Empezar y *comenzar*: diferencias diatópicas, diafásicas y semánticas en corpus digitales

Cuadernos CANELA, 36, pp. 183-199
Recibido: 1-VIII-2024
Aceptado: 7-XII-2024
Publicado, versión impresa: 1-V-2025
ISSN 1344-9109
Publicado, versión electrónica: 1-V-2025
ISSN 2189-9568

canela.org.es

Haakon S. Krohn
Universidad de Costa Rica, San José, Costa Rica

Resumen
En este artículo se analizan los usos de los verbos españoles *empezar* y *comenzar* de acuerdo con distintas variables diatópicas, diafásicas y semánticas en varios corpus digitalizados consultados mediante la herramienta Sketch Engine. En primer lugar, se determina que *comenzar* es más común que *empezar* en la mayoría de los países hispanohablantes, pero que *empezar* es el verbo preferido en los países andinos de Perú, Ecuador y Colombia, así como en Paraguay, Guatemala, Costa Rica y España. En segundo lugar, se concluye que *comenzar* se emplea más en registros formales y escritos, mientras que *empezar* se utiliza más en registros caracterizados por la informalidad y la oralidad. Por último, se encuentran indicios de que *empezar* es el verbo preferido cuando el sujeto presenta mayor grado de agentividad.

Palabras clave
lengua española, sinonimia, lingüística de corpus, dialectología del español, registros lingüísticos, agentividad semántica

Introducción

Los verbos españoles *empezar* y *comenzar* presentan significados y usos tan similares que lo más común es tratarlos como sinónimos totales/absolutos/perfectos. Esto se refleja en las definiciones proporcionadas en obras lexicográficas; por ejemplo, en el *Diccionario de la lengua española* (Real Academia Española, 2023), todas las acepciones de *comenzar* remiten directamente a *empezar*, cuyas definiciones pueden resumirse como 'dar principio a algo, iniciar el uso o consumo de algo, tener principio en un lugar, dar comienzo en el tiempo'. Sin embargo, como señala la mayoría de los autores que ahondan en el tema de la sinonimia (p. ej. Hurford, Heasley y Smith, 2007; Löbner, 2013; García, 2014; Espinal y Mateu, 2020), la sinonimia total entre dos palabras o expresiones en todos los niveles de la lengua es una particularidad muy difícil de encontrar. Por esta razón, es probable que también en el caso de los verbos inceptivos *empezar* y *comenzar*, aunque pueden ser utilizados en los mismos contextos, se manifiesten tendencias de preferencia hacia uno u otro en función de la interacción entre distintas variables.

En el presente estudio se examinan tres tipos de variables: la variación diatópica (geográfica), diversos factores diafásicos (estilísticos) y las propiedades semánticas del sujeto del verbo. Todos los datos empleados en el análisis provienen de una serie

Correspondencia: Haakon S. Krohn, Escuela de Filología, Lingüística y Literatura, Universidad de Costa Rica, Ciudad Universitaria Rodrigo Facio Brenes, San Pedro de Montes de Oca, Provincia de San José, Costa Rica.
Correo electrónico: hkrohn@gmail.com

de corpus digitalizados y se obtuvieron por medio de la herramienta Sketch Engine (Lexical Computing, 2024; Kilgarriff *et al*., 2004; Kilgarriff *et al*., 2014). Los resultados enriquecen la comprensión de por qué dos verbos con significados tan similares conviven en el español actual, por lo que son relevantes para el estudio de la sinonimia en general. Asimismo, son de utilidad en el ámbito de la lingüística aplicada, particularmente para la enseñanza del español como segunda lengua.

1. Tipos de sinonimia

Dicho de forma sencilla, dos expresiones lingüísticas son sinónimas si presentan el mismo significado. Sin embargo, no existe una definición clara de «el mismo significado» y, además, se pueden identificar distintos grados y tipos de similitud o equivalencia semántica.

Las definiciones lexicográficas de *empezar* y *comenzar* sugieren que estas palabras podrían presentar el tipo de relación que se conoce como sinonimia total, absoluta o perfecta. De acuerdo con Löbner (2013, p. 203), la sinonimia total alude a una equivalencia entre todas las variantes de significado de dos lexemas: todas las dimensiones de significado, incluyendo el descriptivo, el social y el expresivo. Similarmente, según Lyons (1995, p. 61), para que dos expresiones sean totalmente sinónimas, (1) todos sus significados deben ser idénticos, (2) tienen que ser sinónimas en todos los contextos y (3) deben ser semánticamente equivalentes en todas las dimensiones, tanto las descriptivas como las no descriptivas.

En consecuencia, como señala García (2014), «es muy difícil encontrar dos palabras en una lengua que sean sinónimas en todos los niveles de análisis» (p. 126). Al respecto, Hurford, Heasley y Smith (2007) comentan que «ejemplos de sinonimia perfecta son difíciles de encontrar, quizá porque hay poco sentido en que un dialecto tenga dos predicados con exactamente el mismo significado» (p. 106).[1] Por estas razones, la gran mayoría de los sinónimos en cualquier lengua son sinónimos parciales, como también concluyen autores como Löbner (2013, p. 204) y Espinal y Mateu (2020, p. 81).

Desde la perspectiva de Cruse (2000), los sinónimos son «palabras cuyas similitudes semánticas son más prominentes que sus diferencias» (p. 156).[2] Dicho autor (pp. 157-159) subdivide la sinonimia en tres categorías: (1) la sinonimia absoluta, donde dos palabras presentan equivalencia completa en cuanto a su significado; (2) la sinonimia proposicional, donde una palabra puede sustituirse por otra en cualquier expresión sin que cambien las propiedades condicionales de veracidad; y (3) la casi-sinonimia, donde dos palabras presentan significados cercanos, pero no equivalentes. El lingüista admite que el linde entre la casi-sinonimia y la no-sinonimia es difícil de definir de manera exacta.

