

Sobre el orden básico de palabras en euskera: un nuevo estudio de corpus¹

LUIS PASTOR SANTAMARÍA

Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU)

Abstract

Basque is classified as an SOV language, like Japanese or Korean. However, unlike these OV languages, which have rigid word order, Basque has free word order and it allows postverbal arguments. Some corpus studies have determined the basic word order is SOV (De Rijk, 1969; Aldezabal *et al.*, 2003) whereas others have claimed it to be SVO (Hidalgo, 1995a; Aske, 1997). I conducted a new corpus study of Basque to disentangle whether the most frequent word order (i.e., the basic word order) is SOV or SVO. Unlike previous Basque corpus studies, this new corpus study analyzes and compares different sources (press, magazine, books and TV scripts) in order to have a heterogeneous corpus. The results indicate that SOV is the most frequent word order in Basque. Furthermore, I compared the data of all corpus studies in Basque (including the data of this chapter) and, once again, SOV emerged as the most frequent word order. All together, these results show that SOV is the basic word order in Basque, contrary to claims by Hidalgo (1995a) and Aske (1997), who defend that SVO is the basic word order. I present a critical discussion of these two studies.

El orden básico de palabras es el orden que aparece en una oración transitiva declarativa con todos los argumentos expresados fonológicamente (Greenberg, 1963). Greenberg (1963) fue el pionero en utilizar el orden básico de palabras como criterio principal para clasificar tipológicamente las lenguas naturales. En su trabajo, titulado *Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements*, observó que de las posibilidades lógicas de la combinación del sujeto (S), el objeto (O) y el verbo (V), que son seis, los órdenes que se encuentran predominantemente en las lenguas del mundo son SOV, SVO y VSO. Posteriormente, Dryer (2013b) ha confirmado, mediante un corpus de 1377 lenguas, la prevalencia de estos tres órdenes principales observados por Greenberg (1963): los órdenes SOV (41%) y SVO (35%) son los más frecuentes, seguidos con a mucha distancia por el orden VSO (7%) (TABLA 1).

Tabla 1

Porcentajes de los órdenes básicos de palabras en el WALS (Dryer, 2013b).

Entre paréntesis aparecen las frecuencias absolutas

SOV		SVO		VSO		VOS		OVS		OSV		Sin orden		TOTAL	
41%	(565)	35,4%	(488)	6,9%	(95)	1,8%	(25)	0,8%	(11)	0,3%	(4)	13,7%	(189)	100%	(1377)

¹ Este trabajo ha recibido financiación del Gobierno Vasco (IT1169-19), MINECO (FFI2015-64183-P) y MICINN (FPI-2012 EEBB-I-13-06977).

El orden básico de palabras se asocia con dos propiedades: es el orden más frecuente y el menos marcado morfológica y pragmáticamente (Hawkins, 1983; Comrie, 1989; Dryer, 1995; Whaley, 1997; Dryer, 2013a, entre otros), aunque como afirma Greenberg (1966) la propiedad de ser el orden menos marcado está correlacionado con ser el orden más frecuente. Por tanto, podemos decir que ser el orden más frecuente es la propiedad principal del orden básico de palabras en la oración. Así, de todos los órdenes posibles que puede tener una oración en una lengua, el más frecuente será su orden básico de palabras. Por ejemplo, el castellano puede hacer uso de diferentes órdenes de palabras dado que tiene cierta libertad en la distribución de los argumentos; sin embargo, de entre todos ellos el orden básico SVO es significativamente el más frecuente (92%) (López, 1997).

En este artículo trato sobre la frecuencia del orden de palabras de las oraciones en euskera. En la siguiente sección 1.1, presento y discuto los diferentes estudios de corpus que he llevado a cabo en euskera sobre las frecuencias de los órdenes de palabras y trato de explicar por qué dichos estudios observan diferencias en los resultados que obtienen para la distribución de los órdenes de palabras. En la sección 1.2, presento un nuevo estudio de corpus en euskera con la intención de disipar la divergencia de los resultados de los estudios de corpus anteriores, y analizo estadísticamente las frecuencias de los órdenes de palabras. El artículo termina con la discusión (sección 1.3) de los resultados y las conclusiones (sección 1.4).

1. Revisión de los estudios de corpus sobre el orden de palabras en euskera

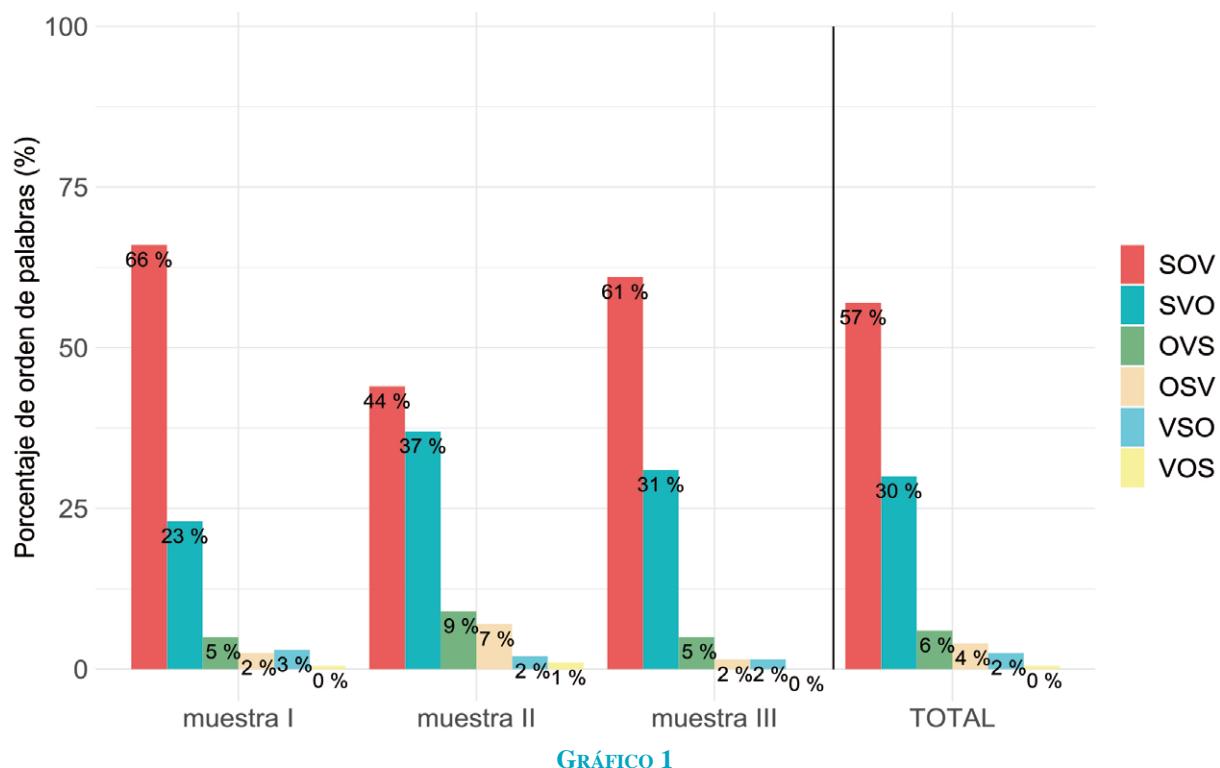
Como he comentado en la introducción, el orden básico o canónico se caracteriza por ser el orden de palabras que se usa con mayor frecuencia. Si el euskera es una lengua OV, el orden de palabras SOV debe de ser el más frecuente. Sin embargo, los estudios de corpus existentes reflejan frecuencias opuestas sobre el orden de constituyentes en las oraciones: unos muestran que SOV es el orden más frecuente (De Rijk, 1969; Aldezabal *et al.*, 2003); y otros que no, porque sería aventajado por el orden SVO (Hidalgo, 1995a, 1995b; Aske, 1997).

