**RASAL**
LINGÜÍSTICA
2021: 91-119
Recibido: 14.08.2020 | Aceptado: 27.01.2021
ARK: http://id.caicyt.gov.ar/ark:/s26183455/zz672wn1e

# PREDICCIÓN DEL ERROR DE CONCORDANCIA PLURAL EN CUATRO APRENDIENTES ITALIANOS DE ELE

# PREDICTION OF PLURAL AGREEMENT ERRORS IN FOUR ITALIAN LEARNERS OF SPANISH FL

*Pablo E. Marafioti*
*Facultad de Lenguas - Universidad Nacional de Córdoba*
https://orcid.org/0000-0001-7299-5324

RESUMEN

Mediante técnicas de minería de datos se predijo la presencia del error de concordancia plural en cuatro aprendientes italianos de español como lengua extranjera. Se crearon atributos provenientes de redes complejas, atributos que caracterizaban los casos de concordancia y atributos que tenían en cuenta una componente temporal. Se aplicaron diferentes estrategias para lidiar con el desbalance de clase y la selección de atributos. Por último, se aplicó un *clustering* por mixtura de distribuciones para distinguir entre sesiones con mayor y menor intensidad de error. Se seleccionaron atributos relevantes mediante un modelo mixto con regularización y se ajustó a los datos de cada *cluster* un modelo mixto generalizado. Se llegó a obtener una precisión balanceada del 80 %. Los atributos derivados de los grafos fueron seleccionados en todos los alumnos. El *cluster* con mayor error reveló que influían en la chance de cometerlo los controladores animados, que la concordancia incluyera un adjetivo y un determinante, y la presencia de posesivos. Fue facilitadora una estrategia para evitar los plurales en *-es*.

PALABRAS CLAVE: español lengua extranjera; corpus de español; concordancia; número; minería de datos.

ABSTRACT

Using data mining techniques we predict the presence of plural agreement error in four Italian learners of Spanish as a foreign language. Predictors were created from complex networks, from attributes characterizing the agreement instances and from attributes which involved a temporal component. A set of strategies was employed to deal with class imbalance and attribute selection. After that, a clustering with mixed distributions was applied. The attribute selection was made using a mixed model with regularization. A generalized linear mixed model was adjusted to each cluster. The maximum balanced

accuracy obtained was 80 %. The attributes derived from graphs were selected in all students. The cluster with more error intensity revealed that the error chance increased with animate controllers, instances of agreement including an adjective or determinant and the presence of possessives. A strategy oriented to avoid plurals in *-es* diminished the chance of error.

KEYWORDS: Spanish foreign language; corpus for Spanish; agreement; number; data mining.

## 1. Concordancia en español: definición y antecedentes en L2

La concordancia se define como una relación entre rasgos sublexicales (pares 'valor: atributo') de los ítems léxicos (O'Grady 2005). En español dichos rasgos son 'persona', 'número' y 'género', junto a sus valores. Corbett (2006) denomina *controlador* al ítem léxico que determina la concordancia y *objetivo* al elemento cuya forma es determinada por aquel. Se denomina *dominio* al entorno sintáctico en el cual ocurre la concordancia. La concordancia se establece por covarianza sistemática de rasgos. En el presente trabajo el controlador será nominal, presente y expresará sus rasgos abiertamente. Por otra parte, los objetivos consistirán en artículos (definidos e indefinidos), adjetivos y pronombres (demostrativos, posesivos, indefinidos). Usarán morfemas ligados (flexión) para expresar concordancia; con marcado obligatorio; usando morfología regular, productiva y con diferentes grados de aliteración. Concuerdan con un solo controlador obligatoriamente. Los dominios relevantes serán el sintagma nominal, el sintagma verbal (predicativo) y la oración subordinada. La concordancia se considerará asimétrica (el género y número de los objetivos dependen del controlador nominal).

La literatura de adquisición ha arrojado los siguientes resultados. En primera instancia, la adquisición del plural tiende a seguir las etapas: plural nulo > plural en -s > plural en -es (Bruhn de Garavito 2008). La concordancia de número parece ser más fácil de adquirir que la de género ya que, en general, los errores en el primero tienden a ser menos que en el segundo. Si bien el nivel de competencia hace que los errores disminuyan, los de género tienden a persistir incluso luego de muchos años de práctica de producción oral (Muñoz Liceras, Díaz & Mongeon 2000; Franceschina 2001; White *et al*. 2004).

La concordancia de género y número del artículo resulta más fácil de adquirir que la del adjetivo. Esto parece ser así para cualquier nivel de competencia, en bilingües tempranos y tardíos, y tanto en producción como en procesamiento (Fernández-García 1999; White *et al*. 2004; Montrul *et al*. 2008; Alarcón 2011; Gillon Dowens *et al*. 2010). Por otra parte, la concordancia plural del cuantificador *mucho* parece más fácil de adquirir que la de *bastante* y *demasiado* (Español-Echevarría & Prévost 2004).

En cuanto al género, resulta más fácil producir y procesar la concordancia: (a) de masculino respecto al femenino, reflejando el hecho de que las formas de masculino se utilizan en contextos femeninos, o sea, como *defaults* (Fernández-García 1999; Bruhn de Garabito & White 2002; White *et al*. 2004; Montrul *et al*. 2008; McCarthy 2008; Alarcón 2011;); (b) con controladores de morfología transparente (*-o*/*-a*, como en *vaso, maestra*) respecto a los menos transparentes (en *-e*, como en *el puente*, *la suerte*; en consonante:

*el camión*, *la canción*; u opuestos, como en *la mano*), estos últimos en el orden de facilidad *consonante* > *-e* > *opuestos* (Fernández-García 1999; Montrul *et al*. 2008; Alarcón 2011); (c) con controladores inanimados (como en *hospital*) respecto de los animados en correspondencia con el sexo biológico, como en *doctor*/*a* (Sagarra & Herschensohn 2013; aunque cfr. también Alarcón (2009), que encontró el efecto contrario en núcleos de SN complejos); (d) en SN complejos del tipo *N1 de N2*, cuando el género de N1 coincide con el N2 (Foote 2015).

En un estudio de González *et al.* (2019) se analizaron errores de concordancia de género y número en el ámbito nominal en composiciones escritas de 23 estudiantes holandeses de español LE. Se encontraron efectos significativos de aumento de errores: (i) del plural respecto del singular, (ii) del femenino respecto del masculino, (iii) en los artículos femeninos (sin importar el rasgo de número).

El aumento de distancia estructural (cantidad de nodos sintácticos entre controlador y objetivo) causa que disminuya la sensibilidad a las violaciones de género y que la concordancia se procese más lentamente ("peor") en el dominio no local (Sagarra 2007; Lichtman 2009; Keating 2009, 2010; Foote 2011; Gillon Dowens *et al.* 2010). Se ha propuesto que dicho efecto de distancia se relaciona con la capacidad de memoria de trabajo, ya que el aprendiente debe mantener el valor del rasgo del controlador (nominal) en la memoria para concordarlo luego con el del objetivo (adjetivo) a larga distancia. Sagarra (2007) y Keating (2010) hallaron correlaciones positivas entre distancia y memoria de trabajo; aunque en Foote (2011) no hubo evidencia de ello.

Por otro lado, hay algunos trabajos enmarcados en el llamado *análisis de errores asistido por computador*. Se realiza un etiquetado semiautomático de un corpus electrónico (oral o escrito) de acuerdo con un sistema de anotación de errores establecido. Luego, se lleva a cabo análisis de frecuencia y/o recurrencia de los tipos de error, junto con una explicación y/o evaluación de la gravedad del error. La taxonomía de errores puede responder a los siguientes criterios (Ferreira Cabrera & Elejalde Gómez 2020). El criterio descriptivo identifica mecanismos para la conformación de la interlengua, como por ejemplo: la adición, omisión, falsa elección, sustitución y elección errónea de formas. El criterio lingüístico delimita el nivel de localización del error: palabra, oración, párrafo. El criterio etiológico infiere las causas de los errores mediante la transferencia negativa de la L1. De este modo la concordancia se clasificaría como un mecanismo de falsa selección, a nivel de palabra/oración/párrafo tanto dentro del sintagma como entre sujeto y atributo con verbos copulativos. En italiano, las causas de transferencia negativa se podrían deber, por ejemplo, a diferencias de especificación de los rasgos de género y número o al mecanismo de sobre-extensión: *las mujeras* (it. *le donne*); *los dineros* (it. *i soldi*); *los trenos* (it. *i treni*); *el leche* (it. *il latte*), etcétera. Ferreira Cabrera, Elejalde Gómez & Vine Jara (2014) presentaron un análisis de un corpus compuesto por 84 resúmenes escritos por 22 estudiantes de nivel B1. La falsa selección de género y número en la concordancia gramatical evidenció una frecuencia de error del 23 %, con más errores de género que de número. Además fue la tercera causa de error, luego de los errores debidos a la ortografía y a las preposiciones. Campillos Llanos (2014) trató el análisis de errores léxicos en español oral como LE, para estudiantes de nivel A2 y B1. Los errores de asignación de género fueron más frecuentes que los de número. Si bien los de género disminuían cuando se pasaba al nivel B1, en el número se mantenían casi sin cambios de frecuencia. Ferreira

Cabrera & Elejalde Gómez (2017) trataron un corpus escrito por 26 sujetos de nivel A2+ y 36 sujetos de nivel B1. Además de investigar la frecuencia de error, se interesaron en su recurrencia, o sea, cuán sistemáticos y repetitivos resultaban los errores dentro de un mismo sujeto. La concordancia sintáctica fue la tercera causa de frecuencia de error, contando con un 18 % de los errores totales. Asimismo, la falsa selección de concordancia gramatical de género fue el segundo subtipo de error más frecuente (luego de las preposiciones). En lo concerniente a la recurrencia, se tomó el número total de errores producidos por al menos cuatro sujetos. Los errores de concordancia sintáctica representaron el 24 % del total. Otra vez, los de falsa selección de concordancia de género estuvieron en el segundo puesto por subcategoría (después de los errores por omisión de acento en palabras esdrújulas). Por ende, los errores de concordancia de género no solo resultaron ser muy frecuentes sino también sistemáticos, manteniéndose en el tiempo en la producción de los sujetos.

## 2. Objetivos

El objetivo de este trabajo consiste en predecir el estatus de error de las concordancias plurales en las últimas cuatro sesiones de cuatro aprendientes italianos de español L2. También se buscará identificar algunos factores influyentes en la chance de error para las sesiones con mayor intensidad de error de concordancia. El material suplementario y el código de R empleado se encuentra en: https://github.com/pablomarafioti/PabloMarafioti/tree/master/prediccion_de_error.

