Creación de un modelo estadístico predictivo para la determinación de las funciones de atenuación en español hablado*

Creating a predictive statistical model to determine mitigation functions in spoken Spanish

Adrián Cabedo Nebot
Departamento de Filología Española
Facultad de Filologia, Traducció i Comunicació
Universidad de Valencia
Av. de Blasco Ibáñez, 32, 3.ª planta, despacho 13. Valencia, 46010
adrian.cabedo@uv.es
Orcid ID 0000-0002-3881-9308

RECIBIDO: 9 DE DICIEMBRE DE 2016
ACEPTADO: 2 DE FEBRERO DE 2017

Resumen: Recientemente, algunos autores han definido distintas variables para caracterizar la atenuación lingüística en el marco de una base de datos multidimensional (Briz/Albelda; Albelda/otros). En este estudio se han seleccionado 982 elementos de atenuación de dieciocho entrevistas de español hablado; todos ellos han sido supervisados por hasta cuatro entrevistadores del proyecto Es.VaG.Atenuación. Finalmente, los datos se han evaluado mediante tres pruebas estadísticas para la clasificación y la reducción de variables: el análisis múltiple de correspondencias, el árbol de clasificaciones y el *Random Forest*. Las variables más determinantes en la discriminación de las funciones de atenuación han sido la imagen de los hablantes, el acto de habla previo, la franja de edad, la posición sintáctica y la variable prosódico-paralingüística.

Palabras clave: Funciones. Atenuación. Estadística. Árbol de decisiones. *Random Forest*.

Abstract: Several authors have established a large group of variables to characterise linguistic mitigation using a multidimensional database as a framework (Briz/Albelda; Albelda/others). In this study, 982 mitigation sequences have been selected from 18 spoken Spanish interviews; all of them have been identified as instances of mitigation by at least four different informants. Finally, these data have been analysed by means of three statistical methods: multiple correspondence analysis, decision trees and *Random Forest*. The results suggest that the variables that better characterise the three functions of mitigation (self-protection, prevention and repair) are (a) image of the speaker(s), (b) the previous speech act, (c) age, (d) syntactic position and (e) the prosodic-paralinguistic component.

Keywords: Functions. Mitigation. Statistics. Decision trees. *Random Forest*.

* Esta investigación ha sido posible gracias a la subvención de dos proyectos de investigación: "La atenuación pragmática en el español hablado: su variación diafásica y diatópica" (Ministerio de Economía y Competitividad de España, ref. FFI2013-40905-P) y Es.VaG.Atenuación ("La atenuación pragmática en su variación genérica: géneros discursivos escritos y orales en el español de España y América"; Ministerio de Economía y Competitividad de España, ref. FFI2016-75249-P).

1. Introduccción[1]

La atenuación, como señala Albelda, entendida como fenómeno pragmático, ha sido foco de interés durante los últimos años y ello ha conllevado su estudio en diferentes lenguas y géneros (Holmes 1984a, 1984b; Bazzanella/Caffi/Sbisà; Briz 1995, 2003, 2007, 2010; Puga; Caffi 1999, 2007; Sbisà; Erlich, entre muchos otros). El problema básico de la atenuación, no obstante, radica en la definición del fenómeno, ya que depende de un contexto interactivo real y, según Briz (2001), todo fenómeno pragmático solo puede determinarse en esa exacta situación.

Planteamientos teóricos más recientes, como los de Briz/Albelda o los de Albelda/otros, han pretendido formalizar esa inherente dificultad identificativa, mediante el establecimiento de patrones o regularidades constantes. Para ello, dado lo complicado de definir un concepto como el de la atenuación, se sigue la definición operativa que se recoge en estas obras:

> La atenuación es una actividad argumentativa (retórica) estratégica de minimización de la fuerza ilocutiva y del papel de los participantes en la enunciación para lograr llegar con éxito a la meta prevista, y que es utilizada en contextos situacionales de menos inmediatez o que requieren o se desea presenten menos inmediatez comunicativa. Así, unas veces, habrá atenuación de hablante y, por tanto, la estrategia tendrá un carácter más monológico; otras, atenuación de hablante y oyente y, por tanto, tendrá un carácter más dialógico y a menudo cortés. (Briz/Albelda 292)

Así mismo, en la ficha de análisis planteada por Briz/Albelda o Albelda/otros, se han definido otros elementos de interés lingüístico, como la inclusión de funciones comunicativas para mejorar la explicación del fenómeno pragmático de la atenuación. En tal sentido, se le da importancia como fenómeno estratégico que responde a la voluntad del hablante; se mantiene así una conexión con la perspectiva establecida por Caffi (1999, 885; 2007, 80), quien sostiene que en la atenuación se dan los *bushes*, que atenúan el contenido proposicional (niegan lo que se dice); los *hedges*, la locución (niegan el acto de decir), pero también los *shields*, que crean distancia entre el emisor y lo dicho (resultado, por tanto, de la negación del yo).

---

1. Doy las gracias a los dos evaluadores anónimos que han leído este artículo. Sus comentarios han resultado ser de gran interés y han permitido esclarecer algunas zonas grises de la investigación.

En esta línea, Briz/Albelda hablan de tres funciones preponderantes de la atenuación, según la repercusión que el fenómeno puede tener tanto en el emisor como en el receptor y, por tanto, en la protección de la imagen personal involucrada en el espacio de interacción concreto.

La función 1 establece una protección de la imagen propia, es decir, mediante ella se realizan labores de autoimagen, como puede observarse en la secuencia *yo por lo menos* del ejemplo (1):

(1) B: ...yo creo que Képler ahí tenía razón// bueno *yo por lo menos* me lo creo (*Corpus Preseea Valencia*, archivo VALE_M23_001)

Por su parte, la función 2 cumple el requisito de proteger la imagen ajena, habitualmente de un participante o interlocutor presente en el momento de la conversación, pero también de personas no presentes. Es el caso del ejemplo (2), en el que el hablante B protege la imagen de un interlocutor ficticio no presente, como es la referencia impersonal a *los jóvenes actuales*. Esta protección se observa mediante un mecanismo paralingüístico (chasquido) y un marcador discursivo *hombre* que antecede la afirmación conflictiva *tienen bastantes problemas*.