Otra tipología de la sinonimia proviene de Apresian (1957). Esta difiere de la anteriormente expuesta por el hecho de que no considera el grado, sino el tipo de equivalencia semántica entre las palabras. Una traducción al español de la clasificación de Apresian, publicada por Cruz (2013), se cita a continuación:

> 1. *Sinónimos absolutos*: Son palabras que tienen idéntico significado y pueden ser sometidas a la prueba de sustitución (*lenguas romances-lenguas neolatinas*). Se incluyen aquí los términos científico-técnicos que, procedentes indistintamente del griego y del latín, coexisten en la lengua (*oculista-oftalmólogo*). También los vocablos que son

préstamos de otras lenguas y que se emplean paralelamente con los vocablos propios del español (*buró-escritorio*, *almanaque-calendario*).

2. *Sinónimos semánticos*: Aquí se incluyen las palabras estilísticamente neutrales, que se diferencian entre sí por los matices de la significación fundamental o general de cada uno de ellos (*afirmar-corroborar*, *edificio-monumento*).

3. *Sinónimos estilísticos*: Son las palabras de igual significación que se diferencian entre sí por el matiz estilístico. Este tipo se divide en dos subgrupos: a) palabras que se emplean solo dentro de los límites de determinada región, llamadas regionalismos (*cubo-balde, silla-balance*, *mamey-zapote*). b) Palabras empleadas en la lengua literaria, escrita y hablada (*ornato-adorno*, *cabello-pelo*, *embriagado-borracho-curda*).

4. *Sinónimos estilísticos-semánticos*: Son las palabras que designan un mismo fenómeno de la realidad objetiva y que se diferencian tanto por su significación como por el matiz estilístico (*agrio-ácido*, *esposo-marido*) (Cruz, 2013, pp. 111-112).

2. Metodología

Todos los datos analizados en esta investigación se obtuvieron por medio de Sketch Engine (Lexical Computing, 2024; Kilgarriff *et al*., 2004; Kilgarriff *et al*., 2014). Dicha herramienta se empleó para acceder a una serie de corpus digitales y extraer información acerca del uso de los dos verbos bajo estudio. Si bien el etiquetado gramatical en los corpus disponibles a menudo es muy limitado, o incluso erróneo, la posibilidad de realizar búsquedas en varios corpus y subcorpus distintos permite adquirir información cuantitativa acerca de la distribución diatópica y diafásica de los verbos en cuestión. Además, la visualización de las colocaciones de los lemas analizados proporciona algunas indicaciones sobre posibles matices semánticos.

Según las diferentes variables consideradas, se presentarán a lo largo del artículo los números absolutos de apariciones de cada verbo, así como los porcentajes correspondientes a la proporción de la cantidad total de ocurrencias de ambos lemas analizados. Por tanto, si ambos fueran equivalentes, se esperaría una frecuencia de aproximadamente 50 % de cada uno en todos los casos. Todas las tablas se ordenarán descendientemente según la proporción de uso de *empezar*, ya que, como se apreciará, es el menos frecuente de los dos en el mundo hispanohablante, por lo que interesa determinar cuáles son los factores que propagan su uso.

La preferencia por uno de los verbos en cada caso se verificará mediante una prueba Z de dos proporciones. Se empleará un valor *p* de 0,01 como umbral para la significancia estadística, lo cual quiere decir que, con base en dicha prueba, se comprobará si las diferencias de distribución entre *empezar* y *comenzar* detectadas son estadísticamente significativas con un nivel de confianza del 99 %. Cuando una diferencia observada no resulte significativa, se señalará explícitamente.

3. Resultados

3.1. Variación diatópica

Para el análisis de la variación geográfica en el uso de *empezar* y *comenzar* dentro del mundo hispanohablante, se emplearon tres corpus: *Spanish web corpus 2023* (*esTenTen23*), *Spanish web corpus 2018* (*esTenTen18*) y *Spanish web corpus 2011* (*esTenTen11*). Estos son los únicos corpus de la lengua española disponibles en Sketch Engine que incluyen subcorpus geográficos. Los tres forman parte de la familia de corpus *TenTen* y han sido recopilados de páginas web mediante el *web crawler* Spiderling. Posteriormente, han sido limpiados, tokenizados y lematizados automáticamente. El corpus *esTenTen23* fue recopilado en el 2023 y contiene 28 652 392 686 palabras, *esTenTen18* proviene del 2018 e incluye 16 953 735 742 palabras, y *esTenTen11* es del año 2011 y consta de 9 497 213 009 palabras. Considerar tres corpus recopilados en tres años distintos ayuda a corroborar las tendencias observadas, aunque, desde luego, es posible que cierto contenido sea compartido por todos, ya que no se especifican las fuentes exactas. En estos corpus, Sketch Engine permite realizar búsquedas en subcorpus clasificados según el dominio de nivel superior geográfico (ccTLD) del que se extrajo el contenido.

Primero, se realizó una búsqueda de *empezar* y *comenzar* en cada subcorpus geográfico disponible en el corpus *esTenTen23* y se registró la cantidad de ocurrencias de cada lema.[3] Los resultados se presentan en la tabla 1. En todos los casos, las diferencias entre los dos verbos son estadísticamente significativas. Los números acumulados no corresponden exactamente a las sumas de los países constituyentes, debido a que no todos los textos están asignados a un subcorpus geográfico. De todas formas, estas cifras sumativas son de menor interés, dada la distribución geográfica tan desequilibrada del corpus.