## 3. Recolección de datos y creación de variables

Se analizan datos de cuatro casos de estudiantes de español como lengua extranjera. Se trató de cuatro alumnos adultos, de lengua nativa italiana, estudiantes del Instituto Cervantes de Milán en el año académico 2008/2009. Cada alumno poseía un nivel distinto de competencia lingüística según el *Marco Común Europeo de Referencia*. Se hicieron entrevistas de 30 minutos entre el alumno y el investigador (autor de este trabajo). La tarea consistió en una conversación no estructurada, sobre temas acordes al nivel de competencia del sujeto. Dichas entrevistas tuvieron lugar aproximadamente cada 20 días, según la disponibilidad de los alumnos. Cada alumno realizaba simultáneamente el curso de español. Hubo entre doce y catorce entrevistas por alumno. El corpus está constituido por los siguientes conjuntos de transcripciones: Sonia (nivel A1/A2): 12 transcripciones; Nati (nivel B1): 14 transcripciones; Jako (nivel B2): 14 transcripciones; Mirka (nivel C1): 12 transcripciones (nombres ficticios). La codificación y la transcripción de los datos se hicieron mediante el formato CHAT, siguiendo a Mac Whinney (2021). Cada concordancia se codificó con dos términos, pero pudiendo haber más términos “objetivo”: por ejemplo, en *los libros azules* se codificaron dos instancias: *los libros* y *libros azules*. Se anotaron a continuación marcadores (‘tags’) en el corpus para realizar el conteo posterior. Son los siguientes:

- [*0] = ausencia de error;
- [*1] = errores en el género; por ej.: *muchos personas* [Sonia, sesión 1, línea 56] (error por: *muchas personas*)
- [*2] = errores debidos al uso de la terminación *-(e)s*: (a) por no tomar en cuenta la última consonante de la raíz léxica, que exige un plural con *e* epentética en *-(e)s*; (b) por uso en contexto incorrecto u omisión en correcto; por ej.: *muchos trenos* [Sonia, sesión 7, línea 148] (error por *muchos trenes*)
- [*3] = errores de plural, o sea ausencia de *-s*; por ej.: *los veneciano* [Sonia, sesión 5, línea 293] (error por *los venecianos*)
- [*4] = errores mixtos por acumulación de los anteriores; por ej.: *les joven* [Sonia, sesión 2, línea 144] (error por *los jóvenes*)

El siguiente constituye un ejemplo de transcripción del aprendiente Sonia (sesión 2):

1 @Begin
2 @Languages: spa
3 @Participants: STU Sonia Participant, INV Pablo Investigator
4 @ID: spa | sonia | | |female| | |Participant| |
5 @ID: spa | sonia | | |male| | |Investigator| |
6 *INV: háblame de tus amigos .
7 *STU: yo tengo muchos amigos [*0] .
8 *STU: Marina es fotógrafa .
9 *STU: ella quiere hacer fotos .
10 *STU: y para hacer fotos ella va a Londres (.) a París (.) a Madrid (.) a Berlino@s:ita a Roma .
11 *STU: por le@s:ita grande ciudades [*3] de Europa .
[...]
240 @End

El conteo se hizo con el programa CLAN. Se crearon variables que caracterizaban cada instancia producida de concordancia. Se las describe a continuación (el primer nivel se considera el de referencia).

- *Target*. Variable respuesta (categórica). Niveles: 1 = comete error [tags “1”, “2”, “3”, “4”]; 0 = no comete error [tag “0”].
- *Instancia*. Instancia de concordancia producida por el hablante.
- *Esp*. Concordancia en español (sin error).
- *Mod*. Tipo de modificador del controlador. Niveles: 0 = artículo definido; 1 = artículo indefinido; 2 = determinante (adjetivos posesivos, indefinidos, demostrativos, interrogativos, exclamativos); 3 = adjetivos (calificativos, numerales, ordinales).
- *Gram*. Si se trataba de una instancia de concordancia de más de dos términos. Niveles: 0 = dos términos; 1 = más de dos términos.
- *LDa*. Si la concordancia era o no a larga distancia: 0 = no, 1 = sí.

- *ES*. Se especificó si en el controlador, en el objetivo, o en ambos, había una desinencia que requería la inserción de *-e-* epentética (*-(e)s*). El razonamiento fue que realizar concordancia con dos operaciones de este tipo resulta más complicado que con una o con ninguna; según ES: 0 = sin *-e-* epentética; 1 = con *-e-* epentética en un término; 2 = con *-e-* epentética en ambos términos.
- *Anim*. Si el controlador era o no animado, según: 0 = inanimado, 1 = animado.
- *Esp1*. Se codificó la vocal final de la palabra más la desinencia de plural en español del primer término. Según: 1 = *-us* (ej.: *sus*); 2 = *-is* (ej.: *mis*); 3 = *-os* (ej.: *rojos*); 4 = *-as* (ej.: *blancas*); 5 = *-es* (ej.: *celestes*); 6 = **-es* (ej.: *útiles* [*-e-* epentética]).
- *Esp2*. Se codificó la vocal final de la palabra más la desinencia de plural en español del segundo término. Mismos niveles que Esp1.
- *Acum*. Cantidad de errores hechos hasta la instancia anterior a la actual dentro de una misma sesión (o sea que el conteo es desde cero cuando cambia la sesión). Por ejemplo, para las primeras seis concordancias observadas de la primera sesión de Nati, el error se produce en la tercera concordancia; por lo tanto, en la cuarta concordancia los errores acumulados hasta la instancia anterior son iguales a uno, y continuará así hasta que se produzca un nuevo error y el conteo crezca.

Los siguientes atributos, sobre rasgos del controlador, se extrajeron de la base de datos BuscaPalabras (Davis & Perea 2005):

- *Concretud* (*Conc*). Índice subjetivo en escala de 1 a 7, que indica cuán concreta es una palabra de menos (+ abstracta) a más (+ concreta).
- *Familiaridad* (*Fam*). Índice subjetivo en escala de 1 a 7, que indica cuán frecuentemente una palabra es oída, leída o producida diariamente.
- *Imaginabilidad* (*Ima*). Índice subjetivo en escala de 1 a 7, que indica la intensidad con la que una palabra evoca imágenes.
- *Frecuencia* (*LEXESP*). Frecuencia de la palabra en el corpus BuscaPalabras, en escala por mil.

A modo de ilustración de los atributos descriptos hasta ahora, considérese el siguiente fragmento de transcripción de Mirka (sesión 6) y su registro de variables en el Cuadro 1:

46 *STU: entonces <lo que> [//] ehh@fp yo creo que ehh@fp los animales ehh@fp tienen derechos.
47 *STU: pero <no> [/] no son lo [*3] mismos derechos [*0] ehh@fp que ehh@fp
48 +...
49 *STU: lo <que deben tener las> [//] que tienen las personas [*0] .
50 %err: los mismos derechos
51 *STU: no sé si me explico .
52 *STU: los [*0] seres humanos [*0] ehh@fp tenemos +...
53 *STU: es algo un poco malo que decir .
54 *STU: pero tenemo [*] más derechos que los animales .

55 %err: tenemos
56 *STU: en el sentido que [*] ehh@fp <me>[/] me doy cuenta que a veces los animales [*0] son muy sensibles [*0] .

| Instancia | Esp | Mod | LDa | ES | Gram | Anim | Fam | Ima | Conc | LEXESP |
|---|---|---|---|---|---|---|---|---|---|---|
| lo derechos (1) | los derechos | 0 | 0 | 0 | 1 | 0 | 6,17 | 3,71 | 3,62 | 130 |
| mismos derechos (2) | mismos derechos | 2 | 0 | 0 | 1 | 0 | 6,17 | 3,71 | 3,62 | 130 |
| las personas | las personas | 0 | 0 | 0 | 0 | 1 | 7 | 6,22 | 5,49 | 171,79 |
| los seres (1) | los seres | 0 | 0 | 1 | 1 | 1 | 5,29 | 4,23 | 2,37 | 82,5 |
| seres humanos (2) | seres humanos | 3 | 0 | 1 | 1 | 1 | 5,29 | 4,23 | 2,37 | 82,5 |
| los animales | los animales | 0 | 0 | 1 | 0 | 1 | 6,63 | 6,31 | 3,54 | 73,04 |
| [animales] <muy> sensibles | animales sensibles | 3 | 1 | 1 | 0 | 1 | 6,63 | 6,31 | 3,54 | 73,04 |

Cuadro 1. Ejemplo ilustrativo de registro de atributos

Además, se crearon dos variables basadas en la distancia de Levenstein (Oakes 1998; Nerbonne *et al*. 2013), con el objetivo de medir la similitud entre las raíces léxicas entre el español y el italiano; y entre los morfemas de género y número plural. Además se las combinó en un índice para representar la distancia total del par de términos de la concordancia. Por último, se crearon siete atributos binarios de estrategia para la formación del plural: cada atributo registraba 1 en aquella instancia donde la estrategia de plural podía ser aplicada en alguno de los dos términos de concordancia (o en ambos). Dichas estrategias buscaron identificar casos que facilitaran o dificultaran la producción de concordancias. Se definieron como sigue:

- *Estrategia 1* (*Est1*): si la palabra plural del italiano termina en *-i*, poner en español plural en *-os*.
- *Estrategia 2* (*Est2*): si la palabra plural del italiano termina en *-e*, poner en español plural en *-as*.
- *Estrategia 3* (*Est3*): si la palabra plural del italiano termina en *-o* o en *-a* no acentuada (*le foto* [‘las fotos’], *le osa* [‘los huesos’]), poner el plural del italiano.
- *Estrategia 4* (*Est4*): si la palabra plural del italiano termina en *-e*, poner en español el plural en *-es*. Por ej.: *vacanze* > *vacaciones*; *strade* > *calles*; *volte* > *veces*.

- *Estrategia 5* (*Est5*): si la palabra singular del italiano termina en *-e*, poner en español el plural en *-es*. Por ej.: la palabra *sole* ('sol') podría ser la base para formar el plural español agregando *-s*: *sole* > *soles*; y el singular también, sacando *-s*: *sole* > *sol*; *istituzione* > *instituciones*. Es decir, casos en los cuales el español coincide con la aplicación del plural con *-e-* epentética.
- *Estrategia 6* (*Est6*): si la palabra singular del italiano termina en *-e*, poner en español el plural en *-es*. Por ej.: la palabra *grande* ('grande') o *studente* ('estudiante') podrían formar plural (y singular) a partir de una base singular en italiano: *grandes, estudiantes.* Otros casos: *fonte* > *fuentes*; *abitudine* > *costumbres*; *dolce* > *dulces*. Son casos que no coinciden con la *-e-* epentética.
- *Estrategia 7* (*Est7*): si la palabra plural del italiano termina en *-a* acentuada (*università* ['universidades']) o es invariante terminada en consonante (*i film* ['las películas']), poner, en general, plural en *-es*.