(2) A: háblanos de los problemas que en tu opinión tienen los jóvenes actuales
B: pues tienen (chasquido)/ ¡hombre! tienen bastantes problemass/ yo creo quee/ mm tienen los problemas típicos de cualquier adolescente/ y ess los problemas de identidaad/ los problemas dee– de cómo se organizan ellos en el muundo/ yy los ti– bueno las– los problemas que cualquier persona que nace crece se desarrolla y muere/ tiene ¿no? (*Corpus Preseea Valencia*, archivo VALE_M23_001)

Finalmente, la función 3 pretende reparar un daño ya realizado a la imagen de alguien presente o ausente en la conversación. En el ejemplo (3) se observa que B ha hecho una afirmación conflictiva sobre el horóscopo de A; al darse cuenta, B introduce la estrategia paralingüística de la risa seguida de un marcador discursivo de corrección, como es *hombre*.

(3) A: y los Escorpiones ¿qué tal lo tenemos?
B: lo del estómago no tan bien los Escorpioonees (risas)// ¡hombre! los Escorpiones sois signos– sois de agua// (chasquido) sois signos de agua/ lo que quiere decir que lo que predomina son los sentimientos y la afectividad (*Corpus Preseea Valencia*, archivo VALE_M23_001)

Además de la definición operativa y del establecimiento de tres funciones básicas, Briz/Albelda y Albelda/otros han definido distintas variables para caracterizar la atenuación lingüística en el marco de una base de datos multidimensional. Esto permite la elaboración de un estudio estadístico de carácter predictivo que establezca cuáles son las variables más importantes en la determinación de las diferentes funciones atenuantes. De este modo, dada la amplia cantidad de variables involucradas, se plantea la necesidad de simplificarlas mediante la aplicación de métodos estadísticos de reducción dimensional, como la prueba del *Random Forest* o los árboles de decisiones.

En este estudio, por tanto, tras una breve introducción teórica al fenómeno y a las funciones de la atenuación, se dedicará la sección 2 a introducir elementos de carácter metodológico, como las características del corpus analizado, las variables de análisis utilizadas y los métodos estadísticos empleados; finalmente, la sección 3 incluirá los resultados más destacables junto con el análisis y valoración de algunos datos de interés.

2. Metodología

La intención prioritaria de este estudio ha sido reducir las dimensiones de análisis de las funciones atenuadoras. Así, ante una ficha de análisis que recoge más de 30 variables (Preseea; Briz/Albelda; Albelda/otros), la pregunta que se realiza el investigador es hasta qué punto son útiles todas ellas para la delimitación del fenómeno.

En tal sentido, cabe tener en cuenta que la atenuación, como fenómeno pragmático y por ende contextual, es un valor interpretativo. Trabajos recientes de corpus han postulado la dificultad que supone delimitar claramente las fronteras de la atenuación pragmática (Villalba 2016a, 2016b). De este modo, mientras un investigador puede observar *atenuación* en determinadas secuencias de la interacción conversacional, otro investigador distinto puede manifestar acuerdo o desacuerdo con esa evaluación.

En la investigación lingüística general resulta habitual, por tanto, que una evaluación subjetiva, basada en lecturas bibliográficas previas, sea la responsable de etiquetar un conjunto de enunciados, intervenciones o fragmentos conversacionales como portadores de un determinado valor semántico-pragmático (López Serena). De cara a minimizar un sesgo interpretativo, se precisa de técnicas que disminuyan esa posibilidad. Por ejemplo, la aplicación de métodos o pruebas estadísticas es una de ellas, como se verá en la sección

de resultados; no obstante, con antelación a esa sistematización de los datos, el investigador puede utilizar otras vías. Entre ellas, la más habitual supone consensuar los datos analizados con otros compañeros del área.

Así, ampliando la metodología que ya se aplicó con el proyecto Fonocortesía (Hidalgo 2012), tanto en los proyectos Esvaratenuación como Esvagatenuación, los datos recogidos se consensuan regularmente por hasta tres investigadores con anterioridad a cualquier análisis cuantitativo o cualitativo. Para el presente estudio, cuatro investigadores han participado en la recogida y etiquetado de los datos. De esta manera, cada elemento incluido en la base de datos ha sido sometido a debate.[2]

2.1 *Corpus seleccionado*

Los datos seleccionados para este estudio proceden del corpus de entrevistas sociolingüísticas PRESEEA (Proyecto para el estudio sociolingüístico del español de España y de América); más concretamente, se han utilizado 18 entrevistas del subcorpus PRESEVAL, que recoge datos para el español hablado en la zona de Valencia.

Con la intención de ser representativos y de cubrir todas las intersecciones de análisis, se han tomado entrevistas de tres niveles sociolingüísticos distintos (alto, medio y bajo), de tres franjas de edad distintas (18-25, 25-55, >55) y de los dos sexos (hombre y mujer). En total, esto ha supuesto más de 9 horas de grabación y de 40 000 palabras.

Sobre la frecuencia de las funciones de la atenuación, de los 982 registros analizados, han sido mayores los casos de función 2, con un 55 % (546 casos), que de función 1, con un 42 % (410 casos). Curiosamente, la denominada función 3 tiene una presencia residual, ya que solo constituye un 3 % (26 casos) de los datos.

En general, debe establecerse una reserva metodológica sobre el corpus utilizado en esta investigación, ya que procede de un proyecto de entrevistas sociolingüísticas y, en cuanto género discursivo, su arquitectura textual y el tipo de interacción que se lleva a cabo puede variar en comparación con otros géneros orales, como los debates parlamentarios (Estellés) o las conversaciones (Cabedo 2016).

