

# Lógica difusa y gramaticalidad para una gramática de las lenguas naturales

Adrià Torrens Urrutia  
Universitat Rovira i Virgili  
adria.torrens@estudiants.urv.cat

## Resumen

---

La gradualidad en el lenguaje natural es un fenómeno que ha sido tradicionalmente muy discutido en lingüística. En este trabajo, se presentan las bases teóricas que nos han llevado a tratar el lenguaje natural de una manera gradual. Definir el lenguaje natural y su gradualidad nos obliga a buscar herramientas formales que puedan definir las bases de una gramática con grados de gramaticalidad. Aunque el uso de modelos matemáticos y formales es habitual en lingüística, la lógica difusa no ha recibido toda la atención que merece como herramienta para poder explicar el procesamiento del lenguaje natural. Aquí se exponen las bases de la lógica difusa como herramienta para lidiar con la formalización de una gramática que sea capaz de recoger valores no discretos, por lo tanto graduales o difusos. Las reglas de la gramática difusa vendrán dadas por una gramática de propiedades del español extraída mediante una metodología de corpus.

**Palabras clave:** lógica difusa, grados de gramaticalidad, gramáticas de propiedades, procesamiento del lenguaje natural

## 1. La lingüística y el procesamiento del lenguaje natural

El lenguaje natural se caracteriza por ser un lenguaje inmediato, espontáneo, con errores y ambiguo. Estas características se manifiestan en las producciones lingüísticas como *noisy text* o texto ruidoso (Baldwin et al. 2013) y suponen un gran inconveniente para el teórico ya que hacen que el análisis del lenguaje sea una tarea extremadamente complicada.

Por ello, a pesar de que las frases perfectas resultan una excepción en el lenguaje natural, la tendencia en lingüística ha sido establecer unos límites muy claros entre las producciones que son gramaticales y agramaticales: correctas e incorrectas. A este tipo de clasificación binaria la llamamos clasificación discreta o categórica.

Algunos autores han sido firmes defensores de una teoría lingüística completamente categórica que rechaza el aspecto gradual del lenguaje, como Hockett (1955), Joos (1957) y Bouchard (1995). En cambio, otros autores aunque reconocen que el aspecto gradual es inherente en el lenguaje natural establecen la concepción categórica del lenguaje para el estudio del lenguaje natural a modo metodológico, como Chomsky (1998) y Bever (1975).

La concepción categórica del lenguaje queda reforzada tras la distinción metodológica entre competencia y actuación (Chomsky 1965). Las teorías generativistas están interesadas en generar lenguaje perfecto, sin errores, por lo tanto no les interesa hablar de grados de gramaticalidad, ya que las reglas deben generar (sin interferencias) *outputs* perfectos desde lo más subyacente (competencia) hasta lo más superficial (actuación). Por lo tanto, las tendencias generativistas, a pesar de aceptar la existencia de grados de gramaticalidad, conocidos como *degrees of grammaticality* (Chomsky 1965: 11, 77), tienden a decantarse por la idealización del lenguaje como herramienta metodológica y, por consiguiente, a obviar o desestimar las tesis que consideran al lenguaje natural como un objeto de naturaleza gradual (Chomsky 1995).

Sin embargo, lo cierto es que en la mayoría de situaciones entendemos y procesamos frases con errores con independencia de la gravedad del error o de la violación de la regla. Si se tiene en cuenta que la capacidad de procesamiento de un *input* lingüístico es parte de la competencia lingüística, queda claro que la competencia lingüística es capaz de lidiar con una realidad difusa y gradual. Por lo tanto, procesamos el lenguaje de manera gradual (Keller 2000) o en términos difusos (Aarts 2004). Es decir, como humanos, no aceptamos y entendemos solamente las frases perfectas y, por consiguiente, rechazamos las frases con errores, sino que nuestra capacidad para procesar las producciones lingüísticas que recibimos (o *inputs*) es directamente proporcional a su gramaticalidad (o a la cantidad de reglas gramaticales respetadas y violadas).