País	***empezar***		***comenzar***	
Perú	129 236	**57,8 %**	94 337	**42,2 %**
Ecuador	32 038	**57,1 %**	24 062	**42,9 %**
Paraguay	22 506	**53,2 %**	19 800	**46,8 %**
Guatemala	9 287	**52,0 %**	8 565	**48,0 %**
Colombia	170 679	**51,9 %**	158 227	**48,1 %**
Costa Rica	12 689	**51,3 %**	12 032	**48,7 %**
España	2 004 282	**51,3 %**	1 902 597	**48,7 %**
Nicaragua	7 086	**46,6 %**	8 113	**53,4 %**
Todo el corpus	9 305 892	**46,6 %**	10 680 235	**53,4 %**
Bolivia	18 662	**42,5 %**	25 275	**57,5 %**
México	415 508	**41,1 %**	595 590	**58,9 %**
Hispanoamérica	1 816 525	**38,3 %**	2 929 579	**61,7 %**

República Dominicana	10 885	**37,1 %**	18 466	**62,9 %**
El Salvador	3 549	**36,5 %**	6 173	**63,5 %**
Argentina	631 453	**35,9 %**	1 126 449	**64,1 %**
Honduras	8 764	**33,0 %**	17 761	**67,0 %**
Uruguay	70 291	**32,1 %**	148 676	**67,9 %**
Chile	186 029	**29,5 %**	444 764	**70,5 %**
Venezuela	19 887	**28,5 %**	49 892	**71,5 %**
Cuba	62 069	**27,4 %**	164 655	**72,6 %**

Tabla 1. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2023* (*esTenTen23*), según país.

Los números revelan diferencias diatópicas relativamente claras. Se puede apreciar que *comenzar* se usa con mayor frecuencia que *empezar* en el corpus en su totalidad, lo cual se debe a la alta preferencia por *comenzar* en muchos países de Hispanoamérica. En solo siete países —Perú, Ecuador, Paraguay, Guatemala, Colombia, Costa Rica y España—, *empezar* es el más frecuente de los dos verbos. El patrón geográfico más sobresaliente es la preferencia por *empezar* en Ecuador, Perú y Colombia, la cual se revela como una característica dialectal compartida por estos países vecinos, que abarcan la parte septentrional de la cordillera de los Andes.

En el otro extremo de la escala, se observa una fuerte preferencia por *comenzar* principalmente en dos áreas geográficas: (1) en los países insulares caribeños de Cuba y República Dominicana, así como en Venezuela, cuya habla (excepto la de la zona andina) también suele considerarse parte del español caribeño, y (2) en Chile, Uruguay y Argentina, que conforman el Cono Sur. En menor grado, el verbo *comenzar* también es el preferido en Bolivia, por lo que este lexema es el más frecuente en toda América del Sur excepto en Perú, Ecuador, Paraguay y Colombia. Adicionalmente, *comenzar* es el verbo inceptivo más usado en México y en los países centroamericanos de Honduras, El Salvador y Nicaragua, aunque las diferencias entre *empezar* y *comenzar* son mucho menos destacadas en Nicaragua y México.

Cabe recordar que existe la posibilidad de que las cifras se encuentren sesgadas debido a diferencias diafásicas en las páginas web de donde se recopilaron los datos de cada país. No obstante, la manifestación de patrones geográficos tan evidentes es un indicio de que tales oblicuidades potenciales no tendrían un gran impacto.

En la figura 1 se visualizan los números de la tabla 1 en un mapa. El porcentaje de oscuridad del color gris de cada país hispanohablante se correlaciona directamente con el porcentaje de aparición de *empezar*, pero las proporciones han sido normalizadas para aumentar los contrastes en el mapa.[4] Los datos de Panamá provienen del corpus *esTenTen11*, cuyas cifras se presentan más adelante. A su vez, los demás países aparecen en color blanco.

Figura 1. Mapa que muestra el uso del verbo *empezar* en comparación con *comenzar* en países hispanohablantes: cuanto más oscuro el color gris, mayor es la preferencia por *empezar*. Fuente: elaboración propia.[5]

Para validar estos hallazgos, se realizó el mismo tipo de análisis en los corpus *esTenTen18* y *esTenTen11* (el cual es el único que incluye Panamá entre los subcorpus), que son significativamente más pequeños. Se reconoce que partes de estos corpus probablemente se incluyen en *esTenTen23* también, por lo que no se trataría de recopilaciones totalmente independientes, pero la distribución de los diferentes registros lingüísticos podría ser distinta en cada corpus. Los resultados de *esTenTen18* se muestran en la tabla 2.

País	***empezar***		***comenzar***	
Ecuador	20 720	**61,7 %**	12 881	**38,3 %**
Perú	49 686	**58,3 %**	35 568	**41,7 %**
Colombia	95 171	**52,8 %**	84 954	**47,2 %**
Costa Rica	8 685	**52,2 %**	7 946	**47,8 %**
Guatemala	801	**50,4 %**	789	**49,6 %**
España	873 688	**50,2 %**	868 122	**49,8 %**
Nicaragua	8 920	**48,8 %**	9 373	**51,2 %**
Todo el corpus	5 307 654	**46,5 %**	6 107 836	**53,5 %**
Paraguay	11 248	**46,4 %**	12 982	**53,6 %**
México	260 905	**43,9 %**	332 926	**56,1 %**
Bolivia	13 272	**42,1 %**	18 253	**57,9 %**
Hispanoamérica	1 077 189	**38,4 %**	1 728 098	**61,6 %**
El Salvador	2 325	**37,9 %**	3 813	**62,1 %**
Honduras	2 325	**37,5 %**	3 870	**62,5 %**
República Dominicana	5 313	**36,7 %**	9 178	**63,3 %**
Argentina	436 574	**35,9 %**	777 854	**64,1 %**
Uruguay	31 700	**29,5 %**	75 827	**70,5 %**
Venezuela	12 663	**29,5 %**	30 228	**70,5 %**
Chile	84 002	**28,1 %**	215 106	**71,9 %**
Cuba	32 879	**25,4 %**	96 550	**74,6 %**

Tabla 2. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2018* (*esTenTen18*), según país.