---

2. Agradezco en estas líneas la participación de Marta Albelda, Amparo Soler y Cristina Villalba, que me facilitaron el acceso a bases de datos ya etiquetadas previamente.

2.2 *Variables de análisis*

Las variables cualitativas han sido establecidas en distintos trabajos hasta la fecha (PRESEEA; Briz/Albelda; Albelda/otros). Aunque en las versiones preliminares de la fecha se establece un número de 30 variables, la versión utilizada para esta investigación ha desarrollado alguna de ellas, con lo que, en total, suman 45 variables. Algunas variables, como el nombre del archivo o el contexto, no han sido incluidas en el análisis estadístico, ya que son solo variables cualitativas de etiquetado. El conjunto de variables puede verse en la tabla 1:

ARCHIVO	ELEMENTO O CONTEXTO	TIPO DE ATENUACIÓN
Modificación morfológica interna	Modificación morfológica externa	Selección léxica
Mod temp verb	Aserción duda I	Aserción duda II
Acotar opinión	Negar lo que pides	Incluir posible rechazo
Pedir disculpas	Modificación acto de habla	Justificaciones
Estilo directo	Elipsis conclusión	Fórmulas fáticas
Impersonalización I	Impersonalización II	Impersonalización III
Md correctores	Md consecutivo-lógico	Movimiento concesivo
Comparación	Rebajar acto habla	Extranjerismos u onomatopeyas
Prosodia / Paralenguaje	Otras variantes fónicas	Cantidad de estrategias
Posición sintáctica	Contenido dicho	Fuerza ilocutiva
Acto habla discurso previo	Temática	Registro
Estilo discursivo	Propósito	Carácter discurso
Función atenuación	Turno de habla	Control temático
Edad	Sexo	Nivel de instrucción

Tabla 1. Variables cualitativas de la ficha de análisis

Por su parte, en la base de datos se incluye una variable de análisis cuantitativa, que incorpora el número de estrategias de atenuación que pueden encontrarse en un mismo fragmento conversacional. En el siguiente ejemplo aparecen varios fenómenos de atenuación marcados en cursiva:

(4) y luego la situación allí *la verdad es que* es *bastantee// bueno* yo cre– para mí dramática *supongo que a lo mejor* ellos son felicísimos

Por defecto, los valores de las distintas variables recogen el de la estrategia de atenuación que los investigadores consideran como más destacable. En el caso de (4) se ha seleccionado la secuencia *supongo que* como mayor representante de la atenuación de la secuencia.

Así pues, teniendo en cuenta que los casos de atenuación recogidos han sido 982, esto ofrece un total de 44 190 datos para el análisis estadístico.

2.3 *Pruebas estadísticas aplicadas*

En primer lugar, un estudio de naturaleza multivariable precisa de reducir el tamaño de las variables definidas inicialmente para mejorar la operatividad de futuras investigaciones (Field; Gries); es decir, si bien hay 45 variables iniciales, se debe abogar por reducir el número de variables explicativas al menor número posible. En este estudio, este hecho se antoja complicado, ya que no hay que determinar características recurrentes para un determinado fenómeno, sino para tres elementos, tres tipos de funciones, que, según la bibliografía (Briz/Albelda), presentan una constitución diferente.

Al mismo tiempo, tal y como se ha comentado en la sección de variables analizadas, todas menos una son variables de tipo cualitativo y, habitualmente, esto limita el tipo de pruebas estadísticas reduccionista que pueden llevarse a cabo. Así, antes de establecer patrones de clasificación y de reducir la cantidad de variables explicativas, se requiere observar cómo los datos se agrupan unos con otros; esto es, se precisa estudiar las asociaciones de las diferentes variantes que participan en la base de datos. Con este fin, se ha utilizado inicialmente una prueba estadística conocida como *análisis múltiple de correspondencias* (Greenacre; Glynn), que, concretamente, delimita un mínimo de dos dimensiones explicativas y sobrepone las variantes según la mayor o menor distancia entre sí. En palabras de Glynn:

> Linguists often wish to find relations between given linguistic forms, between their meanings and in what situations those forms and meanings are used. Correspondence analysis is especially designed for identifying such usage patterning. The visualization of the relations takes the form of configuration biplots, or maps, which depict degrees of correlation and variation through the relative proximity of data points (which represent linguistic usage features and / or the actual examples of use). (133)

Posteriormente a la visualización de esas relaciones, se han utilizado dos pruebas de clasificación estadística: el *Random Forest* (Liaw/Wienner) y el árbol de clasificaciones (Hothorn/Hornik/Zeileis). Estas no son desconocidas en el ámbito lingüístico; por poner solo una muestra, pueden mencionarse trabajos recientes como los de Tanghe, para el análisis de marcadores del discurso, o Roberts, Torreira y Levinson, que analizan los tiempos de transición entre turnos de habla.

En general, *Random Forest* es una prueba similar a la del árbol de clasificaciones, ya que, en realidad, se basa en la generación aleatoria de centenares de árboles simulados; esto es, a partir de los datos observados, se genera un conjunto amplio de árboles de decisiones (normalmente, entre 500 o 1000) y se observa qué variables operan con mayor regularidad ante ese conjunto simulado. Finalmente, con el resultado final, se observa cuál es el porcentaje de error con la clasificación de los datos reales.

Por su parte, el árbol de clasificaciones permite reducir al máximo el número de variables y variantes que participan en la identificación de una variable básica; en el caso de este estudio, esa variable es la función de la atenuación, compuesta por tres variantes distintas.

Todas las pruebas estadísticas de esta investigación se han realizado mediante el paquete de análisis estadístico R (R Core Team), a través del uso específico de los siguientes paquetes: *FactoMiner* (Lê/Josse/Husson), para el análisis múltiple de correspondencias; *Random Forest* (Liaw/Wiener), para la prueba que lleva el mismo nombre; y, finalmente, el paquete *Party* (Hothorn/Hornik/Zeileis), para la realización del árbol de decisiones.