## 2. El término *gradience* y teorías sobre la gradación del lenguaje

La concepción gradual del lenguaje, o el término *gradience*, ha aparecido a lo largo de la historia de la lingüística. En Aarts (2004), *gradience* se define como "un término que sirve para designar el espectro de los fenómenos continuos en el lenguaje, desde las categorías en el nivel de la gramática hasta los sonidos en el nivel de la fonética". De este modo, *gradience* es un término usado en lingüística para designar fenómenos que, debido a su naturaleza aparentemente difusa, encajan mejor en una explicación no discreta, como por ejemplo, con explicaciones de fenómenos lingüísticos en términos de grado.

Aarts (2004) define dos tipos básicos de gradación para las gramáticas:

- Gradación lineal: tenemos dos categorías diferentes claramente definidas:  $\alpha$  y  $\beta$ . La gradación aquí se encuentra en el tercer grupo de elementos que se ajustan al campo intermedio entre estas dos categorías. La idea es clasificar los objetos lingüísticos en la escala lineal en tres grados. Por lo tanto, en una escala de  $[0,1]$ ,  $\alpha$  es 1,  $\beta$  es 0 y el área media-borrosa es 0.5.
- Gradación teórica de conjuntos: Tenemos dos categorías diferentes claramente definidas ( $\alpha$  y  $\beta$ ). Sin embargo, aquí encontramos gradación cuando ocurre una intersección entre las dos categorías y la definimos como:  $\alpha \cap \beta$ . En este caso, ambos los elementos comparten características que se señalan en la intersección.

A primera vista, ambos tipos de gradación muestran algunos problemas. Por un lado, la gradación lineal es demasiado discreta, ya que solo muestra tres grados. Por lo tanto, esta parece meramente una noción discreta con un tercer grado. Por otro lado, la gradación teórica de conjuntos es demasiado vaga. Simplemente explica las funciones compartidas sin señalar ningún tipo de grado.

Algunos detractores de la gradación usan estas características de *gradience* para señalar contradicciones. Muñoz (2006) afirma que es contradictorio definir a un elemento como vago pero aún así clasificarlo de manera discreta, incluso si esta forma discreta tiene tres grados. Este tipo de críticas con relación a la gradación lingüística son frecuentes. Estas se ven motivadas debido a una mala comprensión de lo que es la gradación lingüística. La visión discreta es gradual en sí misma, pero solo tiene dos grados. La principal diferencia entre el enfoque discreto y el enfoque gradual es el siguiente: el enfoque discreto niega la gradualidad en lingüística mientras que *gradience* acepta el enfoque discreto y, además, lo mejora para los casos en los que se necesite establecer definiciones para el lenguaje natural jugando con todas las posibilidades de grado entre 1-0. Donde los enfoques discretos solo ven blanco ( $\alpha$ ) y negro

(β), las teorías de gradación lingüística aceptan blanco, negro, y todos los grises posibles entre esos dos.

### 2.1. Teorías sobre la gradación del lenguaje

Las principales teorías que han intentado explicar de un modo formal el aspecto gradual del lenguaje son la *Optimality Theory* (Prince & Smolensky 1993), y la *Linear Optimality Theory* (Keller 2000). Sin embargo, estas teorías son puramente evaluativas y no ponen de manifiesto cuáles son las normas que deben ser evaluadas a la hora de determinar la gramaticalidad de una producción. Tampoco son capaces de definir la competencia gramatical en términos de grado. Por lo que han sido acusadas de hacer lingüística de la actuación. Por este motivo, la ponderación de la gramaticalidad de una misma frase puede diferir en base a las diferentes teorías de gramáticas que estén evaluando un determinado *input*.

Teniendo en cuenta las limitaciones de teorías como las enumeradas anteriormente, en nuestro trabajo, sentamos las bases de una gramática con lógica difusa y definimos las reglas de su gramática a través de una gramática de restricciones. En este caso, se ha utilizado una *Property Grammars* o Gramática de Propiedades (Blache 2016) como base lingüística para presentar un sistema que resuelva la evaluación de las producciones del lenguaje natural y de sus agramaticalidades.