El Cuadro 2 ejemplifica los casos en italiano, español y la instancia efectivamente producida por el alumno.

| **Italiano plural** | **Italiano singular** | **Español** | **Instancia** | **Est 1** | **Est 2** | **Est 3** | **Est 4** | **Est 5** | **Est 6** | **Est 7** |
|---|---|---|---|---|---|---|---|---|---|---|
| molte volte | (molta) volta | muchas veces | muchas vesas | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| vacanze gradevoli | vacanza gradevole | vacaciones agradables | vacacione agreables | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| uniche moto | unica moto | únicas motos | unicas moto | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| molti cinema | (molto) cinena | muchos cines | muchos cines | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| responsabilitá sociali | responsabilitá sociale | responsabilidades sociales | responsabilidades sociales | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| le abitudini | l'abitudine | las costumbres | los costumbre | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Cuadro 2. Ejemplos de estrategias

Se realizaron las siguientes operaciones de preprocesamiento. En primera instancia, se transformaron al logaritmo los atributos que tienen que ver con las frecuencias y los errores acumulados, sumándoles una unidad, de la forma que sigue: (i) frecuencia de controlador (corpus BuscaPalabras) como LEXESP = log(LEXESP + 1); (ii) Acum = log(Acum+1).

En segundo lugar, había datos faltantes. Se recolectaron 1857 casos de concordancia en total. Sin embargo, los atributos relacionados con el controlador (excepto Anim) a veces no tenían datos registrados en la base de datos de BuscaPalabras. Debido a ello, hubo 161 casos en los que faltaban datos en una o más de estas variables. Los casos faltantes representaron el

8,6 % de la base de datos. Se utilizó el paquete Multivariate Imputation by Chained Equations (*mice*) de R (Van Buuren & Groothuis-Oudshoorn 2011), que realiza imputación múltiple.

En tercer lugar, fue necesario resolver la colinealidad existente entre las variables Imaginabilidad (Ima), Concretud (Conc), Familiaridad (Fam), Frecuencia en escala logarítmica (LEXESP). Se aplicó un análisis de componentes principales (PCA, por su sigla en inglés).

La técnica permite obtener nuevas variables ortogonales llamadas *componentes principales*, que se calculan como combinación lineal de las variables cuantitativas originales (Peña 2002). En el primer componente (llamado *Ima.Conc*) cargaban las variables Imaginabilidad y Concretud; y en el segundo (llamado *Fam.LEX*), Familiaridad y LEXESP. En suma, se resolvió la colinealidad pasando de cuatro variables de controlador con correlación a dos componentes sin correlación entre sí.

Por último, se discretizaron los atributos cuantitativos Morf, Stem, Ima.Conc y Fam.LEX utilizando *clustering* por mezcla de gausianas. Se utilizó el paquete *mclust* de R (Scrucca *et al*. 2016) para hacer el agrupamiento. El atributo Fam.LEX se discretizó en dos categorías poniendo como punto de corte a la mediana, ya que el *clustering* no resultó efectivo.

A modo de resumen, los Cuadros 3 y 4 que siguen muestran las variables creadas.

| **Atributo** | **Descripción** | **Discretización** | **Casos** | **Ejemplos del corpus** |
|---|---|---|---|---|
| Morf.f | Similitud entre terminaciones | 0 = [2,8; 2,2; 2,6; 2,4; 3) | 159 | mis amigos (2,8); vacaciones agradables (2,6) |
| | | 1 = [3; 3,2; 3,4) | 1258 | las personas (3); mujeres jóvenes (3,2) |
| | | 2 = [3,4; 3,6] | 440 | los trenes (3,4); relaciones industriales (3,6) |
| Stem.f | Similitud entre raíces léxicas | 0 = [1,8; 4) | 1499 | todas reglas (1,8); los grupos (2,9) |
| | | 1 = [4; 10,2] | 358 | alemanes fieles (4,5); mujeres guapas (5,8) |
| Ima.Conc.f | PCA1 | 0 = [−3,48; 0,58) | 1163 | nuevos conocimientos (−3,11); los servicios (−1,19) |
| | | 1 = [0,58; 2,35] | 694 | muchas personas (1,31); los hospitales (2,17) |
| Fam.LEX.f (corte: mediana) | PCA2 | 0 = −4,24; 0,17) | 934 | los sultanes (−4,42); las comodidades (−0,64) |
| | | 1 = [0,17; 1,98] | 923 | los años (1,22); los hombres (1,89) |

Cuadro 3. Discretización de atributos utilizando *clustering* por mixtura de gausianas

| Variable | Descripción | Clase | Niveles |
|---|---|---|---|
| Target | Estatus: acierto/error | Cualitativa | 0 = acierto; 1 = error |
| Alumno | Alumno | Cualitativa | - |
| Sesión | Sesión transcripta | Cualitativa | - |
| Línea | Línea en la transcripción en formato CHAT | Cuantitativa | [6; 515] |
| Instancia | Instancia de concordancia observada | Caracteres | - |
| Mod | Tipo de modificador del controlador | Cualitativa | 0 = artículo definido; 1 = artículo indefinido; 2 = determinante; 3 = adjetivo |
| LDa | Concordancia a larga distancia | Cualitativa | 0 = sin larga distancia; 1 = con larga distancia |
| ES | Presencia de *-e-* epentética | Cualitativa | 0 = sin *-e-* epentética; 1 = con «e» epentética en un término; 2 = con *-e-* epentética en ambos términos |
| Gram | Concordancia de dos términos o más | Cualitativa | 0 = dos términos; 1 = más de dos términos |
| Conc | Concretud del controlador | Cuantitativa | Escala de 1 a 7 |
| Fam | Familiaridad del controlador | Cuantitativa | Escala de 1 a 7 |
| Ima | Imaginabilidad del controlador | Cuantitativa | Escala de 1 a 7 |
| LEXESP | Log(Frecuencia del controlador + 1) | Cuantitativa | [2,32, 744,6] |
| Anim | Animicidad del controlador | Cuantitativa | 0 = inanimado; 1 = animado |
| Esp | Instancia en español | Caracteres | - |
| Morf | Similitud entre terminaciones | Cuantitativa | [2,2, 3,6] |
| Stem | Similitud entre raíces | Cuantitativa | [1,8; 10,2] |
| Est1 | Estrategia 1 | Cualitativa | 0 = no aplica; 1 = aplica |
| Est2 | Estrategia 2 | Cualitativa | 0 = no aplica; 1 = aplica |
| Est3 | Estrategia 3 | Cualitativa | 0 = no aplica; 1 = aplica |
| Est4 | Estrategia 4 | Cualitativa | 0 = no aplica; 1 = aplica |
| Est5 | Estrategia 5 | Cualitativa | 0 = no aplica; 1 = aplica |

| | | | |
|---|---|---|---|
| Est6 | Estrategia 6 | Cualitativa | 0 = no aplica; 1 = aplica |
| Est7 | Estrategia 7 | Cualitativa | 0 = no aplica; 1 = aplica |
| Acum | log(errores acumulados + 1) | Cuantitativa | [0, 28] |
| Esp1 | Terminación del primer término en español | Cualitativa | 1 = *-us* (por ej.: *sus*); 2 = *-is* (por ej.: *mis*); 3 = *-os* (por ej.: *rojos*); 4 = *-as* (por ej.: *blancas*); 5 = *-es* (por ej.: *celestes*); 6 = *-*es* (por ej.: *azules*) |
| Esp2 | Terminación del segundo término en español | Cualitativa | 1 = *-us* (por ej.: *sus*); 2 = *-is* (por ej.: *mis*); 3 = *-os* (por ej.: *rojos*]; 4 = *-as* (por ej.: *blancas*); 5 = *-es* (por ej.: *celestes*); 6 = *-*es* (por ej.: *azules*) |

Cuadro 4. Variables creadas

## 4. Predicción de estatus de error

En esta sección se intenta predecir el estatus de error de concordancia (correcto/incorrecto) en las últimas cuatro sesiones de cada aprendiente. Se usan predictores que involucran características de las concordancias, dinámica de error y atributos de los grafos de concordancia. Se crearon cuatro redes complejas acumulando los casos de concordancia de las sesiones. Las redes eran dirigidas, donde la dirección indicaba la relación de asimetría del controlador en la concordancia plural. Cada nodo fue una palabra y las aristas indicaban una relación de concordancia. Específicamente, los predictores fueron los siguientes, calculados con las concordancias pertenecientes al componente gigante de los grafos (Newman 2010, ver también Marafioti 2021).

En primer lugar, se usaron los atributos que describen las concordancias (ya descriptos): Mod, LDa, Grams, Es, Morf.F, Stem.F, Acum, Est1, Est2, Est3, Est4, Est5, Est6, Est7, Ima. Conc.F, Fam.LEX.F, Anim, Esp1, Esp2. En segunda instancia, hubo atributos derivados de los grafos que caracterizan los enlaces del grafo (o sea, las concordancias) en términos de la proximidad/similitud entre los nodos conectados por el enlace. Se calcularon mediante la librería de R *linkprediction* (Bojanowski & Chrol 2019). También se incluyó la frecuencia del enlace (cuántas veces aparece una determinada concordancia). En lo que atañe a los atributos dinámicos, se calculó la cantidad de errores acumulados hasta la instancia de concordancia anterior a la instancia n-ésima. Sobre dicho vector se aplicó una ventana móvil de largo w = 45 (o sea que las primeras 44 observaciones recibieron datos faltantes). Sobre dicha ventana se calcularon predictores que tenían que ver con estadística descriptiva, autocorrelación, entropía de permutación y complejidad (Bandt & Pompe 2002), análisis cuantitativo de recurrencias (Webber & Marwan 2015), onditas (Percival & Walden 2008; Nason 2008), entre otros.

Asimismo, se tomaron las 50 palabras más frecuentes como predictores. En total hubo 126 predictores. El siguiente cuadro muestra el total de observaciones asignadas al conjunto de entrenamiento (70 % de los datos) y al de validación (30 % de los datos), para cada alumno. Los datos faltantes de las primeras 44 observaciones de los predictores dinámicos fueron reemplazados por cero. Los atributos continuos fueron estandarizados ($(x - \mu_x)/\sigma_x$). La variable respuesta fue $y \in \{0,1\}$, donde 1 era concordancia incorrecta y 0, correcta. Entre paréntesis se detalla la cantidad de observaciones por categoría de respuesta.

| | **Sonia** | **Nati** | **Jako** | **Mirka** |
|---|---|---|---|---|
| Entrenamiento | 207<br>(0: 158, 1: 49) | 278<br>(0: 178, 1: 100) | 336<br>(0: 276, 1: 60) | 438<br>(0: 322, 1: 116) |
| Validación | 96<br>(0: 83, 1: 13) | 105<br>(0: 81, 1: 24) | 115<br>(0: 104, 1: 11) | 247<br>(0: 163, 1: 84) |
| Total | 303 | 383 | 451 | 685 |

Cuadro 5. Conjuntos de entrenamiento y validación para cada alumno

El desbalance de clase muestra la relación de la clase minoritaria respecto a la mayoritaria ($IR = \frac{\#\,clase\,1}{\#\,clase\,0} \in [0,1]$ $IR = 1$ si están balanceadas). Para el conjunto de entrenamiento es: (i) Sonia: IR ≈ 0,31; Nati: IR ≈ 0,56; Jako: IR ≈ 0,21; Mirka: IR ≈ 0,36. Ya que el desbalance puede afectar el desempeño de los clasificadores, se usaron las siguientes estrategias para corregirlo (o sea, obtener IR ≈ 0,5 para el conjunto de entrenamiento): (i) *undersampling* (US): tomar al azar un subconjunto de observaciones de la clase mayoritaria; (ii) *oversampling* (OS): duplicar al azar un conjunto de observaciones de la clase minoritaria; (iii) *weighting* (CW): pesar las clases (aumentando la importancia de la categoría minoritaria); (iv) *synthetic minority oversampling* (SMOTE): generar observaciones de la clase minoritaria como combinaciones aleatorias convexas de los vecinos de las observaciones.

