Estrategias facilitadoras del procesamiento en lenguas OV-VO

Estudio comparativo de corpus

Luis Pastor

Directora: Dra. Itziar Laka

DEPARTAMENTO DE LINGÜÍSTICA Y ESTUDIOS VASCOS
UNIVERSIDAD DEL PAIS VASCO / EUSKAL HERRIKO UNIBERTSITATEA
(UPV/EHU)

Esta tesis doctoral se presenta para el grado de
DOCTOR EN LINGÜÍSTICA

2019

Agradecimientos

Dice el refrán que «es de bien nacido ser agradecido». Así que permitidme dedicar las siguientes líneas a agradecer a todas aquellas personas que han estado durante todo este periodo académico. Es posible que me olvide de algunas y pido, de antemano, perdón por ello.

Ante todo, mi eterno agradecimiento a mi profesora, mentora y directora Itziar Laka. Gracias por tu paciencia, tiempo, generosidad y consejos. Ha llegado a su fin nuestra relación doctoral, pero siempre quedaran cosas que nos unan y nos fascinen como la evolución del lenguaje.

Quiero agradecer especialmente a mi familia por todo su apoyo y cariño. A mis padres, mis hermanas, mi amama, mis tíos y primos, por estar siempre ahí. También a los que nos dejaron pero permanecen en nuestro recuerdo, que allá donde estén se sentirán orgullosos de nosotros.

A la *kuadrilla*, mi otra familia: Guillermo, Ane, Nagore, Miguel, Ana, Jonatan, Alex, Paula, Borja, Janire, y Oihan –*el benjamín*–. A Janire Val, por esa amistad que empezó con las simples expresiones "*y un jamón*" y "*orange juice*" –solo nosotros lo entendemos–, y que con los años ha crecido y seguirá creciendo –ya casi eres de la *kuadrilla*–. A Andrei Vázquez, gran amigo y gramático (o *ultracorreccionista*), por todos esos debates gramaticales y existenciales que hemos tenido y seguiremos teniendo –la ortografía castellana siempre nos tendrá enfrentados–. A Idoia Ros y Sergio López-Sancio, mi *kuadrilla predoc*, por toda vuestra ayuda, apoyo, sabiduría, consejos y, obviamente, todos esos momentos predoctorales vividos. No olvidaré que por vosotros migré a *R*, aunque no lo conseguisteis con *LaTeX*.

A Aitor Urrutia, amigo, mentor y compañero de desayunos dominicales. A Naroa Mendizabal, por esos *brunches* y por dar su vida para sacarme una foto frente al *Duomo* de Florencia. A Tamara Torralbo, que se ha convertido en una gran amiga. A Amaia Allende, amiga y gran compañera de viaje (y futuros viajes). A Paula Navarro, compañera doctoral. A Nuria Colás, que siempre se acuerda de mi cada vez que suena *Mi Gran Noche* de Raphael. A Inge Conde, que sigamos encontrándonos en *Aste Nagusia* –siempre ganaré yo–. A Jon Ander Gónzalez –el Jonan, para los amigos–, uno de los miembros fundadores de nuestros *retiros laredanos*. A Yuan Yiming –Simón, para los amigos–, el estudiante chino que maravillado por Bilbao se ha quedado con nosotros. A Milia Mayora, *ene Milia Lasturko*. Y a Marta Konpinska, Ainara Imaz, Ruth Milla, Miguel Ángel Maestre, Alberto

Santolaya, Zoe Nubla, José Luis Pérez, Iraide Talavera, Ana Carrera (Pedroche), Elena Benito, Iñigo Mancisidor, Inazio Álava, Idoia Ocerín, María Losada, Selene Varela… y a otros tantos amigos por estar siempre ahí.

A todos mis compañeros de universidad, pero en especial, a Ziortza Gandarias, porque hicimos un buen tándem académico –ella en literatura y yo en lingüística–. A mis compañeros y amigos de doctorado: Rubén Pérez, Aurora Troncoso, Marina Ortega, Laura Vela, Blanca Arias, Ainhoa Aizpurua y Asier Calzada, por esos cafés en los que desconectábamos de la dura vida doctoral. A Aránzazu López, por su "*frikismo*" compartido de *Doctor Who* y por haberme descubierto, tal vez, el mejor juego de cartas: *La fallera calavera*. A Cesar Benito, por sus charlas de pasillo y momentos de "*radio patio*" entre un *giputzi* y un *bilbaino (con diptongo)*. A mi grupo de amigos de Harvard: Eloi Grasset, José de León, Flora Noel, Thibaut Mihelich… y en especial a Cristina Sanz, por esos momentos en *Starbucks,* esa tradición postal y su enorme amistad –que nunca desaparezca *#HarvardTeam*–; y a Marta Llorente, compañera y amiga gastronómica durante mi estancia en Harvard –con ella tuve el honor de preparar mi primer *Thanksgiving* (americano)–.

Agradecer también a todos los miembros, actuales y pasados, del grupo de investigación *Gogo Elebiduna*. A Irene de la Cruz-Pavía, por ser la primera en acogerme cuando yo era un recién llegado y por todos sus consejos. A los *mosqueteros* Mikel Santesteban, Kepa Erdozia y Adam Zawiszweski, pero especialmente a Adam por darme cada año la oportunidad de mostrarles a sus alumnos de "Lingüística I" el fascinante mundo de la evolución del lenguaje. A Edurne Petrirena y Yolanda Acedo, porque sin su ayuda el grupo entraría en caos. Y a los pasados, presentes y futuros *predocs* del grupo: Ane Berro, Ane Odria, Paolo Lorusso, Nerea Egusquiza, Miren Urteaga, Gillen Martínez de la Hidalga, Itziar Orbegozo, Bea Gómez, Marta De Pedis, Noèlia Sanahuja… y a Victoria Cano, mi *favorita*, por haberme dejado ser su mentor académico en la sombra y, sobre todo, por su gran amistad –me debes una buena *paella valenciana*–.

A todos mis profesores, por haber compartido todo su conocimiento y su amor por la enseñanza. En especial, a Itziar Turrez, por haberme descubierto el fascinante mundo de la lingüística; y a Maria Polinsky, por haberme dado la oportunidad de profundizar mis conocimientos en la Universidad de Harvard. También a Elena Castroviejo, que sin haber sido mi profesora he aprendido con ella cosas sobre semántica y algo de catalán, aunque de igual modo ella ha aprendido algo de euskera.

Y por último agradecer a Raphael, Bunbury, Ludovico Einaudi, Giovanni Allevi, Bruno Bavota… porque su música ha sido la banda sonora de esta tesis doctoral, desde los tediosos momentos de etiquetado de los corpus hasta la escritura de la misma.

Esta tesis es en honor a la *Más Noble Orden del Gorrión.*

"Amistades que son ciertas nadie las puede turbar"

Miguel de Cervantes

Tabla de contenidos

Índice de gráficos

Índice de tablas

Índice de imágenes

Índice de mapas

Lista de abreviaturas

A-P	articulatorio-perceptual
ABS	absolutivo
ACC	acusativo
AUX	auxiliar
C-I	conceptual-intencional
CP	sintagma complementante (oración)
DAT	dativo
DLT	*Dependency Locality Theory*
ERG	ergativo
ERP	*Event Releated Potentials* (eventos potenciales evocados)
EU	unidad de energía
fMRI	*functional magnetic Magnetic Rresonance Imaging* (imágenes de resonancia magnética funcional)
IC	coste de integración
MiD	*Minimize Domains*
N-V	(ratio) nombre-verbo
NP	sintagma nominal
NP$_C$	sintagma nominal corto
NP$_L$	sintagma nominal largo
O	objeto (directo)
O-drop	objeto omitido
O$_{ANI}$	objeto animado
O$_C$	objeto corto
OD	objeto directo
OD$_C$	objeto directo corto
OD$_L$	objeto directo largo
OI	objeto indirecto
OI$_C$	objeto indirecto corto
OI$_L$	objeto indirecto largo
O$_{INA}$	objeto inanimado
O$_L$	objeto largo
OSV	objeto-sujeto-verbo
OV	objeto-verbo (lengua de núcleo final)
OV	objeto-verbo (omisión de sujeto en lenguas OV)
OVS	objeto-verbo-sujeto
P	preposición/posposición

PCD	*Phrasal Combination Domain*
PGCH	*Performance-Grammar Correspondence Hypothesis*
PL	plural
PoS	*part-of-speech* (categoría léxica)
PP	sintagma preposicional
PP$_C$	sintagma preposicional corto
PP$_L$	sintagma preposicional largo
PRES	presente
REL	relativo
S	sujeto
S-drop	sujeto omitido
S$_{ANI}$	sujeto animado
S$_C$	sujeto corto
SG	singular
S$_{INA}$	sujeto inanimado
S$_L$	sujeto largo
SO-drop	sujeto y objeto omitido
SOV	sujeto-objeto-verbo
SV	sujeto-verbo (omisión de objeto)
SVO	sujeto-verbo-objeto
V	verbo
VO	verbo-objeto (lengua de núcleo inicial)
VO	verbo-objeto (omisión de sujeto en lenguas VO)
VOS	verbo-objeto-sujeto
VP	sintagma verbal
VSO	verbo-sujeto-objeto

Capítulo 1

Los corpus y el estudio del lenguaje

La psicolingüística estudia cómo los hablantes de las lenguas procesan el lenguaje mediante estudios experimentales. Para ello se emplean diferentes métodos para investigar los procesos relacionados con el lenguaje durante su producción y comprensión. Uno de ellos es la lectura auto-dirigida (*self-paced reading*) en la que los participantes leen oraciones palabra por palabra, pulsando un botón para continuar con la siguiente palabra y así medir el tiempo (i.e., tiempos de reacción) que tardan en leer cada una de las palabras. Otro método que también registra tiempos de reacción es el seguimiento de movimientos oculares (*eye-tracking*) que registra los movimientos de los ojos, la duración de fijaciones y el número de regresiones que los participantes hacen mientras leen oraciones. Un último grupo de métodos lo forman las técnicas neurofisiológicas que miden la actividad neuronal ante estímulos lingüísticos, como los potenciales evocados (*ERP*s), que registra la actividad eléctrica neuronal; y las imágenes de resonancia magnética funcionales (*fMRI*), que mide el aumento del flujo sanguíneo en una región cerebral. Existen otros métodos que se utilizan a la hora de llevar a cabo los experimentos, pero he mencionado los más utilizados. Y en línea con ello, en los últimos años hay un nuevo método no experimental que está que se está utilizando junto a los métodos antes mencionados: el uso de *corpus*.

En las tres primeras secciones, defino qué es un corpus (sección 1.1), detallo brevemente algunos de los tipos de corpus que se pueden encontrar (sección 1.2) y trato la cuestión de qué tamaño ha de tener un corpus (sección 1.3). A continuación, en las siguientes secciones, muestro que los corpus pueden contribuir en la investigación sobre el lenguaje, en especial en psicolingüística (sección 1.4), y como para ello es inevitable el uso de la estadística (sección 1.5). Por último, en la sección 1.6, describo los objetivos de esta tesis doctoral y las principales cuestiones de investigación abordadas en cada capítulo.

1.1 Corpus

Un *corpus* es una forma específica de datos lingüísticos que está formado por un conjunto de textos escritos (y transcripciones del habla) que representa una muestra de la lengua. En general, los corpus se caracterizan por (a) ser legible para los ordenadores, (b) ser representativo de la lengua, (c) ser equilibrado y (d) que los textos que lo forman hayan sido producidos en situaciones comunicativas naturales (Tognini-Bonelli, 2001; Gilquin y Gries, 2009; Gries, 2016; Desagulier, 2017; Gries y Berez, 2017). El primer rasgo (a) se refiere a que los textos que componen el corpus han de estar en un formato que todos los ordenadores sean capaces de leer, modificar y procesar, como los formatos estándar UTF y XML. El segundo rasgo (b) significa que los textos que lo componen representan, en la medida de lo posible, la variabilidad de la lengua (e.g., diferentes tipos de registros, géneros, modos…). El tercer rasgo (c) implica que todos los textos que lo hacen representativo son proporcionales a lo que representan en la lengua. Por último, el cuarto rasgo (d) hace referencia al hecho de que los textos recogidos en el corpus han sido producidos con algún propósito comunicativo y no con el propósito de ponerlos en un corpus. Sin embargo, es difícil que todos los corpus existentes cumplan estrictamente todas estas características a la vez, por lo que son más un ideal que un objetivo alcanzable (Gilquin y Gries, 2009; Gries, 2009; Desagulier, 2017).

1.2 Tipos de corpus

Existe una gran variedad de tipos de corpus (Gries, 2009; Gries, 2016), según diferentes distinciones. Tal vez, la distinción más importante es la que se basa en la representatividad de la lengua, por la que los corpus pueden ser *generalistas* si representan todas las variedades de la lengua, y *específicos* si solo representan una variedad específica. Dependiendo de si tienen algún tipo de etiquetaje o no, pueden ser *corpus anotados* si están etiquetados (e.g., morfológicamente, sintácticamente y semánticamente) (ver IMAGEN 1.1 Y 1.2) y *corpus brutos/crudos* si no lo están. Otra distinción se basa en el periodo histórico de la lengua, los *corpus diacrónicos* están compuestos de textos de diferentes periodos/épocas y representan cómo cambia una lengua con el tiempo, y los *corpus sincrónicos* están compuestos de textos de un periodo/época concreta y representan un tiempo determinado de una lengua. Una última distinción por mencionar es la de los *corpus monolingües* y los *corpus paralelos*, los primeros solo tienen textos de una sola lengua, mientras que los segundos contienen textos de varias lenguas diferentes.

```
<sentence coord="yes">
  <S>
    <sn arg="arg0" coreftype="ident" entity="entity35" entityref="ne" func="suj" ne="person" tem="agt">
      <grup.nom gen="m" num="s">
        <n lem="Ibarretxe" ne="person" pos="np00000" postype="proper" wd="Ibarretxe"/>
      </grup.nom>
    </sn>
    <grup.verb>
      <v lem="acabar" mood="indicative" num="s" person="3" pos="vmis3s0" postype="main" tense="past" wd="acabó"/>
      <gerundi>
        <v lem="acatar" lss="A21.transitive-agentive-patient" mood="gerund" pos="vmg0000" postype="main" wd="acatando"/>
      </gerundi>
    </grup.verb>
    <sn arg="arg1" entityref="nne" func="cd" homophoricDD="yes" tem="pat">
      <spec gen="f" num="s">
        <d gen="f" lem="el" num="s" pos="da0fs0" postype="article" wd="la"/>
      </spec>
      <grup.nom gen="f" num="s">
        <n gen="f" lem="doctrina" num="s" pos="ncfs000" postype="common" sense="16:04563224" wd="doctrina"/>
        <s.a gen="f" num="s">
          <grup.a gen="f" num="s">
            <a gen="c" lem="oficial" num="s" pos="aq0cs0" postype="qualificative" wd="oficial"/>
          </grup.a>
        </s.a>
        <sp>
          <prep>
            <s contracted="yes" gen="m" lem="del" num="s" pos="spcms" postype="preposition" wd="del"/>
          </prep>
          <sn coreftype="ident" entity="entity3" entityref="spec">
            <grup.nom gen="m" num="s">
              <n gen="m" lem="partido" num="s" pos="ncms000" postype="common" sense="16:06131180" wd="partido"/>
            </grup.nom>
          </sn>
        </sp>
      </grup.nom>
    </sn>
  </S>
[...]
</sentence>
```

IMAGEN 1.1. Ejemplo de una oración en castellana etiquetada morfo-sintácticamente en el corpus AnCora (Martí, Taulé, Bertran y Màrquez, 2008; Taulé, Martí y Recasens, 2008).

```
<terms>
<!-- Juan -->
<term tid="t1" type="entity" lemma="Juan" pos="R.IZE-IZB" netype="Pertsona">
<span>
<target id="w82"/>
</span>
</term>
<!-- Jose -->
<term tid="t2" type="close" lemma="Jose" pos="R.IZE-IZB">
<span>
<target id="w83"/>
</span>
</term>
<!-- Ibarretxe -->
<term tid="t3" type="close" lemma="Ibarretxe" pos="R.IZE-IZB" case="ABS">
<span>
<target id="w84"/>
</span>
</term>
<!-- gaur -->
<term tid="t4" type="open" lemma="gaur" pos="A.ADB-ARR">
<span>
<target id="w85"/>
</span>
</term>
<!-- doa -->
<term tid="t5" type="open" lemma="joan" pos="V.ADT">
<span>
<target id="w86"/>
</span>
</term>
<!-- Galiziara -->
<term tid="t6" type="entity" lemma="Galizia" pos="R.IZE-LIB" case="ALA" netype="Tokia">
<span>
<target id="w87"/>
</span>
</term>
```

IMAGEN 1.2. Ejemplo de una oración en euskera etiquetada morfológicamente en el corpus EPEC (Aldezabal, Aranzabe, Arriola y Díaz de Ilarraza, 2009).

1.3 Tamaño del corpus

Actualmente los corpus existentes son cada vez más grandes, cuantificado su tamaño en número de palabras: por ejemplo, el corpus de euskera ETC (*Egungo Testuen Corpusa*) tiene

un tamaño de 269 millones de palabras, el corpus de castellano CORPES XXI (*Corpus del Español del Siglo XXI*) de 277 millones de palabras, el corpus de inglés COCA (*Corpus of Contemporary American English*) de 560 millones de palabras, o *Google Books Corpora* de 155 billones de palabras (en su versión para inglés), tal vez es más extenso hasta la fecha. Este incremento de tamaño que están teniendo los corpus es en gran parte gracias a la disponibilidad de material digital y a las herramientas automáticas de anotación. Sin embargo, los corpus muy grandes tienen el problema de que tienden a no ser ni representativos ni balanceados, además de tener datos que son ruido (i.e., datos no deseados que son difíciles de filtrar a la hora de consultar) (Desagulier, 2017). Así la pregunta que surge es ¿cómo de grande ha de ser el corpus? No existe una respuesta concreta, ya que el tamaño del corpus depende de la pregunta de investigación que planteemos y del tipo de características lingüísticas que queramos investigar (Desagulier, 2017; Brezina, 2018). Es más, los corpus de tamaño pequeño pueden ser suficientes para estudiar fenómenos lingüísticos que suceden con relativa frecuencia, como por ejemplo el uso de oraciones intransitivas y transitivas, oraciones pasivas, morfemas flexivos y derivativos, etc. (Gries, 2009; Gries y Newman, 2014; Desagulier, 2017; Brezina, 2018).

Por otro lado, a la hora de hacer una comparación de frecuencias entre dos corpus o más lo recomendable es que todos los corpus que se utilicen tengan un tamaño similar. Aun así, muchas veces esto no es posible y hacemos comparaciones con corpus de diferentes tamaños. En estas situaciones, lo recomendable es normalizar la frecuencia relativa del rasgo lingüístico que estamos investigando en los diferentes corpus (Biber, Conrad y Reppen, 1998). La frecuencia normalizada se calcula dividiendo la frecuencia absoluta observada entre el tamaño total del corpus, y multiplicando el resultado por 1.000.000 (que es la base común elegida, aunque puede ser otra). Supongamos, por ejemplo, que queremos comparar si hay diferencia en el uso de pronombres personales en el castellano peninsular y en el castellano rioplatense. Para ello usamos dos subcorpus diferentes (TABLA 1.1) del CORPES (*Corpus del Español del Siglo XXI*) (RAE,): uno para el castellano peninsular (Corpus A) y otro para el castellano rioplatense (Corpus B). El Corpus A (peninsular) consta de 103.600.595 palabras y hay 3.659.385 casos de pronombres personales; mientras, el Corpus B (rioplatense) consta de 29.288.104 palabras y tiene 1.044.420 casos de pronombres personales. A simple vista, podemos concluir que hay más casos de pronombres personales en el castellano peninsular (Corpus A) que en el rioplatense (Corpus B). No obstante, si normalizamos ambos corpus a una base común de 1.000.000 palabras, podemos observar que realmente casi no hay diferencias en los casos de pronombres personales entre las dos variedades de castellano: 35.322 casos por millón en el Corpus A (peninsular) y 35.660 casos por millón en el Corpus B (rioplatense).

	CORPUS A	**CORPUS B**
Tamaño total	103.600.595	29.288.104
pronombres personales (freq.)	3.659.385	1.044.420
pronombres personales (freq. norm. [1 mill.])	35.322	35.660

TABLA 1.1. Ejemplo de la comparación de las frecuencias absolutas y normalizadas de pronombres personales en dos corpus de diferente tamaño.

Aun así, cuando hacemos una comparación con corpus pequeños es recomendable normalizar a una base más pequeña que el tamaño real de los corpus, de lo contrario podemos estar ampliando artificialmente nuestras frecuencias y, por tanto, estar falseando nuestros datos (Brezina, 2018).

1.4 Corpus y el estudio del lenguaje

Como se ha visto, el único tipo de información que nos proporcionan los corpus son frecuencias. Por ello, hasta hace unos años, muchos lingüistas han menospreciado los datos obtenidos de corpus (y, por ende, la *lingüística de corpus*) para responder a las preguntas de investigación que se plantean en lingüística. Chomsky (entrevistado en Andor, 2004) y Newmeyer (2003), por ejemplo, opinan que las frecuencias obtenidas de los corpus no aportan nada al estudio del lenguaje:

(1.1) Corpus linguistics doesn't mean anything. […] People who work seriously in this particular area [linguistics] do not rely on corpus linguistics.

La lingüística de corpus no significa nada. [...] Las personas que trabajan seriamente en esta área en particular [lingüística] no se basan en la lingüística de corpus.

(Chomsky en Andor, 2004:97-99)

(1.2) one needs to be very careful about the use made of corpora in grammatical analysis, and particularly the conclusions derived from the statistical information that these corpora reveal.

se ha de ser muy cuidadoso con el uso que se hace de los corpus en los análisis gramaticales, y en particular con las conclusiones derivadas de la información estadística que estos corpus revelan.

(Newmeyer, 2003:695)

Sin embargo, actualmente cada vez más lingüistas (Wasow, 2002; Biber, Conrad y Cortes, 2004; Meyer y Tao, 2005; Tummers, Heylen y Geeraerts, 2005; Sampson, 2007; Gries, 2011; McEnery y Hardie, 2011, entre otros) abogan por el uso de corpus para testear hipótesis, ya que proporcionan evidencias cuantitativas y empíricas:

(1.3) I think corpus-based confirmations of syntactic claims can be enormously convincing.

Creo que las confirmaciones basadas en corpus sobre afirmaciones sintácticas pueden ser enormemente convincentes.