Por consiguiente, partimos de la idea de que el aspecto gradual del lenguaje (o el fenómeno *gradience*) puede ser explicado mediante dos herramientas: por un lado, el uso de la lógica difusa para definir el marco de una gramática, con Zadeh como base teórica (1965; 1972) y Novák (2015) con su formalismo *Fuzzy Type Theory*. Y, por otro lado, el uso de una gramática de propiedades de una lengua para la definición de *inputs* lingüísticos.

## 3. Las bases conceptuales de la lógica difusa

La concepción difusa y el concepto de *gradience* son bastante similares. Sin embargo, es necesario hablar sobre “lo difuso” para una comprensión completa de la noción lingüística de la gradación. A pesar de que algunos lingüistas como Bolinger (1961) intentaron acercar la concepción de grado a la lingüística, es en matemáticas y en la lógica difusa en donde podemos encontrar enfoques formales serios sobre los fenómenos de grado en relación con muchos incidentes cotidianos, como por ejemplo: la gradación de la temperatura, colores, o expresiones lingüísticas tales como la relación de grado entre alto-bajo, grande-pequeño. Sin embargo, este fenómeno en matemáticas se denomina fenómeno “difuso” y la lógica difusa es la herramienta adecuada para describir formalmente estas relaciones vagas que también se conocen como *fuzzy*.

Si partimos de un criterio como “altura = 2 metros (2 metros=1)”, la noción discreta diría que una persona solo es alta si alcanza los 2 metros, y las personas que no alcancen esos 2 metros, por ejemplo 1.75m, siempre son bajas (siempre sería 0). En cambio, la lógica difusa establece que dado un criterio y un índice, por ejemplo 0.75, en este caso, una persona será más alta cuanto más cercana esté del criterio máximo (1) y será más baja cuanto más lejos se encuentre del criterio máximo.

La lingüística se nutrirá de una mejor comprensión del término *gradience* a través del vasto campo de trabajo de la lógica difusa. Entendiendo que, dado un criterio de gramaticalidad, un *input* será más o menos gramatical cuanto más cerca o lejos esté del “estándar dorado” (*gold standard*) de la lengua natural que estemos trabajando. Por consiguiente, la “norma difusa si-

entonces” (o *fuzzy if-then rule*) será la siguiente: si el valor de gramaticalidad es alto, entonces el valor de agramaticalidad será bajo.

### 3.1. Definición de una gramática difusa

Siguiendo el formalismo *Fuzzy Type Theory* formulado por Novák (2015), aquí se define a una gramática natural en un sentido multimodal: Gramática (GR) = fonética (*ph*) & morfología (*m*) & sintaxis (*x*) & semántica (*s*) & léxico (*l*) & pragmática (*pg*) & prosodia (*pr*). Cada uno de los módulos de la gramática tiene una gramática de propiedades con restricciones lingüísticas que definen el conocimiento lingüístico de cada uno de los módulos. La multimodalidad nos permite ofrecer un grado de gramaticalidad para cada módulo lingüístico al mismo tiempo que un grado de gramaticalidad total.

La *Fuzzy Type Theory* establece la siguiente regla básica para poder obtener gradación:  $A: U \rightarrow L$ . Lo que significa que en un Universo (A) un elemento (U) tiene un grado de verdad (L).

En lingüística, se entiende que el universo (A) es la lengua (por ejemplo el español), el elemento es el *input* lingüístico (U), y el grado de verdad es el grado de gramaticalidad (L). El *input* lingüístico debe ser contrastado con las reglas que definen cada uno de los módulos de la gramática. En este caso, el grado de gramaticalidad *uni-modal* se conocerá a través del número de reglas satisfechas y violadas en el módulo de una gramática para con un *input* (por ejemplo, el grado de gramaticalidad en módulo de la sintaxis de una lengua).