3. Resultados y análisis

3.1 *Análisis múltiple de correspondencias*

Con los 44 190 datos recogidos para los 982 registros de atenuación, se ha aplicado, en primer lugar, una técnica estadística para la identificación de posibles grupos de elementos. Dada la constitución de estos en torno a variables nominales o cualitativas, la prueba del *análisis múltiple de correspondencias*, según la propuesta conjunta establecida por Lê, Josse y Husson para el paquete FactoMiner, permite visualizar en un mapa la distribución de los elementos de una base de datos, según el grado de cercanía o lejanía en cuanto al comportamiento manifestado para cada variable. Téngase en cuenta, en tal sentido, el gráfico 1:

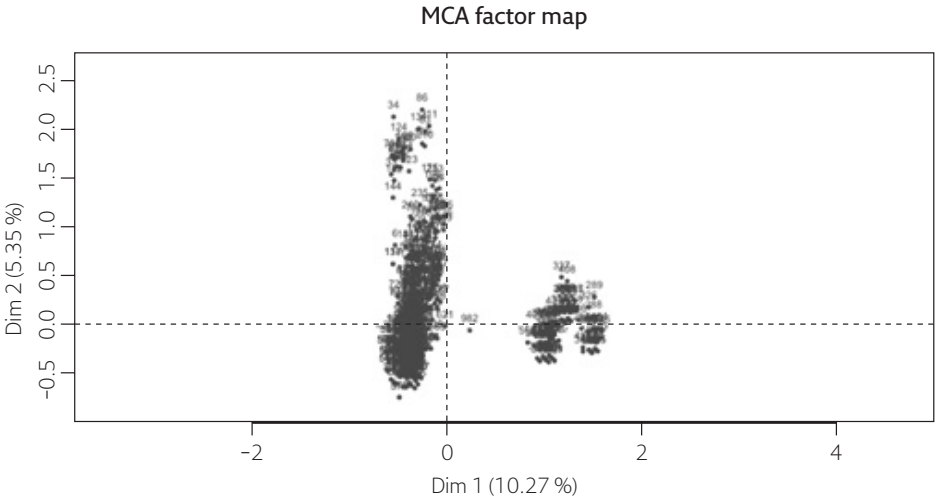

Gráfico 1. Distribución de los registros según el análisis múltiple de correspondencias

En el gráfico 1, los datos se distribuyen en tres grupos. Estos son los datos en crudo, sin el cruce todavía con las variantes correspondientes, así que todavía no pueden determinarse las características de los elementos, sino simplemente que se distribuyen de este modo en un mapa bidimensional. Hay un grupo de elementos que puntúa alto en la primera dimensión, entre 1 y 2, pero puntúa bajo en la segunda dimensión, entre 0 y 0.5. Por su parte, el grupo más numeroso se encuentra entre –0.5 para la segunda dimensión y 1.5. Finalmente, un conjunto menor de elementos se distribuye, también en la segunda dimensión, entre 1.5 y 2.5.

Aunque el análisis múltiple de correspondencias no es una técnica de estadística inferencial, sino simplemente de visualización y análisis de las relaciones, sí se ofrecen las puntuaciones de cada variable y de cada variante para el reparto final de los datos. Estas puntuaciones pueden observarse distribuidas, esta vez proyectadas sobre las variables y variantes correspondientes, en los gráficos 2 y 3.

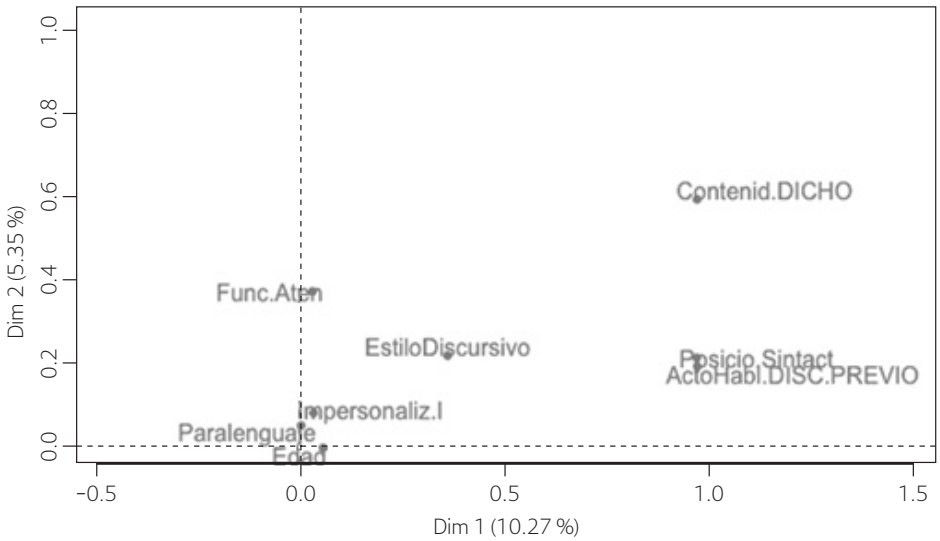

Gráfico 2. Distribución y asociación de las variables según el análisis múltiple de correspondencias