(Pullum, 2007:38)

(1.4) While data from corpora and other naturalistic sources are different in kind from the results of controlled experiments (including introspective judgment data), they can be extremely useful. […] there is no good excuse for failing to test theoretical work against corpora.

Aunque los datos de corpus y otras fuentes naturalistas son diferentes en tipo de los resultados de los experimentos controlados (incluyendo los datos de juicios introspectivos), pueden ser extremadamente útiles. […] no hay buena excusa para no testear el trabajo teórico contra los corpus.

(Wasow, 2002:163)

Gries (2009) y McEnery y Hardie (2011), por ejemplo, ofrecen diferentes ejemplos de cómo los corpus son útiles a la hora de investigar en diferentes áreas lingüísticas: adquisición, fonología, morfología, sintaxis, semántica, pragmática, psicolingüística, etc. Seguramente, es en psicolingüística donde el uso de corpus está aumentando más para testear hipótesis. De hecho, el número de artículos con estudios de corpus en revistas indexadas, como *Journal of Memory and Language* y *Cognitive Linguistics*, está aumentando cada vez más (Gilquin y Gries, 2009; Gries, 2011). Gilquin y Gries (2009) han llevado a cabo una búsqueda en la base de datos *Scopus*[1] para observar el número de artículos en psicolingüística que aúnan estudios de corpus e investigación experimental, y encuentran que de los 85 artículos que componen su búsqueda el 89% de los artículos combinan ambas metodologías. En esos artículos el uso de corpus tiene diferentes funciones: (a) el corpus

[1] *Scopus* es una base de datos de referencias bibliográficas (artículos científicos, libros, actas de conferencias) revisadas por pares.

revalida los resultados de los experimentos, (b) los resultados del corpus son ratificados por experimentos, y (c) el corpus sirve como base de datos para seleccionar los materiales de los experimentos.

El motivo de este auge se debe a que los corpus desempeñan un papel importante en los estudios psicolingüísticos contemporáneos (McEnery y Hardie, 2011). Por un lado, los datos que proporcionan los corpus son más naturales (TABLA 1.2) en comparación a los datos experimentales, dado que son producidos en contextos comunicativos reales. Por otro lado, pueden ayudar en el diseño de materiales que se usan en los experimentos y a incluir información contextual relacionada con el uso del lenguaje en los análisis.

	Fuente de datos
1	corpus con textos escritos
2	colección de ejemplos
3	corpus de lengua hablada grabada en sociedades/comunidades en las que la grabación no es particularmente invasiva.
4	corpus de lengua hablada grabada como resultado de trabajo de campo en sociedades/comunidades en las que la grabación no es particularmente invasiva
5	datos de entrevistas
6	experimentos que requieren participantes hagan algo con el lenguaje que normalmente hacen de todos modos: – producción de oraciones – descripción de imágenes
7	datos provocados haciendo trabajo de campo
8	experimentos que requieren participantes hagan algo con el lenguaje que normalmente no hacen, en unidades con las que normalmente interactúan: – clasificación de oraciones – medición de tiempos de reacción en tareas de decisión léxica – asociación de palabras
9	experimentos que requieren participantes hagan algo con el lenguaje que normalmente no hacen, en unidades con las que normalmente interactúan, y que implican típicos resultados lingüísticos: – mediciones de potenciales evocados [*ERPs*] al ver imágenes – movimientos oculares durante lectura – juicios de aceptabilidad/gramaticalidad

TABLA 1.2. Tipos de datos lingüísticos ordenados por naturalidad (adaptado de Gilquin y Gries, 2009).

Los datos que aportan los corpus son frecuencias y está demostrado que la frecuencia tiene importantes efectos en el procesamiento del lenguaje (para una revisión ver Ellis, 2002; Diessel, 2007; Roland, Dick y Elman, 2007): los elementos (palabras, estructuras sintácticas, oraciones…) más frecuentes se leen, se comprenden, se producen y se adquieren más rápido que los menos frecuente. Dicho de otro modo, la frecuencia se correlaciona con la facilidad en el procesamiento lingüístico y por ello los corpus son una metodología indispensable en la psicolingüística actual.

Por último, hay que tener en cuenta que los corpus no son perfectos y pueden tener alguna que otra desventaja, por lo que su combinación con métodos experimentales (e.g., *self-paced reading*, *eye-tracking, ERPs, fMRI,* etc.) puede ayudar a comprender con más precisión los datos obtenidos de los corpus y a separar "el trigo de la paja" (Gilquin y Gries, 2009). En línea con esto, en los últimos años se están creando lo que se podrían llamar "*corpus psicolingüísticos*", un tipo de corpus que aúna los datos de ambas metodologías: *Dundee Corpus* (Kennedy, Hill y Pynte, 2003), *Postdam Sentence Corpus* (Kliegl, Nuthmann y Engbert, 2006), *UCL corpus* (Frank, Fernandez Monsalve, Thompson y Vigliocco, 2013), *The Natural Stories Corpus* (Futrell et al., 2018), *The Provo Corpus* (Luke y Christianson, 2018), *Russian Sentence Corpus* (Laurinavichyute, Sekerina, Alexeeva, Bagdasaryan y Kliegl, 2019), entre otros. La mayoría de ellos están compuestos por textos que contienen información como la frecuencia, la longitud y la predictibilidad de las palabras combinada con datos de tiempos de lecturas y duración de fijaciones oculares. Tal vez, el más interesante de todos ellos es *The Natural Stories Corpus* porque está compuesto por textos escritos con estructuras sintácticas de baja frecuencia y difíciles de procesar. La finalidad de este corpus es poder testear modelos psicolingüísticos sobre la dificultad de comprensión, dado que generalmente los corpus (i.e., los elaborados a partir de textos periodísticos, literarios, etc.) rara vez contienen estructuras sintácticas que dificultan la comprensión.

1.5 Estudios de corpus y estadística

Como he mencionado en la sección anterior (sección 1.4), los corpus solo proporcionan frecuencias (i.e., datos cuantitativos) y por ello es importante el uso de la estadística para interpretarlos. Hasta hace pocos años, la mayoría de estudios que llevaban a cabo estudios de corpus reportaban las frecuencias mediante estadísticas descriptivas, que básicamente describen de manera resumida (e.g., porcentajes, ratios, medias, medianas, etc.) los datos del corpus que se está utilizando. Sin embargo, si queremos determinar si la frecuencia o distribución de los datos del corpus se debe al azar o está condicionada por otros factores, entonces tenemos que recurrir a la estadística inferencial. El siguiente ejemplo puede ayudar a comprender la diferencia entre ambos tipos de estadística. Imaginemos que

queremos saber si la preferencia de uso de los tiempos verbales presente y pasado depende del aspecto verbal imperfectivo y perfectivo. Como podemos observar en la TABLA 1.2 (los datos son inventados), la estadística descriptiva solo nos proporciona de manera resumida los datos de nuestro corpus: hay un mayor uso de presente cuando el aspecto es imperfectivo y un mayor uso de pasado cuando el aspecto es perfectivo. No obstante, mediante el uso de la estadística inferencial sí que podemos concluir si el aspecto verbal influye en el uso de los tiempos verbales o si el uso de estos se debe al azar. En este caso, una simple regresión logística binomial nos muestra que el aspecto verbal no influye en el uso del tiempo verbal, pues la diferencia no es significativa (*p-value* mayor de .05) a diferencia de lo que podíamos interpretar con la estadística descriptiva.

	Imperfectivo	**Perfectivo**	**TOTAL**
presente	12	6	18
pasado	7	13	20
TOTAL	19	19	38

TABLA 1.3. Distribución ficticia de tiempos y aspectos en un pequeño corpus (tomado de Gries, 2013).

Por tanto, es necesario recurrir al uso de estadísticas inferenciales para poder interpretar mejor las distribuciones observadas en los corpus y explicar los posibles motivos de dichas distribuciones. Pese a todo, hay que tener en cuenta que la significatividad estadística (i.e., el *p-value*) depende del tamaño del corpus, pues puede hacer que incluso un efecto minúsculo sea significativo. Por ello, es recomendable aportar también, junto a la significatividad, el *tamaño del efecto* (*effect size*) porque informa sobre la magnitud del efecto encontrado y así interpretar mejor su significatividad (Jenset, 2008; Biber y Jones, 2009; Gries, 2010; Brezina, 2018).

Dicho todo esto, los corpus son una importante fuente información para tratar de responder preguntas que se plantean sobre el uso del lenguaje y las lenguas, y en concreto sobre la distribución y las frecuencias de diferentes elementos lingüísticos (e.g., silabas, palabras, estructuras sintácticas, oraciones, etc.). Incluso, complementando los datos de corpus con los de estudios experimentales puede ayudar mucho mejor a responder y comprender tales cuestiones, tal y como manifiestan Gilquin y Gries (2009) y Ros (2018):

(1.5) Because the advantages and disadvantages of corpora and experiments are largely complementary, using the two methodologies in conjunction with each other often makes it possible to (i) solve problems that would be encountered if one employed one type of data only and (ii) approach phenomena from a multiplicity of perspectives […].

> Dado que las ventajas y desventajas de los corpus y los experimentos son en gran medida complementarios, el uso de ambas metodologías en conjunto permite a menudo (i) resolver los problemas que se encontrarían si se empleara un solo tipo de datos y (ii) abordar los fenómenos desde una multiplicidad de perspectivas […]
>
> (Gilquin y Gries, 2009:9)

(1.6) it has been shown that experimental evidence alongside data from corpus studies can fine-tune our hypotheses, open up new research questions and provide valuable information about the general mechanisms and representations behind what is believed to be a general cognitive principle and its impact on language form.

> Se ha demostrado que la evidencia experimental, junto con los datos de los estudios de corpus, pueden afinar nuestras hipótesis, abrir nuevas preguntas de investigación y proporcionar información valiosa sobre los mecanismos y representaciones generales detrás de lo que se cree que es un principio cognitivo general y su impacto en la forma del lenguaje.
>
> (Ros, 2018:83)

1.6 Objetivos y esquema de la tesis

En esta tesis doctoral voy a explorar la relación entre la frecuencia de uso de determinadas estructuras gramaticales y el procesamiento del lenguaje. En especial, consideraré la hipótesis de que la facilitación del procesamiento del lenguaje condiciona las preferencias con las que lenguas VO-OV recurren con mayor frecuencia a determinados fenómenos gramaticales. Para tal fin, usaré estudios de corpus como metodología de investigación y espero que esta tesis doctoral contribuya en la idea de que los corpus pueden contribuir en reforzar los hallazgos encontrados en estudios experimentales, así como a que se consideren una fuente de datos más que puede contribuir a responder preguntas que se plantean en campo de la lingüística y la psicolingüística. Esta tesis está organizada de la siguiente manera:

En el CAPÍTULO 2, reviso los estudios de corpus existentes sobre la distribución de los órdenes de palabras en euskera. A su vez, llevo a cabo un nuevo estudio de corpus escrito para observar la distribución de los órdenes de palabras en euskera. Este nuevo corpus, a diferencia de los anteriores, es más representativo y heterogéneo pues está compuesto por textos de diferentes géneros y situaciones comunicativas. Además, el corpus está balanceado, i.e., cada uno de los géneros y subgéneros que lo componen tienen una

extensión similar. Los resultados del estudio de corpus confirman que el orden básico de palabras en euskera es SOV, tal y como muestran algunos estudios de corpus previos (de Rijk, 1969; Aldezabal et al., 2003).

En el CAPÍTULO 3, exploro si el uso de nombres frente a verbos (ratio nombres-verbos) está modulado por el orden básico de palabras de las lenguas, tal y como propone Polinsky (2012). Para ello, llevo a cabo dos estudios de corpus. En el primero, comparo doce lenguas (6 SVO y 6 SOV) y observo que el orden básico de palabras modula el ratio de nombres-verbos, pero no en el sentido de Polinsky (2012). En el segundo estudio de corpus, trato de replicar el resultado del primer estudio mediante un corpus paralelo formado por el mismo texto en cuatro lenguas diferentes (2 lenguas VO y 2 lenguas OV). De esta forma, trato de ver si realmente el ratio de nombres-verbos está modulado por el orden básico de palabras de la lengua o no, ya que puede que al ser textos traducidos todos muestren un ratio nombre-verbos modulado por la lengua del texto original. Los resultados de este segundo estudio de corpus replican los del primer estudio de corpus: el ratio de nombres frente a verbos está modulado por el orden básico de palabras de las lenguas.

En el CAPÍTULO 4, examino si las lenguas OV, que muestran un uso menor de nombres que las lenguas VO, tratan de reducir el número de argumentos preverbales minimizando el coste de procesamiento. Así, en este capítulo, trato de replicar los resultados de Ueno y Polinsky (2009), que encuentran que las lenguas OV muestran una frecuencia mayor de oraciones intransitivas y oraciones transitivas con sujetos omitidos que las lenguas VO. El estudio de corpus de este capítulo replica los resultados obtenidos por Ueno y Polinsky (2009), y proporciona, por tanto, evidencia de que las lenguas muestran una preferencia por minimizar el área preverbal reduciendo el número de argumentos y aligerando así el coste de procesamiento.

En el CAPÍTULO 5, abordo otra de las causas subyacentes que puede modular la frecuencia de omisión de argumentos en la oración: la interferencia de animacidad. Gennari, Mirkovic y MacDonald (2012) observan que la omisión de los argumentos está modulada por la animacidad, dado que tener dos argumentos animados juntos en la linearización aumenta el coste de procesamiento. Por ello llevo a cabo dos estudios de corpus, uno en euskera y otro en castellano, para observar si hay mayor omisión en las oraciones transitivas con ambos argumentos animados (sujeto y objeto). El estudio de corpus revela que sí hay mayor omisión de argumentos en oraciones transitivas cuando ambos argumentos son animado, confirmando así la validez intralingüística de los resultados de Gennari et al. (2012).

Por último, en el CAPÍTULO 6, concluyo y resumo los resultados obtenidos de los estudios de corpus presentados en los capítulos anteriores.

Capítulo 2

Sobre el orden básico de palabras en euskera

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.

2.1 Introducción

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 2.1).

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)

TABLA 2.1. Porcentajes de los órdenes básicos de palabras en el WALS (Dryer, 2013b). Entre paréntesis aparecen las frecuencias absolutas.

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 capítulo trataré sobre la frecuencia del orden de palabras de las oraciones en euskera. En la siguiente sección 2.2, presentaré y discutiré los diferentes estudios de corpus que he llevado a cabo en euskera sobre las frecuencias de los órdenes de palabras y trataré de explicar por qué dichos estudios observan diferencias en los resultados que obtienen para la distribución de los órdenes de palabras. En la sección 2.3, presentaré un nuevo estudio de corpus en euskera con la intención de disipar la divergencia de los resultados de los estudios de corpus anteriores, y analizaré estadísticamente las frecuencias de los órdenes de palabras. El capítulo termina con la discusión (sección 2.4) de los resultados y las conclusiones (sección 2.5).

2.2 Revisión de los estudios de corpus sobre el orden de palabras en euskera

Como he comentado en la sección anterior (sección 2.1), 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 argüir 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[1]: 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 2.2, GRÁFICO 2.1).

[1] muestra I: Barandiaran, J.M. (1921-1935). *Eusko-Folklore*. En *El Mundo en la mente Popular Vasca*, III. San Sebastián: Auñamendi, 27 (Ed. 1962).
muestra II: Echániz, N. (1958). *Euskal-Antzerkiak*. Zarauz: Itxaropena, KulixkaSorta 27-28, 7-132.
muestra III: Echániz, N. (1958). *Euskal-Antzerkiak*. Zarauz: Itxaropena, KulixkaSorta 27-28, 135-159.

	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

TABLA 2.2. Frecuencias absolutas de los órdenes de palabras en euskera en el corpus de de Rijk (1969).

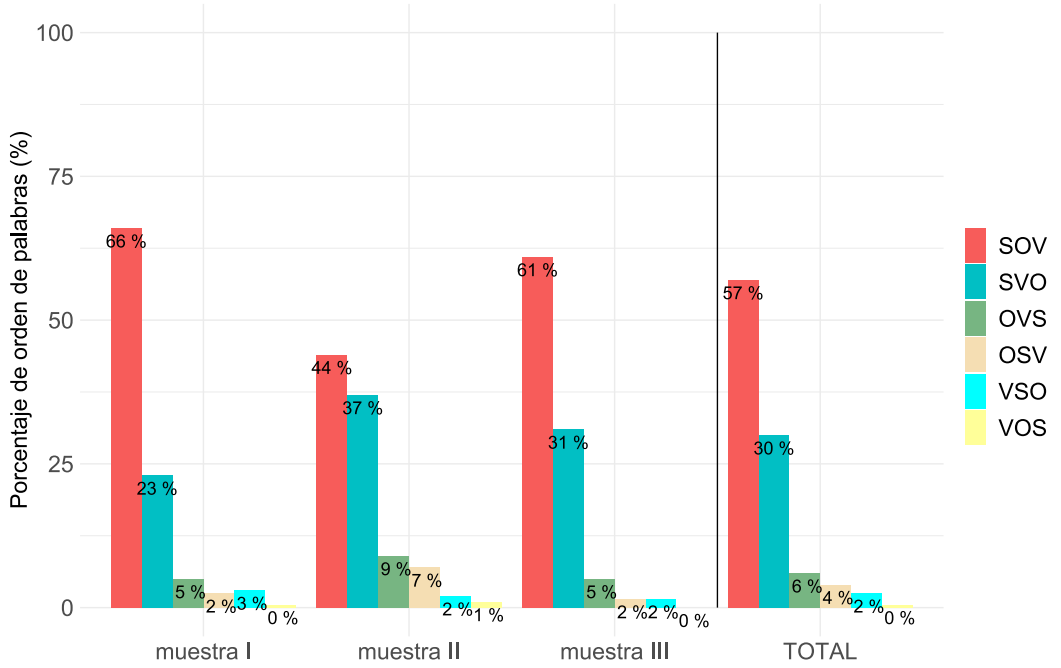


GRÁFICO 2.1. Porcentajes de los órdenes de palabras en euskera en el corpus de de Rijk (1969), por muestras (I, II y III) y en total.

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 etiquetadas 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 2.3, GRÁFICO 2.2).

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

TABLA 2.3. Frecuencias absolutas de los órdenes de palabras en euskera en el corpus de Aldezabal et al. (2003).

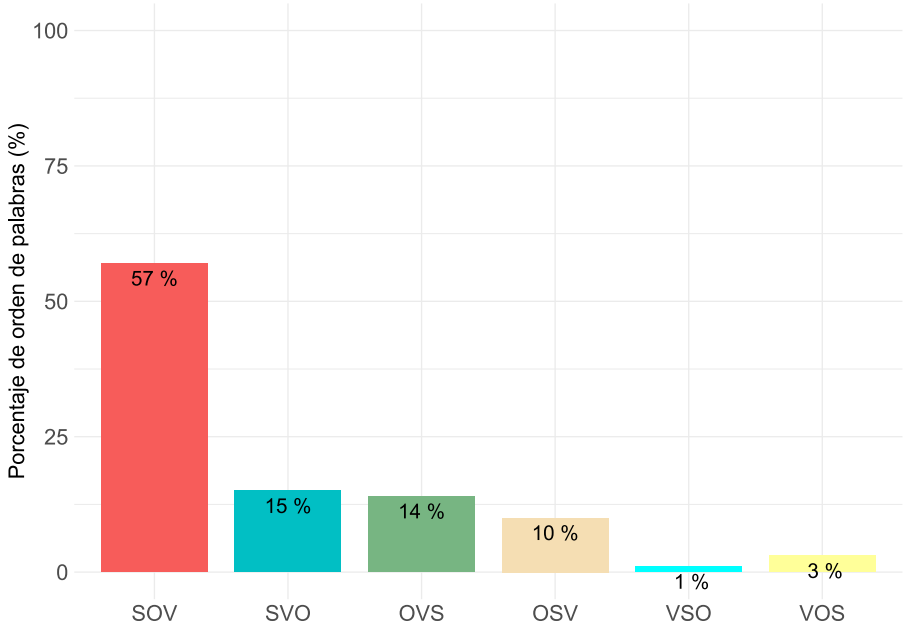


GRÁFICO 2.2. Porcentajes de los órdenes de palabras en euskera en el corpus de Aldezabal et al. (2003).

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[2] de entre los siglos XVI-XX. Para el cómputo de oraciones

[2] Ir al Apéndice A.5 para ver las referencias de los textos utilizados por Hidalgo en su estudio de corpus.

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 2.4, GRÁFICO 2.3). Hidalgo (1995b), basándose en los datos, concluye que es un error decir que el euskera sea una lengua SOV.

	TOTAL
SOV	482
SVO	1081
OVS	204
OSV	70
VSO	136
VOS	40
TOTAL	2013

TABLA 2.4. Frecuencias absolutas de los órdenes de palabras en euskera sumando todos los corpus de Hidalgo (1995a, 1995b).

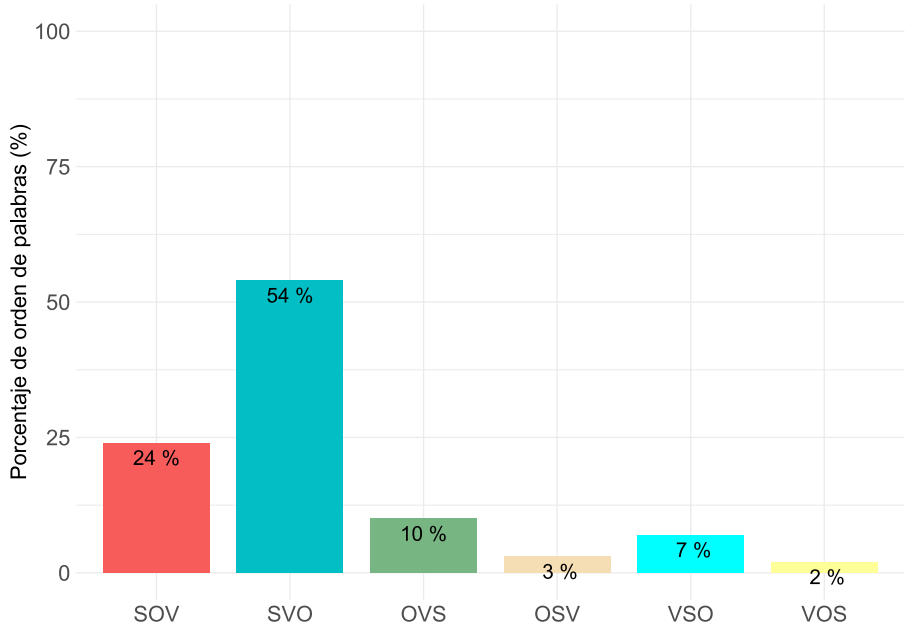


GRÁFICO 2.3. Porcentajes de los órdenes de palabras en euskera en el corpus de Hidalgo (1995a, 1995b).