De manera general, se observan los mismos patrones que en el corpus anteriormente analizado. La única disparidad destacable es el caso de Paraguay, que en este corpus demuestra una clara preferencia por *comenzar*. Por añadidura, la diferencia entre las proporciones de ambos lemas en Guatemala no es estadísticamente significativa en este corpus ($p \approx 0{,}76$).

A continuación, en la tabla 3, se exponen los datos geográficos del corpus *esTenTen11*.

País	***empezar***		***comenzar***	
Perú	87 897	**59,3 %**	60 369	**40,7 %**
Guatemala	8 757	**58,8 %**	6 124	**41,2 %**
Costa Rica	11 230	**55,9 %**	8 867	**44,1 %**
Ecuador	21 340	**55,6 %**	17 035	**44,4 %**
España	630 125	**52,1 %**	580 167	**47,9 %**
Colombia	87 471	**50,8 %**	84 756	**49,2 %**
México	416 381	**47,1 %**	467 580	**52,9 %**
Panamá	3 429	**45,5 %**	4 103	**54,5 %**
Nicaragua	13 760	**45,4 %**	16 555	**54,6 %**
Todo el corpus	2 502 360	**41,2 %**	3 573 568	**58,8 %**
Bolivia	11 817	**39,3 %**	18 281	**60,7 %**
Hispanoamérica	1 864 850	**38,5 %**	2 984 595	**61,5 %**
Paraguay	9 821	**37,4 %**	16 417	**62,6 %**
Argentina	871 189	**36,3 %**	1 525 959	**63,7 %**
Honduras	1 533	**35,4 %**	2 802	**64,6 %**
República Dominicana	7 914	**34,8 %**	14 846	**65,2 %**
El Salvador	5 370	**34,8 %**	10 081	**65,2 %**
Uruguay	45 941	**32,6 %**	95 009	**67,4 %**
Chile	192 910	**30,2 %**	446 299	**69,8 %**
Cuba	37 739	**26,6 %**	103 980	**73,4 %**
Venezuela	30 351	**26,2 %**	85 532	**73,8 %**

Tabla 3. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2011* (*esTenTen11*), según país.

En este corpus, *comenzar* es todavía más frecuente que en los anteriores, con un 58,8 % de las apariciones de los dos lexemas, pero esto podría deberse parcialmente a la menor proporción de textos de España, donde predomina el verbo *empezar*. Todas las diferencias entre ambos verbos son estadísticamente significativas para todos los países. Se observan tendencias muy similares a las de los corpus anteriormente analizados: *empezar* es el verbo preferido en los países andinos de Perú, Ecuador y Colombia, en los países centroamericanos de Guatemala y Costa Rica, y en España, mientras que *comenzar* es el más común en el resto de los países hispanoamericanos, incluyendo Panamá.

3.2. Variación diafásica

La variación diafásica ocurre de acuerdo con el contexto y el propósito de la comunicación, e incluye variables como el medio de comunicación y distintos niveles de formalidad. No existe consenso acerca de la denominación de cada variedad en la dimensión diafásica; Biber y Conrad (2019, p. 21) señalan que las nociones *registro* y *género* han sido utilizadas de manera equivalente para referirse a dicho concepto, mientras que otros autores han planteado una distinción teórica entre ambas. En el presente artículo se emplearán estos términos como sinónimos totales.

Siguiendo el punto de vista de Biber y Conrad (2019), el caso de la variación en el uso de *empezar* y *comenzar* en textos de distintos registros/géneros se estudiaría desde una perspectiva estilística, dado que la variación entre ambos vocablos no sería directamente funcional, pues presentan las mismas propiedades sintácticas, sino que la prevalencia de uno u otro se debería a preferencias estilísticas.

Posibles disparidades entre *empezar* y *comenzar* asociadas con la dimensión diafásica pueden detectarse por medio de su frecuencia en corpus distintos, dado que algunos de los corpus disponibles en Sketch Engine representan registros relativamente homogéneos. Los corpus considerados como representativos de registros específicos se describen en la tabla 4, donde se presentan en el mismo orden que en la tabla 5, para facilitar la comparación entre ambas.