El conjunto de los grados de todos los módulos lingüísticos nos proporcionará el valor de gramaticalidad final, o multimodal, con relación a un *input* lingüístico y a una gramática. Como hablamos de gramaticalidad y, por lo tanto de competencia, los grados de gramaticalidad no tienen por qué corresponderse con los de aceptabilidad. Por consiguiente, los grados de gramaticalidad son puramente teóricos y se establecerá que: el grado de gramaticalidad de una lengua se obtiene de una Gramática de Propiedades; Gramaticalidad Alta (GA: 1-0.8), Gramaticalidad Media (GM: 0.8-0.3), Gramaticalidad Baja (GB: 0.3-0).

## 4. Gramáticas de Propiedades y Gramaticalidad

Con respecto a la gramática difusa, las Gramáticas de propiedades de Blache (2000, 2005, 2016) se han elegido como gramática formal para definir “lo difuso” y la variabilidad en el lenguaje natural. Esta teoría combina: un sistema de restricciones lingüísticas (o propiedades), que son independientes y flexibles; dependencias sintácticas; y la noción de construcción de las Gramáticas de construcción de Goldberg (1995). Las construcciones lingüísticas son descritas en términos de sus propiedades. Las gramáticas de propiedades ofrecen varias restricciones para describir las relaciones sintácticas entre los fenómenos del lenguaje local. Sin embargo, para definir los *inputs* y establecer grados de gramaticalidad nos centramos sobre todo en las siguientes:

- **Precedencia** (*Prec*): Señala que un elemento A siempre debe preceder a un elemento B. Ej: Det *prec* N.
- **Requerimiento** (*Req*): Pone de manifiesto que dos categorías se requieren, de lo contrario ocurriría una violación gramatical. Ej: N *Req* Det en construcción de sujeto.
- **Exclusión** (*Excl*): Marca la no coocurrencia de dos elementos o categorías en un mismo contexto o construcción. Ej: Pron *Excl* N.

- **Dependencia** (*Dep*): Se encarga de identificar las relaciones de dependencias entre constituyentes y/o construcciones. Ej: Adj *Dep* [mod] N.

Por ser altamente descriptivas y por su gran capacidad para señalar relaciones lingüísticas, creemos que las *Property Grammars* son la teoría ideal para dar cuenta del fenómeno gradual del lenguaje. Asimismo, sus restricciones son totalmente autónomas y esto las hace potencialmente evaluables. Además, recogen mucha información lingüística indistintamente de si el *input* es gramatical o agramatical. Esto las diferencia de las otras teorías de restricciones en tanto que, a pesar de describir satisfactoriamente *input* canónico, esas teorías presentan restricciones que no son tan independientes y que ofrecen una muy escasa información acerca de los elementos que detonan violaciones. Por consiguiente, las Gramáticas de Propiedades son una buena herramienta para definir modelos sintácticos que toleren distintos niveles de gramaticalidad, y pueden resultar muy útiles a la hora de establecer mecanismos que permitan el análisis sintáctico de cualquier estructura sintáctica con independencia de su nivel de gramaticalidad.

#### 4.1. Extracción de la Gramática de Propiedades del español para una gramática difusa mediante el Corpus Universal Dependencias

Las propiedades de la sintaxis del español se han extraído automáticamente al aplicar *MarsaGram* a *Universal Dependency Spanish Treebank* (Blache et al., 2016). Este corpus se obtiene del conjunto de datos *Universal Dataset* de Google (versión 2.0) y consta de 16.006 estructuras de árbol y 430.764 *tokens*. Está constituido por artículos periodísticos, blogs y reseñas de consumidores.

El *treebank* de las dependencias del español proporciona relaciones de dependencia, núcleos y clases de palabras; mientras que *MarsaGram* clasifica por grupos a los constituyentes sintácticos, enumera las respectivas dependencias; y establece frecuencias, deduciendo automáticamente las propiedades lingüísticas que serán revisadas por el lingüista.

