aprendizaje diferentes, especial atención a la gramática versus cuidado en el mensaje a transmitir, diferentes motivaciones de estudio, porcentajes de uso de L2, conocimiento de otras lenguas... . Esto se ha llevado a cabo paralelamente al desarrollo de

una base de datos en la que se ha recopilado información relevante al proceso de aprendizaje sobre los alumnos, obtenida mediante entrevistas con los propios alumnos y profesores (Andueza 1996).

El corpus recopilado está compuesto de 50 textos escritos por alumnos de IRALE[1] durante el curso académico 90/91, 150 por alumnos de ILAZKI en el 92/93 y 150 textos de estudiantes en ILAZKI durante el 94/95.

3.2. Construcción del modelo de la competencia gramatical en la Interlengua para el nivel n mas alto de estudio

La construcción del modelo para el nivel n se realizará de forma semi-automática ayudándonos de las herramientas lingüístico-computacionales disponibles para el tratamiento del idioma. En nuestro caso, adecuaremos las siguientes herramientas informáticas: EDBL, Base de Datos Léxica para el Euskara, (Agirre et. al. 1994); el analizador morfológico (Agirre et. al. 1992); el lematizador (Aldeazabal et. al. 1994) y partes del sistema de gramática de restricciones CG (Aduriz et. al. 1996, Karlsson et. al. 1992). Los pasos a seguir para la construcción del modelo de Interlengua en el nivel n son:

1) Diseño y construcción de la base de datos léxica para el estudio de Interlenguas.
2) Selección del corpus de estudio para el nivel de idioma n (CORPUS-N) y subconjuntos de CORPUS-N para los pasos posteriores.
3) Definición de la morfología y morfosintaxis para el nivel n.
4) Determinación del conocimiento compilado y variable para el nivel n según los contextos comentados en el apartado 2.
5) Estudio de la fiabilidad del paso 4) en base a nuevos subconjuntos de corpus.

3.3. Construcción de los modelos de Interlengua para niveles inferiores

La competencia gramatical de la Interlengua está compuesta de un *conocimiento general* independiente del contexto y un *conocimiento específico* dependiente del contexto. Dicha diferencia no aparece explícitamente en el gráfico de construcción que se refleja en la figura 1, donde se representa el proceso que siguen los conocimientos compilados y variables en el paso a niveles inferiores. Supongamos que el concocimiento para el nivel X+1 está definido. A partir de X+1 y el CORPUS-X vamos a obtener el conocimiento correspondiente a X:

---

[1] IRALE e ILAZKI: escuelas especializadas en la enseñanza del euskara

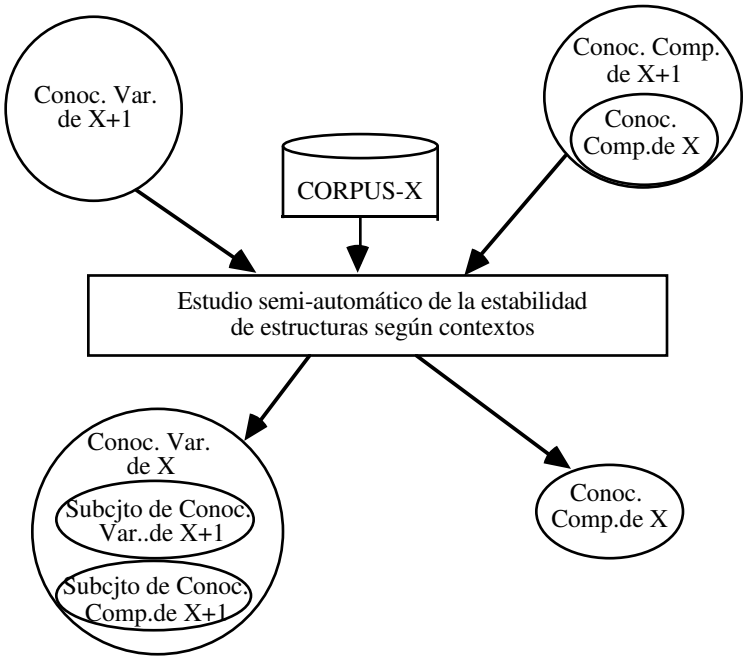

Fig1. Gráfico de construcción de la interlengua en el nivel X a partir de X+1

4. Implementación.

Como adelantamos en la introducción, en este apartado explicaremos diferentes aspectos del sistema cuya implementación se encuentra en desarrollo. Las herramientas informáticas para el análisis de corpus que utilizamos son: un filtro de palabras, un filtro de frases, una base de datos léxica para el estudio del euskara como segunda lengua (L2), un analizador morfológico basado en la morfología de dos niveles adaptado (MORF_H2) y una herramienta CG adaptada para el estudio de la morfosintaxis de L2 (MSIN_H2).