De Rijk (1969) es el primero arguir que el orden básico de palabras en euskera es SOV, y entre los criterios que aporta, uno es ser el orden más frecuentemente usado. Para examinar la frecuencia de los órdenes de palabras posibles en euskera De Rijk (1969) lleva a cabo un estudio pionero de corpus escrito en el que solo tiene en cuenta las oraciones declarativas transitivas en las que aparecen expresados el sujeto (S), el objeto (O) y el verbo (V). Del conjunto total de oraciones transitivas, De Rijk (1969) deja a un lado las transitivas interrogativas y negativas, las transitivas que no tienen el verbo conjugado, y transitivas en las que el sujeto o el objeto son oraciones subordinadas (salvo cuando la subordinada es una oración relativa que modifica el sujeto o el objeto), y las transitivas directas como «*Bihar jatera joango naiz» esan zuen mutilak*» [Mañana iré a comer» dijo el chico] e indirectas como «*Mutilak esan zuen bihar joango zela jatera*» [El chicho dijo que mañana iría a comer]. El corpus está compuesto por tres muestras: La muestra I consiste en textos folklóricos recogidos por José Miguel Barandiaran y contados por hablantes de Gipuzkoa y Bizkaia entre los años 1920-1936; la muestra II está compuesta por pequeñas obras teatrales escritas por Nemesio Echániz; y la muestra III la componen las obras *Mateo Falcone* y *Oillasko Iturri* de Mérimée (traducidas del francés por Nemesio Echániz). Los resultados de su cómputo muestran que de los seis posibles órdenes de palabras que se pueden utilizar en euskera SOV es el orden predominante en las tres muestras (57% en total. Desglosado por muestra: I: 66%; II: 44%; y III: 61%), seguido por SVO (Total: 30% = I: 23%; II: 37%; y III: 31%) (TABLA 2, GRÁFICO 1).

Tabla 2

Frecuencias absolutas de los órdenes de palabras en euskera
en el corpus de De Rijk (1969)

	I	II	III	TOTAL
SOV	138	80	41	259
SVO	48	67	21	136
OVS	11	17	3	31
OSV	5	13	1	19
VSO	6	4	1	11
VOS	1	2	0	3
TOTAL	209	183	67	459

**GRÁFICO 1**

**Porcentajes de los órdenes de palabras en euskera en el corpus de De Rijk (1969),
por muestras (I, II y III) y en total**

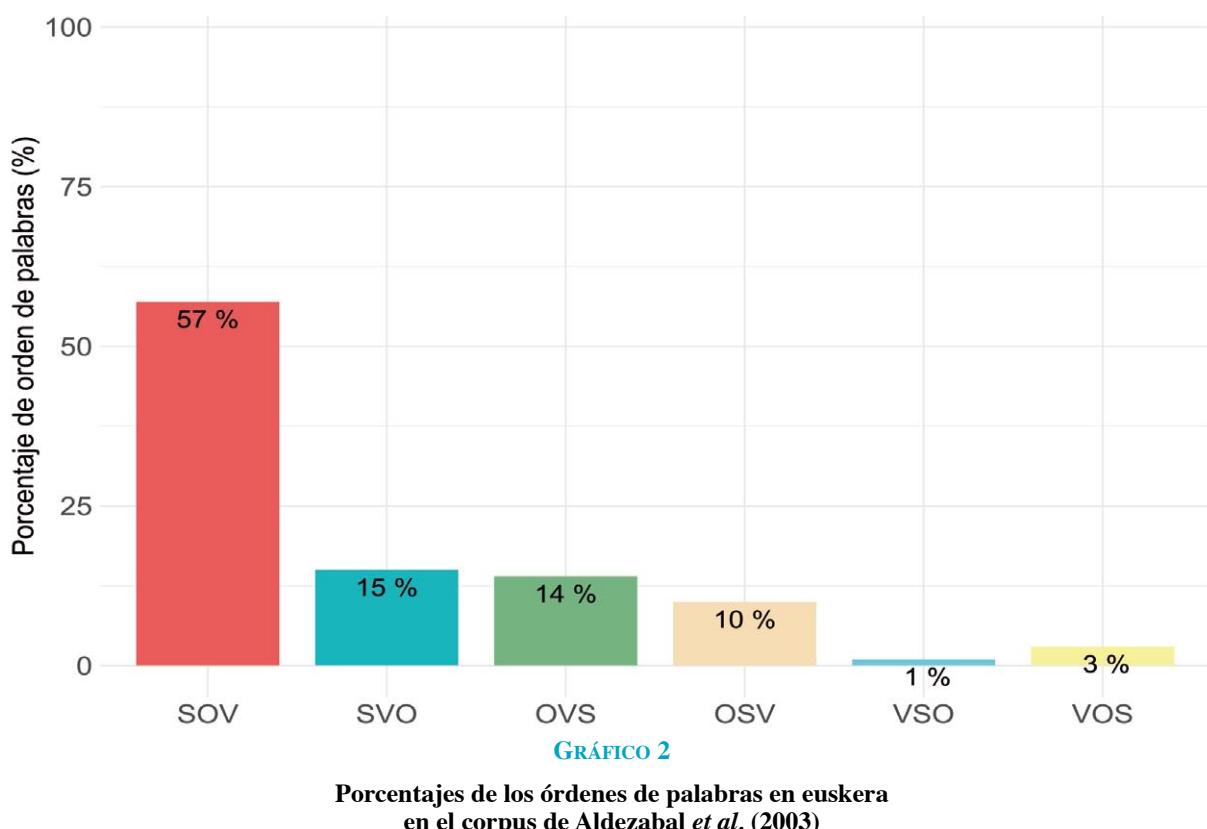
De estos datos, De Rijk (1969) concluye que, pese a la libertad de orden de palabras en la oración, el euskera sí tiene un orden básico o canónico: el orden SOV. Aldezabal *et al.* (2003) encuentran la misma distribución encontrada por De Rijk (1969) mediante un estudio de corpus de mayor tamaño. Este corpus está compuesto por 5.639 oraciones de artículos periodísticos del periódico en euskera *Euskaldunon Egunkaria*, entre los años 1999-2000. Las oraciones han sido eti-

quetadas de manera automática mediante un analizador sintáctico computacional según el orden del sujeto (S), el objeto (O) y el verbo (V) en las oraciones. De total de oraciones etiquetadas, solo en 512 oraciones aparecen los tres argumentos expresados. Los resultados muestran que SOV es el orden de palabras que aparece con mayor frecuencia (56,8%), seguido de SVO (14,8%) y OVS (13,8%) (TABLA 3, GRÁFICO 2).

Tabla 3

Frecuencias absolutas de los órdenes de palabras en euskera
en el corpus de Aldezabal *et al.* (2003)