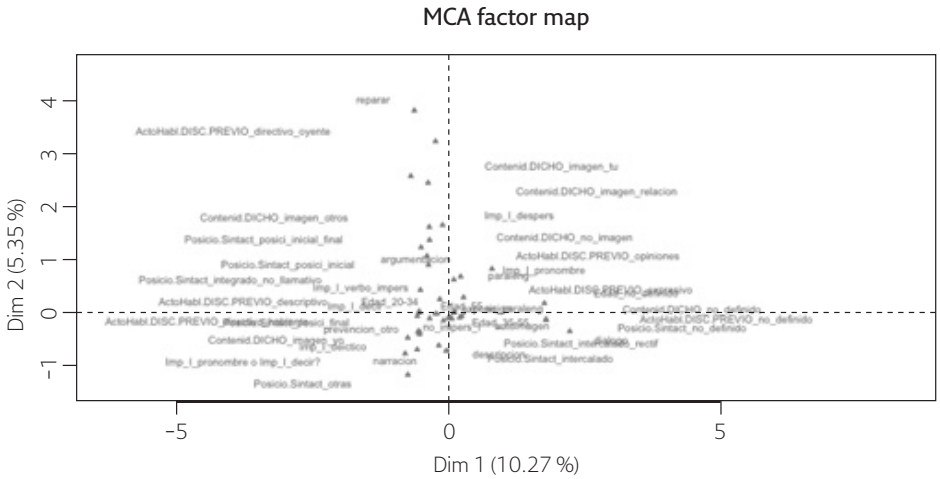

Gráfico 3. Asociación de las variantes según el análisis múltiple de correspondencias

Del gráfico 2 se deduce la importancia de una variable en ambas dimensiones: el *contenido de lo* DICHO (imagen involucrada en el acto de atenuación), mientras que otras dos variables, la *posición sintáctica* y el *acto de habla previo*, son determinantes, sobre todo, en la primera dimensión.

En la segunda dimensión, por su parte, el *estilo discursivo* y el tipo de *impersonalización* acompañan, en cuanto a relevancia, a la función atenuadora utilizada. Finalmente, la variable denominada *paralenguaje*, que integra también los valores prosódicos en la ficha de atenuación, puntúa bajo en ambas dimensiones en conjunción con la *franja de edad*.

En general, por tanto, y a partir de los primeros datos visualizados, puede decirse que estas seis variables son las más determinantes de las 45 del corpus analizado. Ahora bien, queda por ver, tomando en consideración de nuevo el gráfico 3, cuál es el papel de cada variante en el resultado final.

Aunque el gráfico 3 puede no visualizarse correctamente por el número amplio de variantes mapificadas, sí pueden detectarse algunos patrones de interés. El grupo menor de elementos que puntuaba alto en la segunda dimensión en el gráfico 1 queda ahora representado por la función atenuadora de la reparación, por el que el acto de habla previo que origina la afrenta es directivo y dirigido al oyente y por el que en el contenido de lo dicho se integra la imagen del tú. No obstante, a nivel de aparición, como se comentó anteriormente, la función 3 ha resultado ser residual. Como ejemplificación, puede tomarse de nuevo el caso de 3, renombrado ahora como 3':

(3') A: y los Escorpiones ¿qué tal lo tenemos?
B: lo del estómago no tan bien los Escorpioonees (risas)// ¡hombre! los Escorpiones sois signos– sois de agua// (chasquido) sois signos de agua/ lo que quiere decir que lo que predomina son los sentimientos y la afectividad (*Corpus* PRESEEA *Valencia*, archivo VALE_M23_001)

La función 1, de autoimagen, queda en el gráfico 3 emparentada con el estilo discursivo del diálogo, con edades superiores a 35 años y con una indefinición de varias variables, como el *contenido de lo dicho*, el *acto de habla del discurso previo* o la *posición sintáctica*. En último lugar, la función 2, de prevención de la afrenta a la imagen de otros, está relacionada en el gráfico 3 con la narración, con edades entre los 20 y los 34 años, con diferentes tipos de impersonalización I (uso de *decir*, de elementos deícticos...), con la imagen del yo vinculada en el contenido de lo dicho y con distintas posiciones sintácticas. En (5), (6) y (7) pueden observarse diferentes elementos lingüísticos de atenuación: (5) representa un caso de función 1 y ha sido enunciado por un hablante de entre 20 a 34 años; (6) usa el verbo *decir* como elemento de impersonalización, mientras que (7) usa la segunda persona como marca de impersonalidad; tanto (6) como (7), contextualmente, son muestras de función 2:

(5) A: ¿alguna fiesta a la que asististe o algún cuento que recuerdes?// para contárnoslo
B: pues lo que recuerdo/// quizás porque es lo que más gracia me hace yy/ lo que me hizo aborrecer la bebida fue en un cumpleaños cuando tenía/ diez años (*Corpus PRESEEA Valencia*, archivo VAL00231HC096)

(6) entonces/ dos vías son decir bueno pues por un lado/ ee vamos a hacer una prestación// sustitutoria que empieza mal/ porque ya habla de sustitución ¿no?// pero luego está el tema de la– de la insumisión ¿no? (*Corpus PRESEEA Valencia*, archivo VALE00332HC097)

(7) B: ¿qué argumentos usarías tú para convencer pues a los jóvenes de que fuesen un poco más disciplinados?/ si tú tuvieses hijos por ejemplo/ imagina que tienes hijos de catorce a dieciocho años
A: intentaría ser para empezar sus amigo(s)/// si eres amigo le llegarás// si eres padre autoritario no le llegarás (*Corpus PRESEEA Valencia*, archivo VALE_M23_001)

En síntesis, a partir del gráfico 3 puede esbozarse una idea bastante clara: la función 1 y la función 2 se comportan de maneras similares, si bien la función 1 manifiesta menos características demarcativas. En otras palabras, se observa cierta recursividad y jerarquía entre las dos primeras funciones de la atenuación. Así, como ya han comentado algunos autores, la función 2 es más rica en matices porque también muestra comportamientos de la función 1. Toda labor de prevención del daño a la imagen del otro no deja de ser una labor de autoimagen (Estellés/Albelda).

3.2 Random Forest *y árbol de decisiones*

Tanto las técnicas de *Random Forest* como el árbol de decisiones son técnicas de clasificación que, a diferencia del análisis múltiple de correspondencias, sí pueden mostrar valores de estadística inferencial. Además, permiten la inclusión de variables cuantitativas y cualitativas.