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 capítulo de las novelas en euskera *Behi euskaldun baten memoriak* y *Kuba triste dago*[3]. Aske concluye que, en conjunto, SVO es el orden de palabras más frecuente (61%) (TABLA 2.5, GRÁFICO 2.4). 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.

	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

TABLA 2.5. Frecuencias absolutas de los órdenes de palabras en euskera en los corpus oral y escrito (y la suma de estos) de Aske (1997).

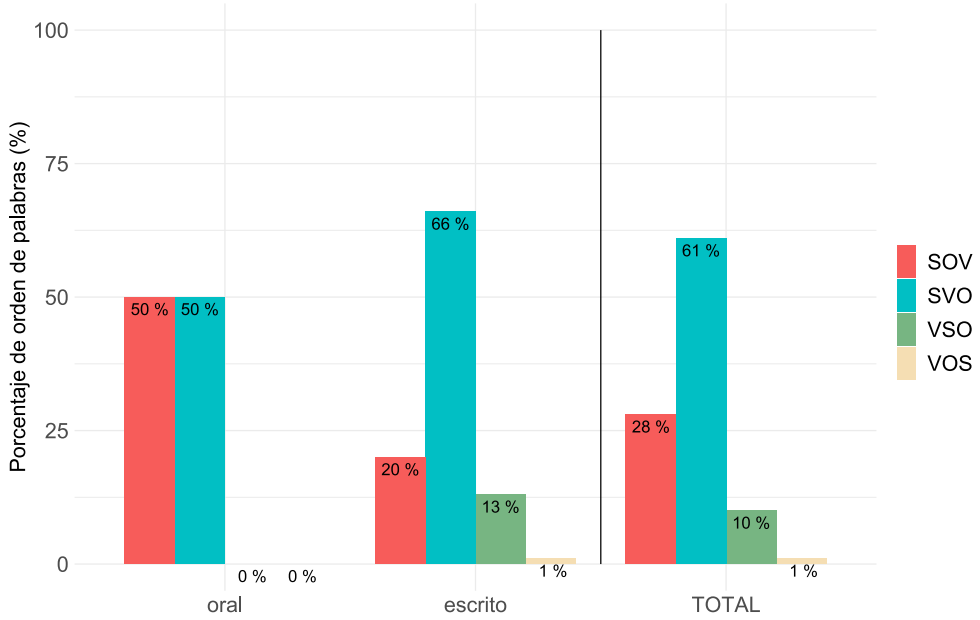


GRÁFICO 2.4. Porcentajes de los órdenes de palabras en euskera en los corpus oral, escrito y la suma de estos (Total) de Aske (1997).

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

[3] Atxaga, B. (1991). Behi euskaldun baten memoriak. Iruñea: Pamiela.
Elexpuru, J. M. (1994). Kuba triste dago. Donostia: Elkar.

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 abordaré con más detalle en la discusión (sección 2.4).

A continuación, presentaré 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, compararé los resultados obtenidos con los de los estudios previos y haré un análisis de todos los datos para determinar cuál es la frecuencia de uso de los posibles órdenes de palabras en euskera.

2.3 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.

2.3.1 Materiales

El corpus escrito de euskera utilizado en este capítulo consta de 4000 oraciones. He obtenido las oraciones del corpus *EPG – Ereduzko Prosa Gaur* (Sarasola, Salaburu, Landa y Zabaleta, 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 oraciones x 7 secciones = 1750 oraciones). En libros, he seleccionado libros de cuatro géneros diferentes: Comedia, Misterio, Histórica y Ensayo (No-ficción) (325 oraciones x 4 géneros = 1300 oraciones). Y en la revista divulgativa *Elhuyar* las secciones Historia, Cultura, Naturaleza, Salud, Tecnología y Ciencia (100 oraciones x 6 secciones = 600 oraciones).

2.3.2 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):

(2.1) a. Nik[S] albaniarrak[O] defenditu ditut[V] — SOV [libros]
"Yo he defendido a los albaneses"

b. Gaizkak[S] irekitzen du[V] atea[O] — SVO [guiones]
"Gaizka abre la puerta"

c. Lana[O] sei ikerketa-taldek[S] egin dute[V] — OSV [revista]
"El trabajo lo han hecho seis grupos de investigación"

d. Lau txanda[O] aurreikusi ditu[V] batzordeak[S] — OVS [periódico]
"Cuatro turnos ha previsto el comité"

e. Berehala irentsi zuen[V] amua[O] katxaloteak[S] — VOS [libros]
"En seguida se ha tragado el anzuelo el cachalote"

f. Eugin jaso zuen[V] Azkuek[S] kanta hau[O] — 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 1054 oraciones transitivas declarativas, descartando las oraciones intransitivas, las ditransitivas y las transitivas negativas, interrogativas, con subordinadas como objeto (objetos CP) y aquellas con argumentos omitidos (NP omitidos) (TABLA 2.6).

Intransitivas	**Transitivas**					**Ditransitivas**	**TOTAL**
989	declarativas	negativas	interrogativas	objetos CP	NP omitidos	122	4000
	1054	80	21	233	1501		

TABLA 2.6. Distribución del tipo de oraciones del corpus escrito de euskera.

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).

2.3.3 Resultados

La TABLA 2.7 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 (χ^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%: χ^2 (5, N = 582) = 726.16, $p < .001$, $V = 0.49$), libros (41%: χ^2 (5, N = 232) = 233.98, $p < .001$, $V = 0.44$), revistas (50%: χ^2 (5 N = 151) = 175.42, $p < .001$, $V = 0.48$) y guiones (69%: χ^2 (5 N = 89) = 181.81, $p < .001$, $V = 0.63$).

	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)

TABLA 2.7. Distribución en euskera de las oraciones transitivas según el orden de palabras y el tipo de género. Entre paréntesis aparecen las frecuencias absolutas.

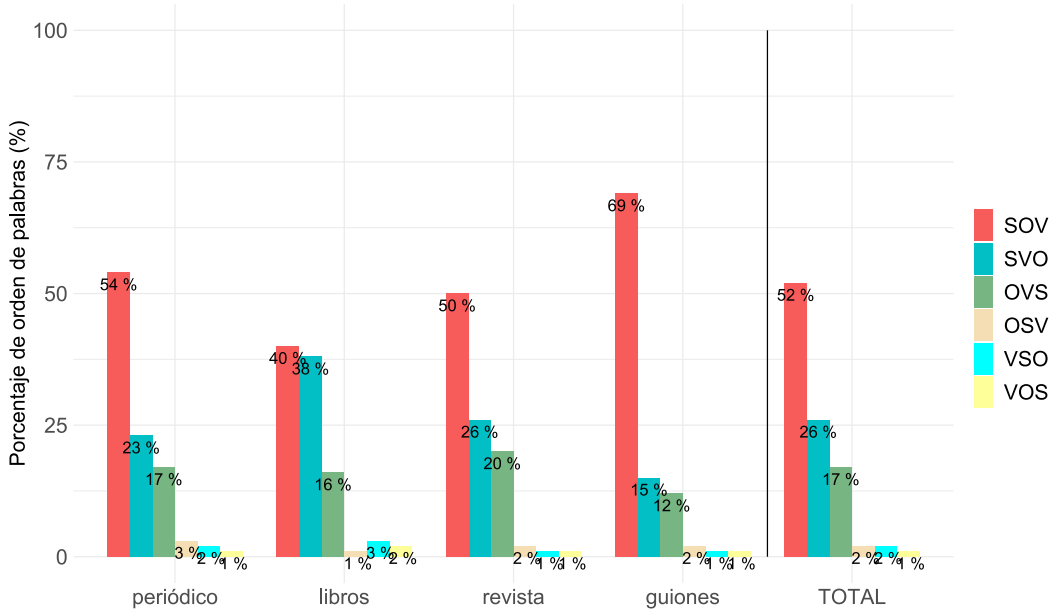


GRÁFICO 2.5. Porcentajes de los órdenes de palabras en euskera en los diferentes géneros (periódico, libros, revista, guiones) y la suma de estos (Total).

Como vemos en el GRÁFICO 2.5, 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 cuatros 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 1000. La prueba chi-cuadrado muestra que la frecuencia de uso de SOV es significativamente similar en cada género [χ^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 [β = 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 [β = 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 2.8).

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)
p-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						

TABLA 2.8. 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).

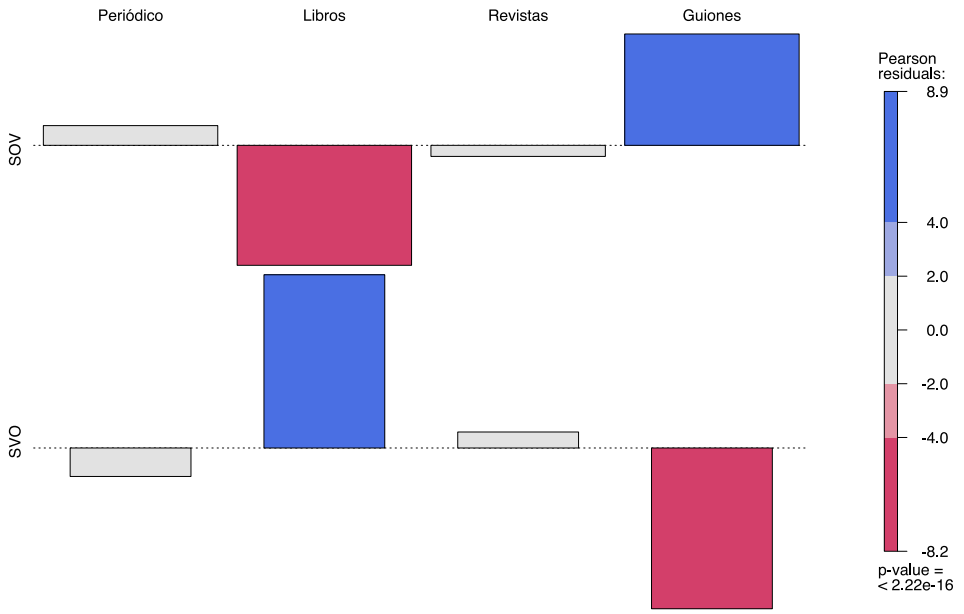


GRÁFICO 2.6. Usos de los SOV y SVO en los cuatro géneros: asociación de residuales. Las barras azules indican residuales positivos y las barras rojas residuales negativos.

El GRÁFICO 2.6 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.

2.3.3.1 Comparación con los estudios de corpus previos

El GRÁFICO 2.7 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 2.3.3). 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).

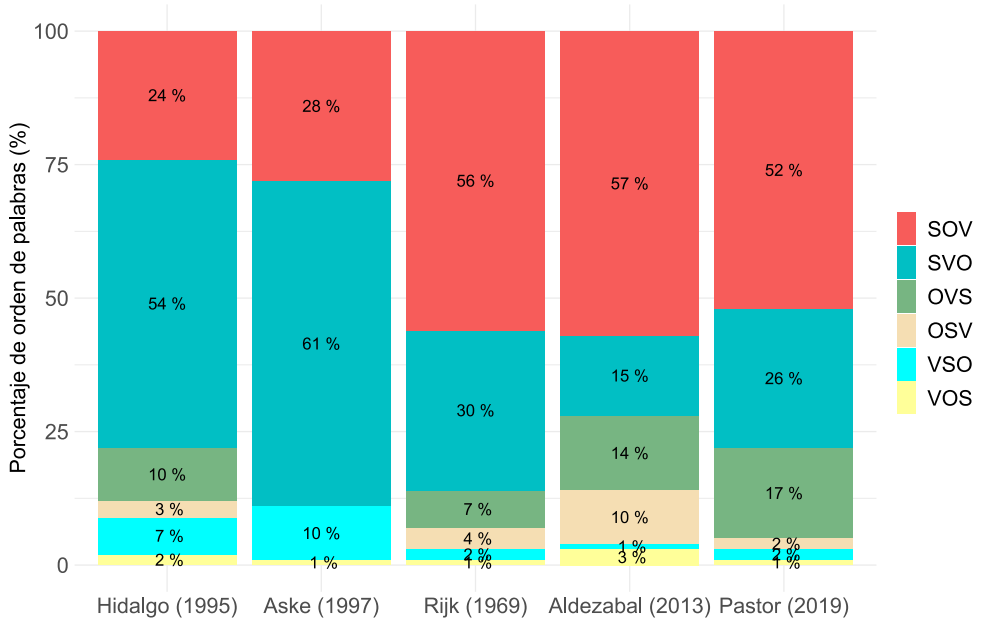


GRÁFICO 2.7. Distribución de la frecuencia de uso de los órdenes de palabras en euskera en los diferentes estudios de corpus (de Rijk, Hidalgo, Aske, Aldezabal, Pastor [esta tesis doctoral]).

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 2.9). 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.

Orden de palabras por autor del corpus – Coeficientes de *estimate*:						
	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)
p-value 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						

TABLA 2.9. 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).

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 ($H(3) = 0.201$, $p = . 0.977$). El orden SOV es el más frecuente, seguido, en orden decreciente, de SVO, OVS, OSV, VSO y VOS (GRÁFICO 2.8).

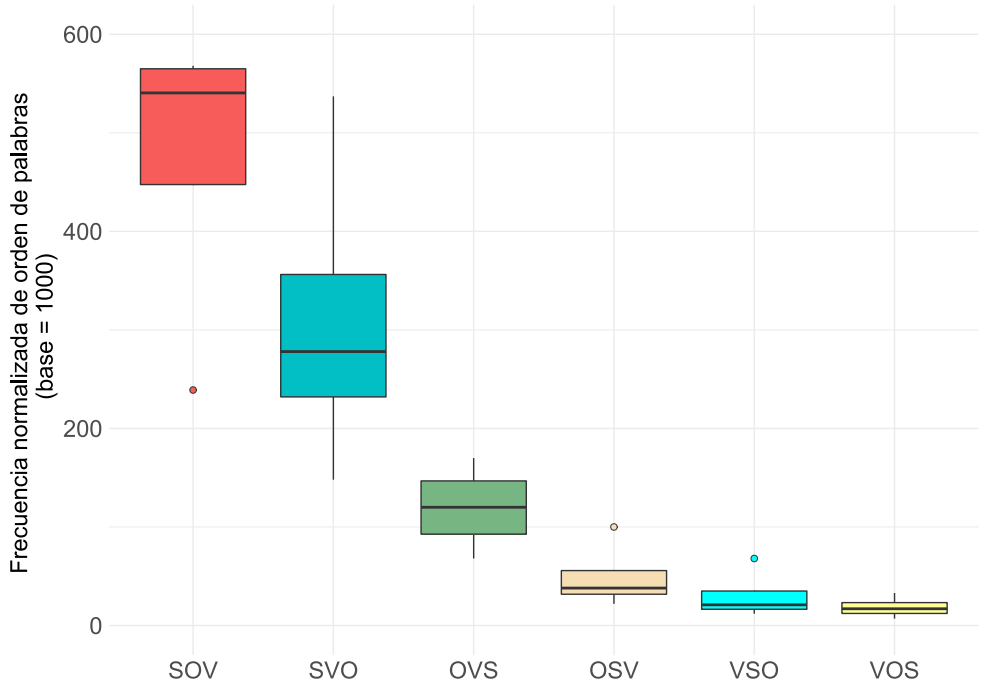


GRÁFICO 2.8. Distribución de la frecuencia de uso de los órdenes de palabras en euskera colapsando los estudios de corpus existentes (de Rijk, Hidalgo, Aldezabal, Pastor).

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.

2.4 Discusión

En este capítulo he llevado a cabo 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 entren 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 cuentan 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, estudios con pacientes afásicos (Arantzeta, Webster, Laka, Martinez-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.

2.5 Conclusiones

El estudio de corpus de este capítulo 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 capítulo 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.

Capítulo 3

Lenguas VO-OV y el ratio de nombres-verbos

ABSTRACT

All languages distinguish the categories "noun" and "verb", and their distinction relies on inflectional morphology, semantic correspondence and syntactic distribution. Based on the universality of nouns and verbs, some studies have observed correlations between the relative distribution of these two categories cross-linguistically and certain linguistic traits, and have tried to explain what factors account for them. Polinsky (2012) investigated the noun-verb ratio in 30 languages and her results revealed a correlation between the noun-verb ratio and the headedness of the language (VO/OV): head-initial languages (VO) have a lower noun-verb ratio than head-final languages (OV). In this chapter, I tested the distribution of nouns vs. verbs in digital corpora in twelve languages in order to replicate the findings in Polinsky (2012). Each digital corpus in this study contains about 300,000 words, obtained from newspaper articles. These corpora, unlike those used by Polinsky (2012), reflect the usage frequency of nouns and verbs in each language, that is, it is not a vocabulary list of nouns and verbs of each language, but a collection of texts/sentences. Results confirm a correlation between the noun-verb ratio and headedness: OV languages have a lower noun-verb ratio than VO languages. I conducted a second corpus study in order to determine whether the language of the original text in a translation could modulate the noun-verb ratios (due to their being translated texts) or, on the contrary, whether each language would show the noun-verb ratios according to its linguistic typology (i.e., word order) regardless of. To this end, I examined and analyzed the first chapter of *The Language Instinct* (Pinker, 1994) book in four languages (English, Spanish, Basque and Korean) and I found that each language showed the distribution of noun-verb ratio in accordance with its headedness: OV languages exhibit a lower noun-verb ratio than VO languages. I conclude that these ratio differences between VO-OV languages could be related to facilitating processing by reducing the number of arguments to process in OV languages.

3.1 Introducción

Las lenguas naturales se componen de combinaciones arbitrarias de sonidos y significados (Saussure, 1916). Estas combinaciones arbitrarias dan lugar a elementos léxicos, sintagmas y oraciones. Los elementos léxicos se dividen en categorías sintácticas y, en particular, todas las lenguas naturales distinguen las *nombres* y *verbos* (Schachter, 1985; Whaley, 1997; Baker, 2003; Schachter y Shopen, 2007; Dixon, 2010; Chung, 2012). Recientemente, Mithun (2007), Seifart (2011) y Polinsky (2012) han investigado la proporción de nombres y verbos en distintas lenguas y han hecho diversas propuestas correlacionando la variabilidad del ratio nombres-verbos con rasgos tipológicos gramaticales. Uno de estos rasgos tipológicos es el orden básico de palabras.

En este capítulo investigaré si la frecuencia de uso de nombres y verbos (i.e., el ratio de nombres-verbos) difiere según el orden básico de palabras que tienen las lenguas. Para ello llevaré a cabo dos estudios de corpus, donde compararé las frecuencias de uso de nombres y verbos en lenguas VO-OV. El capítulo está organizado de la siguiente manera. En la sección 3.2, expondré las diferencias que se han encontrado entre los nombres y verbos. Después, en la sección 3.3, explicaré los estudios de corpus en los que se han propuesto correlaciones entre el ratio de nombres-verbos y rasgos tipológicos, en especial el orden básico de palabras. En las siguientes dos secciones (3.4 y 3.5) presentaré dos estudios de corpus en diferentes lenguas. El capítulo termina con la discusión (sección 3.6) de los resultados de ambos estudios de corpus y las conclusiones (3.7).

3.2 Nombres y verbos

En estudios de adquisición y desarrollo del lenguaje se ha observado que la adquisición de nombres y verbos tiene un patrón evolutivo diferente. Numerosos estudios en diferentes lenguas han demostrado que los niños adquieren y empiezan a usar antes los nombres que los verbos (Gentner, 1982; Verlinden y Gillis, 1988; Jackson–Maldonado, Thal, Marchman, Bates y Gutierrez– Clellen, 1993; Caselli et al., 1995; Poulin–Dubois, Graham y Sippola, 1995; Tardif, Shatz y Naigles, 1997; De Houwer y Gillis, 1998; Sakurai, 1998; Yamashita, 1999; Bassano, 2000; Maital, Dromi, Sagi y Bornstein, 2000; Parisse y Le Normand, 2000; Bornstein et al., 2004; Imai, Haryu, Okabe, Lianjing y Shigematsu, 2006; Casart y Iribarren, 2007; Gentner y Boroditsky, 2009, entre otros).

Estudios de pacientes con lesiones cerebrales también han encontrado diferencias entre nombres y verbos (Goodglass, Klein, Carey y Jones, 1966; Miceli, Silveri, Villa y Caramazza, 1984; Caramazza y Hillis, 1991; Daniele, Giustolisi, Silveri, Colosimo y Gainotti, 1994; Hillis y Caramazza, 1995; Cappa y Perani, 2003; Shapiro y Caramazza, 2003;

Aggujaro, Crepaldi, Pistarini, Taricco y Luzzatti, 2006, y Vigliocco, Vinson, Druks, Barber y Cappa, 2011 para una revisión general): el área fronto-parietal está relacionada con los nombres y el área temporal inferior con los verbos.

Por último, en estudios de corpus se ha observado que la frecuencia de uso de nombres y verbos difiere dependiendo del tipo de texto: en textos más formales (e.g., textos científico-académicos) el uso de nombres es mayor que el de verbos, mientras que en textos menos formales (e.g., textos periodísticos, de ficción, etc.) el uso de nombres es menor (Heylighen y Dewaele, 2002; Hudson, 1994; Bortolini, Tagliavini y Zampolli, 1971; Juilland y Traversa, 1973; Zampolli, 1977; Uit den Boogaert, 1975; Soto, Martínez y Sadowsky, 2005).

3.3 Ratio nombres-verbos y tipología

Mithun (2007) es pionera en proponer una correlación entre el ratio de nombres-verbos y la tipología lingüística. En concreto, Mithun (2007) propone que el ratio de nombres-verbos se correlaciona con el grado de síntesis de la lengua. Su propuesta se basa en el análisis comparativo del inglés y el mohawk (una lengua polisintética). Mithun (2007) observa que el mohawk tiene un ratio de nombres-verbos menor que el inglés, y sugiere que esta diferencia de ratios se debe a que las lenguas polisintéticas, como el Mohawk, recurren a la *incorporación nominal* al verbo, reduciendo así el número de nombres en la oración. Seifart (2011), sin embargo, propone que el ratio nombres-verbos está correlacionado con la concordancia verbal. Una diferencia metodológica entre Mithun (2007) y Seifart (2011) es que el segundo incluye pronombres en la categoría nombres, mientras que la primera los excluye. Seifart (2011) compara las lenguas baure, chintang, bora, N|uu y malayo de Sri Lanka (ver MAPA 3.1), tomadas del corpus DOBES[1], y propone que hay una correlación inversa entre la concordancia verbal y el ratio de nombres-verbos. Es decir, que cuanta más concordancia verbal tenga una lengua menor es su ratio de nombres-verbos y cuanta menos concordancia verbal mayor es su ratio. Seifart (2011) sugiere que esta diferencia se debe a que las lenguas con concordancia verbal (i.e., concordancia verbal de sujeto y objeto) tienen la posibilidad de omitir más argumentos nominales.