Se aplicaron los siguientes clasificadores: (i) *Support Vector Machine* (SVM); (ii) *Random Forest* (RF); (iii) *Recursive Partitioning and Regression Trees* (RPART); (iv) *Gradient Boosting Machine* (GBM); (v) *Extreme Gradient Boosting* (XGB); (vi) *logistic regression* (LogReg). En el cuadro que sigue se indica el rango de los parámetros de cada uno para afinar. También se indica el rango de afinación de los parámetros de los métodos de corrección de desbalance. Cada clasificador se aplicó tanto con dichos métodos de corrección como sin ellos.

| Método | Parámetros. Rango: (inferior, superior) |
|---|---|
| SVM | C = ($2^{-8}$, $2^{15}$); sigma = ($2^{-15}$,$2^{3}$) |
| RF | mtry = (3, 10); ntree = (50, 500); nodesize = (10, 50) |
| RPART | cp = (0,0001, 0,1); minsplit = (1, 10); minbucket = (5, 50) |
| XGB | nrounds = (200, 600); max depth = (3, 20); lambda = ($2^{-10}$, $2^{10}$) |
| GBM | n.trees = (100, 5000); interaction.depth = (2, 10); bag.fraction = (0,7, 1); shrinkage = (0,001, 0,5); n.minobsinnode = (5, 15) |
| LOGREG | - |
| US | rate = (0,6 × IR) |
| OS | rate = (1,5 × $IR^{-1}$) |
| CW | weight class 1 = (1, 10); weigth class 0 = 1 |
| SMOTE | rate = (1,5 × $IR^{-1}$) |

Cuadro 6. Parámetros de los clasificadores empleados

Se llevó a cabo una primera selección de 80 atributos por medio de un *ensemble* de medidas de selección, a saber: área bajo la curva (AUC), *relief score* (Kira & Rendell 1992; cfr. también el cap. 18 de Kuhn & Johnson 2013), pesos de *Random Forest*, información mutua (*praznik_JMI*).

Si se tiene la tabla de confusión, donde: VP = verdaderos positivos; VN = verdaderos negativos; FP = falsos positivos (falsa alarma); FN = falsos negativos (negativos clasificados como positivos); se pueden derivar las medidas descriptas en el Cuadro 7, utilizadas para verificar el desempeño de los modelos entrenados con el conjunto de evaluación. A estas, se agregó el área bajo la curva ROC (AUC) y una medida de costo (C) de clasificación. Esta última multiplica por un peso *w* a cada observación predicha (1 o 0) y luego saca un promedio. Hubo una matriz de costo para cada alumno, donde las filas eran los niveles observados de la variable respuesta y las columnas, aquellos predichos. Observar que la celda (observado = 0, predicho = 0) siempre fue nula:

$$\text{SONIA} = \begin{bmatrix} 0 & 1 \\ 6 & 0 \end{bmatrix};\ \text{NATI} = \begin{bmatrix} 0 & 2 \\ 6 & -2 \end{bmatrix};\ \text{JAKO} = \begin{bmatrix} 0 & 1 \\ 10 & 0 \end{bmatrix};\ \text{MIRKA} = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}$$

| Medidas | Fórmula | Definición |
|---|---|---|
| Precisión balanceada (*Balanced accuracy*, BAc) | $\left[\frac{VP}{FP+FN} + \frac{VN}{FP+VN}\right]/2$ | Proporción de clasificados correctamente del total (para datos desbalanceados) |
| Sensibilidad (*recall*) | $\frac{VP}{VP+FN}$ | Proporción de positivos clasificados correctamente y que lo son en el patrón de referencia |
| Valor predictivo positivo (VPP) | $\frac{VP}{VP+FP}$ | Proporción de positivos clasificados correctamente (VP) de todos los clasificados como positivos (VP+FP) |
| F1 | $2\times\left[\frac{VPP \times recall}{VPP + recall}\right]$ | Media armónica entre VPP y *recall* |
| Tasa de falsos positivos | $\frac{FP}{FP+VN}$ | Probabilidad de falsa alarma |
| Tasa de falsos negativos | $\frac{FN}{FN+VP}$ | Probabilidad de falsos negativos |

Cuadro 7. Medidas derivadas de la tabla de confusión

A continuación, se aplicó para cada método una segunda selección de atributos (en el rango [1; 80]) basado en la medida *praznik_JMI* en conjunto con el entrenamiento de cada modelo de clasificación. El conjunto de entrenamiento se dividió a su vez en dos conjuntos de entrenamiento y validación aplicando validación cruzada con cinco *folds*. Se evaluó el desempeño de los modelos mediante la minimización de: (i) tasa de falsos positivos; (ii) tasa de falsos negativos; (iii) medida de costo de clasificación. Usualmente si P (x = correcto) > 0,5, se le asigna la categoría 1= correcto y, si no, la de 0 = incorrecto; sin embargo, aquí se usó la medida de costo para optimizar el punto de corte de la probabilidad para la decisión de asignar los valores a la categoría correcto u error. Se llevó a cabo el análisis usando la librería *mlr* de R (Bischl *et al*. 2016).

Los Cuadros 8 a 11 muestran los resultados de los mejores clasificadores obtenidos, por alumno. En primer lugar, se muestran los clasificadores: SVM, RF, RPART, GBM, XGB, LogReg, *ensemble*. Seguidamente, los tipos de pesos: *undersampling* (US), *oversampling* (OS), *SMOTE*, *class weight* (CW), base (sin pesos). Por ultimo, están las medidas de desempeño: precisión balanceada (BAc), valor predictivo positivo (VPP), *recall* (Rec), F1, área bajo la curva (AUC), medida de costo (C). La fila denotada *ensemble* se refiere a un clasificador que usa como predictores los valores predichos del resto de los clasificadores de la tabla. Las filas se hallan ordenadas según los valores de precisión balanceada (BAc).

En general, no se logró superar el 80 % de precisión balanceada. En el caso de Sonia, los dos mejores clasificadores resultaron ser SVM con *US* y RF con SMOTE. Para Nati, RPART con *CW* y SVM con *OS*. Respecto de Jako, SVM con *US* y XGB con *CW*. En lo que atañe a Mirka, fueron SVM con *OS* y RPART con SMOTE. Con lo cual SVM resultó el clasificador de mejor desempeño. Sonia y Jako alcanzaron mejores niveles de precisión balanceada que Nati y Mirka. Por otro lado, el *ensemble*, aunque fue el valor más alto de precisión balanceada en el caso de Sonia y Mirka, representó una mejora muy leve respecto del mejor de los clasificadores para ambos casos.

De los clasificadores con mejor desempeño apuntados, RF, RPART y XGB permiten ordenar los predictores seleccionados asignando pesos que indican la contribución, dentro del modelo, de cada predictor para clasificar la variable respuesta. Por lo tanto, es posible ordenar los predictores según su importancia. El Cuadro 12 muestra esto mismo, con los diez predictores más importantes en cada modelo. En lo que respecta a Sonia, resultaron de importancia: la imaginabilidad y concretud altas del controlador, que haya un término en la concordancia con *-e-* epentética, la animicidad del controlador y que esté presente un artículo definido. Las otras variables fueron dinámicas (autocorrelación [lag-2] y *wavelets* [D5]) y de los enlaces del grafo ("rwr", "pa", "cos_l"). En el caso de Nati resultaron de importancia para predecir la respuesta: la frecuencia del artículo *los*; variables dinámicas (autocorrelación [lag-1, lag-2], *wavelets* [D4], curtosis [kurt] y desvío típico [SD]) y variables de enlaces del grafo ("mf", "cos_l", "act_n", "1"). Notar que no hubo variables de atributos de concordancias en este caso. En lo que concierne a JAKO, fueron importantes: la terminación en "as" del primer término de la concordancia, la similitud (de las terminaciones) media y alta con el italiano, los modificadores que son adjetivos, la similitud media con el italiano de la raíz léxica, que haya "e" epentética en un término de la concordancia, la frecuencia de los enlaces del grafo y la frecuencia de las palabras "lugares", "competencias" y "particulares". En lo que atañe a MIRKA, solamente seis atributos fueron importantes: la frecuencia de enlace y los atributos de enlace del grafo ("mf", "ka", "act", "rwr", "act_n"). Nótese que los atributos derivados del grafo fueron seleccionados para todos los alumnos, lo cual aboga por la inclusión de datos sobre la estructura de grafos en tareas de clasificación.

| **Clasificador** | **Tipo** | **BAc** | **VPP** | **Rec** | **F1** | **AUC** | **C** |
|---|---|---|---|---|---|---|---|
| Ensemble | - | 0,800 | 0,417 | 0,769 | 0,541 | 0,841 | 0,104 |
| SVM | US | 0,788 | 0,385 | 0,769 | 0,513 | 0,726 | 0,354 |
| RF | SMOTE | 0,785 | 0,324 | 0,846 | 0,468 | 0,794 | 0,365 |
| GBM | CW | 0,772 | 0,306 | 0,846 | 0,449 | 0,718 | 0,385 |
| XGB | CW | 0,754 | 0,282 | 0,846 | 0,423 | 0,678 | 0,417 |
| LogReg | CW | 0,687 | 0,286 | 0,615 | 0,390 | 0,652 | 0,521 |
| RPART | CW | 0,684 | 0.250 | 0,692 | 0,367 | 0,652 | 0,531 |