Corpus	**Descripción**	**Cantidad de palabras**	**Referencia**
CELEN: Learner Corpus of Spanish in Japan	Textos escritos producidos por aprendices de español que tienen el japonés como lengua materna	658 467	Valverde (2024)
OpenSubtitles 2018 parallel	Subtítulos de películas y series televisivas traducidos	753 235 853	Lison y Tiedemann (2016)
Europarl spoken parallel	Transcripciones de habla oral de las Actas del Parlamento Europeo	54 302 284	Koehn (2005)
Gutenberg Spanish 2020	Libros electrónicos de dominio libre	37 202 233	Project Gutenberg (s.f.)
Spanish parliamentary debates (ParlaMint 2.1, CoNLL format)	Transcripciones de debates parlamentarios orales de España	12 930 879	Erjavec *et al*. (2023)
Timestamped JSI web corpus 2014-2021	Artículos de flujos de noticias escritos	16 358 148 966	Bušta *et al*. (2017)
DGT-Translation Memory Parallel	Traducciones de documentos legislativos de la Unión Europea	57 311 149	Steinberger *et al*. (2013)

EUR-Lex 2/2016 parallel	Traducciones de leyes y otros documentos públicos de la Unión Europea	635 187 126	Lexical Computing (2023)
United Nations Parallel Corpus (UNPC)	Documentos escritos oficiales de las Naciones Unidas	692 809 915	Tiedemann (2012)

Tabla 4. Corpus en español disponibles en Sketch Engine que se asocian con un registro/género específico.

Seguidamente, en la tabla 5 se presentan los números absolutos y relativos de apariciones de *empezar* y *comenzar* en estos corpus. Las diferencias entre los dos verbos son estadísticamente significativas en todos los casos. Es importante recordar que cualquier desviación de la distribución 50/50 sugiere una preferencia por uno de los verbos en el registro representado en el corpus.

Corpus	***empezar***		***comenzar***	
CELEN: Learner Corpus of Spanish in Japan	477	**88,0 %**	65	**12,0 %**
OpenSubtitles 2018 parallel	407 679	**68,1 %**	190 579	**31,9 %**
Europarl spoken parallel	10 116	**55,9 %**	7 972	**44,1 %**
Gutenberg Spanish 2020	9 963	**45,3 %**	12 025	**54,7 %**
Spanish parliamentary debates (ParlaMint 2.1, CoNLL format)	4 092	**43,8 %**	5 243	**56,2 %**
Timestamped JSI web corpus 2014-2021	4 897 431	**41,2 %**	6 997 349	**58,8 %**
DGT-Translation Memory Parallel	1 538	**31,0 %**	3 426	**69,0 %**
EUR-Lex 2/2016 parallel	17 419	**30,1 %**	40 391	**69,9 %**
United Nations Parallel Corpus (UNPC)	35 333	**28,5 %**	88 836	**71,5 %**

Tabla 5. Ocurrencias de los lemas *empezar* y *comenzar* en diferentes corpus.

Con base en estos datos, se pueden identificar dos factores que influyen en el uso de estos verbos. En primer lugar, *empezar* parece ser el preferido en la comunicación informal, mientras que *comenzar* se asocia más con el registro formal. Por un lado, los dos corpus más estrechamente relacionados con un estilo informal —el de aprendices japoneses y el de subtítulos de películas y series— son los que presentan el mayor porcentaje de uso de *empezar*. Incluso, la proporción tan alta de *empezar* en el *CELEN* sugiere que este es el primero de los dos verbos que es adquirido por los japoneses, pero el tamaño del corpus es muy reducido. En contraste, *comenzar* es predominante en los corpus de comunicación formal. En relación con esto, también cabe resaltar las altas frecuencias de

uso de *comenzar* en los corpus que representan comunicación escrita formal provenientes de España, pues el verbo inceptivo que resultó ser el más frecuente en los subcorpus de España en *esTenTen23*, *esTenTen18* y *esTenTen11* fue *empezar*.

En segundo lugar, el registro oral parece favorecer *empezar*, mientras que *comenzar* es el preferido en textos escritos. El único corpus que se podría considerar de un registro formal donde *empezar* presenta la mayor cantidad de usos es *Europarl spoken parallel*, el cual consiste justamente en transcripciones de habla oral. En el otro extremo, se observa que *comenzar* es más común en textos formales escritos, particularmente en documentos oficiales de las Naciones Unidas y documentos legislativos, pero también en los flujos de noticias.

El corpus *esTenTen18*, además de los dominios nacionales, contiene subcorpus divididos en tres géneros específicos: blog, discusión y noticias. La cantidad de ocurrencias de *empezar* y *comenzar* en estos subcorpus se presenta en la tabla 6, donde todas las diferencias entre los lemas son estadísticamente significativas.

Género	***empezar***		***comenzar***	
Blog	64 083	**66,0 %**	33 062	**34,0 %**
Discusión	37 441	**54,3 %**	31 456	**45,7 %**
Noticias	53 705	**44,7 %**	66 314	**55,3 %**

Tabla 6. Ocurrencias de los lemas *empezar* y *comenzar* en el Spanish web corpus 2018 (esTenTen18), según género textual.

Se puede apreciar que *empezar* es el verbo inceptivo más usado en dos de estos géneros, a saber, blog y discusión, los cuales pueden considerarse informales. En contraste, *comenzar* es el preferido en el género de noticias, el cual es más formal. Por lo tanto, estos números confirman las asociaciones de estos dos lexemas con niveles de formalidad diferentes.

3.3. Variación semántica

Es limitado lo que se puede indagar acerca de los matices semánticos de los verbos en corpus que no proporcionan información semántica detallada, como ocurre en el caso de los que se analizan en el presente estudio.[6] Además, aunque Sketch Engine proporciona listados de los sujetos y objetos directos de los distintos verbos, estas listas no se generan de forma completa, lo cual dificulta el análisis de la información, y los verbos no están etiquetados según su valencia gramatical, lo que significa que la herramienta no ofrece ninguna manera directa de comparar la cantidad de usos transitivos e intransitivos de los verbos. Sin embargo, algunos de los datos obtenidos por medio de esta herramienta sí son de utilidad para un análisis de la semántica de *empezar* y *comenzar*, particularmente los listados de sujetos y objetos pronominales de cada verbo. A continuación, se analiza dicha información del corpus *esTenTen23*.