[1] DOBES (http://dobes.mpi.nl/) es un archivo que contiene grabaciones de audio y vídeo con anotaciones morfosintácticas de lenguas que se encuentran en peligro de extinción.

MAPA 3.1. Situación geográfica de las lenguas utilizadas en Seifart (2011). De color azul las lenguas con solo concordancia verbal de sujeto, de color naranja las lenguas con concordancia de sujeto y objeto, y de color verde las lenguas sin concordancia verbal.

Por último, y más relevante para esta tesis doctoral es la propuesta de Polinsky (2012) de que existe una correlación entre el ratio de nombres-verbos y el orden básico de palabras. Polinsky (2012) examina vocabularios de 28 lenguas con diferentes órdenes básicos de palabras: V-inicial (VSO/VOS), VO y OV (ver MAPA 3.2). Para llevar a cabo esta comparación, Polinsky (2012) utiliza varias bases de datos léxicas etiquetadas (como WordNet y CELEX[2]). Polinsky (2012) propone que hay una correlación entre el orden básico de palabras y el ratio de nombres y verbos: las lenguas VO tienden a un ratio de nombres-verbos menor que las lenguas OV y, a su vez, las lenguas VSO/VOS a un ratio menor que las lenguas VO (GRÁFICO 3.1). Polinsky (2012) sugiere que esta correlación se debe a la diferente manera que tienen las lenguas VO y OV para crear nuevos verbos. Las lenguas VO muestran un ratio de nombres-verbos menor porque crean, mayormente, nuevos verbos mediante la conversión de nombres en verbos sin usar ningún morfema derivativo (e.g., *milk* > *(to) milk*) o mediante derivación morfológica (e.g., *señal* > *señal* + *-ar* > *señalar*). Las lenguas OV, por el contrario, muestran un ratio de nombres-verbos mayor porque tienden a crear nuevos verbos usando predicados complejos, que están formados por un nombre y un verbo ligero (e.g., *koosõo* [*negociación*] > *koosõo suru* [lit., *hacer negocio* = *negociar*]).

[2] WordNet (Miller, Beckwith, Fellbaum, Gross y Miller, 1990; Beckwith, Fellbaum, Gross y Miller, 1991) es una base de datos léxico-conceptual, compuesta de unidades léxicas y las relaciones semánticas entre ellas. CELEX es otra base de datos léxica con información ortográfica, fonológica, morfológica y sintáctica. Los datos de las lenguas que no están en WordNet y CELEX los obtuvo de diccionarios o de diferentes publicaciones

MAPA 3.2. Situación geográfica de las lenguas utilizadas en Polinsky (2012). De color azul las lenguas OV, de color naranja las lenguas VO, y de color verde lenguas VSO/VOS.

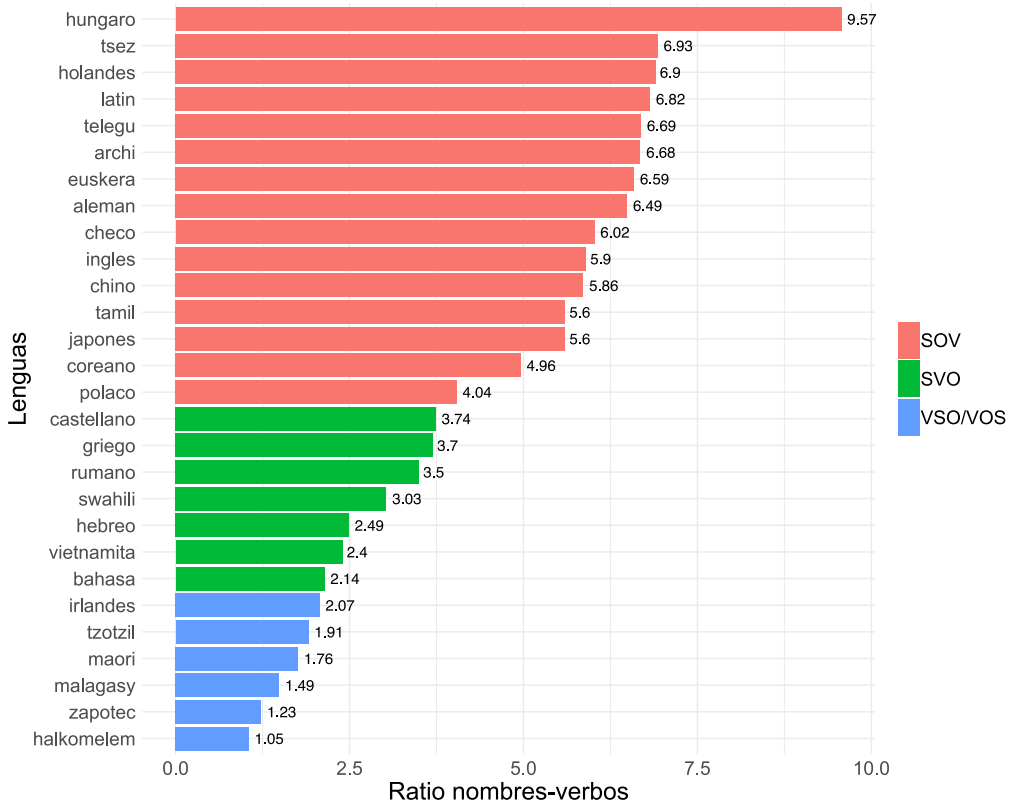


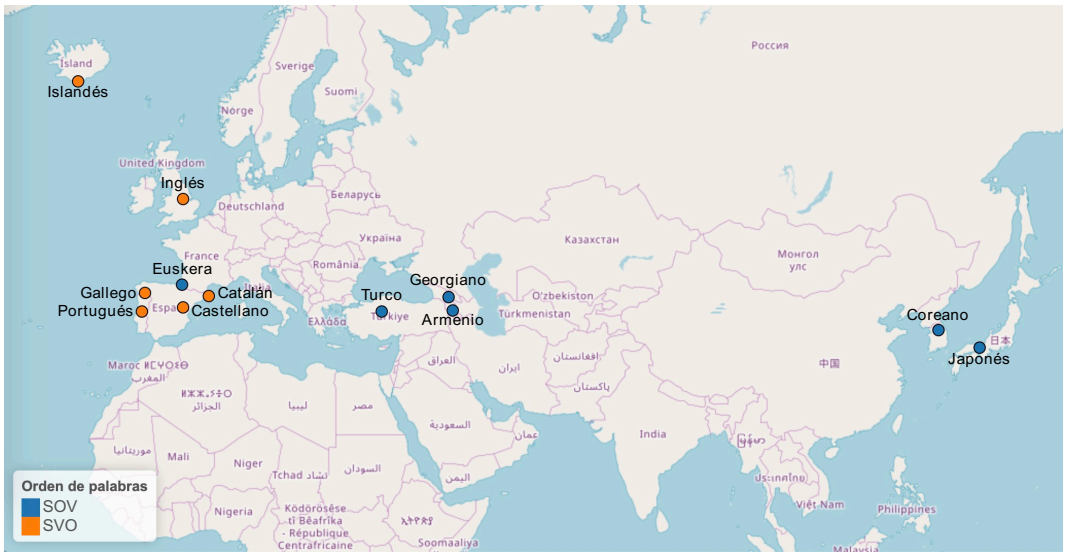
GRÁFICO 3.1. Ratio de nombres-verbos en el estudio de Polinsky (2012).

Sin embargo, y tal y como señala Polinsky (2012), algunas lenguas muestran un ratio de nombres-verbos que no es acorde al que debería de tener de acuerdo a su orden básico de palabras. Por ejemplo, el polaco, el chino, el inglés, el checo, el alemán y el holandés tienen un ratio de nombres-verbos parecido al que muestran las lenguas OV, en vez de tener un

ratio más acorde al de su orden básico de palabras, i.e., VO. Polinsky (2012) sugiere que esto puede deberse a la manera en que se han etiquetado los nombres y los verbos en los diferentes corpus. Por ello, llevaré a cabo un nuevo estudio de corpus para examinar si el orden básico de palabras influye en el ratio de nombres-verbos, en el que todos los corpus tendrán una extensión similar y el etiquetaje de los nombres y verbos tengan un criterio similar.

3.4 Estudio de corpus 1

En el presente estudio de corpus he buscado replicar el efecto encontrado en el estudio de corpus de Polinsky (2012) de que el ratio de nombres-verbos varía en función del tipo de orden básico de palabras que tiene la lengua. Además, otros dos motivos que me llevan a replicarlo son que los corpus que utiliza varían en tamaño (número de palabras total) y en la manera en que están etiquetadas las palabras por categorías léxicas. Para ello he llevado a cabo un estudio de corpus en el que he comparado doce lenguas VO-OV (MAPA 3.3). El grupo SVO está compuesto por las lenguas castellano, catalán, gallego, inglés, portugués e islandés; y el grupo SOV por el euskera, japonés, turco, coreano, armenio y georgiano.



MAPA 3.3. Situación geográfica de las lenguas utilizadas en este estudio de corpus. De color azul las lenguas OV y de color naranja las lenguas VO.

Recordemos que Polinsky (2012) encuentra que las lenguas VO tenían un ratio de nombres-verbos menor que las lenguas OV. Sin embargo, yo espero encontrar la asociación opuesta, es decir, las lenguas VO mostrarán un ratio de nombres-verbos mayor que las lenguas OV. Dicha predicción se basa en el hecho de que las lenguas VO-OV difieren en el número de argumentos que pueden omitir. Casi todas las lenguas utilizadas en este estudio de corpus pueden omitir argumentos, salvo el inglés, el islandés y el

armenio (cf. 3.1e', f' y 3.2c') que no pueden. Sin embargo, mientras que las lenguas VO solo pueden omitir el sujeto (cf. 3.1a', b', c'); las lenguas OV pueden omitir tanto el sujeto como el objeto (cf. 3.2a', b', d', e', f'). Así, espero que las lenguas OV muestren un ratio de nombres-verbos menor que las VO, por el hecho de que estas pueden omitir ambos argumentos.

(3.1) Lenguas VO

a. Yo leí el libro	a'. (yo) Leí el libro	[castellano]
b. Jo legí el llibre	b'. (jo) Legí el llibre	[catalán]
c. Eu lin o libro	c'. (eu) Lin o libro	[gallego]
d. Eu li o livro	d'. (eu) Li o livro	[portugués]
e. I read the book	e'. *(I) Read the book	[inglés]
f. Ég las bókina	f'. *(ég) Las bókina	[islandés]

(3.2) Lenguas OV

a. Nik liburua irakurri nuen	a'. (nik) (liburuak) irakurri nuen	[euskera]
b. Ben kitap okudum	b' (ben) (kitap) okudum	[turco]
c. Es girk'ĕ kardac'i	c'. *(es) girk'ĕ kardac'i	[armenio]
d. Watashiwa hono yonda	d'. (watashiwa) (hono) yonda	[japonés]
e. Nanun chaykul ilkessta	e'. (nanun) (chaykul) ilkessta	[coreano]
f. Me tsigni ts'avik'itkhe	f'. (me) (tsigni) ts'avik'itkhe	[georgiano]

3.4.1 Metodología

He comparado doce corpus escritos de diferentes lenguas. Seis de esas lenguas son VO: castellano, catalán, gallego, inglés, portugués, e islandés. Las otras seis lenguas son OV: euskera, turco, japonés, coreano, armenio y georgiano. En total la suma de los corpus consta aproximadamente de 3.600.000 palabras (~300.000 palabras x 12 lenguas) (TABLA 3.1), compuestos generalmente por textos periodísticos y etiquetados por categorías léxicas (i.e., *part-of-speech* – PoS). Para el castellano y el catalán he utilizado el corpus *AnCora* (Martí et al., 2008; Taulé et al., 2008); para el gallego, el *Corpus de Referencia do Galego Actual (CORGA)* (Rojo, López, Domínguez y Barcala, 2010); para el inglés, el *Corpus of Contempary American English (COCA)* (Davies, 2008); para el portugués, el *Corpus de Extractos de Textos Electrónicos Público (CETEMPúblico)* (Rocha y Santos, 2000; Santos y Rocha, 2001); para el islandés, el *Tagged Icelandic Corpus (MÍM)* (Helgadóttir,); para el euskera, *EPEC* (Aranzabe, 2008; Aldezabal et al., 2009); para el japonés, *JEITA Public Morphologically Tagged Corpus* (Hagiwara,); para el turco, *METU-Sabanci Turkish* Treebank (Oflazer, Say, Hakkani-Tür y Tür, 2003); para el coreano, *High quality morpho-syntactically annotated corpus (HQMSAC)* (Lee y Choi, 1999); para el armenio, *Eastern Armenian National*

Corpus (EANC) (Khurshudian et al., 2009); y para el georgiano, el *Georgian National Corpus (GNC)* (Meurer, 2011).

	corpus	**lenguas**	**nº de palabras**
VO	AnCora	castellano	302.017
	AnCora	catalán	302.927
	CORGA	gallego	300.257
	COCA	inglés	301.043
	CETEMPúblico	portugués	300.034
	MIM	islandés	307.628
OV	EPEC	euskera	300.000
	METU	turco	300.189
	JEITA	japonés	301.789
	HQMSAC	coreano	299.532
	EANC	armenio	653 documentos[3]
	GNC	georgiano	300.000[4]

TABLA 3.1. Tamaño total en número de palabras de los corpus usados para cada lengua.

A la hora de contabilizar el ratio de nombres y verbos, he contado todas las apariciones de palabras etiquetadas como nombres y verbos. Sin embargo, y siguiendo el criterio de Polinsky (2012), he excluido de la muestra de nombres todos aquellos que hacían referencia a nombres propios, de lugar y siglas; de igual modo, dentro de la muestra de verbos, los auxiliares han sido excluidos. El cálculo del ratio de nombres-verbos lo hecho dividendo la frecuencia de nombres entre la de verbos:

(3.3) ratio de nombres y verbos = nombres / verbos

Los ratios de nombres-verbos de los diferentes corpus los he analizado con los siguientes análisis estadísticos: la prueba de Wilcoxon y el modelo de regresión lineal. Mediante la prueba de Wilcoxon he comparado si los ratios de nombres-verbos de las lenguas VO y OV son significativamente diferentes o no. La regresión lineal, por su parte, la he utilizado para modelar y explicar la relación entre el orden de palabras (VO-OV) y el ratio de nombres-verbos, es decir, comprobar si el orden de palabras contribuye a que el ratio de nombres-verbos sea menor o mayor. Ambos análisis los he computado mediante el

[3] No he podido acceder a los archivos, por lo que he obtenido los datos limitando la búsqueda en el EANC a artículos de prensa escritos durante el año 2002.

[4] No he podido acceder a los archivos, por lo que he obtenido los datos limitando la búsqueda en el GNC a artículos de prensa escritos del periódico *Georgian Times* (40.622 palabras) y después he normalizado las frecuencias a 300.000 palabras.

programa estadístico R (R Core Team, 2017). Los resultados los he considerado significativos a un nivel $p < .05$. Los gráficos los he realizados con el paquete *ggplot2* (Wickham, 2009).

3.4.2 Resultados

La TABLA 3.2 muestra la frecuencia de nombres y verbos, y el ratio de nombres-verbos de cada una de las lenguas. El GRÁFICO 3.2 ilustra la distribución de los ratios nombres-verbos entre las lenguas VO (castellano, catalán, gallego, inglés, islandés y portugués) y lenguas OV (armenio, coreano, euskera, georgiano, japonés y turco). La prueba de Wilcoxon revela que la distribución de los ratios de nombres-verbos de las lenguas VO y OV son significativamente diferentes [$Z = -2.8924$, $p < .002$, r = 83]: las lenguas VO muestran un ratio de nombres-verbos mayor que las lenguas OV.

	lenguas	**nº de nombres**	**nº de verbos**	**ratio N-V**
VO	castellano	53.787	33.191	1,62
	catalán	54.641	31.562	1,73
	gallego	56.138	31.527	1,78
	inglés	57.347	33.899	1,69
	portugués	56.274	32.723	1,72
	islandés	73.756	42.613	1,73
OV	euskera	86.910	60.671	1,43
	turco	37.863	25.815	1,47
	japonés	50.216	36.209	1,39
	coreano	85.995	61.901	1,39
	armenio	900.040	624.927	1,44
	georgiano	62.323	47.184	1,32

TABLA 3.2. Frecuencia de nombres y verbos en las diferentes lenguas analizadas y su correspondiente ratio de nombres-verbos.

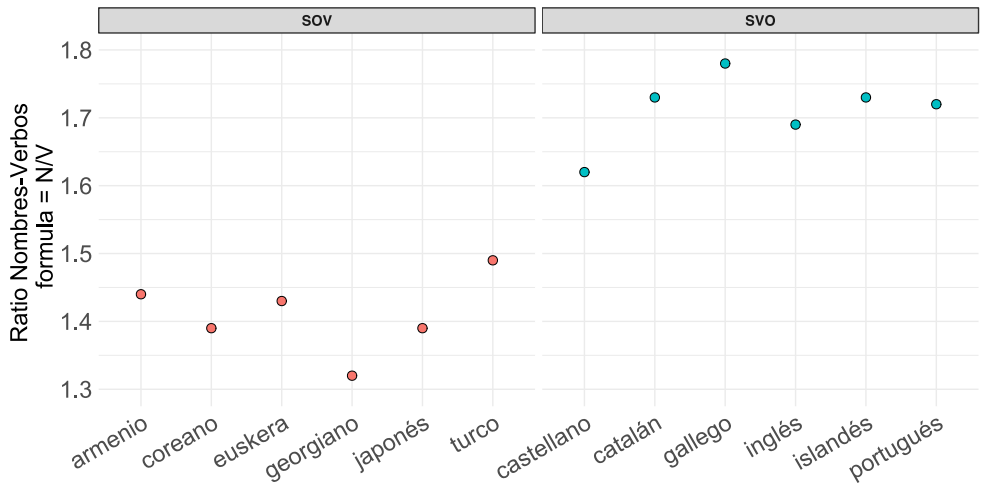


GRÁFICO 3.2. Ratio de nombres-verbos de las lenguas analizadas según el orden de palabras.

En el modelo de regresión lineal (TABLA 3.3) la variable dependiente ha sido el ratio de nombres-verbos. El predictor SVO del modelo compara la media de los ratios de las lenguas OV, que es el nivel de referencia, con la media de ratios de las lenguas VO. El modelo revela que existe una relación el orden básico de palabras y el ratio de nombres-verbos: el ratio de nombres-verbos tiende a ser mayor cuando la lengua es VO que si es OV [$\beta = 0.30167$, $p < .001$, $R^2 = 0.89$] (TABLA 3.3; GRÁFICO 3.3). Es decir, el modelo predice que cuando una lengua es VO el ratio de nombres-verbos es 1,32 veces más probable de ser mayor que cuando es OV.

Ratio N-V – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	1.41000	0.02269	62.149	0.001 ***
SVO	0.30167	0.03208	9.402	0.001 ***
Multiple R-squared: 0.8984			Adjusted R-squared: 0.8882	

TABLA 3.3. Resultados del modelo de regresión lineal para el ratio de nombres-verbos y el orden de palabras (VO vs. OV).

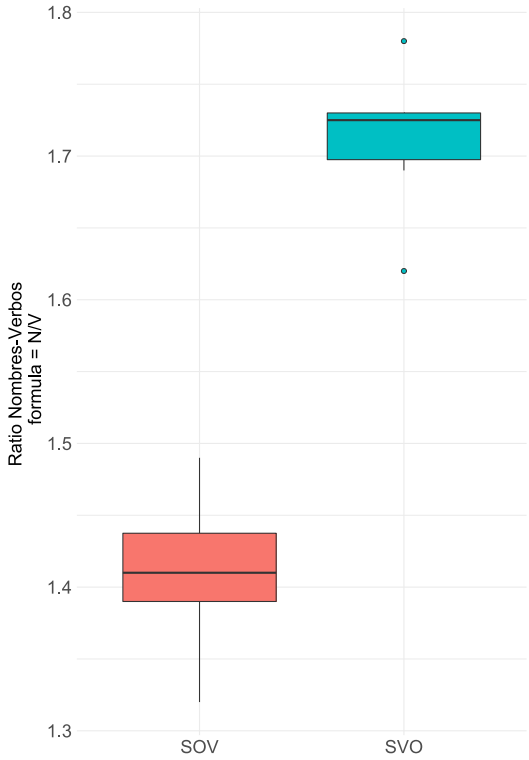


GRÁFICO 3.3. Ratio de nombres-verbos según el orden de palabras.

Los resultados del presente estudio de corpus revelan que existe una tendencia general a que las lenguas con orden de palabras VO tengan un ratio de nombres-verbos mayor que las lenguas con orden OV, que tienden a tener un ratio menor. Esto va en dirección opuesta a lo observado por Polinsky (2012): las lenguas VO mostraban un ratio de nombres-verbos menor que las lenguas OV. Para garantizar que la diferencia de ambos estudios de corpus no se debe al tipo de corpus utilizado, he llevado a cabo un segundo estudio de corpus basado en la comparación de un mismo texto en diferentes lenguas.

3.5 Estudio de corpus 2

En este segundo estudio de corpus he tratado de replicar los resultados del estudio de corpus anterior (sección 3.4), mediante el uso de un corpus paralelo de cuatro lenguas (castellano, euskera, inglés y coreano). Un corpus paralelo consiste en una colección del mismo texto (o textos) en diferentes lenguas, es decir, en textos en su lengua original y sus traducciones en otras lenguas. De esta forma, puedo comprobar, por una parte, si la tendencia del ratio de nombres-verbos en las lenguas VO-OV del estudio de Polinsky (2012) y del estudio de corpus 1 (sección 3.4) se debe al tipo de corpus utilizado o no. Por otra parte, si el orden de palabras de la lengua del texto original influye en el ratio de nombres-verbos que las otras lenguas pueden mostrar en sus textos o, por el contrario, cada lengua mostrará el ratio de nombre-verbos que corresponde a su orden básico de palabras.