Cuadro 8. Resultados de Sonia

| Clasificador | Tipo | BAc | VPP | Rec | F1 | AUC | C |
|---|---|---|---|---|---|---|---|
| RPART | CW | 0,671 | 0,395 | 0,625 | 0,484 | 0,654 | 0,667 |
| SVM | OS | 0,664 | 0,328 | 0,833 | 0,471 | 0,598 | 0,629 |
| XGB | CW | 0,624 | 0,302 | 0,792 | 0,437 | 0,553 | 0,762 |
| LogReg | Base | 0,621 | 0,284 | 0,958 | 0,438 | 0,596 | 0,724 |
| GBM | CW | 0,621 | 0,326 | 0,625 | 0,429 | 0,592 | 0,819 |
| RF | CW | 0,618 | 0,297 | 0,792 | 0,432 | 0,614 | 0,781 |
| Ensemble | - | 0,590 | 0,267 | 0,958 | 0,418 | 0,706 | 0,400 |

Cuadro 9. Resultados de Nati

| Clasificador | Tipo | BAc | VPP | Rec | F1 | AUC | C |
|---|---|---|---|---|---|---|---|
| SVM | US | 0,784 | 0,257 | 0,818 | 0,391 | 0,768 | 0,400 |
| XGB | CW | 0,760 | 0,225 | 0,818 | 0,353 | 0,685 | 0,443 |
| RF | SMOTE | 0,736 | 0,292 | 0,636 | 0,400 | 0,701 | 0,496 |
| LogReg | SMOTE | 0,717 | 0,250 | 0,636 | 0,359 | 0,646 | 0,530 |
| Ensemble | - | 0,710 | 0,316 | 0,545 | 0,400 | 0,786 | 0,148 |
| RPART | US | 0,702 | 0,151 | 1,000 | 0,262 | 0,725 | 0,539 |
| GBM | OS | 0,674 | 0,155 | 0,818 | 0,261 | 0,629 | 0,600 |

Cuadro 10. Resultados de Jako

| Clasificador | Tipo | BAc | VPP | Rec | F1 | AUC | C |
|---|---|---|---|---|---|---|---|
| Ensemble | - | 0,674 | 0,467 | 0,845 | 0,602 | 0,715 | 0,146 |
| SVM | OS | 0,647 | 0,447 | 0,810 | 0,576 | 0,674 | 0,194 |
| RPART | SMOTE | 0,644 | 0,449 | 0,786 | 0,571 | 0,644 | 0,206 |
| XGB | SMOTE | 0,602 | 0,416 | 0,738 | 0,532 | 0,536 | 0,279 |
| LogReg | Base | 0,579 | 0,385 | 0,881 | 0,536 | 0,530 | 0,259 |
| RF | OS | 0,559 | 0,371 | 0,940 | 0,532 | 0,581 | 0,263 |
| GBM | US | 0,554 | 0,371 | 0,857 | 0,518 | 0,524 | 0,300 |

Cuadro 11. Resultados de Mirka

| Sujeto | Importancia |
|---|---|
| Sonia (RF) | Ima.Conc = 1 (6); rwr (3,71), l (2,35), es = 1 (2,25), ACF_2 (2,24), D5 (2,04), cos_l (2), Anim = 1 (1,87), pa (1,80), Mod = 0 (1,73) |
| Nati (RPART) | mf (29,75), l (27,15), cos_l (20,28), act_n (20,07), *los* (19,15), D4 (18,73), SD (16,71), ACF_1 (16,17), ACF_2 (15,47), kurt (13,58) |
| Jako (XGB) | Esp1 = *-as* (0,16); Morf.f = 1 (0,09); *lugares* (0,07), *competencias* (0,066), Morf.f = 2 (0,063), Mod = 3 (0,055), Stem.f = 1 (0,051), ES = 1 (0,048), *particulares* (0,044), Freq_enlace (0,043) |
| Mirka (RPART) | mf (33,81), Freq_enlace (6,0433), ka (6,0434), act (4,41), rwr (3,92), act_n (2,94) |

Cuadro 12. Los diez primeros predictores más importantes (pesos entre paréntesis)

## 5. Variables relacionadas con la intensidad del error

En primera instancia, se recategorizó la variable Esp1 con los siguientes *niveles*; 1 = *-us*, *-is*; 3 = *-os*, 4 = *-as*; 5 = *-es*; 6 = **-es*, porque la segunda categoría (2 = *-is*) incluía pocos casos, lo cual generaba problemas de estimación. Se aplicó un clustering de distribuciones Bernoulli, con datos agrupados por sesión. Es decir que los clusters agrupan las observaciones dentro de las sesiones. Se utilizó la implementación del paquete flexmix de R (Grün & Leisch 2007). La Figura 1 muestra los dos clusters obtenidos. Se observa que el primero agrupa las sesiones en torno a una media de error del 34 % (rango [0,32; 0,62]); por otro lado, el segundo las agrupa entorno a un error del 17 % (rango [0; 0,27]). Por ende, se obtuvo un conjunto de error alto y otro de error medio-bajo. Para denotar las sesiones, la letra es la inicial del alumno y el número es la sesión (por ej., *s4* es la sesión 4 de Sonia).

Luego, se aplicó un modelo mixto de regularización a cada conjunto de datos. Se usó el paquete *glmmLasso* de R (Groll & Tuntz 2014). Allí se implementa un modelo mixto generalizado con penalización $L_1$ que permite realizar una selección de variables mediante *shrinkage* de coeficientes (Lasso mixto). El parámetro de regularización $\lambda$ se obtuvo mediante validación cruzada. Los factores aleatorios fueron las sesiones. El Cuadro 13 muestra las primeras doce variables seleccionadas para cada *cluster*.

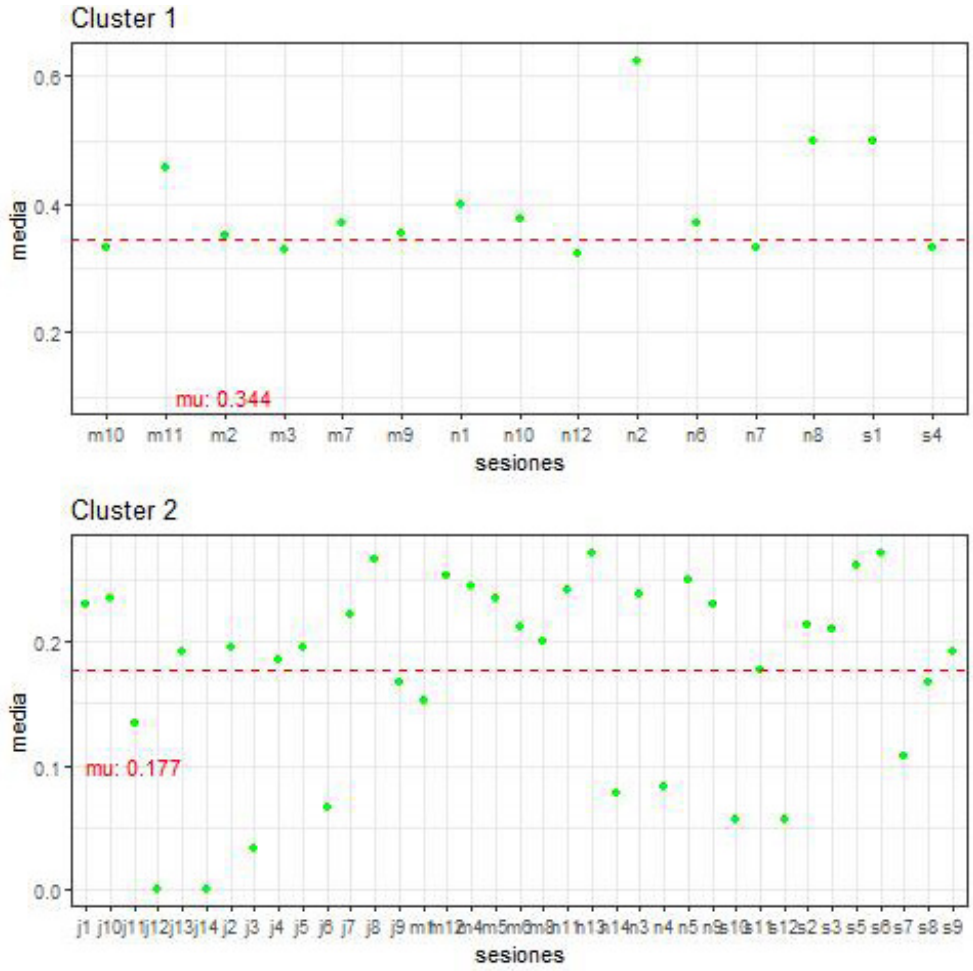


Figura 1. *Clusters* obtenidos mediante mixturas de Bernoulli

| ***Cluster*** | **λ** | **Selección** |
|---|---|---|
| 1 | 15 | Esp1, Morf.f, Mod, Est5, LDa, Est2, C, Anim, Est6, Esp2, SKEW, STEM |
| 2 | 5 | ES, Esp1, Morf.f, Est6, Fam.LEX, LDa, Stem, Est4, Est1, Est7, Mod, Esp2 |

Cuadro 13. Selección de variables mediante regularización mixta

A continuación, se empleó un modelo logístico mixto condicional para la concordancia *i* en el grupo *j* (el grupo está definido como la sesión *k* anidada en el alumno *g*). Se practicó una selección de modelos basada en medidas de información (Burnham & Anderson 2010). Primero, se ajustaron modelos logísticos mixtos con las predictoras significativas seleccionadas. Fueron $2^{12} - 1 = 4095$ modelos, jerarquizados mediante la medida de información AIC (como: $\frac{n_{c.1}}{p} = \frac{631}{12} \approx 52 > 40$ y $\frac{n_{c.2}}{p} = \frac{1191}{12} \approx 99 > 40$, no se usó la versión AICc corregida por tamaño muestral). Luego, se examinó la frecuencia de las predictoras en el conjunto completo de modelos, que da un panorama de la incerteza por la selección. A continuación, se redujo la cantidad de modelos al subconjunto de confianza con la regla $\frac{W(i)}{W(1)} > \frac{1}{8}$. Sobre dicho subconjunto se llevó a cabo un promedio de coeficientes con la varianza calculada con *full average* (ver Apéndice).

Las gráficas de las Figuras 2 y 3 muestran, para ambos *clusters*, los *odds ratio* del mejor modelo hallado. En lo que respecta al primer *cluster* (error alto), el cuadro de coeficientes promediados (ver Apéndice) indica que resultan significativas las variables: Anim = 1, C (Complejidad), Esp1 = *-os*, Esp1 = *-as*, Esp1 = *-es*, Esp1 = **-es*, Est5 = 1, Mod = 2, Mod = 3, Skew (asimetría). En lo que atañe al segundo *cluster*, las variables significativas fueron: Esp1 = *-os*, Fam.LEX = 1, Esp2 = *-as*. Nótese que los *odds ratio* significativos del primer *cluster* coinciden con los coeficientes promediados. En cambio, el mejor modelo encontrado para el *cluster* 2 señala que también son significativas las variables: Esp2 = *-as*, Est1, Mod = 2, Stem = 1. O sea que este modelo en particular no coincide con el promedio de los coeficientes de los modelos del subgrupo de confianza.