	TOTAL
SOV	291
SVO	76
OVS	71
OSV	51
VSO	6
VOS	17
TOTAL	512



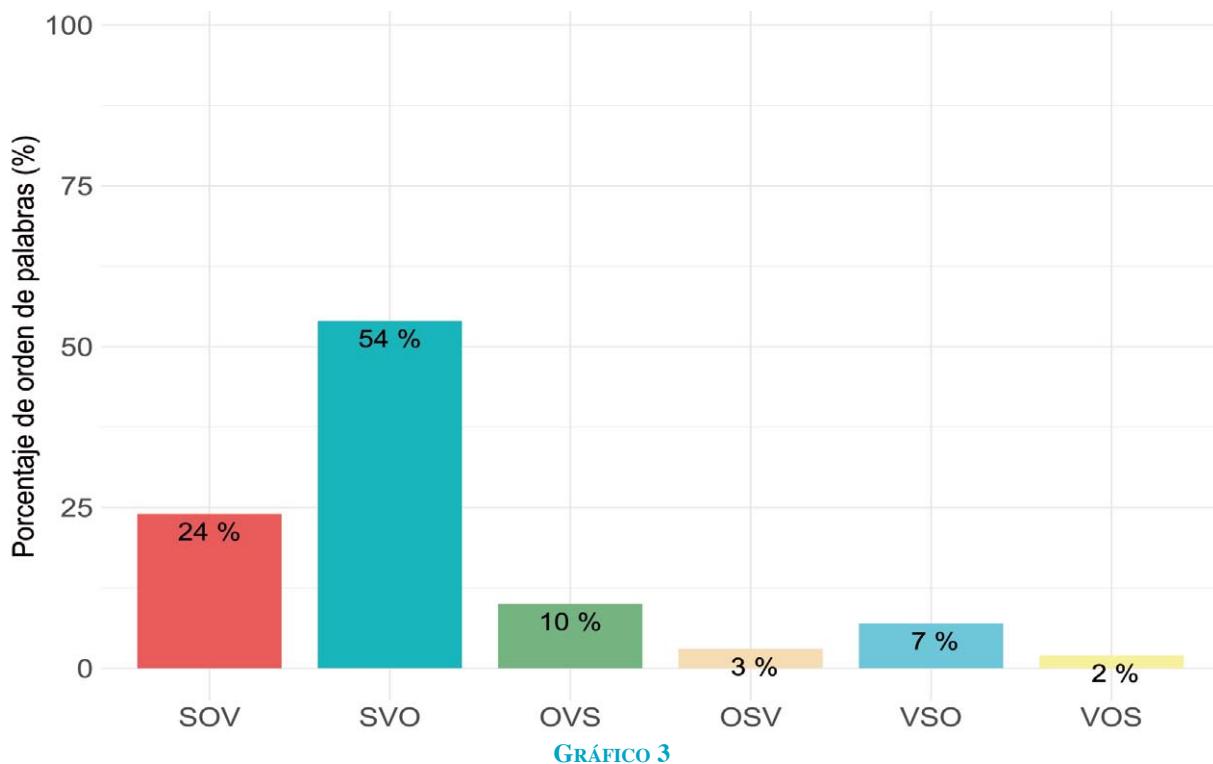
Como vemos, la distribución obtenida por Aldezabal *et al.* (2003) es similar a la de De Rijk (1969): SOV es el orden de palabras más frecuente en euskera y, por tanto, es el orden básico de palabras.

Hidalgo (1995a, 1995b) y Aske (1997), sin embargo, sostienen que el orden básico de palabras en euskera es SVO. Hidalgo (1995a, 1995b) lleva a cabo un estudio con un corpus escrito de euskera de mayor tamaño que los anteriores. Aunque no detalla el número total de oraciones que componen su corpus, sí puede calcularse el total de oraciones que usa para la distribución de los órdenes de palabras. Su corpus consta de oraciones obtenidas de 19 textos de diferentes autores de entre los siglos XVI-XX. Para el cómputo de oraciones Hidalgo excluyó menos tipos de oraciones de las requeridas. Solamente excluyó las oraciones sin verbo, con verbo no conjugado, sin complementos, oraciones interrogativas, imperativas y negativas. No excluyó las oraciones ditransitivas e intransitivas, que constituyen una parte significativa de las oraciones a computar. Sus resultados muestran que SVO (54%) es el orden de palabras más frecuencia en euskera, seguido por el orden SOV (24%) (TABLA 4, GRÁFICO 3). Hidalgo (1995b), basándose en los datos, concluye que es un error decir que el euskera sea una lengua SOV.

Tabla 4

Frecuencias absolutas de los órdenes de palabras en euskera
sumando todos los corpus de Hidalgo (1995a, 1995b)

	TOTAL
SOV	482
SVO	1.081
OVS	204
OSV	70
VSO	136
VOS	40
TOTAL	2.013



Porcentajes de los órdenes de palabras en euskera en el corpus de Hidalgo (1995a, 1995b)

Al igual que Hidalgo (1995a, 1995b), Aske (1997) encuentra que el orden SVO es el más frecuente en euskera. Mediante dos estudios de corpus, uno oral y otro escrito, compara únicamente la frecuencia de los órdenes de palabras SOV, SVO, VSO y VOS en euskera. No reporta la frecuencia de los ordenes OSV y OVS. El corpus oral consta de relatos de 46 participantes (niños y adultos) de dos películas mudas *Pearn Film* y *Modern Times* y el corpus escrito constaba del primer artículo de las novelas en euskera *Behi euskaldun baten memoriak* y *Kuba triste dago*. Aske concluye que, en conjunto, SVO es el orden de palabras más frecuente (61%) (TABLA 5, GRÁFICO 4). Separándolos por tipo, el corpus escrito muestra esta misma predominación del orden SVO (66%), pero en el corpus oral no hay diferencias entre los órdenes SOV (50%) y SVO (50%). Aske (1997) sugiere que esta preferencia por SVO puede deberse a la influencia del castellano de los participantes.

Tabla 5

Frecuencias absolutas de los órdenes de palabras en euskera en los corpus oral y escrito (y la suma de estos) de Aske (1997)

	Corpus oral	corpus escrito	TOTAL
SOV	12	14	26
SVO	12	46	57
VSO	0	9	9
VOS	0	1	1
TOTAL	24	70	94

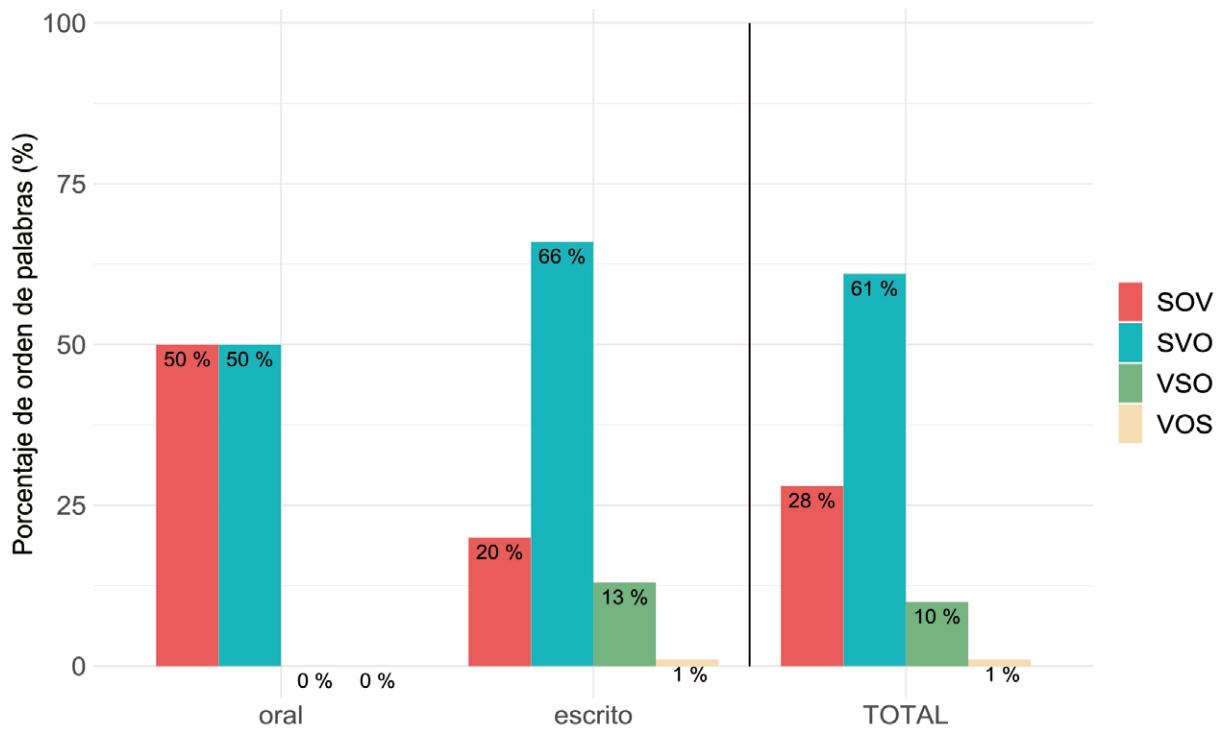


GRÁFICO 4

Porcentajes de los órdenes de palabras en euskera en los corpus oral, escrito y la suma de estos (Total) de Aske (1997)

Así pues, las distribuciones observadas por Hidalgo (1995a, 1995b) y Aske (1997) sugerirían que SVO es el orden básico en euskera, en contra de la distribución encontrada por De Rijk (1969) y Aldezabal *et al.* (2003), quienes observan que SOV es el orden más frecuente. Esta discrepancia en las frecuencias encontradas se debe a la metodología utilizada por Hidalgo (1995a, 1995b) y Aske (1997) frente a los otros dos estudios. Mientras que De Rijk (1969) y Aldezabal *et al.* (2003) etiquetan correctamente como objeto (O) solo los objetos directos, Hidalgo (1995a, 1995b) y Aske (1997) incluyen también en esta categoría cualquier otro tipo de complemento verbal (e.g., objetos indirectos, complementos circunstanciales, etc.), desvirtuando así el cómputo de la categoría O. Esta cuestión la abordo con más detalle en la discusión (sección 1.3).