3.5.1 Metodología

El corpus paralelo está compuesto por el primer capítulo del libro *The Language Instinct* (*El Instinto del Lenguaje*) de Steven Pinker en cuatro idiomas diferentes: inglés (versión original), castellano, euskera y coreano. El corpus tiene una extensión total de unas 14.500 palabras (TABLA 3.4).

	lenguas	**nº de palabras**	**corpus**
VO	inglés	3.861	Pinker (1994)
	castellano	4.477	Pinker (1995a)
OV	euskera	3.158	Pinker (2010)
	coreano	3.046	Pinker (1995b)

TABLA 3.4. Tamaño de los corpus de las lenguas VO (inglés y castellano) y lenguas OV (euskera y coreano).

Los textos han sido etiquetados manualmente según el tipo de categoría léxica que es cada palabra. Las versiones de castellano y euskera han sido etiquetadas por mí mismo, y las versiones de inglés y coreano por dos alumnos nativos de cada una de las lenguas de la Universidad de Harvard. Se han etiquetado como nombres solo los nombres comunes. En coreano, los nombres denominados "nombres dependientes" no se han etiquetado como nombres comunes dado que estos no pueden aparecer por si solos; sino que requieren siempre de un determinante (Chang, 1996; Kim y Yang, 2007; Yeon y Brown, 2011): por ejemplo, el nombre dependiente "*kes*" (것) ["cosa/hecho"] no puede aparecer solo y requiere de los demostrativos "*i*" (이), "*ku*" (그), "*ce*" (저) ["este", "eso", "aquello", respectivamente], convirtiéndose el conjunto en pronombres demostrativos (Chang, 1996). En el caso de los verbos, se han etiquetado como verbos todas las formas verbales personales y las formas verbales de infinitivo (cf. 3.4a,b) y gerundio (cf. 3.4c,d), dado que estas formas verbales no personales pueden tomar objetos (Croft, 1991):

(3.4) a. … es[V] posible *imaginar*[V] la vida[N] sin él[PP].
b. … there is[V] something[N] *to write*[V] about it[PP].
c. …*coordinando*[V] sus esfuerzos[N]…
d. …insects[N] *using*[V] Doppler sonar[N].

En cuanto a las formas de participio, se han etiquetado como verbos en los casos que van acompañados de auxiliares (cf. 3.5a,b), mientras que en el resto de casos se han considerado como adjetivos o como una oración reducida de relativo (cf. 3.5c,d) (Croft, 1991).

(3.5) a. Chomsky has[AUX] *puzzled*[V] many readers[N]…
b. Chomskyk irakurle[N] asko *harritu*[V] ditu[AUX]…
c. …el lenguaje[N] *hablado*[ADJ]…
d. Pertsona[N] *eskolatu*[ADJ] gehienek badituzte[V]…

Al igual que en el estudio de corpus anterior (sección 3.4.1), he calculado el ratio de nombres-verbos dividendo la frecuencia de nombres entre la de verbos. De igual modo, los ratios de nombres-verbos de las diferentes lenguas los he analizado mediante la prueba de Wilcoxon y el modelo de regresión lineal. Ambos análisis los he llevado a cabo con el programa estadístico R (R Core Team, 2017) y los gráficos los he realizados con el paquete *ggplot2* (Wickham, 2009). Los resultados los he considerado significativos a un nivel $p < .05$.

3.5.2 Resultados

La TABLA 3.5 muestra la frecuencia de nombres y verbos, y el ratio de nombres-verbos de cada una de las lenguas. Al igual que en el estudio de corpus anterior (sección 3.3.2), las lenguas VO muestran un ratio de nombres-verbos mayor que las lenguas OV (GRÁFICO 3.4). Sin embargo, esta diferencia de ratios no es significativa [$Z = -1.549$, $p = .121$, $r = 0.77$], debido al tamaño de la muestra pues solo hay dos lenguas dentro de cada tipo de orden de palabras (VO-OV).

	lenguas	**nº de nombres**	**nº de verbos**	**ratio N-V**
VO	inglés	878	520	1,69
	castellano	968	599	1,62
OV	euskera	872	594	1,47
	coreano	1.057	742	1,43

TABLA 3.5. Frecuencia de nombres y verbos en las diferentes lenguas analizadas y su correspondiente ratio de nombres-verbos.

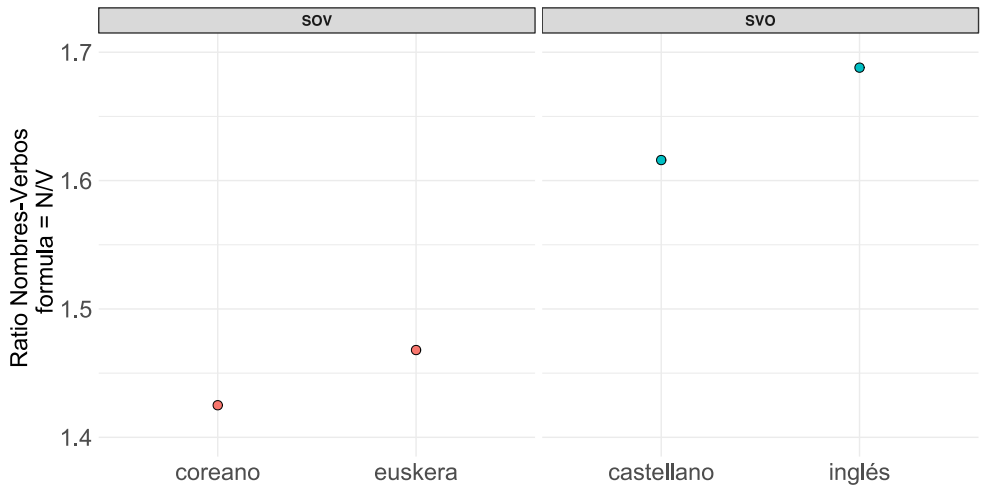


GRÁFICO 3.4. Ratio de nombres-verbos de las lenguas analizadas, según el orden de palabras (VO-OV).

El modelo de regresión lineal, por su parte, muestra una relación entre el orden básico de palabras y el ratio de nombres-verbos: las lenguas VO tiende a tener un ratio mayor que las lenguas OV [$\beta = 0.20050$, $p < .049$, $R^2 = 0.85$] (TABLA 3.6; GRÁFICO 3.5). Así, el modelo predice que el ratio de nombres-verbos será 1,22 veces mayor si la lengua es VO que si es OV.

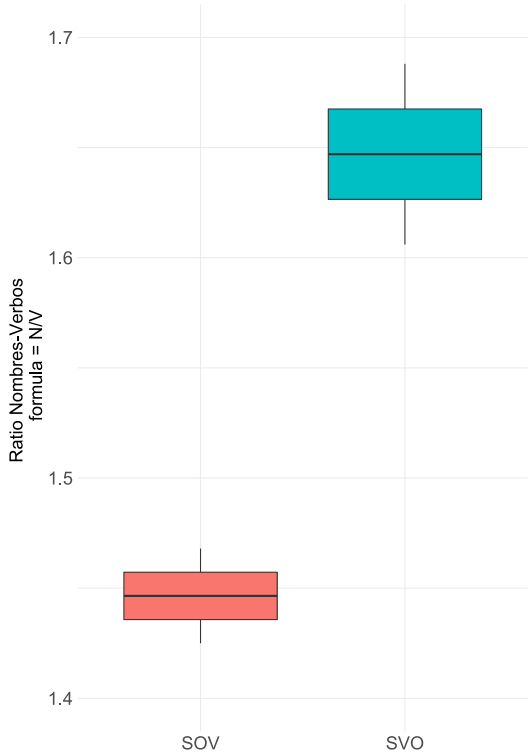


GRÁFICO 3.5. Ratio de nombres-verbos según el orden de palabras.

Ratio N-V – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	1.44650	0.02965	48.786	0.001 ***
SVO	0.20550	0.04193	4.901	0.039 *
Multiple R-squared: 0.9231			Adjusted R-squared: 0.8847	

TABLA 3.6. Resultados del modelo de regresión lineal para el ratio de nombres-verbos y orden de palabras (SVO-SOV).

Finalmente, para asegurar que el efecto del ratio de nombres-verbos no se debe a un subconjunto de oraciones dentro de cada lengua, he examinado los ratios por oración y lengua. Como puede observase en el GRÁFICO 3.6, la mayoría de las oraciones de euskera y coreano (ambas OV) tienen una mayor proporción de ratio de nombres-verbos menor que las oraciones de inglés y castellano (ambas VO) [franja azul del gráfico]. A su vez, cuando se observa las oraciones con un ratio de nombres-verbos mayor [franja roja del gráfico], el inglés y el castellano tienen una proporción mayor de oraciones con este ratio que el euskera y el coreano.

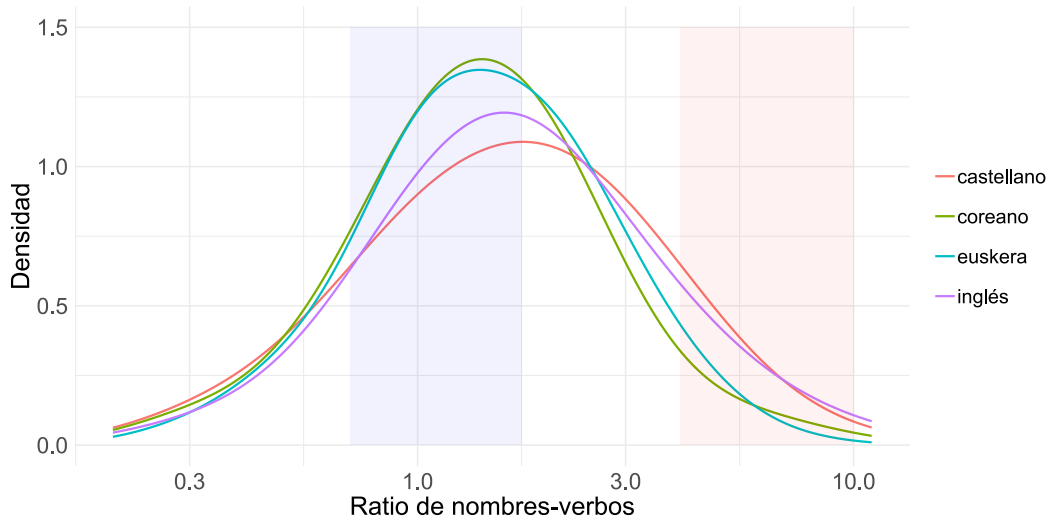


GRÁFICO 3.6. Distribución de los ratios de nombres-verbos de todas las oraciones de cada una de las lenguas.

En este segundo estudio de corpus, he replicado los resultados encontrados en el primer estudio de corpus de este capítulo (sección 3.4.2). Existen diferencias en la frecuencia de uso de nombres frente a verbos (i.e., en el ratio de nombres-verbos) entre las lenguas VO y OV. Más concretamente, existe una tendencia a que las lenguas VO tengan un ratio de nombres-verbos mayor que las lenguas OV. Dicho de otra forma, las lenguas VO tienden a usar más nombres que las lenguas OV.

3.6 Discusión

En este capítulo he llevado a cabo dos estudios de corpus con la intención de verificar la correlación propuesta por Polinsky (2012) de que la proporción de nombres es menor en las lenguas VO que en las lenguas OV. Los dos estudios de corpus de este capítulo indican una correlación inversa: las lenguas OV muestran un ratio de nombres-verbos menor que las lenguas VO. A continuación, compararé y discutiré los resultados de este capítulo con los del estudio de corpus de Polinsky (2012).

Los resultados de los dos corpus en este capítulo revelan que las lenguas OV tienen un ratio de nombres-verbos significativamente menor que las lenguas VO, que muestran un ratio mayor. En particular, los resultados del segundo estudio de corpus (sección 3.5) replican esta correlación, independientemente de si el texto es escrito originalmente en dicha lengua o es una traducción de un texto en otra lengua. Polinsky (2012) reporta que las lenguas VO tienden a mostrar un ratio de nombres-verbos menor que las lenguas OV, las cuales tienden a tener un ratio mayor. Sin embargo, los resultados de este capítulo muestran la tendencia contraria: las lenguas VO tienen un ratio mayor y las lenguas OV un ratio menor. Polinsky (2012) propone que el motivo de la diferencia de ratios entre las lenguas VO y SOV se debe al modo en que estas crean nuevos verbos. Según ella, las lenguas VO muestran un ratio de nombres-verbos menor porque estas tienden a crear verbos mediante la conversión de nombres en verbos (e.g.: *milk*$_{[n]}$ > *to milk*$_{[v]}$) y la derivación morfológica usando sufijos derivativos (e.g.: *señal*$_{[n]}$ > *señal*$_{[n]}$ + *-izar*$_{[suf.v]}$ > *señalizar*$_{[v]}$). De esa forma, los nombres se transforman en verbos y eso hace que la frecuencia de los nombres se reduzca y aumente la de los verbos. Las lenguas OV, por el contrario, muestran un ratio de nombres-verbos mayor porque tienen un preferencia de crear verbos uniendo un nombre con un verbo ligero, creando un predicado complejo (e.g.: *emayru*$_{[n]}$ *"e-mail"* > *emayru suru*$_{[v]}$ *"crear un e-mail = 'e-mailear'"*). Este argumento que da Polinsky (2012) para las lenguas OV no se sustenta si nos fijamos en datos de corpus. Por ejemplo, el euskera puede recurrir a ambos mecanismos para crear verbos: usar predicados complejos y la derivación. Sin embargo, el uso de predicados complejos "nombre + verbo ligero" no es productivo, ya que la forma más frecuente para crear verbos en euskera es mediante el uso del sufijo derivativo *-tu* (Azkue, 1905:XVIII). Una simple búsqueda en el corpus EPEC[5] confirma esta observación: aparecen 646 verbos que son predicados complejos (nombre + verbo ligero) frente a 8.963 verbos formados con el sufijo *-tu*.

En mi opinión, esta diferencia de ratios entre las lenguas VO-OV se debe al uso de argumentos omitidos. Como mostraré en el siguiente capítulo (CAPÍTULO 4), es más frecuente que sean las lenguas OV las que pueden omitir argumentos por ello las lenguas

[5] La búsqueda está limitada a 155.654 palabras.

OV tienden a usar oraciones con menos argumentos que las lenguas VO. La omisión de argumentos facilitaría el coste de procesamiento si es cierto que en las lenguas OV necesitan mantener más argumentos activos en la memoria hasta procesar el verbo que las lenguas VO (Hawkins, 1994, 2004, 2014; Gibson, 1998, 2000). Sin embargo, casi todas las lenguas VO-OV utilizadas en los estudios de corpus de este capítulo pueden omitir argumentos y, a pesar de ello, se observan diferencias de ratio de nombres-verbos entre ellas. Por tanto, lo que sí puede influir no es la omisión de argumento *per se*, sino el número de argumentos que se pueden omitir: solo el sujeto, solo el objeto o ambos. Curiosamente, todas las lenguas OV (excepto el armenio) de ambos estudios de corpus pueden omitir tanto el sujeto como el objeto. Pero las lenguas VO que pueden omitir argumentos (todas salvo el inglés y el islandés) solo omiten el sujeto. Por tanto, que las lenguas OV tengan un ratio de nombres-verbos menor que las lenguas VO también puede estar correlacionado con el número de argumentos que estas pueden omitir. Así, el menor ratio de nombres-verbos observado en las lenguas OV es consistente con la idea de reducir el número de argumentos y aligerar el coste de procesamiento (Hiranuma, 1999; Ueno y Polinsky, 2009). Ueno y Polinsky (2009) proponen que las lenguas OV son más difíciles de procesar porque necesitan mantener en la memoria más argumentos hasta que llegar a procesar el verbo, pues en ese punto donde los argumentos son interpretados e integrados (Pickering y Barry, 1991; Gibson y Hickok, 1993; Pickering, 1993; Trueswell, Tanenhaus y Kello, 1993; Garnsey, Pearlmutterb, Myersa y Lotockyc, 1997).

En línea con la anterior, creo que otra explicación del efecto observado puede ser que las lenguas OV muestran un ratio de nombres-verbos menor para reducir el coste de planificación de la oración. Seifart et al. (2018) encuentran, en un corpus oral de nueve lenguas[6], que los hablantes tienden a reducir la velocidad de articulación, así como a hacer una pausa antes de articular un nombre que un verbo. Seifart et al. (2018) (aunque en contra, Szekely et al., 2002; Vigliocco et al., 2011) sugieren que ello se debe a que los nombres requieren más tiempo de planificación que los verbos. De esta forma, las lenguas OV requerirían más tiempo de planificación de una oración que las lenguas VO, ya que las primeras deben de mantener durante la planificación dos nombres antes del verbo y las segundas solo uno. Esto explica también el menor ratio de nombres-verbos observado en los corpus de este capítulo: las lenguas OV recurren a reducir el número de nombres mostrando un ratio de nombres-verbos menor que las lenguas VO, mediante la omisión de argumentos, para reducir el coste de planificación de oraciones con ambos argumentos en posición preverbal y, por tanto, reducir su coste de procesamiento.

[6] Lenguas que forman el corpus oral de Seifart et al. (2018): baure, bora, chintang, even, hoocak, n||ng, texistepec, holandés e inglés.

En lo referente a la discrepancia de los resultados de este capítulo con los de Polinsky (2012), creo que se deben fundamentalmente a diferencias metodológicas utilizadas. En primer lugar, las frecuencias de Polinsky (2012) se basan en *lemas* y las de este capítulo en *tokens*. El termino *lema* hace referencia a cada una de las palabras únicas en un corpus y el termino *token*, por el contrario, hace referencia a cualquiera de las palabras que aparecen en un corpus (Martín Herrero, 2009). Por ejemplo, la siguiente frase "*un buen libro es un libro que te entretiene*" contiene siete *lemas* (*un, buen, libro, es, que, te, entretiene*) pero nueve *tokens*, que son cada una de las palabras que componen la oración. Como ya he explicado anteriormente, Polinsky (2012) obtiene las frecuencia de nombres y verbos de los WordNet de cada una de las lenguas, que son listas de palabras organizadas como diccionarios, por lo que la frecuencia que observa se basa en *lemas* y no en *tokens*. En los estudios de corpus de este capítulo, por el contrario, he analizado la frecuencia de los *tokens* etiquetados como nombres y verbos en el corpus, es decir, la frecuencia de uso de todos los nombres y verbos que aparecen. Así, un cálculo basado en *lemas* refleja la disponibilidad de nombres y verbos que tiene cada lengua en su lexicón y uno basado en *tokens* el uso que hace cada lengua de los nombres y verbos. En el caso de las lenguas que no tienen WordNet, Polinsky (2012) no detalla si las frecuencias están basadas en *lemas* o *tokens*. En segundo lugar, también hay diferencias entre las frecuencias reportadas por Polinsky (2012) y las reportadas por las publicaciones originales empleadas como fuente para su estudio. Haciendo un repaso a dichas publicaciones, he encontrado que algunas de ellas no reportan ningún tipo de frecuencia o que los datos que reportan y los utilizados por Polinsky (2012) son diferentes. Por ejemplo, en el caso del latín, Polinsky (2012) reporta 4777 nombres y 700 verbos (ratio de nombres-verbos = 6,82); sin embargo, en la publicación original (Minozzi, 2009) la frecuencia de los verbos es mayor: 2609 verbos. Teniendo en cuenta las frecuencias de nombres y verbos de la publicación, se obtiene un ratio de nombres-verbos de 1,83, es decir, un ratio menor al de Polinsky (2012) y este nuevo ratio sitúa al latín en el grupo de lenguas V-inicial (VSO-VOS). En tercer lugar, Polinsky (2012) explica que en algunos casos, sin detallar cuáles, ha excluido del cómputo de frecuencias los verbos formados por un nombre y un verbo ligero. Esta exclusión no está motivada, dado que argumenta que la diferencia de ratio de nombres-verbos en las lenguas VO-OV se basa crucialmente en el modo en que estas crean nuevos verbos: las lenguas VO mediante conversión o derivación y las lenguas OV mediante el uso de predicados complejos (i.e., nombre + verbo ligero).

3.7 Conclusiones

Los estudios de corpus de este capítulo tenían como objetivo verificar si el orden básico de palabras se correlaciona con la frecuencia de uso de nombres frente a verbos. Los resultados han mostrado que sí existe una correlación: el ratio de nombres-verbos de una

lengua depende del orden básico de palabras que esta tenga. En particular, las lenguas OV tienden a usar menos nombres que las lenguas VO. Además, esta correlación entre el orden de palabras y el ratio de nombres-verbos es independiente del que el texto esté escrito en versión original o sea la traducción de otra lengua. En resumen, estos resultados sugieren que las lenguas OV recurren a un menor uso de nombres que las lenguas VO para reducir el coste de procesamiento de los argumentos requeridos por el verbo.

Capítulo 4

Argumentos preverbales y la minimización del coste de procesamiento

ABSTRACT

Several studies argue that the basic word order influences processing strategies. Hawkins (2003, 2004, 2009) argued that a mother node and all its constituents (e.g. V and NP for VP) have to be processed as soon as possible to reduce memory load. Yamashita and Chang (2001) showed that the position of long and short constituents depends on the type of VO-OV languages, so that OV languages have a preference to move long constituents in front of short ones ("long before short") while VO languages have a preference to move a long constituents behind a short one ("short before long "). Ueno and Polinsky (2009) argued that VO-OV languages use certain grammatical resources with different frequencies to reduce the number of preverbal arguments and thus facilitate real-time processing. They observed that OV languages use intransitive predicates more frequently than VO languages, while both types have a similar frequency of omission of subjects. In this chapter, I aimed to examine whether Basque (OV) resorts more often than Spanish (VO) to certain grammatical operations, in order to minimize the number of arguments to be processed before the verb. I conducted a comparative corpus study of Spanish and Basque. Results showed that (a) the frequency of use of subject pro-drop was higher in Basque than in Spanish; (b) Basque did not use more intransitive sentences than Spanish; both languages had a similar frequency of intransitive sentences; and (c) Basque placed arguments in postverbal position as an equivalent strategy to minimize the number of preverbal arguments. Based on these findings, I conclude that OV languages tend to reduce the preverbal area resort to certain grammatical resources to facilitate the processing and that the frequency of use of them does not depend on a single typological trait (VO-OV) but it is modulated by the concurrence of other grammatical features.