Con *Random Forest*, como se ha comentado en la sección 2.3, se determina cuáles son las variables más importantes para la clasificación de un conjunto simulado de entre 500 y 1 000 árboles de clasificación distinto. Para cada árbol cambia algún dato en la simulación efectuada. Las variables que manifiestan una mayor importancia, de nuevo en la línea de lo expresado en la sección 3.1, pueden contemplarse en el gráfico 4.

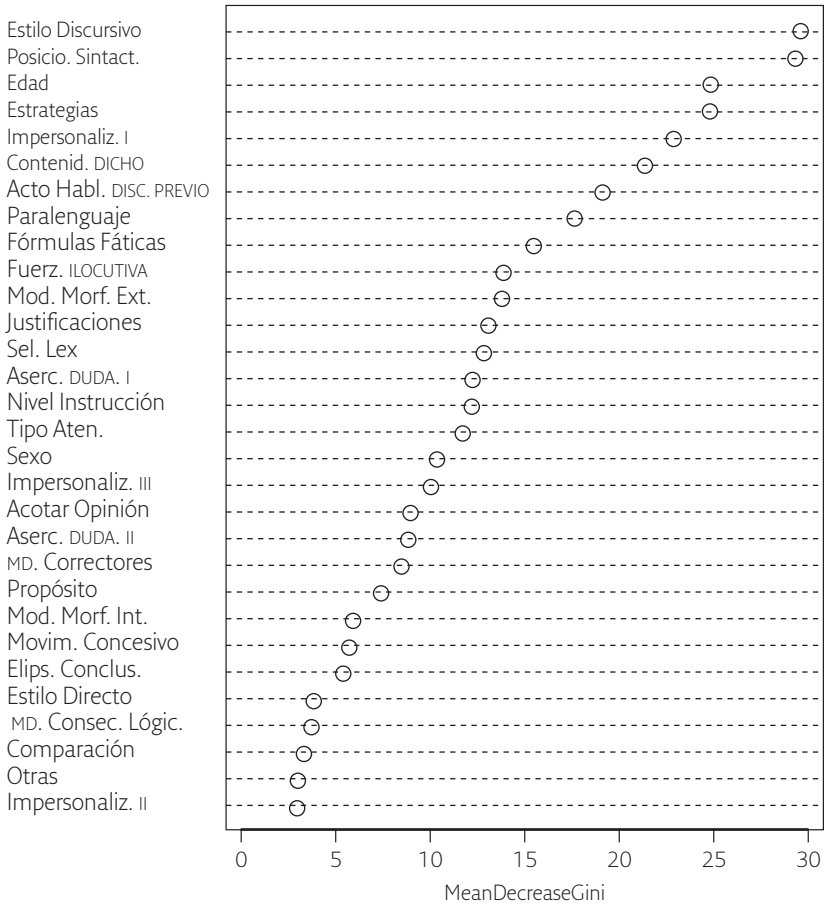

Gráfico 4. Importancia de las variables en la técnica *Random Forest*

Así pues, a partir del gráfico 4, se visualiza que las cinco variables más relevantes para la clasificación de los datos simulados han sido el *estilo discursivo*, la *posición sintáctica*, la *edad*, el *número de estrategias empleadas* y el tipo de *impersonalización I* utilizada. Con estos resultados, el programa elabora clasificaciones automáticas de los datos simulados y se arrojan los datos de la tabla 2:

	Autoimagen	Prevención	Reparación
Autoimagen	234	174	2
Prevención	119	427	0
Reparación	18	8	0

Tabla 2. Clasificación de los datos según *Random Forest*

En la tabla 2 se observa que la clasificación efectuada por la técnica *Random Forest* comete una tasa de error del 32.69 % de los datos. En general, los casos de reparación, función 3, se clasifican automáticamente como pertenecientes a la función 1 o a la función 2. Este dato podría cuestionar la operatividad de mantener tres funciones de la atenuación, ya que una de ellas, al menos según *Random Forest*, no muestra suficiente solidez constitutiva como para independizarse del resto.

También es curioso, si se observan los datos, que los valores de autoimagen o de prevención se clasifican bastante bien, aunque con un 20-30 % de intercambiabilidad, es decir, de elementos de función 1 clasificados como función 2 y viceversa.

Finalmente, el árbol de decisiones, que se visualiza en el gráfico 5, sin simulación de datos, determina tres variables como las más cruciales en el camino de clasificación de las funciones: el *contenido de lo dicho (imagen involucrada)*, la *posición sintáctica* y la variable *prosódica-paralingüística*.