A continuación, presento el diseño y los resultados de un nuevo estudio de corpus escrito en euskera que he llevado a cabo. Este nuevo corpus es más heterogéneo que los anteriores, ya que incluye textos de diferentes géneros textuales. A su vez, comparo los resultados obtenidos con los de los estudios previos y hago un análisis de todos los datos para determinar cuál es la frecuencia de uso de los posibles órdenes de palabras en euskera.

2. Estudio de corpus: orden de palabras en euskera

En el presente estudio de corpus examino la frecuencia de distribución de los seis posibles órdenes de palabras en euskera. Los propósitos de este nuevo estudio de corpus en euskera son dos: por un lado, usar un corpus amplio, similar en tamaño a los de Hidalgo (1995a, 1995b) y Aldezabal *et al.* (2003), pero que incluye diferentes géneros para ser lo más representativo posible

de los diferentes estilos discursivos en euskera; por otro lado, analizar las frecuencias encontradas en el corpus mediante análisis estadísticos para testear la significatividad de las diferencias observadas. Los estudios de corpus en euskera previos utilizan estadísticas descriptivas básicas, i.e., la frecuencia de aparición de un orden de palabras en el corpus; pero no analizan si las frecuencias observadas son simplemente debidas al azar o no.

2.1. *Materiales*

El corpus escrito de euskera utilizado en este artículo consta de 4000 oraciones. He obtenido las oraciones del corpus *EPG-Ereduzko Prosa Gaur* (Sarasola, Salaburu, Landa y Zabala, 2009) y he seleccionado diferentes géneros para tener una muestra heterogénea. El criterio de selección ha sido el siguiente: 1750 oraciones del periódico *Berria*, 1300 oraciones de diferentes libros, 300 oraciones de los guiones de la serie de televisión *Goenkale*. A estas he añadido 600 oraciones de la revista de divulgación *Elhuyar* y 50 oraciones de la revista científica *Uztaro*. Con la intención de tener un corpus aún más heterogéneo, he obtenido las oraciones de diferentes subapartados en cada género (salvo las oraciones de los guiones de la serie televisiva *Goenkale* y la revista científica *Uztaro*), y con un número de oraciones igual en cada una de ellos. En el periódico de lengua vasca *Berria* he utilizado las secciones de Economía, Sociedad, Mundo, Deportes, Cultura, Política y Nacional ($250 \text{ oraciones} \times 7 \text{ secciones} = 1.750 \text{ oraciones}$). En libros, he seleccionado libros de cuatro géneros diferentes: Comedia, Misterio, Histórica y Ensayo (Noficción) ($325 \text{ oraciones} \times 4 \text{ géneros} = 1.300 \text{ oraciones}$). Y en la revista divulgativa *Elhuyar* las secciones Historia, Cultura, Naturaleza, Salud, Tecnología y Ciencia ($100 \text{ oraciones} \times 6 \text{ secciones} = 600 \text{ oraciones}$).

2.2. *Procedimiento*

He etiquetado las oraciones transitivas manualmente y las he clasificado según el orden de palabras, i.e., teniendo en cuenta el orden lineal en el que aparecen el sujeto (S), el objeto directo (O) y el verbo (V):

- | | | |
|-----|--|-----------------|
| (1) | a. <u>Nik_[S]</u> <u>albaniarrak_[O]</u> <u>defenditu_[V]</u> | SOV [libros] |
| | «Yo he defendido a los albaneses» | |
| | b. <u>Gaizkak_[S]</u> <u>irekitzen_[V]</u> <u>atea_[O]</u> | SVO [guiones] |
| | «Gaizka abre la puerta» | |
| | c. <u>Lana_[O]</u> <u>sei ikerketa-taldek_[S]</u> <u>egin dute_[V]</u> | OSV [revista] |
| | «El trabajo lo han hecho seis grupos de investigación» | |
| | d. <u>Lau txanda_[O]</u> <u>aurreikusi ditu_[V]</u> <u>batzordeak_[S]</u> | OVS [periódico] |
| | «Cuatro turnos ha previsto el comité» | |
| | e. <u>Berehala irentsi zuen_[V]</u> <u>amua_[O]</u> <u>katxaloteak_[S]</u> | VOS [libros] |
| | «En seguida se ha tragado el anzuelo el cachalote» | |
| | f. <u>Eugin jaso zuen_[V]</u> <u>Azkuek_[S]</u> <u>kanta hau_[O]</u> | VSO [periódico] |
| | «En Eugi recogió Azkue esta canción» | |

Del total de oraciones que componen el corpus, solo he tenido en cuenta para el análisis estadístico las 1.054 oraciones transitivas declarativas, descartando las oraciones intransitivas, las di-

transitivas y las transitivas negativas, interrogativas, con subordinadas como objeto (objetos CP) y aquellas con argumentos omitidos (NP omitidos) (TABLA 6).

Tabla 6

Distribución del tipo de oraciones del corpus escrito de euskera

Intransitivas	Transitivas					Ditransitivas	TOTAL
989	declarativas	negativas	interrogativas	objetos CP	NP omitidos	122	4.000
	1.054	80	21	233	1.501		

A la hora de analizar los datos del corpus he usado los análisis estadísticos prueba de bondad de ajuste chi-cuadrado (*chi-square goodness of fit test*) y el modelo de regresión logística multinomial. La prueba de bondad de ajuste chi-cuadrado la he usado para analizar si la distribución de la frecuencia de los órdenes de palabras en el corpus se debe al azar o no. La regresión logística multinomial, por su parte, la he utilizado para analizar si el tipo de género del corpus influye en la frecuencia de uso de los diferentes órdenes de palabras. Los análisis estadísticos los he computado mediante el programa estadístico R (R Core Team, 2017) y usando el paquete *polytomous* (Arppe, 2013). He tomado como nivel de referencia la media de las medias de los cuatro géneros (periódico, libros, revistas y guiones), dado que el *intercept* es una media no ponderada, i.e., las variables tienen diferentes frecuencias. Los resultados los he considerado significativos a un nivel $p < .05$. Los gráficos han sido realizados con el paquete *ggplot2* (Wickham, 2009).

2.3. Resultados

La TABLA 7 muestra la clasificación de las 1054 oraciones transitivas que componen el corpus, según el orden de palabras y el tipo de género. De acuerdo con la prueba de bondad de ajuste chi-cuadrado (*chi-square goodness of fit test*) la distribución de las frecuencias de los seis posibles órdenes de palabras en euskera no es idéntica, i.e., no son igual de frecuentes ($\chi^2 (5, N = 1054) = 1249.9, p < .001, V = 0.48$): el orden que aparece significativamente con mayor frecuencia es SOV (52%), doblando la frecuencia de SVO (26%), que es el segundo orden más frecuente. Esta misma frecuencia mayoritaria de SOV se observa también en cada género: periódico (54%: $\chi^2 (5, N = 582) = 726.16, p < .001, V = 0.49$), libros (41%: $\chi^2 (5, N = 232) = 233.98, p < .001, V = 0.44$), revistas (50%: $\chi^2 (5, N = 151) = 175.42, p < .001, V = 0.48$) y guiones (69%: $\chi^2 (5, N = 89) = 181.81, p < .001, V = 0.63$).

TABLA 7

Distribución en euskera de las oraciones transitivas según el orden de palabras y el tipo de género.