4.1 Introducción

Una pregunta central de la teoría lingüística contemporánea es hasta qué punto las propiedades del lenguaje son resultado de las condiciones impuestas por la forma externa y el significado. Así, por ejemplo, el Programa Minimalista (Chomsky, 1995) plantea estudiar la arquitectura del lenguaje bajo la hipótesis de que gran parte de las propiedades del lenguaje se derivan necesariamente de las condiciones impuestas por la interfaz conceptual-intencional (C-I) y la interfaz articulatorio-perceptual (A-P).

(4.1) The language is embedded in performance systems that enable its expressions to be used for articulating, interpreting, referring, inquiring, reflecting, and other actions. […] The performance systems appear to fall into two general types: articulatory-perceptual and conceptual-intentional. If so, a linguistic expression contains instructions for each of these systems. Two of the linguistic levels, then, are the interface levels A-P and C-I, providing the instructions for the articulatory-perceptual and conceptual-intentional systems.

El lenguaje está encapsulado entre sistemas de actuación que permiten que sus expresiones se utilicen para articular, interpretar, referir, preguntar, pensar y otras acciones. […] Los sistemas de actuación parecen dividirse en dos tipos generales: el articulatorio-perceptual y el conceptual-intencional. Si es así, una expresión lingüística contiene instrucciones para cada uno de esos niveles. Dos de los niveles lingüísticos son, entonces, los niveles de interfaz A-P y C-I, que proporcionan las instrucciones para los sistemas articulatorio-perceptual y el conceptual-intencional respectivamente.

(Chomsky, 1995:168)

En el campo de la psicolingüística, la idea de que los requerimientos del procesamiento del lenguaje influyen en la forma de las oraciones fue postulada originalmente por Yngve (1960) en su "*Depth Hypothesis*", que observa que las estructuras se expanden hacia la izquierda (*left-branching*) y hacia la derecha (*right-branching*) por restricciones del procesamiento y la memoria (Miller, 1956). El "*Depth Hypothesis*" predice que las estructuras que se expanden hacia la derecha requieren menos capacidad de memoria para almacenar los nodos sin completar que las estructuras que se expanden hacia la izquierda, ya que estas últimas superarían el límite de nodos que puede almacenar la memoria[1]. Más tarde, Bever (1970) arguyó que la estructura del lenguaje refleja procesos cognitivos

[1] Yngve (1960) asume que hasta un máximo de 7 palabras pueden almacenarse al mismo tiempo en la memoria.

generales, de forma que los mecanismos de adquisición y procesamiento del lenguaje son algunos de los factores que determinan la forma del lenguaje:

(4.2) Many aspects of adult language derive from the interaction of grammar with the child's processes of learning and using language. Certain ostensibly grammatical structures may develop out of other behavioral systems rather than being inherent in grammar. That is, linguistic structure is itself partially determined by the learning and behavioral processes that are involved in acquiring and implementing that structure.

Muchos aspectos del lenguaje adulto derivan de la interacción de la gramática con los procesos de aprendizaje y uso del lenguaje del niño. Ciertas estructuras aparentemente gramaticales pueden desarrollarse a partir de otros sistemas conductuales en lugar de ser inherentes en la gramática. Es decir, la estructura lingüística está determinada en parte por los procesos de aprendizaje y de comportamiento que están involucrados en la adquisición e implementación de esa estructura.

(Bever, 1970:280)

Como he mencionado en el capítulo anterior (CAPÍTULO 3), las lenguas muestran diferencias en la frecuencia del uso de las categorías léxicas (nombres y verbos) según su orden básico de palabras (SVO-SOV). En este capítulo exploraré algunos aspectos de la relación entre la estructura gramatical y las condiciones impuestas a la externalización de las expresiones lingüísticas; en particular consideraré la hipótesis de que la facilitación del procesamiento del lenguaje en tiempo real condiciona la frecuencia de uso de ciertas construcciones gramaticales que hacen las lenguas VO-OV. Este capítulo está organizado de la siguiente manera. La sección 4.2 introduce algunas teorías que proponen que el orden de palabras que exhiben las lenguas son el resultado de las restricciones del procesamiento y cómo estás recurren a mecanismo que mejoren dicho procesamiento. La sección siguiente (4.3), se presenta un estudio de corpus comparativo entre euskera y castellano, y donde se analiza la frecuencia con que las lenguas VO-OV usan ciertas estrategias (la omisión y la intransitividad) para reducir el coste de procesamiento. El capítulo termina con la discusión (sección 4.4) de los resultados del estudio de corpus y las conclusiones (4.5).

4.2 Estrategias de procesamiento en las lenguas VO-OV

Seis son los posibles órdenes de palabras básicos documentados en lenguas naturales: SVO, SOV, VSO, VOS, OSV, OVS; aunque la gran mayoría de las lenguas humanas son de tipo SOV (41%) o SVO (35%) (Dryer, 2013b). Greenberg (1963) reveló que estos órdenes básicos de palabras en la oración se corresponden fuertemente con otras propiedades de orden de palabras: si una lengua es del tipo *x*, entonces esa lengua tendrá una característica *y*. En especial, Greenberg (1963) observó que las diferentes posiciones del verbo (VSO, SVO y SOV)[2] convergen con otros órdenes de palabras. Por ejemplo, si una lengua es de tipo VSO tiende a usar preposiciones (universal 3)[3], mientras que si es de tipo SOV usará posposiciones (universal 4)[4]; o si una lengua es del tipo VSO el auxiliar siempre precede al verbo, pero si es del tipo SOV el auxiliar sucede al verbo (universal 16)[5].

4.2.1 Procesamiento y estrategias de linearización

Varias teorías de procesamiento sostienen que estas correlaciones se deben a que son órdenes que implican una complejidad mínima y, por tanto, son más eficientes y más fáciles de procesar (Yngve, 1960; Miller y Chomsky, 1963; Frazier, 1979, 1985; Hawkins, 1994; Gibson, 1998, 2000; Hawkins, 2004, 2014). Frazier (1979, 1985) arguye que estas correlaciones se deben a que facilitan el reconocimiento de los sintagmas que constituyen la oración durante el procesamiento. Frazier (1979, 1985) asume que el procesamiento tiene un alcance muy limitado, es decir, que solo es capaz de procesar a la vez 5 o 6 palabras; por lo que requiere de elementos que faciliten la identificación de cada uno de los sintagmas. Esto elementos son las categorías funcionales (e.g., determinantes, auxiliares, preposiciones, posposiciones, etc.). Así, las lenguas OV tienden a situar las categorías funcionales al final de los sintagmas (como en el caso de las posposiciones) para segmentar la oración en sintagmas después de cada categoría funcional. Las lenguas VO, por el contrario, las sitúan al comienzo de los sintagmas (como en el caso de las preposiciones) para segmentar la oración en sintagmas antes de cada categoría funcional.

Hawkins (1983, 1994, 2004, 2014), en cambio, propone su *Performance-Grammar Correspondence Hypothesis* (en adelante, PGCH) que el orden de palabras en las lenguas OV y SVO viene determinada por su eficiencia a la hora de ser procesadas, i.e., por ser menos

[2] VSO = lenguas con verbo en posición inicial; SVO = lenguas con verbo en posición intermedia; y SOV = lenguas con verbo en posición final.

[3] Universal 3. Languages with dominant VSO order are always prepositional (Greenberg, 1963:45).

[4] Universal 4. With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional (Greenberg, 1963:45).

[5] Universal 16. In languages with dominant order VSO, an inflected auxiliary always precedes the main verb. In languages with dominant order SOV, an inflected auxiliary always follows the main verb (Greenberg, 1963:50).

complejas. La eficacia de los órdenes viene determinada por el principio "Minimizar Dominios" (*Minimize Domains* – MiD). El MiD maximiza la rapidez de procesamiento y predice una preferencia por estructuras con dominios más pequeños. Algunos órdenes de constituyentes reducen el número de palabras necesarias para reconocer la estructura sintáctica y el MiD refleja la preferencia por aquel orden que permite el reconocimiento de la estructura lo antes posible. Así, mediante este principio Hawkins explica, por ejemplo, por qué las lenguas VO tienden a utilizar preposiciones (cf. 4.3a) y las lenguas OV posposiciones (cf. 4.3b): cada tipo de lengua trata de que la distancia lineal entre el verbo (V) y la preposición/posposición (P) sea la más corta posible:

(4.3) a. VP[fue PP[a la universidad]] b. [[unibertsitatera]PP joan zen]VP
universidad-a-la ir AUX
c. *VP[fue [la universidad a]PP] d. *[PP[-ra unibertsitate] joan zen]VP
-a-la universidad ir AUX

Del mismo modo, el MiD predice que ante dos órdenes alternativos que expresan una misma proposición, los hablantes tenderán a preferir el que más eficiente y menos complejo sea. Un orden es considerado más eficiente si permite reconocer y producir lo antes posible la estructura de la oración. Hawkins determina la eficiencia de un orden mediante el *ratio sintagma-palabra*, que calcula dividiendo el número de sintagmas por el número total de palabras dentro del *Dominio de Combinación Sintagmática* (PCD), y el resultado obtenido se expresa mediante porcentaje. Cuanto más alto sea el porcentaje obtenido en un orden de palabras, mayor será su eficiencia y, por tanto, mayor también su preferencia de uso. Por ejemplo, en (4.4) las dos oraciones del inglés (VO) son gramaticales, pero una de ellas es preferible a la otra según predice el MiD. En (4.4a), los sintagmas VP, PP_1 y PP_2 pueden reconocerse tras procesar cuatro palabras: "*went*", "*to*", "*London*", "*in*", mientras que en (4.4b) son necesarias seis palabras: "*went*", "*in*", "*the*", "*late*", "*afternoon*", "*to*", "*London*". El ratio sintagma-palabra para reconocer el PCD de VP del ejemplo (4.4a) es 3/4 = 75% (se necesitan 4 palabras para reconocer los 3 sintagmas), mientras que en (4.4b) es 3/6 = 50% (se necesitan 6 palabras para reconocer los 3 sintagmas).

(4.4) a. John VP[went PP[to London] PP[in the late afternoon]]
nº palabras 1 2 3 4
nº sintagmas 1 2 3
ratio sintagma-palabra = 3/4
MiD = 75%

"John fue a Londres al final de la tarde"

b.	John	VP[went	PP[in	the	late	afternoon]	PP[to	London]]			
nº palabras		1	2	3	4	5	6				
nº sintagmas		1			2			3			
ratio sintagma-palabra = 3/6											
MiD = 50%											

"John fue al final de la tarde a Londres"

En japonés (OV), las oraciones de (4.5) también son gramaticales, aunque una de ellas es preferible a la otra tal y como predice el MiD. En (4.5a) –similar a (4.4a) en inglés– los sintagmas VP, PP_1 y PP_2 pueden reconocerse tras procesar cuatro palabras: "*London-e*", "*yugata*", "*osoku-ni*", "*it-ta*", mientras que en (4.5b) –similar a (4.4b) en inglés– son necesarias tres palabras: "*yugata*", "*osoku-ni*", "*London-e*". El ratio sintagma-palabra para reconocer el VP del ejemplo (4.5a) es 3/4 = 75% (se necesitan 4 palabras para reconocer los 3 sintagmas), mientras que en (4.5b) es 3/3 = 100% (se necesitan 3 palabras para reconocer los 3 sintagmas).

(4.5)

a.	John-ga	[[London-e]PP	[yugata	osoku-ni]PP	it-ta]VP					
nº palabras		1	2	3	4					
nº sintagmas		1		2		3				
ratio sintagma-palabra = 3/4										
MiD = 75%										

"John fue a Londres al final de la tarde"

b.	John-ga	[yugata	osoku-ni]PP	[[London-e]PP	it-ta]VP					
nº palabras		1	2	3						
nº sintagmas		1		2		3				
ratio sintagma-palabra = 3/3										
MiD = 100%										

"John fue al final de la tarde a Londres"

Gibson (1998, 2000), por su parte, sostiene que la preferencia de las lenguas OV y VO por un orden de palabras u otro viene determinado por los costes de memoria a la hora de procesar las oraciones. Tal y como defiende en su *Dependency Locality Theory* (en adelante, DLT), la capacidad de memoria es limitada (Miller, 1956a, 1956b; Miller y Chomsky, 1963; Gibson, 1991) por lo que durante el procesamiento de una oración la integración de referentes discursivos (i.e., nombres y verbos) supone un coste de recursos en la memoria. Este coste de integración es calculado en unidades de energía (EU), de tal forma que cada nuevo referente discursivo consume una EU y otra EU más por cada nuevo referente discursivo que interviene hasta que se integra dentro de su dependencia. Así, en lenguas

VO como el inglés, la DLT predice que la preferencia de «*sintagmas cortos delante de largos*» se debe a que tiene un menor coste de integración que colocar los sintagmas largos delante de los cortos. En (4.6a), la integración del NP "*the beautiful pendant that was in the jewelry*" [*el precioso colgante que estaba en la joyería*] consume 2EUs: 1EU por ser referente discursivo y otro 1EU por el referente discursivo ("*girl*") que interviene entre el NP y el verbo "*gave*" [dar]; y la integración del PP "*to the girl*" [*a la chica*] solo consume 1EU por ser referente discursivo. En total la oración (4.6a) tiene un coste de integración de 7EUs. Por el contrario, en (4.6b) la integración del NP "*the beautiful pendant that was in the jewelry*" consume solo 1EU por ser referente discursivo; y la integración del PP "*to the girl*" consume 4EUs: 1EU por ser referente discursivo y 3EUs más por los referentes discursivos ("*pendant*", "*was*", "*jewelry*") que intervienen entre el PP y el verbo "*gave*". En total la oración (4.6b) tiene un coste de integración de 9 EUs. Por lo tanto, en las lenguas VO las oraciones con sintagmas cortos delante largos (cf. 4.6a) son más fáciles de procesar porque suponen un coste de integración menor que las oraciones con sintagmas largos delante de cortos (cf. 4.6b).

(4.6)[6] a. the boy gave [to the girl] [the beautiful pendant that was in the jewelry]
IC 0 1 1 +0 0 1 +1 0 1 0 1 0 0 1
EUs = 7
"el chico le dio a la chica el precioso colgante que estaba en la joyería"

b. the boy gave [the beautiful pendant that was in the jewelry] [to the girl]
IC 0 1 1 +0 0 1 0 1 0 0 1 +3 0 1
EUs = 9
"el chico le dio el precioso colgante que estaba en la joyería a la chica"

En las lenguas OV como el euskera, por el contrario, la DLT predice una preferencia de anteponer los sintagmas largos a los cortos por tener un menor coste de integración y de memoria. En (4.7a), la integración del NP "*bitxi-dendan zegoen lepoko polita*" [*el precioso colgante que estaba en la joyería*] consume 2EUs: 1EU por la integración del NP y otro 1EU por el referente discursivo ("*neskari*") que interviene entre el NP y el verbo "*eman*" [dar]; y la integración del NP "*neskari*" [*a la chica*] solo consume 1EU por ser referente discursivo. En total la oración (8a) tiene un coste de integración de 11EUs. Por el contrario, en (4.7b) la integración del NP "*bitxi-dendan zegoen lepoko polita*" consume solo 1EU por ser referente discursivo; y la integración del NP "*neskari*" consume 4EUs: 1EU por ser referente discursivo y 3EUs más por los referentes discursivos ("*bitxi-dendan*", "*zegoen*", "*lepokoa*") que

[6] Los ejemplos originales de Gibson (1998) son los siguientes:
(a) *The young boy gave [to the girl] [the beautiful green pendant that had been in the jewelry store window for weeks]*
(b) *The young boy gave [the beautiful green pendant that had been in the jewelry store window for weeks] [to the girl]*

intervienen entre el NP y el verbo "*eman*". En total la oración (4.7b) tiene un coste de integración de 13 EUs. Así, en las lenguas OV las oraciones con sintagmas largos delante de cortos (cf. 4.7a) se procesan más fácil que las oraciones con sintagmas cortos delante de largos (cf. 4.7b), porque tienen un coste de integración menor.

(4.7) a.

	mutilak	[bitxi-dendan	zegoen	lepoko	polita]	[neskari]	eman	zion
IC	1+4	1	1	1	+1	1	1	0

EUs = 11

"el chico le dio el precioso colgante que estaba en la joyería a la chica "

b.

	mutilak	[neskari]	[bitxi-dendan	zegoen	lepoko	polita]	eman	zion
IC	1+4	1+3	1	1	1	+0	1	0

EUs = 13

"el chico le dio el precioso colgante que estaba en la joyería a la chica "

En definitiva, tanto la PGCH de Hawkins (1994, 2004, 2014) como la DLT de Gibson (1998, 2000) predicen las mismas preferencias de orden de palabras en las lenguas VO y OV, a pesar de que en la primera la preferencias no están correlacionadas con el coste de memoria y en la segunda sí; de que ambas calculan la complejidad de los órdenes de diferente manera. Ambas teorías predicen que las lenguas tratan de minimizar las dependencias recurriendo con mayor frecuencia al uso de un orden u otro, de tal forma que las lenguas VO prefieren tener sintagmas cortos delante de los largos y las lenguas OV largos delante de los cortos. Estas preferencias se han confirmado en estudios de corpus y experimentales: tanto en lenguas VO (polaco: Siewierska, 1993; inglés: Wasow, 1997b, 1997a; Stallings, MacDonald y O'Seaghdha, 1998; Arnold, Wasow, Losongco y Ginstrom, 2000; Hawkins, 2000; ruso: Kizach, 2012), como en lenguas OV (japonés: Yamashita y Chang, 2001; Yamashita, 2002; Kondo y Yamashita, 2011; coreano, Choi, 2007; Dennison, 2008; euskera, Ros et al., 2015). En lenguas VO, por ejemplo, Stallings et al. (1998) examinan la producción de oraciones de inglés con sintagmas nominales largos (NP_L) y cortos (NP_C) y sintagmas preposicionales cortos (PP_C), y encuentran que los hablantes de inglés producían con mayor frecuencia sintagmas cortos delante de largos (PP_C-NP_L) que largos delante de cortos (NP_L-PP_C). Además, los participantes tardan más tiempo en empezar a producir oraciones con el orden sintagmas largos-cortos (NP_L-PP_C = 744 ms) que con el orden sintagmas cortos-largos (PP_C-NP_L = 699 ms). Arnold et al. (2000), en otro experimento, examinan la producción de oraciones ditransitivas en inglés, y encuentran que los hablantes producen con mayor frecuencia el objeto indirecto (OI) delante del objeto directo (OD) cuando el indirecto es más corto que el directo (OI_C-OD_L = *give* $_{OI}$*[to the rabbit]* $_{OD}$*[the large orange carrot]* [*da* $_{OI}$*[al conejo]* $_{OD}$*[la zanahoria naranja grande]*]). Del mismo modo, los participantes producen con mayor frecuencia el objeto directo

delante del indirecto, cuando este último es más largo (ODC-OIL = *give* OD*[the carrot]* OI*[to the small white rabbit]* [*da* OD*[la zanahoria]* OI*[al pequeño conejo blanco]*]).

En cuanto a lenguas OV, Yamashita y Chang (2001) han explorado en japonés la producción de oraciones transitivas y ditransitivas en las que la longitud del sujeto y los objetos variaban. En las oraciones transitivas, encuentran que los hablantes de japonés tienden a posicionar el objeto directo delante del sujeto cuando es largo (e.g., OD*[se-ga takakute gassiri sita hannin-o]* S*[keezi-ga] oikaketa* [OD*[el sospechoso que el es alto y huesudo]* S*[el detective] atrapó*]) que cuando es corto (e.g., S*[keezi-ga]* OD*[se-ga takakute gassiri sita hannin-o] oikaketa* [S*[el detective]* OD*[el sospechoso que el es alto y huesudo] atrapó]*]). En las oraciones ditransitivas, los participantes también producen con mayor frecuencia los objetos directos e indirectos a principio de la oración cuando estos son largos (ODL-S-OIC-V o OIL-S-ODC-V) que cuando son cortos (ODC-S-OIL-V o OIC-S-ODL-V). Ros et al. (2015) han replicado el trabajo de Yamashita y Chang (2001) en euskera, que a diferencia del japonés tiene un orden de palabras más libre y permite constituyentes postverbales. Sus resultados muestran que los hablantes de euskera también prefieren posicionar los sintagmas largos delante de los cortos. En las transitivas, los participantes producen más oraciones con el objeto delante del sujeto cuando este es largo (OL-SC = 19,7%) que cuando es corto (OC-SL = 5,5%). En las ditransitivas, de igual modo, los participantes producen más oraciones con el objeto indirecto delante del directo cuando este es largo (OIL-ODC = 52,4%) que cuando es corto (OIC-ODL = 37,3%); del mismo modo, producen más oraciones con el objeto directo delante del indirecto cuando este es largo (ODL-OIC = 60,1%) que cuando es corto (ODC-OIL = 33%).

4.2.2 Procesamiento y construcciones gramaticales

Ueno y Polinsky (2009) abordan la interacción entre el orden básico de palabras (VO-OV) y el procesamiento desde una perspectiva diferente. Mientras Hawkins (1994, 2004, 2014) y Gibson (1998, 2000) se centran en el orden de los sintagmas según su peso (i.e., en su longitud en número de palabras), Ueno y Polinsky (2009) tienen como objetivo investigar si las lenguas VO y las lenguas OV recurren con distinta frecuencia al empleo de ciertas construcciones sintácticas como estrategia para facilitar el procesamiento. Ueno y Polinsky (2009) parten de la hipótesis de que las relaciones argumentales de la oración se resuelven al procesar el verbo (Trueswell et al., 1993; MacDonald, Pearlmutterb y Seidenberg, 1994; Garnsey et al., 1997; Vosse y Kempen, 2000), de forma que en las lenguas OV serían más difíciles de procesar porque han de mantener dos argumentos (i.e., el sujeto (S) y el objeto (O)) en la memoria hasta procesar el verbo, mientras que las lenguas VO solo deben de mantener un argumento, el sujeto (S). Así, las lenguas OV presentarían un mayor coste de procesamiento en comparación a las lenguas VO por retener más argumentos en la

memoria de trabajo. Ueno y Polinsky (2009) sugieren que las lenguas VO-OV recurren con mayor frecuencia a estructuras sintácticas como oraciones con sujeto omitido y oraciones intransitivas para reducir el número de argumentos preverbales y así minimizar el tiempo para poder procesar cuanto antes el verbo (Lindsley, 1975). Basándose en esos dos tipos de estructuras (i.e., el uso sujetos omitidos y oraciones intransitivas), Ueno y Polinsky (2009) predicen que (a) tanto las lenguas OV como las lenguas VO recurrirán a la omisión del sujeto para reducir el número de argumentos ya que en ambos tipos de lenguas el sujeto está en posición preverbal, y (b) que las lenguas OV recurrirán con mayor frecuencia que las lenguas VO a los predicados intransitivos, porque al no tener O, se reduce el número de argumentos preverbales, mientras que en una lengua VO la reducción del O no afecta a la carga de memoria requerida previa al procesamiento del V.