En resumen, la chance de cometer un error de concordancia (Y = 1) respecto de no cometerlo (Y = 0) disminuye cuando se pasa de la categoría de referencia a la k-ésima categoría de la predictora para: (i) el primer término de la concordancia terminadas en *-os* (respecto de la referencia *-is*, *-us*) para ambos *clusters*; (ii) las palabras que funcionan como controlador de familiaridad y frecuencia léxica alta (*cluster* 2); (iii) las palabras objetivo de la concordancia terminadas en *-as*, *-es*, **-es* (*cluster* 1); (iv) la estrategia 5 (*cluster* 1). Por otro lado, la chance de cometer un error de concordancia (Y = 1) respecto de no cometerlo (Y = 0) aumenta cuando se pasa de la categoría de referencia a la k-ésima categoría de la predictora para: (i) los controladores animados, (ii) los modificadores que son adjetivos o determinantes (ambos en el *cluster* 1). Además, la chance de error disminuye a medida que aumenta la complejidad de errores acumulados precedentes medido por C (*cluster* 1). La chance aumenta a medida que se incrementa la asimetría de la distribución de errores acumulados precedentes (Skew) (*cluster* 1). Resulta de interés notar que las terminaciones de los primeros términos (Esp1) constituyeron siempre factor de protección; y los tipos de modificador (Mod), siempre factor de riesgo. La animicidad del controlador fue factor de riesgo para las concordancias de error alto pero dicha variable no resultó seleccionada en el caso de error medio/bajo.

Ya que para el caso de las variables categóricas, los p-valores no están ajustados por tests múltiples, las Figuras 4 y 5 muestran las comparaciones entre los niveles ajustando los p-valores por el método de Tukey. En lo que atañe al *cluster* 1, resultan significativas las palabras del primer término terminadas en *-os* y **-es* (*-e-* epentética) respecto de la

referencia *-us*, *-is* (la chance de error disminuye). También hay diferencia significativa entre las palabras terminadas en *-as* con respecto a las terminadas en *-os* (la chance de error aumenta). En cuanto a la variable Mod del primer *cluster*, solo resultó significativa la diferencia entre las concordancias con modificador adjetivo y el nivel de referencia (artículo definido) (la chance de error aumenta). En lo que respecta al segundo *cluster*, se observaron las siguientes diferencias significativas: (i) Esp1: las palabras terminadas en *-os* respecto de la referencia (la chance de error disminuye) y aquellas terminadas en *-as* y *-es* respecto de las terminadas en *-os* (la chance de error aumenta); (ii) Esp2: las palabras del segundo término de la concordancia terminadas en *-as* respecto de aquellas terminadas en *-os* (referencia) (la chance de error disminuye) y aquellas terminadas en *-es* respecto de aquellas terminadas en *-as* (la chance de error aumenta); (iii) Mod: ninguna diferencia resultó significativa.

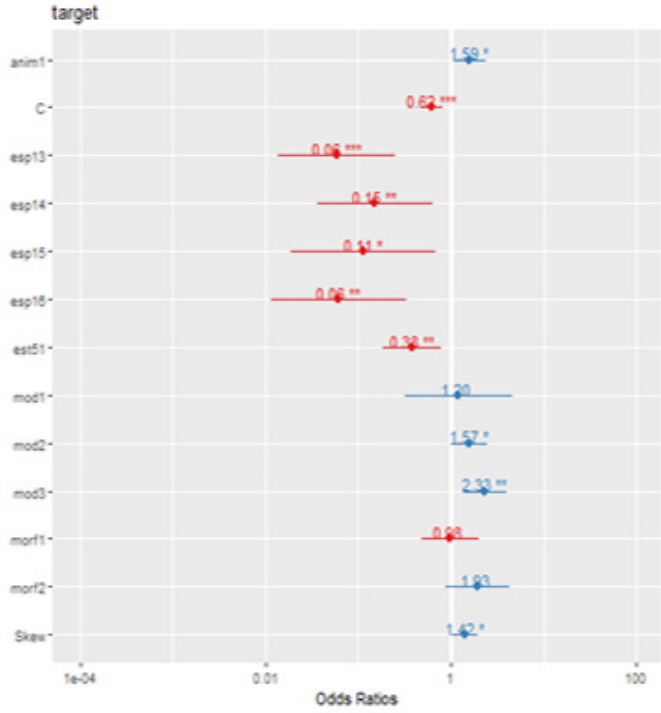


Figura 2. *Odds ratio* de los mejores modelos para el *cluster* 1

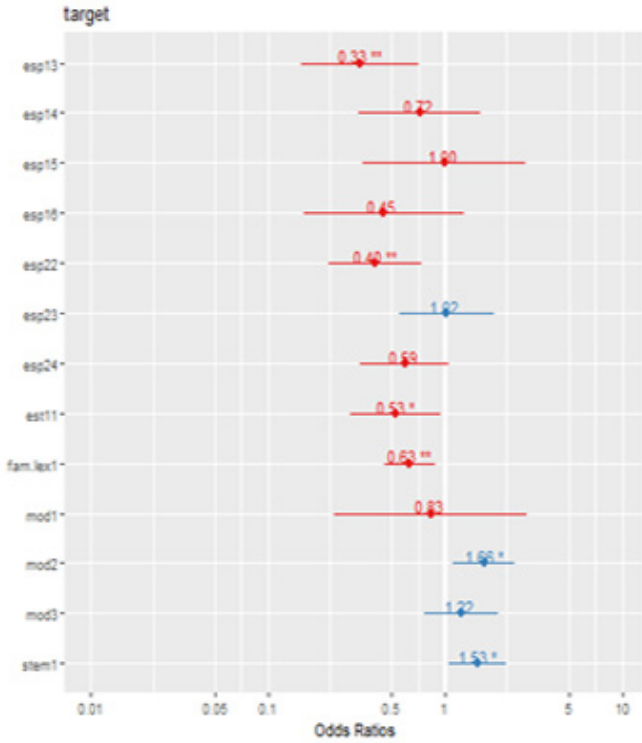


Figura 3. *Odds ratio* de los mejores modelos para el *cluster* 2

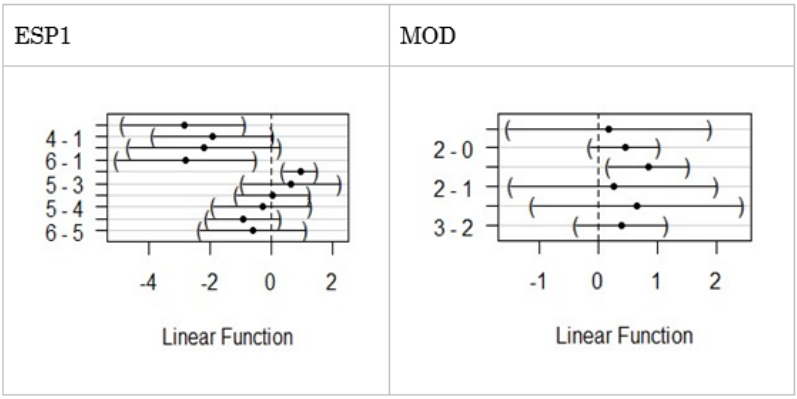


Figura 4. Cluster 1: Comparaciones múltiples (p-valores ajustados por Tukey)

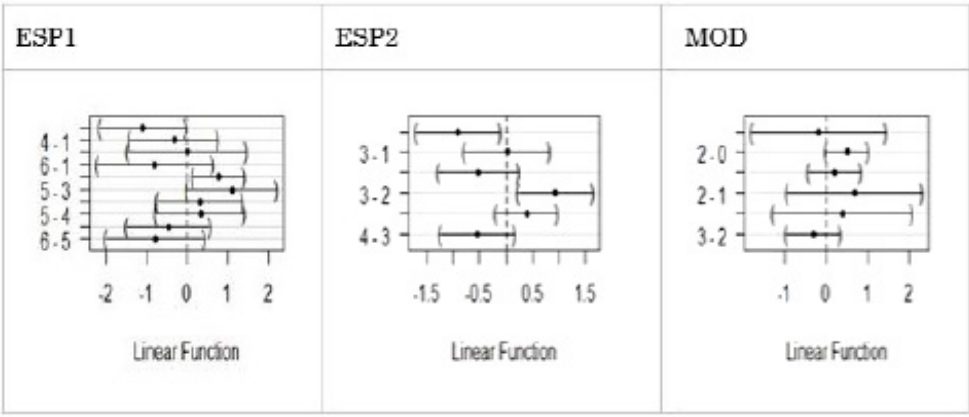


Figura 5. *Cluster* 2: comparaciones múltiples (p-valores ajustados por Tukey)

## 6. Discusión

Un hallazgo bastante establecido en la literatura sobre adquisición de la concordancia en L2 es que la concordancia de género y número del artículo es más fácil de adquirir que la del adjetivo. El Lasso mixto generalizado seleccionó la variable Mod tanto para el grupo de error alto (media 34 %) como para el de error medio/bajo (media 17 %). Sin embargo, en el segundo grupo Mod no resultó significativo. En el primero, de error medio más alto, se observó un aumento de la chance de error para los adjetivos respecto de los artículos definidos. En lo que respecta a las variables más influyentes en la predicción de estatus de error, los artículos definidos se seleccionaron en el caso de Sonia y los adjetivos en el caso de Jako. Además, para Nati influyó el artículo *los* para predecir el error.

Respecto del género, en la literatura se halló más facilidad de adquisición de la concordancia para los nombres de morfología transparente (terminados en *-o*, *-a*) respecto a los de morfología no transparente (terminados en consonante, *-e* y opuestos tipo *la mano*). En cuanto al plural, el morfema en *-es* se adquiere después del morfema en *-s*. La variable ES fue codificada como 0 si no hay necesidad de insertar una *-e-* epentética (controladores y objetivos de morfología transparente para el género, terminados en *-e* y géneros opuestos); y como 1 y 2, para aquellos terminados en consonante que requieran plural en *-es*, en un único término de la concordancia o en ambos. El Lasso mixto generalizado seleccionó la variable ES para el grupo de error medio/bajo pero *no* resultó significativa. En lo que atañe al primer *cluster* (error alto), resultaron significativas las palabras del primer término (Esp1) terminadas en *-os* y **-es* (*-e-* epentética) respecto de la referencia *-us*, *-is* (la chance de error disminuye). También hubo diferencia significativa entre las palabras terminadas en *-as* con respecto a las terminadas en *-os* (la chance de error aumenta). O sea que, en el grupo de error alto, los plurales masculinos y los de palabras terminadas en consonante generan

menos error que las terminadas en *-is*/*-us*, que en su mayoría eran posesivos (apuntando a la dificultad de estos). Pero también los femeninos plurales son más complicados que los masculinos (en los atributos, que tienden a estar en el primer término Esp1). Esto coincide con la literatura: el femenino es más difícil que adquirir que el masculino. En lo que respecta al segundo *cluster* (error medio/bajo), se observaron las siguientes diferencias significativas. En primer lugar, con Esp1, las palabras del primer término terminadas en *-os* respecto de la referencia en *-is*/*-us* (la chance de error disminuye) y aquellas terminadas en *-as* respecto de las terminadas en *-os* (la chance de error aumenta) (estos resultados coinciden con el primer grupo). En segunda instancia, con Esp2, las palabras del segundo término de la concordancia terminadas en *-as* respecto de aquellas terminadas en *-os* (referencia) (la chance de error disminuye) y aquellas terminadas en *-es* respecto de aquellas terminadas en *-as* (la chance de error aumenta). Aquí las femeninas parecen ser más fáciles que las masculinas en el segundo término de la concordancia, que tiende a coincidir con el controlador. Además, hubo un efecto de morfología *no transparente* para el plural: las palabras del segundo término terminadas en *-e* son más difíciles que las terminadas en *-a*. En cuanto a las variables seleccionadas para la predicción de estatus de error, el atributo ES = 1 se eligió para Sonia y Jako. En suma, si bien se observaron efectos inhibidores para los términos individuales de la concordancia terminados en *-e* y en consonante, la variable ES no resultó significativa.