Entre paréntesis aparecen las frecuencias absolutas

	periódico	libros	revista	guiones	TOTAL
SOV	54% (314)	40% (94)	50% (76)	69% (61)	52% (545)
SVO	23% (133)	38% (88)	26% (39)	15% (13)	26% (273)
OSV	17% (101)	16% (37)	20% (30)	12% (11)	17% (179)
OVS	3% (15)	1% (2)	2% (4)	2% (2)	2% (23)
VSO	2% (10)	3% (7)	1% (1)	1% (1)	2% (19)
VOS	1% (9)	2% (4)	1% (1)	1% (1)	1% (15)
TOTAL	100% (582)	100% (232)	100% (151)	100% (89)	100% (1054)

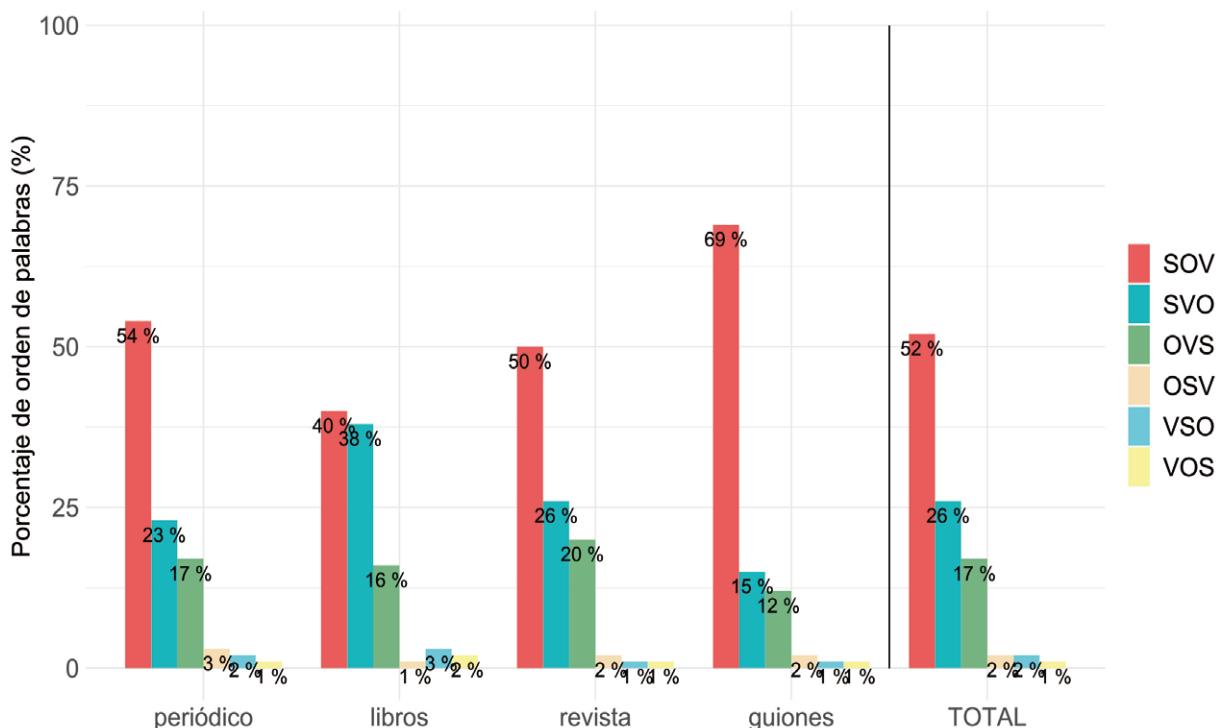


GRÁFICO 5

Porcentajes de los órdenes de palabras en euskera
en los diferentes géneros (periódico, libros, revista, guiones) y la suma de estos (Total)

Como vemos en el GRÁFICO 5, el orden SOV es el más frecuente en general y en todos los géneros; pero su frecuencia es mayor en los guiones (69%) que en el resto de géneros (periódico: 54%; libros: 41%; revistas: 50%). Para analizar si esta diferencia de frecuencia de SOV entre los cuatro géneros es significativa he llevado a cabo una prueba chi-cuadrado (*Pearson's chi-square test*), comparando su frecuencia con la de SVO en los cuatro géneros, que es el siguiente orden

más frecuente. Dado que los subcorpus de cada género son de diferente tamaño no se refleja con precisión las frecuencias relativas de ambos órdenes en cada género; por ello, he normalizado la frecuencia de ambos órdenes en cada género y he calculado sus frecuencias por 1.000. La prueba chi-cuadrado muestra que la frecuencia de uso de SOV es significativamente similar en cada género [$\chi^2 (3, N = 4000) = 220.91, p < .001, V = 0.24$]. Aún así, el modelo de regresión logística multinomial muestra que el género guiones influye significativamente en que SOV se utilice con mayor frecuencia comparado con el resto de géneros [$\beta = 0.6371, p < .001$, odds ratio = 1.89]: la probabilidad de uso de SOV es 1,9 veces mayor en los guiones que en el resto de géneros. El orden SVO, por el contrario, se ve favorecido por el género libros [$\beta = 0.64, p < .001$, odds ratio = 1.89], donde es 1,9 veces más probable que se use con mayor frecuencia que en el resto de géneros (TABLA 8).

Tabla 8

Resultados del modelo de regresión logística para el orden de palabras en euskera según el tipo de género.

El Intercept es la media de las medias de los cuatro géneros (periódico, libros, revistas, guiones)

Orden de palabras por género-Coeficientes de estimate:						
	SOV	SVO	OVS	OSV	VSO	VOS
(Intercept)	(0.1416)	-1.132	-1.644	-3.939	-4.251	-4.421
periódico	(0.01682)	(-0.08421)	(0.08331)	(0.3062)	(0.2046)	(0.2675)
libros	-0.5255	0.64	(-0.01802)	(-0.8064)	(0.781)	(0.3781)
revistas	(-0.1283)	(0.07753)	(0.2495)	(0.3344)	(-0.7595)	(-0.5895)
guiones	0.6371	-0.6333	(-0.3148)	(0.1658)	(-0.2262)	(-0.05617)
<i>p</i> -value de los coeficientes						
(Intercept)	0.078	0.001***	0.001***	0.001***	0.001***	0.001***
periódico	0.866	0.483	0.533	0.373	0.641	0.555
libros	0.001***	0.001***	0.914	0.164	0.092	0.472
revistas	0.361	0.636	0.167	0.468	0.344	0.465
guiones	0.001***	0.007**	0.212	0.776	0.779	0.945
R2.likelihood: 0.015						