En la tabla 7 se observa que, a pesar de que *comenzar* ocurre con mayor frecuencia en este corpus en general, *empezar* es mucho más usado con los sujetos pronominales *yo* y *tú*.[7] Se puede añadir que, considerando el número total de apariciones de ambos

verbos en el corpus, *empezar* se emplea con alguno de estos sujetos pronominales en un 0,4 % de los casos, mientras que *comenzar* solo en un 0,1 %. Esto es relevante porque la primera y la segunda persona gramaticales se ubican en las dos primeras posiciones de la jerarquía de animicidad (Dixon, 1994, p. 85; Foley, 1999, p. 210) y casi siempre hacen referencia a seres humanos, mientras que la tercera persona gramatical muchas veces alude a un referente inanimado sin posibilidad de ser agentivo, tal como un objeto, un evento o un proceso. Por consiguiente, estos números sugieren que *empezar* es el verbo preferido cuando el sujeto presenta mayor grado de agentividad, mientras que *comenzar* se emplea predominantemente con sujetos de agentividad más baja.

Sujeto pronominal	***empezar***		***comenzar***	
yo	33 336	**78,2 %**	9 289	**21,8 %**
tú	1 645	**68,7 %**	751	**31,3 %**

Tabla 7. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2023* (*esTenTen23*) con los sujetos pronominales *yo* y *tú*.

La cantidad de apariciones de los verbos con objetos pronominales tiene un uso más indirecto, pero igualmente útil, para el análisis semántico, ya que la presencia de un objeto directo (sea pronominal o no) indica que el verbo se emplea como transitivo. Sketch Engine únicamente cuenta las veces que los verbos ocurren con los objetos pronominales *lo* y *le*. Con *lo*, el verbo siempre es transitivo (a menos que haya casos de loísmo), mientras que con *le*, es transitivo solo en los casos de leísmo o cuando coocurre con un objeto directo expresado mediante un sintagma nominal no pronominal en una oración ditransitiva. De esta manera, las apariciones de cada verbo con *lo* constituyen un subconjunto relativamente representativo de todos sus usos transitivos en el corpus.

La transitividad se relaciona con la agentividad en el sentido de que el sujeto de un verbo transitivo prototípicamente es más agentivo que el de un verbo intransitivo (Kittilä, 2000; Kibort, 2008), lo cual, por ejemplo, es una causa de la existencia de sistemas ergativo-acusativos en muchas lenguas del mundo. Por consiguiente, la preferencia por *empezar* o *comenzar* con el objeto pronominal *lo* (y parcialmente con *le*) ofrece una indicación de la asociación entre estos verbos y el grado de agentividad. Los datos en cuestión se presentan en la tabla 8.[8]

Sujeto pronominal	***empezar***		***comenzar***	
le	38 779	**78,6 %**	10 550	**21,4 %**
lo	57 308	**71,4 %**	23 002	**28,6 %**

Tabla 8. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2023* (*esTenTen23*) con los objetos pronominales *lo* y *le*.

Se aprecia que *empezar*, con un gran margen, es el verbo inceptivo preferido tanto con el objeto pronominal *le* como con *lo*. Lo último apunta a que se utiliza mucho más

frecuentemente como transitivo que *comenzar*. Esto, a su vez, es otra indicación de que *empezar* se vincula en mayor medida con actantes que presentan un alto grado de agentividad. No obstante, parece haber una relación todavía más fuerte entre la existencia de un objeto indirecto y el uso de *empezar*, lo cual requerirá un análisis más comprehensivo, ya que la relación entre la agentividad del sujeto y la presencia de un objeto indirecto ha sido poco explorada. Incluso, es posible que la bivalencia o trivalencia gramatical del verbo sea un factor que propaga el uso de *empezar*, independientemente de las propiedades semánticas de los actantes.

Discusión y conclusiones

En la presente investigación de corpus, se han identificado varios patrones en el uso de los verbos *empezar* y *comenzar* en el mundo hispanohablante. Con respecto a la dimensión diatópica, se observan preferencias claras por uno de los dos verbos en distintas regiones. Según los hallazgos en el corpus más actualizado y extenso disponible, *empezar* predomina en los países andinos de Perú, Ecuador y Colombia, así como en Paraguay, Guatemala, Costa Rica y España. En contraste, *comenzar* es más común en el resto de Hispanoamérica, especialmente en la zona caribeña (Cuba, Venezuela y República Dominicana) y en el Cono Sur (Chile, Uruguay y Argentina). Un posible sesgo que podría afectar estos resultados es la distribución diafásica de los textos de cada país en el corpus. Sin embargo, la presencia de patrones similares en países de una misma región sugiere que este factor no tiene una influencia significativa. A pesar de ello, podría ser más relevante en los países con subcorpus más pequeños, como los centroamericanos, lo que podría explicar heterogeneidad observada en los datos de esa región.

En cuanto a la dimensión diafásica, se ha encontrado una clara relación entre los dos verbos analizados y distintos registros: *empezar* se emplea más en registros informales y orales, mientras que los registros formales y escritos favorecen el uso de *comenzar*. La combinación de ambos factores asociados con uno de los verbos propicia aún más su uso.