Con la intención de testear sus predicciones, Ueno y Polinsky (2009) llevan a cabo dos estudios: un estudio de corpus y un estudio de producción. En el estudio de corpus compraran dos corpus escritos: uno en inglés (VO) y otro en japonés (OV). El corpus consta de 2400 oraciones (300 oraciones x 4 géneros x 2 lenguas). A su vez, el corpus está compuesto por textos de diferentes géneros: revistas de decoración, novelas de misterio, libros de política y oraciones infantiles tomadas de CHILDES[7]. A la hora de etiquetar las oraciones solo tienen en cuenta las oraciones principales y las etiquetan manualmente por tipo de predicado y tipo de omisión. Según el tipo de predicado, diferencian dos tipos de oraciones: intransitivas y transitivas; y según el tipo de omisión tres tipos de oraciones: con sujeto omitido, con objetos omitidos y con sujeto y objetos omitidos. A pesar de que el inglés no tiene omisión de sujeto, consideran las elisiones de sujeto del habla de los niños del CHILDES como casos de omisión basándose en los estudios de Bloom (1970); Braine (1976); Mazuka, Lust, Wakayama y Snyder (1986); Bloom (1990). Sus resultados indican que en ambas lenguas la frecuencia de omisión de sujeto es mayor en las oraciones transitivas (inglés infantil: 19%; y japonés: 38%) que en las oraciones intransitivas (inglés infantil: 3%; y japonés: 21%). En lo referente al uso de tipo de oraciones, observan una mayor frecuencia de uso de intransitivas en japonés (73%) en comparación con el inglés (51%). Los resultados son consistentes con sus dos predicciones: (a) la omisión de sujeto se usa con mayor frecuencia en oraciones transitivas que en intransitivas en ambas lenguas (inglés (VO) y japonés (OV)), y (b) el japonés (OV) usa con mayor frecuencia predicados intransitivos que el inglés (VO). En el segundo estudio, el de producción, tratan de replicar los resultados del estudio anterior en más lenguas VO y OV. Para ello, comparan las oraciones producidas por hablantes nativos de inglés (VO), español (VO), japonés (OV) y

[7] CHILDES es un corpus oral de transcripciones de habla de niños de diferentes lenguas que se utiliza normalmente en estudios de adquisición del lenguaje.

turco (OV) a la hora de describir las viñetas del libro infantil *Frog, where are you?*[8] (Mayer, 1969). En suma, analizan 1211 oraciones producidas por 30 hablantes: 473 oraciones de inglés (10 hablantes), 275 oraciones de japonés (10 hablantes), 198 oraciones de castellano (5 hablantes) y 265 oraciones de turco (5 hablantes). Al igual que en el estudio de corpus, etiquetan las oraciones manualmente según el tipo de predicado (oraciones intransitivas y transitivas) y el tipo de omisión de argumentos (de sujeto, de objeto y, de sujeto y objeto). Los resultados de producción son consistentes con los del estudio de corpus: los hablantes de castellano, japonés y turco producen con mayor frecuencia oraciones con argumentos omitidos en oraciones transitivas (castellano: 67%; japonés: 33%; y turco: 51%) que en oraciones intransitivas (castellano: 38%; japonés: 12%; y turco: 31%); y los hablante de lenguas OV omitían con mayor frecuencia los argumentos en oraciones intransitivas (japonés: 64%; y turco: 56%) que los hablantes de lenguas VO (inglés: 50%; y castellano: 45%). Teniendo en cuenta los resultados de ambos estudios, Ueno y Polinsky (2009) concluyen que las lenguas OV son más costosas de procesar cuando tienen todos los argumentos expresados, por lo que tratan de reducir el número de argumentos preverbales usando con mayor frecuencia oraciones intransitivas. El caso del de la omisión de argumentos, sin embargo, sugieren que su uso no exclusivo de las lenguas OV, sino que también de las lenguas VO; por lo que sugieren que es un principio universal de economía utilizado en todas las lenguas para producir oraciones más cortas.

Los resultados de Ueno y Polinsky (2009) son bastante curiosos en dos casos. El primero, el hecho de encontrar en castellano (VO) más casos de argumentos omitidos en oraciones transitivas que en japonés y turco (ambas OV), cuando el uso de omisión de argumentos ayudaría a aligerar el área preverbal en estas dos lenguas OV. El segundo, el hecho de que la diferencia de uso de oraciones intransitivas entre el inglés y el turco no sea significativa, mientras que entre el inglés y el japonés sí lo es. Ueno y Polinsky (2009) sugieren que ello puede deberse a que el turco, a diferencia del japonés, tiene un orden de palabras menos rígido, pudiendo mover argumentos a posición postverbal. Sin embargo, no hacen ninguna predicción de si el uso de argumentos postverbales ayudaría a reducir el área preverbal. Por ende, a continuación, voy a investigar si las dos estrategias propuestas por Ueno y Polinsky (2009) se observan en euskera (OV) y castellano (VO).

4.3 Estudio de corpus en euskera y castellano

En el presente estudio de corpus he examinado si en euskera se recurre a reducir el número de argumentos con mayor frecuencia que en castellano, mediante la omisión de

[8] El libro *Frog, where are you?* consta solo de dibujos y narra la historia de cómo un niño y su perro tratan de encontrar su rana perdida.

argumentos, el uso de oraciones intransitivas y el uso de argumentos postverbales. Como he explicado en la sección anterior (sección 4.3.2), Ueno y Polinsky (2009) encuentran que las lenguas OV como el euskera tienden a usar argumentos omitidos y oraciones intransitivas con mayor frecuencia para reducir el número de argumentos preverbales. Por tanto, voy a tratar de replicar los resultados de Ueno y Polinsky (2009) utilizando un corpus más extenso en castellano (SVO) y euskera (SOV). Los motivos que me llevan a considerar estas dos lenguas es que tanto el castellano (cf. 4.8) como el euskera (cf. 4.9), siendo dos lenguas VO y SOV respectivamente, tienen una mayor libertad de orden de palabras que el inglés y japonés, lo cual les permite tener sujetos y objetos postverbales (castellano: Zagona, 2002; Batchelor y San José, 2010; Hualde, Olarrea, Escobar y Travis, 2010; euskera: Euskaltzaindia, 1991; Laka, 1996; Hualde y Ortiz de Urbina, 2003). Esta característica es interesante, en especial en euskera, porque podría ser otra estrategia para reducir el número de argumentos a retener en la memoria de trabajo hasta poder procesar el verbo moviendo uno de ellos después del verbo.

(4.8) castellano

a. [S]el hombre [V]ha visto [O]a la mujer	SVO [orden básico]
b. [S]el hombre [O]a la mujer [V]ha visto	SOV
c. [O]a la mujer [S]el hombre [V]ha visto	OSV
d. [O]a la mujer [V]ha visto [S]el hombre	OVS
e. [V]ha visto [S]el hombre [O]a la mujer	VSO
f. [V]ha visto [O]a la mujer [S]el hombre	VOS

(4.9) euskera

a. gizonak[S] emakumea[O] ikusi du[V]	SOV [orden básico]
b. gizonak[S] ikusi du[V] emakumea[O]	SVO
c. emakumea[O] gizonak[S] ikusi du[V]	OSV
d. emakumea[O] ikusi du[V] gizonak[S]	OVS
e. ikusi du[V] gizonak[S] emakumea[O]	VSO
f. ikusi du[V] emakumea[O] gizonak[S]	VOS

Además, ambas lenguas pueden omitir argumentos: el castellano solo el sujeto (cf. 4.10); y el euskera tanto el sujeto como el objeto (cf. 4.11) (castellano: Zagona, 2002; Hualde et al., 2010; euskera: Laka, 1996; Hualde y Ortiz de Urbina, 2003). Esta característica permite comparar si el euskera recurre con mayor frecuencia al uso de argumentos omitidos en oraciones transitivas en comparación al castellano, dado que tiene más opciones de omisión de argumentos.

(4.10) castellano

a. el profesor ha visto al alumno
b. *pro* ha visto al alumno [omisión de sujeto]
c. *el profesor ha visto *pro* [omisión de objeto]

(4.11) euskera

a. irakasleak ikaslea ikusi du
b. *pro* ikaslea ikusi du [omisión de sujeto]
c. irakasleak *pro* ikusi du [omisión de objeto]
d. *pro pro* ikusi du [omisión de sujeto y objeto]

De acuerdo con las predicciones de Ueno y Polinsky (2009) ambos tipos de lenguas VO y SOV recurren con mayor frecuencia al uso de argumentos omitidos en oraciones transitivas que en intransitivas, pero solo las lenguas OV utilizan más oraciones intransitivas que las lenguas VO. Así, en el caso de del euskera (OV) predigo que recurra con mayor frecuencia que el castellano (VO) al uso (a) de argumentos omitidos en oraciones transitivas y (b) de oraciones intransitivas. Además, también predigo que el euskera recurra con mayor frecuencia en oraciones transitivas al uso de argumentos postverbales (SVO-OVS) en lugar de tener todos los argumentos en posición preverbal (SOV-OSV).

4.3.1 Materiales

He comparado dos corpus escritos, uno de castellano (VO) y otro de euskera (OV). Estos corpus son equivalentes en naturaleza y tamaño a los corpus utilizados en el estudio de corpus de Ueno y Polinsky (2009). El corpus consta de 4000 oraciones (2000 oraciones x 2 lenguas), tomadas de tres géneros diferentes (periódicos, libros y revistas) para tener una muestra heterogénea: 1400 oraciones de periódicos (700 oraciones x 2 lenguas), 1400 oraciones de libros (700 oraciones x 2 lenguas), y 1200 oraciones de revistas (600 oraciones x 2 lenguas). Para el género periódicos, he utilizado *El Correo* para castellano y *Berria* para euskera. A su vez, he extraído las oraciones de diferentes secciones para tener un corpus más heterogéneo. Cada sección tiene el mismo número de oraciones (100 oraciones x 7 secciones): Economía, Sociedad, Mundo, Deportes, Cultura, Política y Nacional. En libros, he tomado en cuenta cuatro diferentes géneros para tener también un corpus lo más representativo posible de diferentes estilos discursivos. El número de oraciones es igual en todos ellos (175 oraciones x 4 géneros): Comedia, Misterio, Histórica y Ensayo (No-ficción). Por último, en revistas, he utilizado la revista *Muy Interesante* para el castellano y *Elhuyar Aldizkaria* para euskera. Al igual que en el género prensa, las oraciones han sido

tomadas de diferentes secciones para tener un corpus más heterogéneo: Historia, Cultura, Naturaleza, Salud, Tecnología y Ciencia (100 oraciones x 6 secciones).

4.3.2 Procedimiento

He etiquetado las oraciones manualmente según los criterios de Ueno y Polinsky (2009), i.e., según el tipo de predicado y el tipo de omisión. De acuerdo al tipo de predicado he clasificado las oraciones en "oraciones intransitivas" y "oraciones transitivas". Dentro del grupo "intransitivas" están incluidas oraciones con predicados intransitivos (cf. 4.12a), con predicados copulativos (cf. 4.12b) y con predicados impersonales, voces medias y pasivas (cf. 4.12c). El grupo "transitivas" está formado por oraciones con predicados transitivos con objeto NP (cf. 4.13a), con predicados transitivos con objeto CP (cf. 4.13b) y con predicados ditransitivos (cf. 4.13c). A continuación, se muestran ejemplos de oraciones intransitivas en castellano y euskera:

(4.12) Oraciones intransitivas

a. Predicados intransitivos

[S]El segundo encuentro [V]transita por los mismos derroteros [castellano]

Errepide mapa bat lantzeko unea[S] etorriko da[V] [euskera]
"Llegará el momento para elaborar un mapa de carreteras"

b. Predicados copulativos

[S]Thor no [pred.cop]es un simple superhéroe [castellano]

Aste honetako lanaren ardatza[S] defentsa izan da[Pred.Cop] [euskera]
"El eje del trabajo de esta semana ha sido la defensa"

c. Predicados impersonales, voces medias y pasivas

[S]Estas rocas [pasiva]son transportadas en dos grandes depósitos situados en la cubierta de la nave [castellano]

Zelan diagnostikatzen zen[imper] orain arte epilepsia[S] [euskera]
"Cómo se diagnosticaba hasta ahora la epilepsia"

En (4.13) aparecen ejemplos de oraciones transitivas en castellano y euskera. En el análisis estadístico los objetos directos nominales (NP) y oracionales (CP) no los he diferenciado, y ambos los he etiquetado como "objetos":

(4.13) Oraciones transitivas

a. Predicados transitivos con NP como objeto

[S]El edificio [V]tiene [O]una antigüedad aproximada de 70 años [castellano]

Herri honek[S] premia izugarria[O] du[V] [euskera]
"Este pueblo tiene mucha urgencia"

b. Predicados transitivos con CP como objeto

[S]El presidente Mohamed Hosni Mubarak [V]ha decidido [castellano]
[O]renunciar a su cargo de presidente de la República

Nik[S] beti esaten dut[V] hau opari gisa hartzen dugula[O] [euskera]
"Yo siempre digo que esto lo tomamos como regalo"

c. Predicados ditransitivos

[S]el Pontífice quiso [V]quiso dedicar [OD]el rezo del Ángelus [castellano]
[OI]al tema del pecado

Alemaniarrek[S] 20 milioi euro[OD] zor dizkiote[V] lantegi honi[OI] [euskera]
"Los alemanes deben 20 millones de euros a esta fábrica"

A la hora de etiquetar las oraciones de euskera, he etiquetado las construcciones de tipo "nombre + *egin*" como predicados verbales, i.e., como verbos, y no como verbo y objeto. Laka (1993) explica que, aunque el sujeto vaya marcado en caso ergativo, la construcción "nombre + *egin*" es tomada como una construcción inergativa en sus formas equivalente de otras lenguas. Por ello, he decido etiquetar dicha construcción como oración intransitiva (dentro de la subcategoría "predicado intransitivo"):

(4.14) ...Florentino Perezek herenegun hitz egin zuen gai honi buruz...
"...Florentino Perez habló anteayer sobre este tema..."

Según el criterio de omisión de argumentos, he clasificado las oraciones teniendo en cuentan si tienen o no argumentos omitidos. A su vez, he diferenciado tres grupos según el tipo de argumento que se omite: omisión de sujeto (S-drop), omisión de objeto (O-drop) y omisión de sujeto y objeto (SO-drop). A la hora de etiquetar los tipos de argumentos omitidos en euskera, los NPs con caso dativo los he etiquetado como objetos omitidos (cf. 4.15b).

(4.15) a. Omisión de sujeto (S-drop)
pro examinó las lecturas de los diversos monitores [castellano]

pro kontuan eduki ditut euskalkietako eta garaian garaiko bereizgarriak. [euskera]
"He tenido en cuenta las características de cada época de los dialectos"

b. Omisión de objeto (O-drop)
arropa eta aulkia ekarri zizkidan *pro* amak. [euskera
"la ropa y la silla me trajo (mi) madre"

c. Omisión de sujeto y objeto (SO-drop)
Eta orduan *pro pro* ikusi nuen. [euskera]
"Y entonces lo vi"

A la hora de analizar los datos del corpus he usado los análisis estadísticos prueba χ^2 (*Pearson chi-square*) y el modelo de regresión logística binomial. La prueba χ^2 (*Pearson chi-square*) la he usado para observar, por un lado, si la distribución del uso de argumentos omitidos entre las oraciones intransitivas y transitivas en ambas lenguas es significativamente diferente y, por el otro, si existe diferencias significativas en el uso de oraciones intransitivas frente a transitivas en ambas lenguas. Concretamente, he querido testear si ambas lenguas usan con mayor frecuencia argumentos omitidos en oraciones transitivas que en oraciones intransitivas, y si el euskera recurre con mayor frecuencia al uso de oraciones intransitivas que el castellano. El modelo de regresión logística binomial, por su parte, lo he utilizado para analizar (a) si el uso de argumentos omitidos está influenciado por el tipo de oración (intransitiva vs. transitiva) y por el tipo de lengua (VO vs. OV) y (b) si el uso de oraciones intransitivas está influenciado por el tipo de lengua (VO vs. OV). La finalidad de este análisis es observar si el hecho de que la lengua sea OV influye en un mayor uso de argumentos omitidos en oraciones transitivas y en un mayor uso de oraciones intransitivas frente a transitivas. Los análisis los he computado mediante el programa estadístico R (R Core Team, 2017) y usando el paquete *lme4* (Bates, Maechler, Bolker y Walker, 2015). Los resultados los he considerado significativos a un nivel $p < .05$. Los gráficos los he realizado con el paquete *ggplot2* (Wickham, 2009).

4.3.3 Resultados

La TABLA 4.1 muestra la clasificación de las 4000 oraciones que componen el corpus, según la lengua, el tipo de oración, y si las oraciones tienen argumentos omitidos o no.

		castellano						euskera					
		periódico		**libros**		**revista**		**periódico**		**libros**		**revista**	
intransitivas	sin omisión	43%	(304)	31%	(216)	47%	(281)	45%	(314)	40%	(277)	33%	(200)
	con omisión	8%	(53)	22%	(152)	3%	(21)	9%	(66)	11%	(80)	9%	(52)
transitivas	sin omisión	31%	(218)	23%	(161)	41%	(245)	24%	(166)	16%	(109)	31%	(186)
	con omisión	18%	(125)	24%	(171)	9%	(53)	22%	(154)	33%	(234)	27%	(162)
TOTAL		100%	(700)	100%	(700)	100%	(600)	100%	(700)	100%	(700)	100%	(600)

TABLA 4.1. Distribución de las oraciones del corpus según la lengua (castellano y euskera), el género (periódico, libros, revista), el tipo de oración (intransitivas y transitivas), y la omisión de argumentos.

4.3.3.1 Omisión de argumentos preverbales

Los resultados de este estudio comparativo de corpus muestran que, en lo referente a la distribución de argumentos omitidos en oraciones intransitivas y transitivas, ambas lenguas hacen un mayor uso de argumentos omitidos en oraciones transitivas que en intransitivas. En castellano, la omisión de argumentos se da en el 35,86% de las oraciones transitivas frente al 22% en las oraciones intransitivas [χ^2 (1, N = 2000) = 46.198, p< .001, ϕ = 0.153] (GRÁFICO 4.1). Este uso mayoritario de argumentos omitidos en oraciones transitivas también ocurre en cada uno de los géneros: periódico [intra(nsitivas) vs. tran(sitivas): 14,84% vs. 36,44%, χ^2 (1, N = 700) = 41.898, p< .001, ϕ = 0.248]; libros [intra vs. tran: 41,3% vs. 51,5%, χ^2 (1, N = 700) = 6.9047, p< .008, ϕ = 0.102]; y revista [intra vs. tran: 6,95% vs. 17,78%, χ^2 (1, N =600) = 15.289, p< .001, ϕ = 0.165].

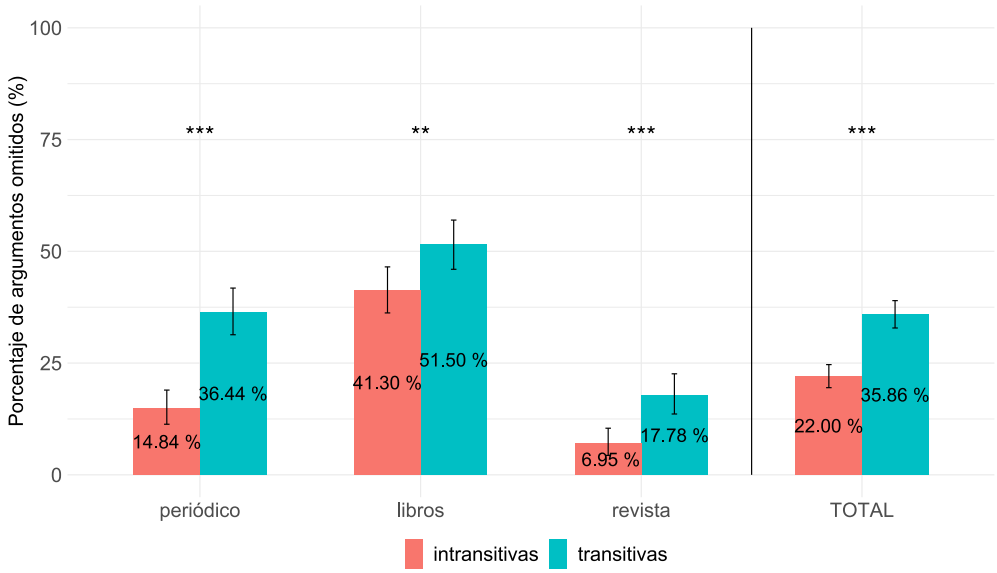


GRÁFICO 4.1. Frecuencia de oraciones intransitivas y transitivas en castellano con argumentos omitidos, agrupadas por géneros. Las barras de error muestran los intervalos de confianza de 95%.

En euskera, al igual que en castellano, el uso de argumentos omitidos también es significativamente mayor en las oraciones transitivas (54%) que en las oraciones intransitivas (20%) [χ^2 (1, N = 2000) = 245.72, p< .001, ϕ = 0.352] (GRÁFICO 4.2). De igual modo, en cada uno de los géneros la omisión de argumentos se da con mayor frecuencia en las oraciones transitivas que en las oraciones intransitivas: periódico [intra vs. tran: 17,4% vs. 47,5%, χ^2 (1, N = 700) = 72.15, p< .001, ϕ = 0.324], libros [intra vs. tran: 22,4% vs. 67,9%, χ^2 (1, N = 700) = 144.8, p< .001, ϕ = 0.458] y revista [intra vs. tran: 20,6% vs. 46,3%, χ^2 (1, N = 600) = 40.818, p< .001, ϕ = 0.264].