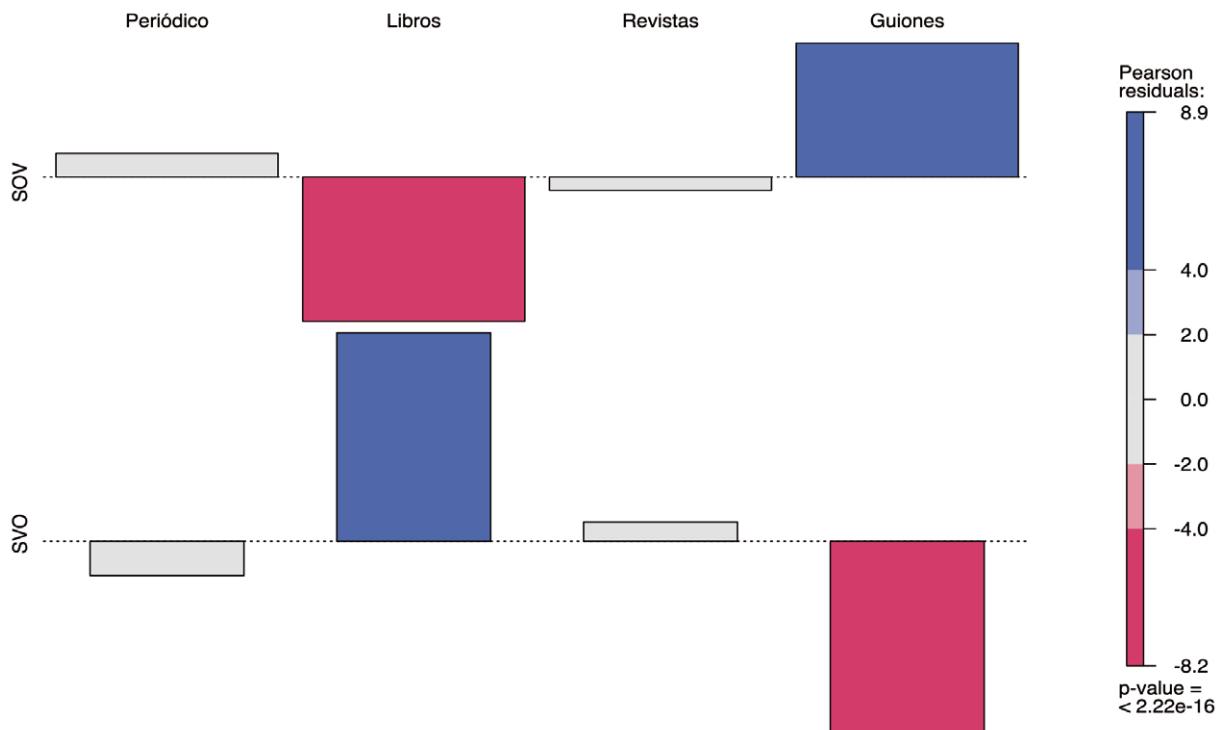


GRÁFICO 6

Usos de los SOV y SVO en los cuatro géneros: asociación de residuales.
Las barras azules indican residuales positivos y las barras rojas residuales negativos

El GRÁFICO 6 enseña que en el género guiones el orden SOV es relativamente más frecuente al resto de los géneros, mientras que en el género libros sucede lo mismo con el orden SVO. De este dato podemos concluir que cuanto más se acerca un corpus al registro oral mayor es la frecuencia de SOV. Por el contrario, en el género libros que puede argüirse es el que más se aleja del registro oral encontramos mayor frecuencia del orden SVO.

2.3.1. COMPARACIÓN CON LOS ESTUDIOS DE CORPUS PREVIOS

El GRÁFICO 7 muestra la distribución de la frecuencia de uso de los órdenes de palabras en euskera en los cinco estudios de corpus existentes (De Rijk (1969), Hidalgo (1995a, 1995b), Aske (1997) y Aldezabal *et al.* (2003); ver sección 2.3) y los resultados del presente estudio de corpus (ver sección 1.2.3). Como puede observarse, en todos los corpus SOV y SVO son los órdenes que aparecen con mayor frecuencia: SOV en los corpus de De Rijk (1969), Aldezabal *et al.* (2003) y el presente estudio de corpus; y SVO en los corpus de Hidalgo (1995a, 1995b) y Aske (1997).

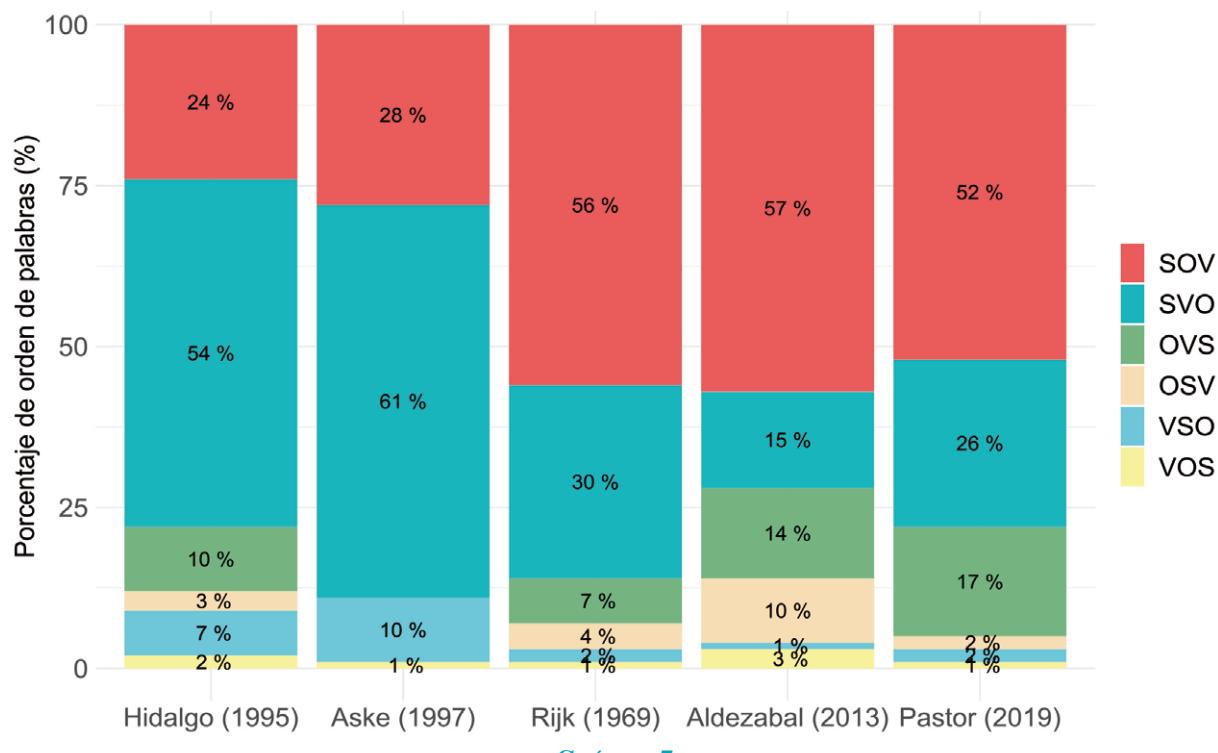


GRÁFICO 7

**Distribución de la frecuencia de uso de los órdenes de palabras en euskera en los diferentes estudios de corpus
(De Rijk, Hidalgo, Aske, Aldezabal, Pastor)**

He llevado a cabo una regresión logística multinomial para observar si el uso de SOV y SVO se ve favorecido por los diferentes estudios de corpus (TABLA 9). Al hacer el análisis estadístico he omitido los datos de Aske (1997) ya que no reporta las frecuencias de los órdenes OVS y OSV. El análisis de regresión logística multinomial muestra que SOV se ve favorecido significativamente en los corpus de De Rijk (1969) [$\beta = 0.3974, p < .001$, odds ratio = 1.49], Aldezabal *et al.* (2003) [$\beta = 0.4141, p < .001$, odds ratio = 1.51] y el presente corpus [$\beta = 0.2053, p < .001$, odds ratio = 1.23]: la probabilidad de SOV es 1,4 veces mayor en estos estudios de corpus que en el de Hidalgo (1995a, 1995b). Por el contrario, el orden SVO se ve favorecido en el corpus de Hidalgo (1995a, 1995b) [$\beta = 1.026, p < .001$, odds ratio = 2.79] y su probabilidad de uso es 2,8 veces mayor que en el resto de corpus.

Tabla 9

Resultados del modelo de regresión logística para el orden de palabras en euskera según el autor del corpus.

El Intercept es la media de las medias de los cuatro autores de corpus

(De Rijk, Hidalgo, Aldezabal, Pastor [esta tesis doctoral]).