Por lo tanto, este método combina dependencias, constituyentes y restricciones sintácticas. En este sentido, *MarsaGram* ofrece importantes ventajas para la revisión lingüística; es decir, la posibilidad de analizar frases y simultáneamente extraer sus constituyentes y dependencias junto con sus frecuencias. Esto nos permite definir y caracterizar las construcciones españolas y sus propiedades de acuerdo con un criterio de datos objetivos. Una vez que se ha realizado una revisión lingüística adecuada, esta información lingüística se utiliza para definir tanto las relaciones graduales en una lengua como los fenómenos difusos en su sintaxis.

## 5. Conclusiones

El uso de la lógica difusa y de sus reglas aplicadas a la lingüística con una gramática de propiedades proporciona interesantes descripciones formales que sientan las bases para una gramática que pretenda recoger valores difusos.

Este método consigue poder explicar la gramaticalidad con una formulación y una base matemática, así como especificar en dónde se encuentran los grados de gramaticalidad, y, de este modo, poder defender su existencia de un modo formal. De la combinación de lógica difusa y gramáticas de restricción se obtendrá un nuevo e interesante método teórico con el que dar cuenta de las variabilidades del lenguaje de un modo más preciso. Además, las formulaciones lingüísticas con razonamiento lógico-difuso nos dan la oportunidad de realizar implementaciones informáticas para el desarrollo de nuevas tecnologías con base lingüística.

## Referencias bibliográficas

- Aarts, B. (2004). *Conceptions of gradience in the history of linguistics*. *Language Sciences*, 26(4):343–389.
- Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). “How noisy social media text, how different social media sources?”. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 1(1): 356-364.
- Blache, P. (2000). “Property Grammars and the Problem of Constraint Satisfaction”. *Proceedings of ESSLLI 2000 workshop on Linguistic Theory and Grammar Implementation*, 1(1):47–56.
- Blache, P. (2005). “Property Grammars: A Fully Constraint-based Theory”. *Constraint Solving and Language Processing*, 3438(1):1–16.
- Blache, P. (2016). “Representing Syntax by Means of Properties: a Formal Framework for Descriptive Approaches”. *Journal of Language Modelling*, 4(2):183–224.
- Blache, P., Rauzy, S., & de Montcheuil, G. (2016). “MarsaGram: an excursion in the forests of parsing trees”. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 1(1): 2336–2342.
- Bever, T. (1975). “Functional Explanations Require independently motivated functional theories”. In Grossman, R.E. et al (eds.), *Papers from the Parasession on Functionalism*. Chicago Linguistic Society, 1(1): 580-609.
- Bolinger, D. L. M. (1961). *Generality: Gradience and the All-or-none*. Mouton & Company, 14 edition.
- Bouchard, D. (1995). *The Semantics of Syntax: A Minimalist Approach to Grammar*. Chicago: University of Chicago Press.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.
- Chomsky, N. (1995). *The minimalist program (current studies in linguistics 28)*. Cambridge et al.
- Chomsky, N. (1998). “Some observations on economy in generative grammar”. *Is the best good enough*, 1(1): 115-127.
- Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.
- Hockett, C. F. (1955). *A Manual of Phonology*. Waverly Press, Baltimore.
- Joos, M. (1957). “Description of Language Design”. *Journal of the Acoustical Society of America*, 22: 701-708.
- Keller, F. (2000). *Gradience in grammar: experimental and computational aspects of degrees of grammaticality*. PhD thesis, University of Edinburgh.
- Muñoz, R. (2006). “Hacia una ¿nueva? lingüística: reflexiones sobre la llamada alternativa no-discreta”. *Pandora: revue d'études hispaniques*, 6(1):37–58.
- Novák, V. (2015). “Fuzzy Natural Logic: Towards Mathematical Logic of Human Reasoning”. In *Towards the Future of Fuzzy Logic*, 1(1): 137-165.
- Prince, A., Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report, New Brunswick: Rutgers University.
- Zadeh, L. A. (1965). “Fuzzy sets”. *Information and control*, 8(3):338–353.
- Zadeh, L. A. (1972). “A fuzzy-set-theoretic interpretation of linguistic hedges”. *Journal of Cybernetics*, 2(3):4–34.