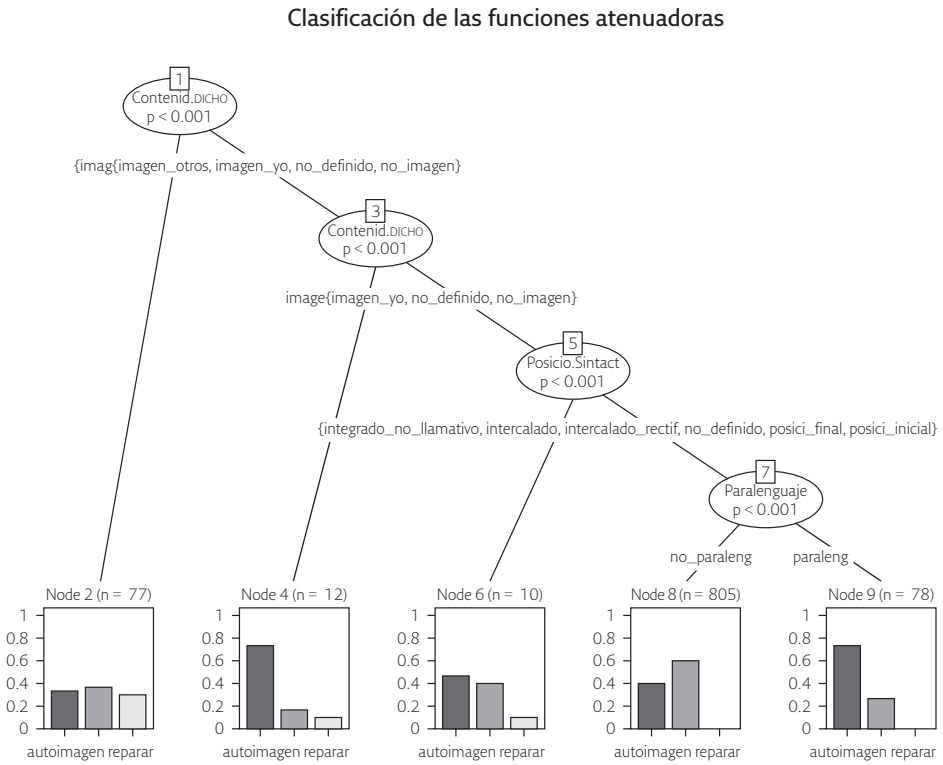

Gráfico 5. Árbol de decisiones para delimitar las funciones de la atenuación

Del gráfico 5 pueden elaborarse algunos comentarios. Gran parte de los datos (805) se clasifican en el nodo 8 como pertenecientes en un 60 % a la función 2 si la imagen afectada en el contenido no es la del tú u otros, sino la del yo o si no hay imagen directamente implicada; también se recoge en el nodo 8 si pertenece a prácticamente cualquier posición sintáctica y no se observan datos paralingüísticos influyentes. Este último dato puede cuestionarse por la tipología misma de la ficha, que no valoriza debidamente la variable prosódica y por el tipo de género discursivo analizado (Cabedo 2016).

Por otro lado, cuando la imagen involucrada es la del tú o la de otros, la clasificación puede corresponder a cualquiera de las funciones de la atenuación, como se observa en el nodo 2. Finalmente, en el nodo 9, que recoge 78 datos (13 % del total), hay una clasificación casi recurrente como función 1 si se activa algún elemento de la variable prosódica-paralingüística.

4. Conclusiones

En este artículo se ha observado la posibilidad de reducir un conjunto amplio de variables a aquellas más relevantes en la discriminación de un determinado fenómeno de análisis. En el caso concreto de la investigación realizada, se ha manifestado que las más de 30-40 variables de análisis establecidas por la ficha de la atenuación (Briz/Albelda; Albelda/otros) pueden reducirse a 5 o 6 si el foco de interés son las funciones comunicativas de la atenuación. En concreto, las variables más determinantes han resultado ser las que introducen el concepto de imagen (del yo, del tú, de otros no presentes...), la posición sintáctica del elemento lingüístico en cuestión, el acto de habla previo, la franja de edad y la variable prosódica-paralingüística.

En tal sentido, cabe indagar aún más en variables como la prosódica (Hidalgo 1997, 2006; Hidalgo/Cabedo; Cabedo 2013) o, también, en el posible influjo generacional que se observa en los datos. Así, mientras la generación de 20 a 34 años parece manifestar mayor predisposición a labores de autoimagen, las generaciones de más edad muestran continuamente casos de función 2 o preventiva. Este es un dato que debería abordarse con mayor detenimiento en el futuro.

Otro resultado interesante ha sido descubrir que, mediante la aplicación de técnicas estadísticas de clasificación, como la prueba *Random Forest*, la llamada función 3, que repara la posible imagen afectada de un interlocutor, se clasifica automáticamente como función 1 o función 2. Este dato puede hacer replan-

tearse la inclusión de la reparación como un elemento de atenuación, ya que, como algunos estudios recientes parecen mostrar (Cabedo 2016), la función 3 no atenúa el mensaje en el que se inserta, sino que precisamente lo intensifica.

Ahora bien, todos los resultados obtenidos deben tomarse con cautela, ya que proceden de un único género discursivo: la entrevista sociolingüística. Por lo tanto, deben analizarse otros formatos orales, como otros tipos de entrevistas de carácter más mediático, conversaciones cotidianas, etc.

En síntesis, las fichas de análisis de fenómenos lingüísticos pretenden recoger el mayor número posible de datos, si bien las pruebas estadísticas de reducción dimensional indican que, con valor predictivo, solo un número reducido, habitualmente cuatro o cinco, son las realmente importantes para determinar e identificar esos valores.

Obras citadas

Albelda, Marta. "¿Cómo se reconoce la atenuación?: una aproximación metodológica basada en el español peninsular hablado". *(Des)cortesía en español: espacios teóricos y metodológicos para su estudio*. Eds. Franca Orletti y Laura Mariottini. Roma: Università di Roma/Estocolmo: Tre/EDICE, 2010. 47-70.

Albelda, Marta, Antonio Briz, Ana María Cestero, Dorota Kotwica y Cristina Villalba. "Ficha metodológica para el análisis pragmático de la atenuación en corpus discursivos del español (ES.POR.ATENUACIÓN)". *Oralia* 17 (2014): 7-62.

Bazzanella, Carla, Claudia Caffi y Marina Sbisà. "Scalar Dimension of Illocutionary Force". *Speech Acts: Fiction or Reality?* Ed. Igor Z. Zagar. Ljubljana: IPRA Distribution Center for Jugoslavia, 1991. 63-76.

Briz, Antonio. *La conversación coloquial: materiales para su estudio*. València: Universitat de València, 1995.

Briz, Antonio. *El español coloquial en la conversación: esbozo de pragmagramática*. Barcelona: Ariel, 2001.

Briz, Antonio. "La estrategia atenuadora en la conversación cotidiana española". *La perspectiva no etnocentrista de la cortesía: identidad sociocultural de las comunidades hispanohablantes*. Ed. Diana Bravo. Estocolmo: Universidad de Estocolmo, 2003. 17-46.

Briz, Antonio. "Para un análisis semántico, pragmático y sociopragmático de la cortesía atenuadora en España y América". *Lingüística Española Actual* 29 (2007): 5-40.