Ya que los corpus disponibles en Sketch Engine carecen de etiquetado semántico, no se ha podido indagar detalladamente en la semántica de estos verbos, pero sí se han identificado indicaciones de una preferencia por *empezar* cuando el sujeto presenta mayor grado de agentividad, pues dicho verbo se emplea más con los sujetos pronominales *yo* y *tú*, que prototípicamente representan actantes con un alto nivel de agentividad, así como con el objeto pronominal *lo*, indicio de que *empezar* es el verbo preferido en construcciones transitivas, cuyo sujeto tiende a ser más agentivo. Asimismo, se ha observado que la presencia de un objeto indirecto también favorece el uso de *empezar* frente a *comenzar*. Con la existencia de corpus etiquetados con funciones sintácticas y semánticas específicas, hay una excelente oportunidad para realizar análisis semánticos mucho más detallados de estos verbos en el futuro.

En lo que concierne al tipo de sinonimia que existe entre *empezar* y *comenzar*, no se ha identificado ningún factor ni combinación de factores que excluya uno de los verbos analizados, por lo que el grado de sinonimia entre ambos es alto. Por esta razón, se hallan cerca de ser sinónimos totales, y por lo menos cumplen con los requisitos para ser considerados sinónimos proposicionales en la tipología de Cruse (2000). Sin embargo, las pautas expuestas en los párrafos anteriores sugieren que distintas variables favorecen

uno de los dos. Por consiguiente, se puede afirmar que se trata de dos sinónimos tendencialmente estilísticos-semánticos, según la tipología de Apresian (1957), con una compleja interacción entre los factores estilísticos (tanto diatópicos como diafásicos) y semánticos.

Los resultados de este análisis son rudimentarios en varios aspectos, pero al mismo tiempo novedosos, puesto que incluyen hallazgos nunca antes publicados acerca del uso de los verbos inceptivos *empezar* y *comenzar*. Se espera que los datos presentados sirvan como fundamento para futuras investigaciones de este tema. Por ejemplo, para corroborar los patrones semánticos identificados en la presente investigación, se recomienda hacer uso de corpus con etiquetado semántico y sintáctico más detallado. Asimismo, las tendencias diafásicas podrán ser corroboradas en corpus de registros diferentes de los analizados aquí. Por último, cabe agregar que otros pares de sinónimos casi totales también podrán ser estudiados de manera similar, con el fin de descubrir los factores que los distingan.

Referencias bibliográficas

Ahearn, L. M. (2001). Language and agency. *Annual Review of Anthropology*, *30*, 109-137.

Apresian, Iu. D. (1957). Problema sinonima. *Voprosy Iazykoznaniia*, *6*(6), 84-88.

Biber, D. y Conrad, S. (2019). *Register*, *Genre*, *and Style* (2ª ed.). Cambridge: Cambridge University Press.

Bušta, J., Herman, O., Jakubíček, M., Krek, S. y Novak, B. (2017). JSI Newsfeed Corpus. *The 9th International Corpus Linguistics Conference*, University of Birmingham, 25 a 28 de julio de 2017. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper382.pdf

Cruse, D. A. (2000). *Meaning in Language: An Introduction to Semantics and Pragmatics.* Oxford: Oxford University Press.

Cruz Palacios, Y. (2013). La sinonimia y la antonimia: problemas en torno a su definición. *ISLAS*, *55*(172), 107-119.

Dixon, R. M. W. (1994). *Ergativity*. Cambridge: Cambridge University Press.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Calzada Pérez, M., de Macedo, L. D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darģis, R., Ring, O., van Heusden, R., Marx, M., Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57*, 415-448. https://doi.org/10.1007/s10579-021-09574-0

Espinal, M. T. y Mateu, J. (2020). Palabras y significado. En M. T. Espinal (Coord.), *Semántica* (pp. 59-109). Madrid: Ediciones Akal.

Foley, W. A. (1999). Information structure. En K. Brown y J. Miller (eds.), *Concise Encyclopedia of Grammatical Categories* (pp. 204-213). Ámsterdam: Elsevier.

García Murga, F. (2014). *Semántica*. Editorial Síntesis.

Hurford, J. R., Heasley, B. y Smith, M. B. (2007). *Semantics*: *a coursebook* (2ª ed.). Cambridge: Cambridge University Press.

Kibort, A. (2008). Transitivity. *Grammatical features*. http://www.grammaticalfeatures.net/features/transitivity.html

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. y Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*, 7-36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Rychlý, P., Smrž, P. y Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, 105-116. https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2004.pdf

Kittilä, S. (2000). Problems in defining a prototypical transitive sentence typologically. *Proceedings of the 14th Pacific Asia Conference on Language*, *Information and Computation*, 189-194. https://aclanthology.org/Y00-1019

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X*: *Papers*, Phuket, Tailandia, 79-86. https://aclanthology.org/2005.mtsummit-papers.11.pdf

Lexical Computing. (2024). Sketch Engine [software]. https://www.sketchengine.eu

Lison, P. y Tiedemann, J. (2016). OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (*LREC'16*), Portorož, Eslovenia, 923-929. http://www.lrec-conf.org/proceedings/lrec2016/pdf/947_Paper.pdf

Löbner, S. (2013). *Understanding Semantics* (2ª ed.). Londres y Nueva York: Routledge.

Lyons, J. (1995). *Linguistic Semantics*: *An Introduction*. Cambridge: Cambridge University Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3ª ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Næss, Å. (2007). *Prototypical Transitivity*. Ámsterdam: John Benjamins.