Los coeficientes que aparecen entre paréntesis no son significativos (ver la tabla: p-value de los coeficientes)

Orden de palabras por autor del corpus-Coefficientes de <i>estimate</i> :						
	SOV	SVO	OVS	OSV	VSO	VOS
(Intercept)	-0.1389	-0.8778	-2.055	-3.118	-3.691	-4.133
De Rijk (1969)	0.3974	(0.01276)	-0.5697	(-0.02455)	(-0.01561)	-0.8907
Hidalgo (1995)	-1.017	1.026	(-0.1269)	(-0.2057)	1.067	(0.2347)
Aldezabal (2003)	0.4141	-0.8691	0.2291	0.9162	-0.7435	0.7618
Pastor [tesis]	0.2053	-0.1697	0.4675	-0.686	(-0.3074)	(-0.1058)
<i>p-value</i> de los coeficientes						
(Intercept)	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
De Rijk (1969)	0.001***	0.881	0.001***	0.897	0.952	0.045*
Hidalgo (1995)	0.001***	0.001***	0.120	0.103	0.001***	0.260
Aldezabal (2003)	0.001***	0.001***	0.037*	0.001***	0.021*	0.002**
Pastor [tesis]	0.001***	0.012*	0.001***	0.001***	0.156	0.677
R2.likelihood: 0.062						

Por último, he llevado a cabo un análisis con los datos de todos los corpus existentes en euskera para obtener una estimación de la frecuencia de los órdenes de palabras. La prueba de Kruskal-Wallis muestra que no hay diferencias significativas en la distribución de los diferentes órdenes de palabras, i.e., todos los corpus revelan una distribución similar en la frecuencia de uso de los órdenes de palabras ($H(3) = 0.201, p = .977$). El orden SOV es el más frecuente, seguido, en orden decreciente, de SVO, OVS, OSV y VOS (GRÁFICO 8).

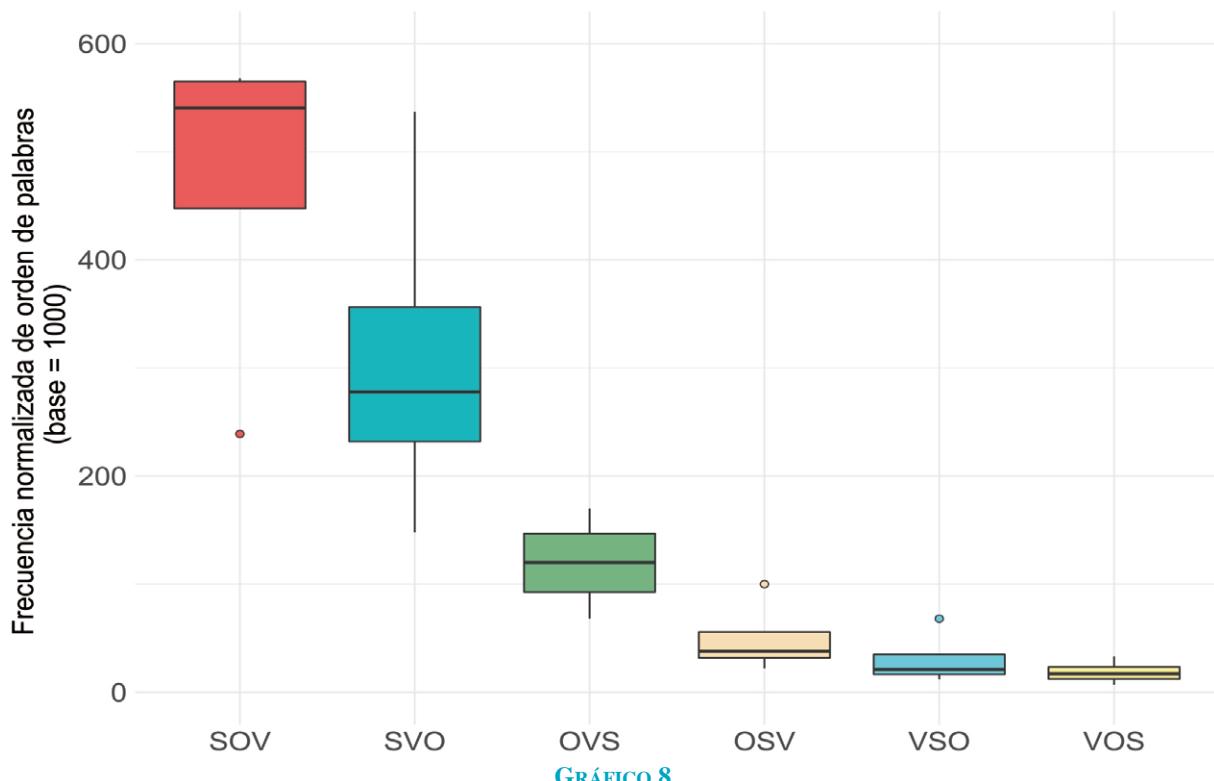


GRÁFICO 8

Distribución de la frecuencia de uso de los órdenes de palabras en euskera colapsando los estudios de corpus existentes (De Rijk, Hidalgo, Aldezabal, Pastor)

En resumen, los resultados de ese estudio de corpus revelan una preferencia por el orden básico SOV, siendo este el orden de palabras que aparece con mayor frecuencia en comparación con el resto de órdenes (SVO, OVS, OSV, VSO y VOS). Además, el orden SOV es también el orden más frecuente en cada uno de los géneros, aunque se ve más favorecido en los guiones de la serie de televisión *Goenkale*. En línea con esto, el análisis conjunto de los estudios de corpus existentes (De Rijk, 1969; Hidalgo, 1995a, 1995b; Aldezabal *et al.*, 2003) también muestran que SOV es el orden más frecuente en euskera.

3. Discusión

En este artículo he presentado un nuevo estudio de corpus escrito formado por diferentes géneros en euskera para explorar la distribución y frecuencia de uso de los seis posibles órdenes de palabras en euskera (SOV, SVO, OVS, OSV, VSO, VOS). Los resultados muestran que SOV es el orden dominante, i.e., el orden de palabras que se usa con mayor frecuencia, independientemente del género del texto. Estos resultados convergen con los estudios de corpus de De Rijk (1969) y Aldezabal *et al.* (2003), que también encuentran que SOV es el orden más frecuente en euskera. Sin embargo, estos resultados contrastan con los de Hidalgo (1995a, 1995b) y Aske (1997), que observan que SVO es el orden que aparece con mayor frecuencia seguido de SOV.

La discrepancia en la frecuencia entre los órdenes SOV y SVO entre los estudios de corpus de De Rijk (1969), Aldezabal *et al.* (2003) y esta tesis doctoral por un lado, y los de Hidalgo

(1995a, 1995b) y Aske (1997) por otro, se debe exclusivamente a los criterios de selección y etiquetado de las oraciones. Mientras que los primeros estudios solo tienen en cuenta oraciones transitivas declarativas afirmativas en las que ambos argumentos aparecen expresados, los segundos incluyen también oraciones intransitivas: «...en vez de observar las oraciones que muestran externamente solamente el Sujeto, Objeto y Verbo conjugado, tener en cuenta junto a estas todas las oraciones principales declarativas que junto al verbo conjugado muestran cualquier otro tipo de complemento (llamémosle a cada uno «X»), y no solo los elementos S y O.» (Hidalgo, 1995a:499). De esta manera, para Hidalgo (1995a, 1995b) y Aske (1997) las siguientes dos oraciones tienen el mismo orden de palabras (SOV): «Sarak_[S] liburua_[O] irakurri du» [Sara ha leído el libro] y «Sara_[S] etxean_[O] gelditu da» [Sara se ha quedado en casa]. Pero ambas oraciones no son iguales: la primera es una oración transitiva, porque el verbo «*irakurri*» pide dos argumentos: un sujeto (*Sarak*) y un objeto directo (*liburu*). La segunda oración, por el contrario, es una oración intransitiva porque el verbo «*gelditu*» solo pide un argumento —el sujeto (*Sara*)— de tal forma que «*etxean*» no es un objeto. Ejemplos similares se pueden encontrar en el corpus de Hidalgo (1995a, 1995b) (e.g., «*hori jin da goizan*_[O]» [Ese se ha ido a la mañana]). También etiquetan como oraciones transitivas, aunque son intransitivas, aquellas oraciones que tienen construcciones «nombre + egin» (e.g., «*baina nik alde egingo nuke*» [pero yo huiría] (Aske, 1997:927)). Las oraciones con construcciones formadas por «nombre + egin» llevan el sujeto en caso ergativo, como el sujeto de una oración transitiva; pero a pesar de ello, este tipo de construcción son inergativas, i.e., intransitivas (Laka, 1993).