En la literatura también se reporta para la concordancia de género un efecto facilitador de los controladores inanimados respecto de los animados. Fue una variable seleccionada por Lasso en el grupo de error alto, la cual arrojó aumento de chance de error. Además, en Sonia resultó ser influyente para predecir el estatus de error. En suma, con Anim se observó el efecto ya reportado.

Otro factor crucial para la concordancia parece ser la distancia (lineal o estructural) entre controlador y objetivo. No obstante ello, la variable LDa (larga distancia) no resultó asociada al error en ninguno de los modelos aplicados. Tampoco tuvo efecto alguno el hecho de que la concordancia estuviera conformada por más de dos términos (Gram).

Se crearon siete estrategias que se supuso que podrían estar poniendo en acto los alumnos cuando creaban las concordancias. Hubo un efecto con la estrategia Est5: si la palabra singular del italiano termina en *-e*, poner en español el plural en *-es*, como en *camion-e* > *camion-es* (casos en los cuales el español coincide con la aplicación del plural con *-e-* epentética). Fue seleccionada por Lasso para el grupo de error alto resultando en una disminución de la chance de error. La interpretación de este resultado es que los hablantes sacan provecho del parecido de las palabras singulares en español e italiano y forman el plural agregando una *-s* a la palabra singular italiana. Como en español estas coinciden con palabras terminadas en consonante que requieren plural en *-es*, sobrepasan de esta forma la dificultad de insertar plural con *-e-* epentética. Este factor quizás debe haber influido en el hecho de no haber hallado efecto de la variable ES.

Además de la animicidad del controlador, se tuvieron en cuenta otras características, a saber: la concretud, familiaridad, frecuencia léxica e imaginabilidad. Debido a problemas de colinealidad se crearon dos variables mediante PCA: (i) Ima.Conc: un índice de imaginabilidad más concretud; (ii) Fam.LEX: un índice de familiaridad más frecuencia del

controlador. El modelo Lasso seleccionó Fam.LEX para el grupo de error medio/bajo, que redujo significativamente la chance de error. Por lo tanto, la familiaridad y frecuencia léxica del controlador protegen contra la comisión de errores. Respecto de la predicción de error, Ima.Conc resultó seleccionado como atributo influyente en el caso de Sonia.

Se crearon dos variables cuantitativas, Stem y Morf, basadas en la distancia de Levenstein, para medir la distancia entre las raíces y las desinencias de cada palabra en español y en italiano. Luego se discretizaron en los atributos Morf.f (distancia entre terminaciones con tres niveles: alto, medio, bajo) y Stem.f (distancia entre raíces con dos niveles: alto y bajo). Stem.f y Morf.f fueron seleccionadas como atributo influyente para predecir el error en Jako. Sin embargo, no hubo efecto para la intensidad del error en los coeficientes promediados.

## 7. Conclusiones

A continuación, se comentan los hallazgos de los diferentes métodos aplicados. En general no se logró superar el 80 % de precisión balanceada. SVM resultó el clasificador de mejor desempeño. Sonia y Jako alcanzaron mejores niveles de precisión que Nati y Mirka. Utilizar el *ensemble* redundó en una mejora muy leve. Los atributos derivados del grafo fueron seleccionados para todos los alumnos. En lo que respecta a la selección de los atributos de las concordancias y a la frecuencia del enlace en el grafo y de las palabras, se consideraron de importancia los siguientes. Respecto de Sonia, la imaginabilidad y concretud altas del controlador, que haya un término en la concordancia con *-e-* epentética, la animicidad del controlador y que esté presente un artículo definido. En lo que atañe a Nati, está la frecuencia del artículo *los*. En lo concerniente a Jako, la terminación en *-as* del primer término de la concordancia, la similitud (de las terminaciones) media y alta con el italiano, los modificadores que son adjetivos, la similitud media con el italiano de la raíz léxica, que haya *-e-* epentética en un término de la concordancia, la frecuencia de los enlaces del grafo y la frecuencia de las palabras *lugares*, *competencias* y *particulares*. Con respecto a Mirka, la frecuencia de enlace. En el caso de Sonia y Nati también resultaron importantes para predecir los atributos dinámicos. Parece interesante señalar que ni Nati ni Mirka tuvieron atributos de concordancia seleccionados y con ellos los clasificadores tuvieron peor desempeño.

Se aplicó un *clustering* por mezcla de distribuciones para hallar dos grupos de probabilidad de error diferentes. El primero agrupaba las sesiones en torno a una media de error del 34 % (error alto) y el segundo, en entorno a un error del 17 % (error medio/bajo). Luego de aplicar un GLMM con penalización $L_1$, se eligieron las primeras doce variables y se procedió a la selección de un conjunto de confianza de modelos GLMM para cada *cluster*. Según los coeficientes promediados en dicho conjunto, a excepción de Skew y C, el resto de las variables seleccionadas fueron atributos de las concordancias. En lo que respecta al primer *cluster* (error alto) resultaron significativas las variables: Anim = 1, C, Est5 = 1, Mod = 3, Skew. También fue significativa la diferencia entre: Esp1 = *-os*, *-as*, *-es*, **-es* y la referencia (Esp1 = *-us*, *-is*). En lo que atañe al segundo *cluster*, las variables significativas fueron: Esp1 = *-os*, Fam.LEX = 1. El cuadro 14 resume lo señalado.

| | Sonia | Nati | Jako | Mirka |
|---|---|---|---|---|
| Clasificadores | Ima.Conc; ES = 1; Anim; Mod = 0 | - | Esp1 = *-as*; Morf.f = 1; Morf.f = 2; Mod = 3; ES = 1; Stem.f = 1 | - |
| Mixtura y GLMM | *CLUSTER* 1 | Anim; Est5; Esp1 = *-os -as*, *-es*, **-es*; Mod = 2, 3 | | |
| | *CLUSTER* 2 | Fam.LEX; Esp1 = *-os* | | |

Cuadro 14. Atributos de concordancias seleccionados según los alumnos y los métodos empleados

## Referencias


Alarcón, I. 2009. "The processing of gender agreement in L1 and L2 Spanish: Evidence from Reaction Time Data", en: *Hispania* 92(4). 814-828.

Alarcón, I. 2011. "Spanish grammatical gender under complete and incomplete acquisition: early and late Bilinguals' linguistic behavior within the noun phrase", en: *Bilingualism: Language and Cognition* 14(3). 332-350.

Bandt, C. & B. Pompe. 2002. "Permutation Entropy: A Natural Complexity Measure for Time Series", en: *Physical review letters* 88(17).

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. & Z. M. Jones. 2016. "mlr: Machine Learning in R", en: *Journal of Machine Learning Research* 17(170). 1-5.

Bojanowski, M. & B. Chrol 2019. "Proximity-based Methods for Link Prediction in Graphs with R package 'linkprediction'". Disponible en: http://recon.icm.edu.pl/wp-content/uploads/2019/05/linkprediction.pdf.

Bruhn de Garavito, J. 2008. "Acquisition of the Spanish plural by French L1 speakers: the role of transfer", en: Liceras, J., Zobl, H. & H. Goodluck (eds.) *The role of features in second language acquisition*, 270-298. Nueva Jersey: Lawrence Erlbaum Associates.

Bruhn de Garavito, J. & L. White. 2002. "The second language acquisition of Spanish DPs: the status of grammatical features", en: Pérez-Leroux, A. T. & J. Muñoz Liceras (eds.) *The Acquisition of Spanish Morphosyntax: The L1/L2 Connection*, 153-178. Dordrecht: Kluwer.

Burnham, K. P. & D. R. Anderson. 2010. *Model selection and multimodel inference: a practical information-theoretic approach*. Nueva York: Springer.

Campillos Llanos, L. 2014. "Errores léxicos en el español oral no nativo: análisis de la interlengua basado en corpus", en: *Revista ELUA* 28. 85-124.

Corbett, G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Davis, C. J. & M. Perea. 2005. "BuscaPalabras: a program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish", en: *Behavior Research Methods* 37(4). 665-671.

Español-Echevarría, M. & P. Prévost. 2004. “Acquiring number specification on L2 Spanish quantifiers: evidence against the rich agreement hypothesis”, en: Van Kampen, J. & S. Baauw (eds.) *Proceedings of the 2003 conference on generative approaches to language acquisition*, 151-162. Utrecht: LOT.

Fernández-García, M. 1999. “Patterns of gender agreement in the speech of second language learners”, en: Gutiérrez-Rexach, J. & F. Martínez-Gil (eds.) *Advances in hispanic linguistics: papers from the 2nd Hispanic Linguistics Symposium*, 3-15. Somerville, MA: Cascadilla Press.

Ferreira Cabrera, A. & J. Elejalde Gómez. 2017. “Análisis de errores recurrentes en el Corpus de Aprendices de Español como Lengua Extranjera, CAELE”, en: *Revista Brasileira de Linguística Aplicada* 17(3). 509-537.

Ferreira Cabrera, A. & J. Elejalde Gómez. 2020. “Propuesta de una taxonomía etiológica para etiquetar errores de interlengua en el contexto de un corpus escrito de aprendientes de ELE”, en: *Forma y Función* 33(1). 115-146.

Ferreira Cabrera, A., Elejalde Gómez, J. & A. Vine Jara. 2014. “Análisis de errores asistido por computador basado en un corpus de aprendientes de Español como Lengua Extranjera”, en: *Revista Signos: estudios de lingüística* 47(86). 385-411.