4.1. Implementación de la base de datos léxica

En el diseño de la base de datos léxica para el estudio de la Interlengua añadimos nuevos campos en las entradas de EDBL (Base de Datos Léxica para el Euskara): valor de estabilización, contexto lingüístico, contexto textual, contexto temático y lista de niveles. Para la construcción de la base de datos utilizamos el lematizador (Aldeazabal et. al. 1994) junto con un entorno de detección de niveles y contextos en los que los lemas y morfemas aparecen. En la base de datos para L2 encontraremos morfemas con más de una entrada; por ejemplo, el morfema *-a* que en EDBL existía sólo como absolutivo, en nuestro caso tendrá una segunda entrada para el ergativo (ver ejemplo del apartado 2).

4.2. Implementación del analizador morfológico adaptado

MORF_H2 esta compuesto de dos módulos principales: analizador y desambiguador.

4.2.1 Un analizador morfológico para L2

Es una adaptación del analizador general (Agirre et. al. 1992) con las siguientes diferencias:

1) Multiplicidad de los autómatas con el objeto de detectar reglas morfofonológicas que se aplican en cada análisis y contextos en los que se activan.

Veamos un ejemplo comparativo:

Reglas para el analizador general:

h:0 => R:= (+:=)+ _	(Regla del lenguaje estándar)
h:0 => [V:V | (0:*) $:$] _ V:V	(Regla para un error típico de la lengua)

[La aplicación de la regla estándar no se detecta aunque el error de competencia se identifica. El contexto de aplicación es ambiguo.]

Reglas para el analizador adaptado:

h:0 => R:= (+:=)+ _	(Regla perteneciente a la Interlengua)
h:0 => V:V _ V:V	(Regla perteneciente a la Interlengua)
h:0 => (0:*) $:$ _ V:V	(Regla perteneciente a la Interlengua)

[Se detecta la aplicación de las tres reglas, identificando así el contexto de aplicación. La detección se realiza duplicando el mismo autómata 3 veces y marcando con un valor negativo los estados correspondientes al éxito de cada regla]

2) Identificación del lugar de aplicación de la regla (lema/morfema) en la interpretación morfológica. Esto nos permitirá, mediante el desambiguador, descartar interpretaciones morfológicas no viables, producidas por la activación de una regla en un morfema cuando sólo es probable su aplicación en el lema, y viceversa.

4.2.2 Un desambiguador morfológico para L2

El objetivo del desambiguador es el de descartar las interpretaciones no viables en el nivel de Interlengua de estudio. Hemos desarrollado un módulo de desambiguación morfológica basado en Constraint Grammar (Aduriz et. al. 1996, Karlsson et. al. 1992) aplicado en tres fases diferentes:

1ª) Desambiguación en el texto de aquellas entradas de la base de datos no pertenecientes al nivel de estudio.

Descartamos las interpretaciones no correspondientes al nivel de estudio, utilizando el valor de la lista de niveles en las entradas de la base de datos. En caso de que una palabra no contenga ninguna interpretación para dicho nivel, quedará al menos una interpretación morfológica. Esto nos permite detectar nuevas entradas del nivel a añadir en la base de datos.

2ª) Desambiguación de interpretaciones que incluyan reglas morfofonológicas no pertenecientes a la Interlengua del nivel de estudio.

3ª) Desambiguación de contextos lingüístico-textuales. Tendremos un módulo para cada nivel.

Consideremos por ejemplo la desviación *analisis (=> analisi)* en la que se ha podido aplicar: la regla para añadir una s en el lema, o bien, la sustitución de la letra *z* por *s* en el morfema. Esta última posibilidad, en ciertos niveles, no es viable cuando se aplica en los morfemas. En esta tercera fase descartaremos esta segunda interpretación. Lo mismo sucede en la activación de la regla de desviación para borrado de la e, sólo válida para nombres y adjetivos.

4.3. Implementación del analizador morfosintáctico adaptado

La herramienta para el estudio de la morfosintaxis en L2 (MSIN_H2) nos permite desambiguar interpretaciones morfosintácticas (Aduriz et. al. 1996, Karlsson et. al. 1992). Veamos un ejemplo:

Sean las dos siguientes entradas para el mismo morfema en la base de datos:

("-a" ... ABS ... @SUBJ @OBJ)
("-a" ... ERG ... @SUBJ)

Retomemos el ejemplo del apartado 2 : *haurra ekarri du*

Tras aplicar MSIN_H2 tendríamos para *haurra* las dos siguientes interpretaciones, válidas ambas en L2:

((lema "haur") (morf "-a") ... ABS ... @SUBJ @OBJ)
((lema "haur") (morf "-a")... ERG ... @SUBJ)

Sin embargo si la frase fuera: *haurra etorri da (= el niño ha venido)*

Tras aplicar MSIN_H2 sólo tendríamos la primera interpretación como consecuencia de la aplicación de la siguiente regla de desambiguación:

(@w =0 (ERG) (0 ABS) (*1C CLB *LC)
(*LC IZAN))

[Hemos descartado (=0) la segunda interpretación (ERG) ya que a la derecha de *haurra* aparece el verbo intransitivo*da* (IZAN) dentro de la misma claúsula (CLB)]

5. Conclusiones

Mediante el sistema que hemos presentado, el grado de precisión de los modelos va aumentando según bajamos de nivel. Por una parte, utilizando el lematizador (Aldeazabal et. al. 1994) construimos la base de datos léxica para el estudio de Interlenguas a partir de EDBL y del CORPUS.