Palmer, M., Gildea, D. y Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, *31*(1), 71-106. https://doi.org/10.1162/0891201053630264

Project Gutenberg (s.f.). Project Gutenberg [sitio web]. https://www.gutenberg.org

Real Academia Española (2023). *Diccionario de la lengua española* (versión 23.7 en línea). https://dle.rae.es

Steinberger, R., Eisele, A., Klocek, S., Pilos, S. y Schlüter, P. (2013). DGT-TM: A freely available translation memory in 22 languages [sitio web]. https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en

Taulé, M., Martí, M. A. y Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC'08*), Marrakech, Marruecos. http://www.lrec-conf.org/proceedings/lrec2008/pdf/35_paper.pdf

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (*LREC'2012*), 2214-2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Valverde, P. (2024). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. TEISEL. *Tecnologías para la investigación en segundas lenguas*, *3*, 1-31. http://doi.org/10.1344/teisel.v3.42898

Notas

[1] Cita original en inglés: «Examples of perfect synonymy are hard to find, perhaps because there is little point in a dialect having two predicates with exactly the same sense».

[2] Cita original en inglés: «words whose semantic similarities are more salient than their differences».
[3] Este corpus no incluye un subcorpus de Panamá. Además, ninguno de los corpus contiene subcorpus de Guinea Ecuatorial, donde el español también es oficial, ni de países como Estados Unidos, Belice y Filipinas, los cuales cuentan con proporciones de hispanohablantes significativas.
[4] Los grados de oscuridad del gris en el mapa corresponden a una normalización de los porcentajes de uso de *empezar* donde X'_{min} = 22 % y X'_{max} = 82 %.
[5] Fuente: elaborado por el autor con base en un mapa en blanco creado por el usuario Cocoloi en Wikimedia Commons titulado «Latin America - First level political divisions.svg» (https://commons.wikimedia.org/wiki/File:Latin_America_-_First_level_political_divisions.svg). La imagen base es utilizada bajo la Licencia Creative Commons Attribution-Share Alike 3.0 Unported (https://creativecommons.org/licenses/by-sa/3.0/deed.es).
[6] A diferencia de corpus como AnCora (Taulé, Martí y Recasens, 2008) del español y PropBank (Palmer, Gildea y Kingsbury, 2005) del inglés.
[7] Esta información fue obtenida de la tabla de sujetos pronominales de la función «Word Sketch Difference». Si bien sería útil comparar estas cifras con las de otros sujetos pronominales, *yo* y *tú* son los únicos que aparecen en la lista generada por dicha función.
[8] Esta información fue obtenida de la tabla de objetos pronominales de la función «Word Sketch Difference». Ningún otro objeto pronominal es mostrado en la tabla.

Perfil del autor
Haakon S. Krohn es profesor de la Escuela de Filología, Lingüística y Literatura de la Universidad de Costa Rica. Se especializa tanto en el español como en las lenguas autóctonas de Costa Rica. Realiza sus investigaciones en las áreas de fonética, fonología, morfosintaxis, semántica y lexicografía.

Title
Empezar and *comenzar*: diatopic, diaphasic and semantic differences in digital corpora

Abstract
In this paper, I analyze the uses of the Spanish verbs *empezar* and *comenzar* according to different diatopic, diaphasic, and semantic variables in various digitalized corpora that are accessed using the Sketch Engine tool. Firstly, I discover that *comenzar* is more common than *empezar* in most Spanish-speaking countries, but that *empezar* is the preferred verb in the Andean countries of Peru, Ecuador and Colombia, as well as in Paraguay, Guatemala, Costa Rica and Spain. Secondly, I conclude that *comenzar* is used more in formal and written registers, whereas *empezar* is employed more in registers characterized by informality and orality. Finally, I find indications of *empezar* being the preferred verb when the subject presents a higher degree of agentivity.

Keywords
Spanish language, synonymy, corpus linguistics, Spanish dialectology, linguistic registers, semantic agentivity

タイトル
Empezarとcomenzar：デジタルコーパスにおける地理的、言語的、意味的差異

要旨
本論文は、Sketch　Engineツールを通じてアクセスされるさまざまなデジタル化されたコーパスを用いて、スペイン語の動詞「empezar」と「comenzar」の使用法を地理的、言語的、意味的変数に基づいて分析する。まず、多くのスペイン語圏国で「comenzar」が「empezar」よりも一般

的であるが、ペルー、エクアドル、コロンビアのアンデス地方、パラグアイ、グアテマラ、コスタリカ、スペインでは「empezar」を好むことが明らかになった。次に、「comenzar」は形式ばった文書でよく使用されるのに対し、「empezar」は非公式かつ口頭でのコミュニケーションで頻繁に用いられることが示された。最後に、行為主体がより高いエージェンシーを持つ場合には「empezar」が好まれる傾向がみられる。

キーワード
スペイン語、同義語、コーパス言語学、スペイン語方言学、言語登録、意味的エージェンシー

Lista de Tablas
Tabla 1. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2023* (*esTenTen23*), según país.
Tabla 2. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2018* (*esTenTen18*), según país.
Tabla 3. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2011* (*esTenTen11*), según país.
Tabla 4. Corpus en español disponibles en Sketch Engine que se asocian con un registro/género específico.
Tabla 5. Ocurrencias de los lemas *empezar* y *comenzar* en diferentes corpus.
Tabla 6. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2018* (esTenTen18), según género textual.
Tabla 7. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2023* (esTenTen23) con los sujetos pronominales *yo* y *tú*.
Tabla 8. Ocurrencias de los lemas *empezar* y *comenzar* en el *Spanish web corpus 2023* (esTenTen23) con los objetos pronominales *lo* y *le*.

Lista de Tablas
Figura 1. Mapa que muestra el uso del verbo *empezar* en comparación con *comenzar* en países hispanohablantes: cuanto más oscuro el color gris, mayor es la preferencia por *empezar*. Fuente: elaboración propia con base en los datos del corpus *esTenTen23*.