En cuanto al etiquetaje de las oraciones transitivas, Hidalgo (1995a, 1995b) y Aske (1997) tampoco siguen el criterio marcado por De Rijk (1969), Aldezabal *et al.* (2003) y esta tesis doctoral, que solo computan aquellas oraciones transitivas que solo tienen objetos nominales (aunque pueden estar modificados por una oración relativa). Hidalgo (1995a, 1995b) y Aske (1997) incluyen dentro de su muestra oraciones transitivas en las que el objeto directo es una oración: «*Zuc erraiten duzu nic emaiten dudala*_[O]» [Tú dices que yo doy]. Estos autores también incluyen oraciones ditransitivas: «*Ama Virgiña erremediotakuak emanen digu osasona*» [La Virgen María de los remedios nos dará salud] (Hidalgo, 1995a:428) y «*berak adieraziko dizu aukerarik onena*» [él te indicará el mejor sitio] (Aske, 1997:926). Estas oraciones son ditransitivas, porque los verbos «*eman*» y «*adierazi*», en esos contextos, pide tres argumentos (sujeto, objeto directo e indirecto), aunque en ambas el objeto indirecto esté omitido.

Los estudios que hacen un etiquetado correcto de las oraciones de la muestra (De Rijk, 1969, Aldezabal *et al.*, 2003 y esta tesis) encuentran que el orden más frecuente es SOV; en los estudios con un etiquetado incorrecto de la muestras (Hidalgo, 1995a, 1995b y Aske, 1997), sin embargo, SVO es el orden más frecuente. Por tanto, podemos concluir y confirmar que el orden más frecuente en euskera es SOV y, por ende, el orden básico.

Además, este resultado converge con los de estudios experimentales de procesamiento oracional en euskera (Erdocia, Laka, Mestres-Misse y Rodriguez-Fornells, 2009; Erdocia, Laka y Rodriguez-Fornells, 2012; Ros, Santesteban, Fukumura y Laka, 2015). Estos estudios hallan que SOV es el orden de palabras que menor coste de procesamiento presenta en euskera. En comprensión, Erdocia *et al.* (2009) y Erdocia *et al.* (2012) observan que los hablantes de euskera muestran una negatividad N400, una positividad P600 y mayores tiempos de lectura en las oraciones con otros órdenes (SVO, OVS, OSV) en comparación con oraciones con el orden SOV. En producción, Ros *et al.* (2015) encuentran que los participantes prefieren producir el orden SOV frente al resto de órdenes, independientemente de la longitud del sujeto y el objeto. En línea con esto, es-

tudios con pacientes afásicos (Arantzeta, Webster, Laka, Martínez-Zabaleta y Howard, 2016 y Arantzeta *et al.*, 2017) también muestran una preferencia por el uso del orden SOV: las personas con afasia comprenden mejor las oraciones con el orden básico frente a las que no tiene el orden básico. Toda esta evidencia va en línea con los modelos *frequency-based accounts* (Hale, 2001; Levy, 2008), que predicen que las oraciones con órdenes más frecuentes son más fáciles de procesar que las menos frecuentes.

4. Conclusiones

El estudio de corpus de este artículo es el primero que analiza la frecuencia de los diferentes órdenes posibles en euskera en diferentes géneros. Puede considerarse también el estudio de corpus más amplio (en número de oraciones) en euskera etiquetado manualmente según la distribución del sujeto (S), objeto directo (O) y verbo (V) dados los errores de etiquetaje encontrados en el estudio de corpus de Hidalgo (1995a, 1995b), que era hasta la fecha el estudio de corpus de euskera más amplio conocido. Los resultados de este artículo sustentan la idea de que el orden de palabras más frecuente en euskera es SOV y por tanto es el orden básico, independientemente del tipo de texto y estilo.

Referencias

- Aldezabal, I., Aranzabe, M. J., Atutxa, A., Gojenola, K., Sarasola, K., y Zabala, I. (2003). *Hitz-hurrenkaren azterketa masiboa corpusean*.
- Arantzeta, M., Bastiaanse, R., Burchert, F., Wieling, M., Martínez-Zabaleta, M., y Laka, I. (2017). Eye-tracking the effect of word order in sentence comprehension in aphasia: evidence from Basque, a free word order ergative language. *Language, Cognition and Neuroscience*, 32(10), 1320-1343. doi: 10.1080/23273798.2017.1344715
- Arantzeta, M., Webster, J., Laka, I., Martínez-Zabaleta, M., y Howard, D. (2016). Cross-linguistic asymmetries in sentence comprehension deficits in bilingual Basque-Spanish aphasia. Evidence from eye-tracking and behavioural data. *Stem-, Spraak- en Taalpathologie*, 21(Suppl.), 28-31.
- Arppe, A. (2013). polytomous: Polytomous logistic regression for fixed and mixed effects [Computer software manual]. Disponible en <https://CRAN.R-project.org/package=polytomous>
- Aske, J. (1997). *Basque Word Order and Disorder Principles, Variation, and Prospects*. (PhD Dissertation), University of California.
- Comrie, B. (1989). *Language Universals and Linguistic Typology* (2 ed.). Chicago: The University of Chicago Press.
- De Rijk, R. P. G. (1969). Is Basque an SOV Language? *Fontes Linguae Vasconum*, 1, 319-351.
- Dryer, M. S. (1995). Frequency and pragmatically unmarked word order. In P. Downing & M. Noonan (Eds.), *Word order in discourse* (pp. 105-135). Amsterdam: John Benjamins.
- Dryer, M. S. (2013a). Determining Dominant Word Order. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
- Dryer, M. S. (2013b). Order of subject, object, and verb. In M. Haspelmath & M. S. Dryer (Eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library.
- Erdocia, K., Laka, I., Mestres-Misse, A., y Rodríguez-Fornells, A. (2009). Syntactic complexity and ambiguity resolution in a free word order language: behavioral and electrophysiological evidences from Basque. *Brain and Language*, 109(1), 1-17. doi: 10.1016/j.bandl.2008.12.003

- Erdocia, K., Laka, I., y Rodriguez-Fornells, A. (2012). Processing Verb Medial Word Orders in a Verb Final Language. In M. Lamers & P. de Swart (Eds.), *Case, Word Order and Prominence* (Vol. 40). Netherlands: Springer.
- Greenberg, J. (1966). *Language Universals (with special reference to feature hierarchies)*. The Hague: Mouton.
- Greenberg, J. H. (1963). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In J. H. Greenberg (Ed.), *Universals of Language* (pp. 40-70). Cambridge, MA: MIT Press.
- Hale, J. (2001). *A probabilistic earley parser as a psycholinguistic model*. Trabajo presentado en the Conference NameL, Conference LocationL.
- Hawkins, J. A. (1983). *Word Order Universals*. New York: Academic Press.
- Hidalgo, V. (1995a). *Hitzen ordena euskaraz*. (PhD Dissertation), Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU).
- Hidalgo, V. (1995b). Ohar estatistiko garrantzitsuak euskararen ordenaren inguru Euskara S.V.O.? *Fontes Linguae Vasconum*, 70, 401-420.
- Laka, I. (1993). Unergatives that assign ergative, unaccusatives that assign accusative. *MIT Working Papers in Linguistics*, 18(Papers on Case and Agreement), 149-172.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126-1177. doi: 10.1016/j.cognition.2007.05.006
- López, B. (1997). Aportaciones de la tipología lingüística a una gramática particular: el concepto de orden básico y su aplicación al castellano. *Verba. Anuario Galego de Filología*, 24, 45-82.
- R Core Team. (2017). R: A language and environment for statistical computing (Version 3.2.1). Vienna, Austria: R Foundation for Statistical Computing. Disponible en www.R-project.org
- Ros, I., Santesteban, M., Fukumura, K., y Laka, I. (2015). Aiming at shorter dependencies: the role of agreement morphology. *Language, Cognition and Neuroscience*, 30(9), 1156-1174. doi: 10.1080/23273798.2014.994009
- Sarasola, I., Salaburu, P., Landa, J., y Zabaleta, J. (2009). Ereduzko Prosa Gaur (EPG). Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU). Disponible en <http://www.ehu.eus/euskara-orria/euskara/ereduzkoa/>
- Whaley, L. J. (1997). *Introduction to typology: The unity and diversity of language*. Thousand Oaks, CA: SAGE.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.