Briz, Antonio. "El registro como centro de la variedad situacional: esbozo de la propuesta del grupo VAL.ES.CO. sobre las variedades diafásicas". *Perspectivas dialógicas en estudios del lenguaje*. Eds. Irene Fonte y Lidia Rodríguez Alfano. Universidad Autónoma de Nuevo León: México, 2010. 1-56.

Briz, Antonio, y Marta Albelda. "Una propuesta teórica y metodológica de la atenuación lingüística en español y portugués: la base de un proyecto en común (ES.POR.ATENUACIÓN)". *Onomázein* 28 (2013): 288-319.

Cabedo, Adrián. "Sobre prosodia, marcadores del discurso y unidades del discurso en español: evidencias de un corpus oral espontáneo". *Onomázein* 28 (2013): 201-13.

Cabedo, Adrián. "La función de la atenuación y la configuración prosódica: un estudio a partir de un corpus de español coloquial". *Revista Internacional de Lingüística Iberoamericana* 27 (2016): 55-74.

Caffi, Claudia. "On mitigation". *Journal of Pragmatics* 31 (1999): 881-909.

Caffi, Claudia. *Mitigation*. Oxford: Elsevier, 2007.

Erlich, Frances. "El discurso político venezolano actual: ¿atenuación o refuerzo del conflicto?". *La perspectiva no etnocentrista de la cortesía: identidad sociocultural de las comunidades hispanohablantes. (Actas del Primer Coloquio del Programa EDICE)*. Ed. Diana Bravo. Estocolmo: Universidad de Estocolmo/Programa EDICE, 2003. 227-39.

Estellés, Maria. "What Genres Tell Us about Evidentials and Vice Versa: A Study of *al parecer* in Spanish Parliamentary Debates (1979-2015)". *Pragmatics and Society* 9.3 (2018): 402-28.

Estellés, Maria, y Marta Albelda. "Evidentials, Politeness and Prosody in Spanish: A Corpus Analysis". *Journal of Politeness Research* 10.1 (2014): 29-62.

Field, Andrew. *Discovering Statistics Using SPSS: (and Sex and Drugs and Rock 'n' Roll)*. 3.ª ed. Londres: SAGE, 2009.

Glynn, Dylan. "Correspondence Analysis: Exploring Data and Identifying Patterns". *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Eds. Dylan Glynn y Jutyna A. Robinson. Amsterdam: John Benjamins, 2014. 443-85.

Greenacre, Michael J. *Correspondence Analysis in Practice*. London: Academic Press, 1993.

Gries, Stephan. *Statistics for Linguistics with R: A Practical Introduction*. Berlin/Boston: De Gruyter Mouton, 2013.

Hidalgo, Antonio. "Notas para el estudio de la entonación como factor integrador y delimitador de unidades sintácticas (enunciados) en el habla es-

pontánea". *Panorama de la investigació lingüística a l'estat espanyol. (Actes del I Congrés de Lingüística General)*. Valencia: Universitat de València, 1997. 14-27.

Hidalgo, Antonio. "Estructura e interpretación en la conversación coloquial: el papel del componente prosódico". *Revista de Filología de La Universidad de La Laguna* 24 (2006): 129-51.

Hidalgo, Antonio. "Fonocortesía: el estudio de la (des)cortesía a través del componente fónico de la lengua". *Miradas multidisciplinares a los fenómenos de cortesía y descortesía en el mundo hispánico*. Eds. Julio Escamilla y Henry Vega. Barranquilla: Universidad del Atlántico/Estocolomo: Universidad de Estocolmo, 2012. 473-97.

Hidalgo, Antonio, y Adrián Cabedo. "On the Importance of the Prosodic Component in the Expression of Linguistic Im/politeness". *Journal of Politeness Research* 10.1 (2014): 5-27.

Holmes, Janet. "Modifying Illocutionary Force". *Journal of Pragmatics* 8.3 (1984a): 345-65.

Holmes, Janet. "Hedging your Bets and Sitting on the Fence: Some Evidence for Hedges as Support Structures". *Te Reo* 27 (1984b): 47-62.

Hothorn, Torsten, Kurt Hornik y Achim Zeileis. "Unbiased Recursive Partitioning: A Conditional Inference Framework". *Journal of Computational and Graphical Statistics* 15.3 (2006): 651-74.

Lê, Sébastien, Julie Josse y François Husson. "FactoMineR: An R Package for Multivariate Analysis". *Journal of Statistical Software* 25.1 (2008): 1-18.

Liaw, Andy, y Matthew Wiener. "Classification and Regression by RandomForest". *R News* 2-3 (2002): 18-22.

López Serena, Araceli: "Selección natural, explicación racional y cambio lingüístico: hacia una fundamentación epistemológica no evolucionista de la teoría de la gramaticalización". *RILCE* 30.3 (2014): 724-75.

PRESEEA. *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá. 25 de noviembre de 2016. <http://preseea.linguas.net>.

Puga, Juana. *La atenuación en el castellano de Chile*. Valencia: Universitat de València, 1997.

R Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna: Austria. 25 de noviembre de 2016. <http://www.R-project.org/>.

Roberts, Seán G., Francisco Torreira y Stephen C. Levinson. "The Effects of Processing and Sequence Organisation on the Timing of Turn Taking: A Corpus Study". *Frontiers in Psychology* 6 (2015): 1-16.

Sbisà, Marina. "Illocutionary Force and Degrees of Strength in Language Use". *Journal of Pragmatics* 33 (2001): 1791-814.

Tanghe, Sanne. *Marcadores derivados de verbos de movimiento: una aproximación cognitiva a su polifuncionalidad*. Gante: Universidad de Gante, 2015.

Villalba, Cristina. "Las formas de tratamiento nominales y fórmulas apelativas convencionalizadas en los juicios orales: ¿elementos ritualizados o estrategias de atenuación?". *Revista Internacional de Lingüística Iberoamericana* 27 (2016a): 91-105.

Villalba, Cristina. *Actividades de imagen*, *atenuación e impersonalidad en los juicios orales*. Tesis doctoral. València: Universitat de València, 2016b.