Dicha base de datos contiene información sobre los niveles de uso que se irá enriqueciendo conforme apliquemos la fase 1ª del desambiguador en MORF_H2. Por otra parte, la aplicación de las fases 2ª y 3ª nos proporciona una definición de la morfología de la Interlengua más exacta.

El trabajo presentado demuestra que la implementación de un sistema semi-automático para el estudio de Interlenguas es viable a partir de la adaptación de las herramientas lingüístico-computacionales disponibles para el estudio automático del euskara. La reutilización de dichas herramientas y la implementación incremental de la competencia gramatical son una buena garantía para continuar con el análisis de corpus en el estudio de segundas lenguas.

Previamente a la implementación de MORF_H2 y MSIN_H2 se han realizado estudios de campo (Maritxalar et. al. 1993; Andueza et. al. 1996). Actualmente estamos estudiando la morfosintaxis y sintaxis para L2 basándonos en los resultados obtenidos en Andueza et. al. (1996), donde en una prueba final con los alumnos se contrastaron las hipótesis obtenidas del estudio realizado durante el curso 94/95.

Bibliografia.

Aduriz, I., Arriola, J.M., Artola, X. Díaz de Ilarraza A., Gojenola, K., Maritxalar, Urkia, M. (1996). *Euskararako murriztapen-gramatika: lehen urratsak*. (Barne-txostena. UPV/EHU/LSI/TR 2-96). Informatika Fakultatea (UPV/EHU).

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K. & Urkia, M. (1992). *XUXEN: A spelling checker/corrector for Basque based on Two-Level Morphology*. Proceedings of the Third Conference ANLP (ACL) (pp. 119-125). Trento, Italy.

Agirre, E., Agirre, X., Arriola, J.M., Artola X., Insausti, J.M. (1994) *Euskararen Datu-Base Lexikala (EDBL)*, (Barne-txostena. UPV/EHU/LSI/TR 8-94). Informatika Fakultatea (UPV/EHU).

Aldezabal I., Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K. Aduriz I., Urkia M. (1994) *EUSLEM: Un lematizador/etiquetador de textos en euskara*, Actas del X.Congreso de la SEPLN Córdoba.

Andueza, A., Díaz de Ilarraza, A., Maritxalar, M., Martiarena, J., Pikabea, I. (1996). *Hizkuntza baten Ikaskuntza Prozesuari buruzko landa lana. Sistema Informatiko adimendun baten oinarria*. (Barne-txostena. UPV/EHU/LSI/TR 8-96). Informatika Fakultatea (UPV/EHU).

Corder, S. (1971). *Idiosyncratic dialects and error analysis.* International Review of Applied Linguistics 9: 147-59.

Karlsson, F., Voutilainen, A., Heikkilä, J. & Anttila, A. (1992). *Constraint Grammar: a language-independent system for parsing unrestricted text.*

Maritxalar, M., & Díaz de Ilarraza, A. (1993). *Integration of Natural Language Techniques in the ICALL Systems Field: The treatment of incorrect knowledge*. (Barne-txostena. UPV/EHU/LSI/TR 9-93). Informatika Fakultatea (UPV/EHU).

Maritxalar, M., & Díaz de Ilarraza, A. (1994). *An ICALL System for Studying the Learning Process*. Computers in Applied Linguistics Conference. Iowa State University.

Maritxalar, M., & Díaz de Ilarraza, A. (1996). *Hizkuntza baten Ikaskuntza-Prozesuan zeharreko Tartehizkuntz Osaketa: Sistema Informatiko baten Diseinurako Azterketa Psikolinguistikoa*. (Barne-txostena. UPV/EHU/LSI/TR 7-96). Informatika Fakultatea (UPV/EHU).

Nemser, W. (1971). *Aproximate Systems of Foreign Language Learners.* International Review of Applied Linguistics, 9, pp. 115-23; reprinted in Richards (1974)

Selinker, L. (1972). *Interlanguage*. International Review of Applied Linguistics 10: 209-31.

Selinker, L. (1992). *Rediscovering Interlanguage*. London: Longman.

Selinker, L.,& Douglas, D. (1994) Interlanguage-draft.