Foote, R. 2011. “Integrated knowledge of agreement in early and late English-Spanish bilinguals”, en: *Applied Linguistics* 32. 187-220.

Foote, R. 2015. “The production of gender agreement in native and L2 Spanish: the role of morphophonological form”, en: *Second Language Research* 31. 343-373.

Franceschina, F. 2001. “Morphological or syntactic deficit in near-native speakers? An assessment of some current proposals”, en: *Second Language Research* 17. 213-247.

Gillon Dowens, M., Vergara, M., Barber, H. & M. Carreiras. 2010. “Morphosyntactic processing in late second language learners”, en: *Journal of Cognitive Neuroscience* 22(8). 1870-1887.

González, P., Mayans, D. & H. Van der Bergh. 2019. “Nominal agreement in the interlanguage of Dutch L2 learners of Spanish”, en: *International Review of Applied Linguistics in Language Teaching*. 1-20.

Groll, A. & G. Tutz. 2014. “Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation”, en: *Statistics and Computing* 24(2). 137-154.

Grün, B. & F. Leisch. 2007. “Flexmix: An R package for finite mixture modelling”, en: *R News* 7(1). 8-13.

Keating, G. D. 2009. “Sensitivity to violations of gender agreement in native and nonnative Spanish: an eye-movement investigation”, en: *Language Learning* 59(3). 503-535.

Keating, G. D. 2010. “The effects of linear distance and working memory on the processing of gender agreement in Spanish”, en: VanPatten, B. & J. Jegerski (eds.) *Research in Second Language Processing and Parsing*, 113-134. Filadelfia: John Benjamins.

Kira, K. & L. A. Rendell. 1992. “The Feature Selection Problem: Traditional Methods and a New Algorithm”, en: *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI’92, 129-134. San José, California: AAAI Press. Disponible en: https://www.aaai.org/Library/AAAI/1992/aaai92-020.php

Kuhn, M. & K. Johnson. 2013. *Applied Predictive Modeling*. Nueva York: Springer.

Lichtman, K. 2009. “Acquisition of Attributive and Predicative Adjective Agreement in L2 Spanish”, en: Bowles, M., Ionin, T. , Montrul, S. & A. Tremblay (eds.) *Proceedings of the 10th Generative Approaches to Second Language Acquisition Conference*, 231-247. Somerville, MA: Cascadilla Proceedings Project.

Mac Whinney, B. 2021. *The Childes Project: Tools for Analyzing Talk. Part 1 & 2* [en línea]. Disponible en: https://talkbank.org/manuals/CHAT.pdf.

Marafioti, P. E. 2021. “Análisis de la evolución de errores de concordancia en cuatro aprendientes italianos de ELE usando redes complejas”, en: *Lingüística y Literatura* 42(79). 181-199.

McCarthy, C. 2008. “Morphological variability in the comprehension of agreement: an argument for representation over computation”, en: *Second Language Research* 24(4). 459-486.

Montrul, S., Foote, R. & S. Perpiñan. 2008. “Gender agreement in adult second language learners and Spanish heritage speakers: the effects of age and context of acquisition”, en: *Language Learning* 58(3). 503-553.

Muñoz Liceras, J., Díaz Rodríguez, L. & C. Mongeon. 2000. “N-drop and determiners in native and non-native Spanish: more on the role of morphology in the acquisition of syntactic knowledge”, en: Leow, R. P. & C. Sanz (eds.) *Current research on the acquisition of Spanish*, 67-96. Somerville, MA: Cascadilla Press.

Nason, G. P. 2008. *Wavelet methods in statistics with R*. Nueva York: Springer.

Nerbonne, J., Van Ommen, S., Gooskens, C. & M. Wieling. 2013. “Measuring socially motivated pronunciation differences”, en: Borin, L. & A. Saxena (eds.) *Approaches to Measuring Linguistic Differences*, 107-140. Berlín/Boston: De Gruyter Mouton.

Newman, M. E. J. 2010. *Networks: An Introduction*. Oxford: Oxford University Press.

Oakes, M. P. 1998. *Statistics for Corpus Linguistics*. Edimburgo: Edinburgh University Press.

O’Grady, W. 2005. *Syntactic Carpentry: An Emergentist Approach to Syntax*. Nueva Jersey: Lawrence Erlbaum Associates.

Peña, D. 2002. *Análisis de Datos Multivariantes.* Madrid: McGraw Hill.

Percival, D. B. & A. T. Walden. 2008. *Wavelet methods for time series analysis*. Cambridge: Cambridge University Press.

Sagarra, N. 2007. “Online processing of gender agreement in low proficient English-Spanish late bilinguals”, en: Camacho, J., Flores-Ferrán, N., Sánchez, L., Déprez, V. & M. J. Cabrera (eds.) *Current Issues in Linguistic Theory Series*, 240-253. Ámsterdam: John Benjamins.

Sagarra, N. & J. Herschensohn. 2013. “Processing of gender and number agreement in late Spanish bilinguals”, en: *International Journal of Bilingualism* 17(5). 607-627.

Scrucca, L., Fop, M., Murphy, T. B. & A. E. Raftery. 2016. “Mclust 5: clustering, classification and density estimation using gaussian finite mixture models”, en: *The R Journal* 8(1). 205-233.

Van Buuren, S. & K. Groothuis-Oudshoorn. 2011. “Mice: multivariate imputation by chained equations in R”, en: *Journal of Statistical Software* 45(3). 1-67.

Webber, C. L. (Jr.) & N. Marwan. (2015). *Recurrence Quantification Analysis*. Cham: Springer.

White, L., Valenzuela, E., Kozlowska-Macgregor, M. & Y. K. I. Leung. 2004. “Gender and number agreement in nonnative Spanish”, en: *Applied Psycholinguistics* 25(2). 105-133.

## Apéndice

| Atributo | Estimación | Error estándar | Error estándar ajustado | Valor Z | VPP(> \|z\|) |
|---|---|---|---|---|---|
| cond((Int)) | 1,0120 | 0,6891 | 0,6905 | 1,4657 | 0,1427 |
| cond(anim1) | 0,4388 | 0,2209 | 0,2213 | 1,9832 | 0,0473 |
| cond(C) | −0,4866 | 0,1443 | 0,1446 | 3,3658 | 0,0008 |
| cond(esp13) | −2,8435 | 0,7231 | 0,7245 | 3,9246 | 0,0001 |
| cond(esp14) | −1,8624 | 0,7240 | 0,7254 | 2,5673 | 0,0102 |
| cond(esp15) | −2,1784 | 0,9125 | 0,9143 | 2,3826 | 0,0172 |
| cond(esp16) | −2,6827 | 0,8484 | 0,8500 | 3,1561 | 0,0016 |
| cond(est51) | −0,9484 | 0,3816 | 0,3823 | 2,4807 | 0,0131 |
| cond(mod1) | 0,1902 | 0,6779 | 0,6792 | 0,2800 | 0,7795 |
| cond(mod2) | 0,4560 | 0,2290 | 0,2295 | 1,9871 | 0,0469 |
| cond(mod3) | 0,8497 | 0,2739 | 0,2744 | 3,0962 | 0,0020 |
| cond(morf1) | 0,0025 | 0,2973 | 0,2979 | 0,0085 | 0,9932 |
| cond(morf2) | 0,4115 | 0,4556 | 0,4561 | 0,9023 | 0,3669 |
| cond(Skew) | 0,3514 | 0,1671 | 0,1674 | 2,0993 | 0,0358 |
| cond(est61) | −0,1008 | 0,2990 | 0,2993 | 0,3368 | 0,7362 |
| cond(esp22) | 0,0352 | 0,1743 | 0,1746 | 0,2014 | 0,8404 |
| cond(esp23) | −0,0663 | 0,2140 | 0,2144 | 0,3093 | 0,7571 |
| cond(esp24) | 0,1974 | 0,3402 | 0,3404 | 0,5800 | 0,5619 |
| cond(est21) | −0,0912 | 0,2354 | 0,2357 | 0,3869 | 0,6988 |
| cond(lda1) | 0,0376 | 0,2093 | 0,2096 | 0,1794 | 0,8576 |
| cond(stem1) | −0,0141 | 0,1062 | 0,1063 | 0,1330 | 0,8942 |

Cuadro 15. *Cluster* 1: coeficientes promediados

| Atributo | Estimación | Error estándar | Error estándar ajustado | Valor Z | VPP (> \|z\|) |
|---|---|---|---|---|---|
| cond((Int)) | −0,3006 | 0,5530 | 0,5533 | 0,5433 | 0,5869 |
| cond(esp13) | −1,2090 | 0,4321 | 0,4324 | 2,7960 | 0,0052 |
| cond(esp14) | −0,4047 | 0,4232 | 0,4236 | 0,9555 | 0,3393 |
| cond(esp15) | −0,0989 | 0,5344 | 0,5350 | 0,1849 | 0,8533 |
| cond(esp16) | −1,2133 | 1,0524 | 1,0531 | 1,1521 | 0,2493 |
| cond(esp22) | −0,7216 | 0,4030 | 0,4032 | 1,7900 | 0,0735 |
| cond(esp23) | 0,0180 | 0,3308 | 0,3311 | 0,0543 | 0,9567 |
| cond(esp24) | −0,8257 | 0,9948 | 0,9953 | 0,8296 | 0,4068 |
| cond(est11) | −0,4401 | 0,3655 | 0,3657 | 1,2033 | 0,2289 |
| cond(fam.lex1) | −0,4528 | 0,1662 | 0,1663 | 2,7223 | 0,0065 |
| cond(mod1) | −0,1097 | 0,4853 | 0,4858 | 0,2258 | 0,8214 |
| cond(mod2) | 0,2717 | 0,2885 | 0,2886 | 0,9413 | 0,3466 |
| cond(mod3) | 0,1132 | 0,2122 | 0,2124 | 0,5331 | 0,5940 |
| cond(stem1) | 0,3272 | 0,2248 | 0,2249 | 1,4547 | 0,1458 |
| cond(est61) | 0,3250 | 0,4425 | 0,4427 | 0,7342 | 0,4628 |
| cond(es1) | 0,4645 | 0,9514 | 0,9519 | 0,4880 | 0,6255 |
| cond(es2) | 0,4471 | 1,6911 | 1,6925 | 0,2641 | 0,7917 |
| cond(lda1) | 0,0920 | 0,2547 | 0,2549 | 0,3611 | 0,7180 |
| cond(est71) | 0,0842 | 0,2439 | 0,2441 | 0,3449 | 0,7301 |
| cond(est41) | −0,0054 | 0,1524 | 0,1525 | 0,0357 | 0,9715 |
| cond(morf1) | −0,0231 | 0,1275 | 0,1276 | 0,1814 | 0,8560 |
| cond(morf2) | −0,0093 | 0,1093 | 0,1093 | 0,0852 | 0,9321 |

Cuadro 16. *Cluster* 2: coeficientes promediados