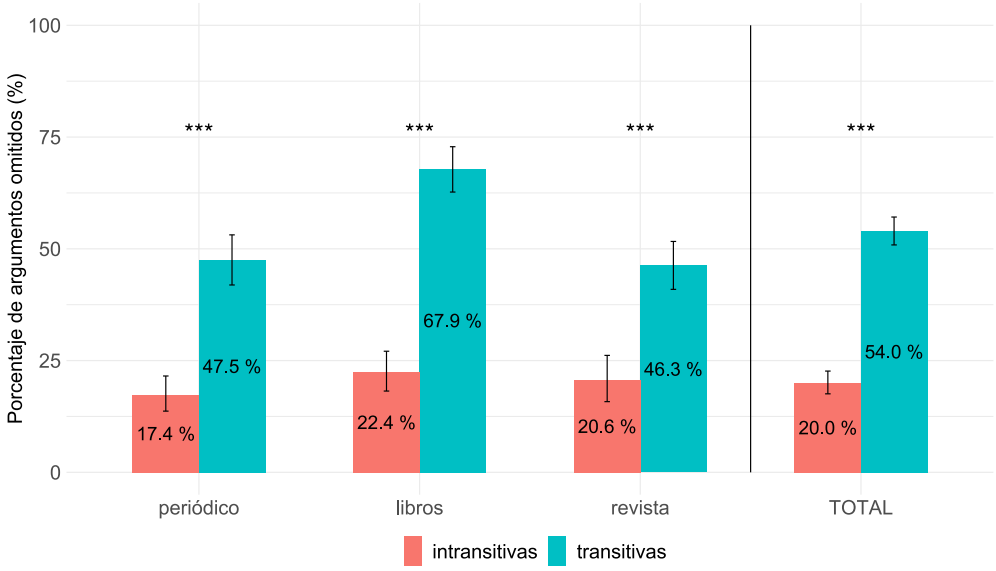


GRÁFICO 4.2. Frecuencia de oraciones intransitivas y transitivas en euskera con argumentos omitidos, agrupadas por géneros. Las barras de error muestran los intervalos de confianza de 95%.

En lo referente al tipo de argumentos omitidos, la TABLA 4.2 muestra que en ambas lenguas la omisión de sujeto es la más frecuente (castellano: 100%; y euskera: 85%). En euskera, que también puede omitir el objeto, la omisión de este es mucho menos frecuente (7%) respecto a la del sujeto y, de igual modo, la omisión de ambos argumentos también es menos frecuente (8%) frente a la omisión del sujeto[9]. Estas mismas frecuencias se observan en cada uno de los géneros en ambas lenguas.

[9] Contabilizando los casos de SO-drop en euskera dentro de los argumentos omitidos de sujeto (S-drop) y de objeto (O-drop), el sujeto seguiría siendo el tipo de argumento que más se omite:

	S-drop + SO-drop	O-drop + SO-drop
periódico	89% (206 = 194 + 12)	11% (26 = 14 +12)
libros	79% (284 = 238 + 46)	21% (76 = 30 + 46)
revista	95% (208 = 204 + 4)	5% (10 = 6 + 4)
TOTAL	86% (698 = 636 + 62)	14% (112 = 50 + 62)

	castellano				euskera			
	S-drop	**O-drop**	**SO-drop**	**Total**	**S-drop**	**O-drop**	**SO-drop**	**Total**
periódico	100% (178)	0% (0)	0% (0)	100% (178)	88% (194)	6% (14)	6% (12)	100% (220)
libros	100% (323)	0% (0)	0% (0)	100% (323)	76% (238)	10% (30)	15% (46)	100% (314)
revista	100% (74)	0% (0)	0% (0)	100% (74)	95% (204)	3% (6)	2% (4)	100% (214)
TOTAL	100% (575)	0% (0)	0% (0)	100% (575)	85% (636)	7% (50)	8% (62)	100% (748)

TABLA 4.2. Distribución de los tipos de argumentos omitidos (omisión de sujeto [S-drop]; de objeto [O-drop]; de sujeto y objeto [SO-drop]) en castellano y euskera por género.

Por tanto, como se observa en los GRÁFICOS 4.1 y 4.2, las oraciones transitivas tienen más casos de argumentos omitidos que las intransitivas en ambas lenguas. Si se compara el uso de argumentos omitidos entre ambas lenguas, se observa que en euskera (54%) la omisión de argumentos en oraciones transitivas es significativamente mayor que en castellano (35,9%) [castellano vs. euskera: 35,9% vs. 54%, χ^2 (1, N = 1323) = 24.002, $p < .001$, $\phi = 0.136$] (GRÁFICO 4.3a). Esta mayor frecuencia observada en euskera podría deberse a que el euskera tiene tres tipos de argumentos omitidos (de sujeto, de objeto y, de sujeto y objeto), mientras que el castellano solo tiene de sujeto. Computar tres tipos de argumentos omitidos en euskera frente a un solo tipo en castellano hace que la comparación no sea equiparable. Por ello, he llevado a cabo una nueva comparación más equiparable teniendo en cuenta solo los casos de argumentos omitidos de sujeto en oraciones transitivas en ambas lenguas. En euskera, los casos de omisión de sujeto y objeto (SO-drop) los he contabilizado también como de sujeto (S-drop). Los resultados de este nuevo análisis muestran la misma preferencia encontrada anteriormente: en euskera (51,9%) la frecuencia de sujetos omitidos en oraciones transitivas es significativamente mayor que en castellano (35,9%) [castellano vs. euskera: 35,9% vs. 51,9%, χ^2 (1, N = 1273) = 18.378, $p < .001$, $\phi = 0.122$] (GRÁFICO 4.3b).

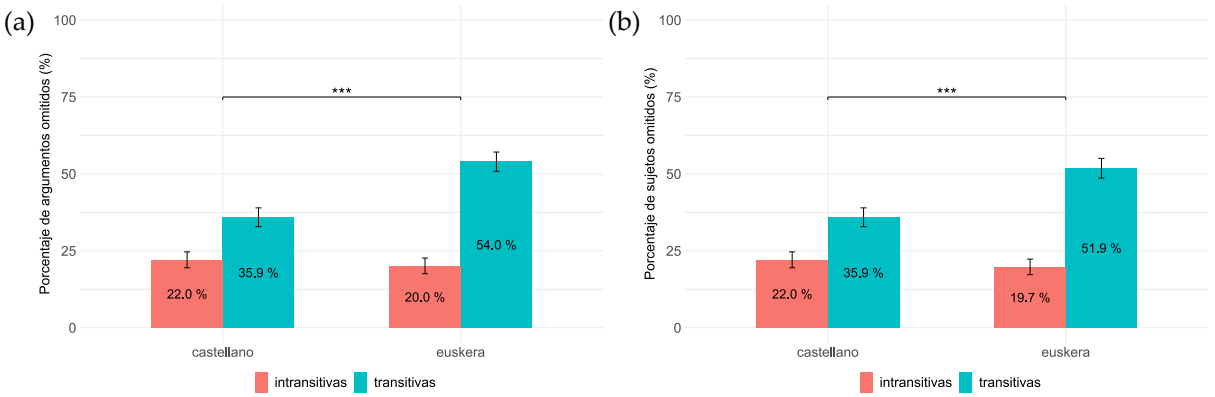


GRÁFICO 4.3. Comparación entre castellano y euskera de la frecuencia de (a) argumentos omitidos en oraciones intransitivas y transitivas, y (b) sujetos omitidos (S-drop) en oraciones intransitivas y transitivas. Las barras de error muestran los intervalos de confianza de 95%.

El modelo de regresión logística, por su parte, sustenta la predicción de que la omisión de argumentos es mayor en las oraciones transitivas que en las oraciones intransitivas [β = 0.684, z = 0.100, p < .001, odds ratio = 1.98, 95% CI = 1.62 – 2.41] (TABLA 4.3), i.e., la probabilidad de argumentos omitidos es 1,98 veces mayor en las oraciones transitivas que en las intransitivas. Además, existe una interacción entre el tipo de oración y lengua: cuando las oraciones son transitivas hay una mayor probabilidad (2,4 veces) de que los argumentos se omitan en euskera que en castellano [β = 0.877, z = 0.142, p < .001, odds ratio = 2.40, 95% CI = 1.81 – 3.18]. Por tanto, el modelo de regresión logística predice que tanto el tipo de oración como el tipo de lengua influye significativamente en la probabilidad de omitir argumentos (pro-drop).

Omisión – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	-1.26533	0.07532	-16.799	0.001 ***
transitivas	0.68425	0.10070	6.795	0.001 ***
euskera	-0.11970	0.10949	-1.093	0.274
transitivas:euskera	0.87730	0.14298	6.136	0.001 ***
index *C*: 0.55				

TABLA 4.3. Resultados del modelo de regresión logística para el uso de argumentos omitidos según tipo de oración y lengua.

En resumen, ambas lenguas tienden a omitir argumentos del área preverbal. En euskera, a diferencia del castellano, el uso de argumentos omitidos es significativamente mayor. Ello no se debe a que el euskera pueda omitir todos los argumentos (i.e., el sujeto y el objeto), ya que restringiendo la comparación entre ambas lenguas a la omisión de sujeto, el euskera sigue mostrando un mayor uso de omisión de sujeto que el castellano. Por tanto,

replico los resultados en castellano, japonés y coreano del estudio de Ueno y Polinsky (2009).

4.3.3.2 Uso de oraciones intransitivas

Respecto al uso de oraciones intransitivas (TABLA 4.5), en castellano (GRÁFICO 4.4a) la frecuencia de uso de ambos tipos de oraciones no es significativamente diferente [intra(nsitivas) vs. tran(sitivas): 51,4% vs. 48,7%, χ^2 (1, N = 2000) = 1.458, p = .227, V = 0.027]. En cada uno de los géneros por separado tampoco se encuentran diferencias significativas: periódico [intra vs. tran: 51% vs. 49%, χ^2 (1, N = 700) = 0.28, p = .596, V = 0. 02]; libros [intra vs. tran: 52,6% vs. 47,4%, χ^2 (1, N = 700) = 1.851, p = .173, V = 0.051]; y revista [intra vs. tran: 50,3% vs. 49,7%, χ^2 (1, N =600) = 0.026, p = .870, V = 0.006]. De igual modo, en euskera (GRÁFICO 4.4b) no se observan diferencias significativas en la frecuencia de uso de ambos tipos de oraciones [intra vs. tran: 49,5% vs. 50,5%, χ^2 (1, N = 2000) = 0.242, p = .622, V = 0.011]. En los géneros por separado se encuentran diferencias significativas en dos de los géneros: en el género periódico las oraciones intransitivas se usan con mayor frecuencia que las oraciones transitivas [intra vs. tran: 54,3% vs. 45,7%, χ^2 (1, N = 700) = 5.142, p < .023, V = 0. 08]; en el género revista, por el contrario, las oraciones transitivas se usan significativamente con mayor frecuencia que las oraciones intransitivas [intra vs. tran: 42% vs. 58%, χ^2 (1, N =600) = 15.36, p < .001, V = 0.16]. En el género libros no hay diferencia entre el uso de oraciones intransitivas y transitivas [intra vs. tran: 51% vs. 49%, χ^2 (1, N = 700) = 0.28, p = .596, V = 0.02].

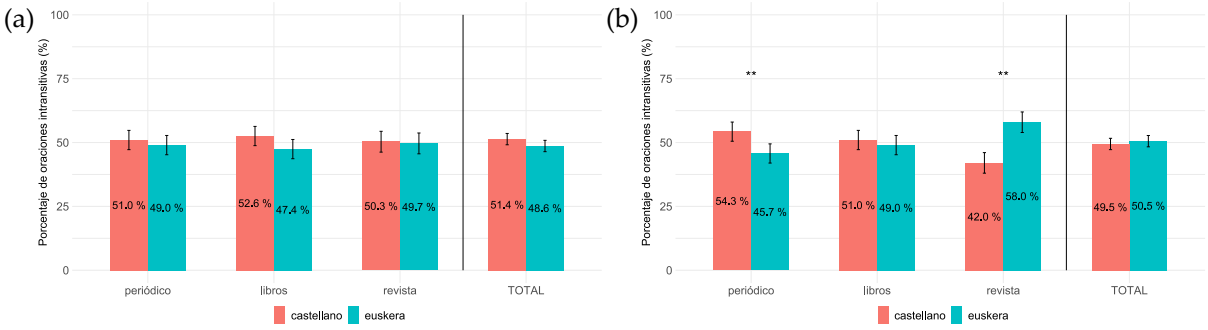


GRÁFICO 4.4. Frecuencia de oraciones intransitivas frente a oraciones transitivas en (a) castellano y (b) euskera, agrupadas por géneros. Las barras de error muestran los intervalos de confianza de 95%.

En la comparación del uso de oraciones intransitivas entre ambas lenguas no existen diferencias significativas: tanto el castellano (51,4%) como el euskera (49,5%) utilizan con una frecuencia similar oraciones intransitivas [castellano vs. euskera: 51,4% vs. 48%, χ^2 (1, N = 4000) = 1.3691, p = .242, ϕ = 0.019] (GRÁFICO 4.4). Sin embargo, en la comparación por géneros se observa una diferencia significativa en el género revista: el castellano (50,3%)

muestra una mayor frecuencia de oraciones intransitivas que el euskera (42%) [castellano vs. euskera: 50,3% vs. 42%, χ^2 (1, N = 1200) = 8.050, p = .004, ϕ = 0.084]. En el resto de géneros no hay diferencias significativas entre el castellano y el euskera (periódico [castellano vs. euskera: 51% vs. 54,3%, χ^2 (1, N = 1400) = 1.3867, p = .239, ϕ = 0.033] y libros [castellano vs. euskera: 52,6% vs. 51 %, χ^2 (1, N = 1400) = 0.28608, p = .592, ϕ = 0.016]). En un segundo análisis, he realizado una regresión logística para evaluar si el euskera, por ser una lengua SOV, afecta en la frecuencia de uso de oraciones intransitivas. Los resultados del modelo muestran que la probabilidad de uso de oraciones intransitivas es menor en euskera que en castellano [β = -0.07601, z = 0.06326, p = .230, odds ratio = 0.92, 95% CI = 0.81 – 1.04] (TABLA 4.6), i.e., el modelo predice que el castellano tienen 1,05 veces más de probabilidades de recurrir al uso de oraciones intransitivas que el euskera, aunque tal probabilidad no es significativa.

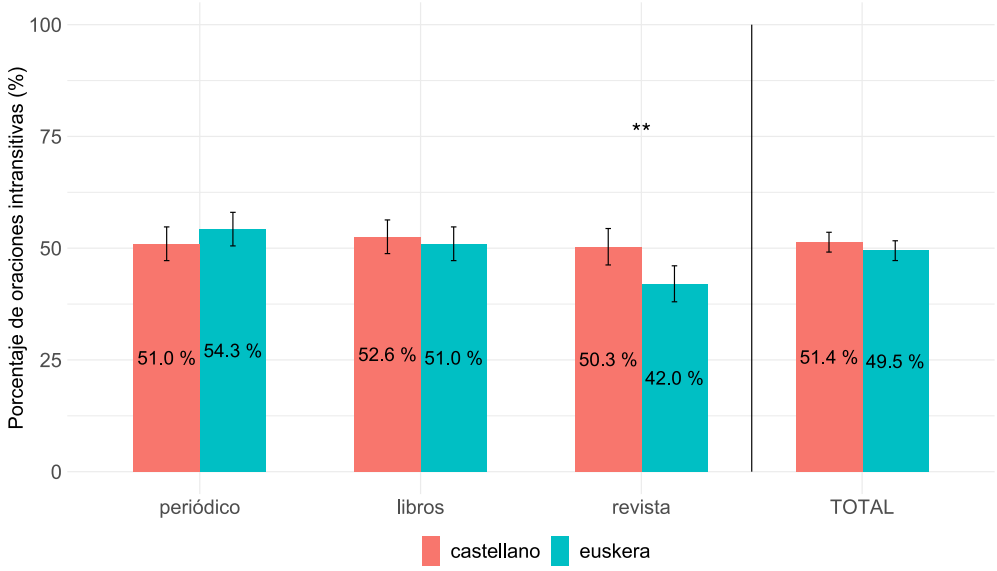


GRÁFICO 4.5. Comparación entre castellano y euskera de la frecuencia de oraciones intransitivas y transitivas. Las barras de error muestran los intervalos de confianza de 95%.

Intransitivas – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	0.05401	0.04474	1.207	0.227
euskera	-0.07601	0.06326	-1.202	0.230
index *C*: 0.26				

TABLA 4.4. Resultados del modelo de regresión logística para el uso de oraciones intransitivas según la lengua.

Teniendo en cuenta los resultados, en general la frecuencia de oraciones intransitivas es similar en castellano y en euskera. La única excepción se observa en el género revista, en

el que el castellano muestra un mayor uso de oraciones intransitivas que el euskera. Por consiguiente, estos resultados no son consistentes con los de Ueno y Polinsky (2009), quienes encuentran una mayor frecuencia de oraciones intransitivas en japonés y turco, ambas leguas SOV como el euskera.

4.3.3.3 El uso de argumentos postverbales

Como puede observarse en el GRÁFICO 4.5, las oraciones transitivas en euskera tienden a tener menos de dos argumentos expresados. Casi la mitad de las oraciones transitivas tienen solo un argumento en posición preverbal (48,6%) y un 20% sin ningún argumento preverbal. Es decir, que más de la mitad de las oraciones transitivas tienen el área preverbal reducida. La diferencia entre oraciones transitivas con menos de dos argumentos preverbales (0 o 1 argumentos) y con más de dos argumentos preverbales (2 o 3 argumentos) es significativa [χ^2 (3, N = 872) = 566.39, p< .001, V = 0.47].

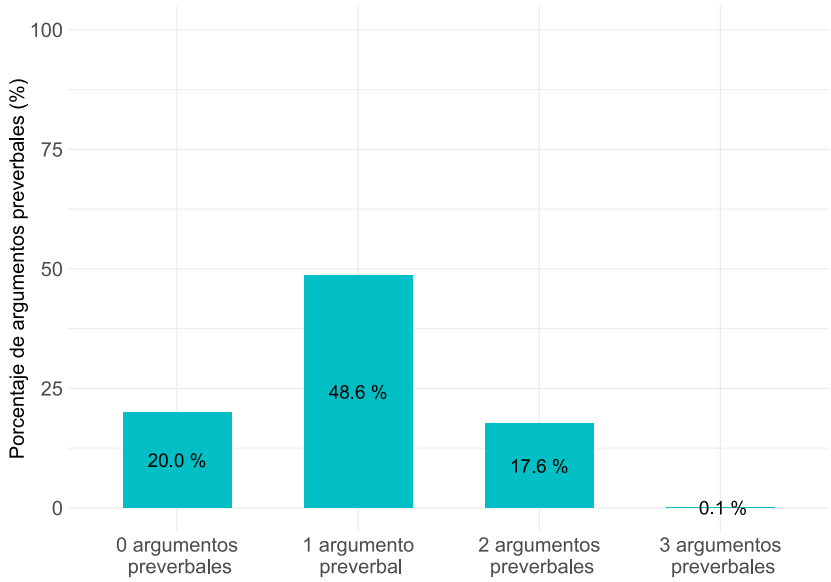


GRÁFICO 4.6. Frecuencia de oraciones transitivas en euskera por número de argumentos preverbales.

Como se ha visto, uno de los mecanismos que ayudan a reducir el área preverbal de las oraciones es el uso de argumentos omitidos. Otro podría ser el uso de argumentos postverbales. Como he mostrado en la sección 4.3 (cf. 4.9), el euskera puede tener argumentos después del verbo. Debido a la libertad que tiene el euskera de mover los argumentos de la oración, cosa que no pueden hacer el japonés (Akiyama y Akiyama, 2002; Kaiser, Ichiwaka, Kobayashi y Yamamoto, 2010), esta podría ser otra estrategia que podría reducir el área preverbal en las oraciones transitivas. Para determinar si la reducción de argumentos preverbales se debe también a la posibilidad de mover argumentos a posición postverbal he llevado a cabo un análisis de regresión logística. La

reducción de los argumentos la he analizado en cuatro posibles condiciones: SOV-OSV (i.e., no hay reducción), omisión (de sujeto, objeto o ambos), postverbales (sujeto, objeto o ambos postverbales) y la combinación de argumentos omitidos y postverbales. El análisis muestra que la probabilidad de reducir el número de argumentos preverbales en las oraciones transitivas mediante el uso de argumentos postverbales es similar a la de argumentos omitidos (TABLA 4.5: 3.89 veces y 4.88 veces, respectivamente), aunque no difieren significativamente [χ^2 (1, N = 587) = 0.16332, p= .686, V = 0.03]. Por tanto, el modelo de regresión logística muestra que existe una preferencia por reducir los argumentos preverbales en oraciones transitivas mediante el uso de argumentos omitidos como el uso de argumentos postverbales (GRÁFICO 4.6).

Oraciones transitivas – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	1.473	525.857	0.003	0.998
SOV-OSV	-22.039	1150.831	-0.019	0.985
Pro-drop	1.587	525.857	0.003	0.998
Postverbales	1.360	525.857	0.003	0.998
Pro-drop + postverbales	19.093	1200.352	0.016	0.987
index *C*: 0.90				

TABLA 4.5. Resultados del modelo de regresión logística para el tipo de reducción de argumentos preverbales en oraciones.

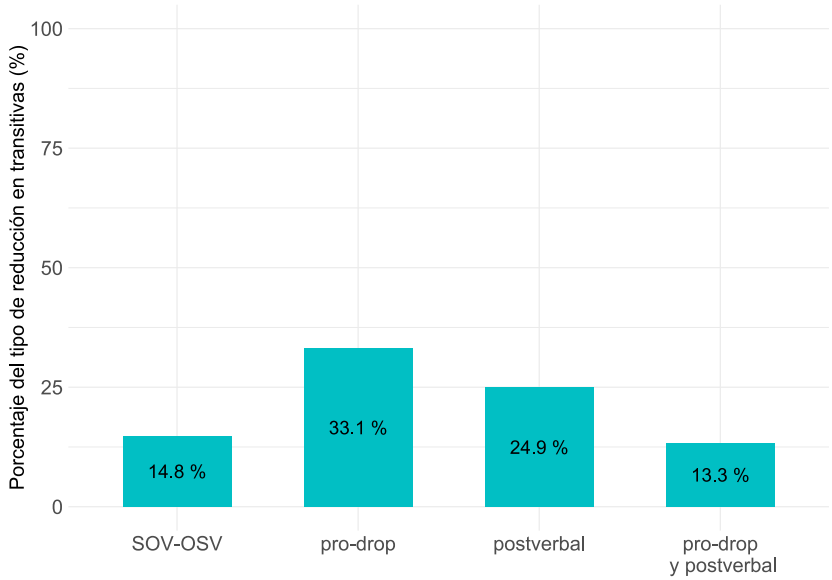


GRÁFICO 4.7. Frecuencia de tipos de reducción de argumentos preverbales en oraciones transitivas en euskera.

Los resultados de esta sección muestran claramente que el euskera tiene una preferencia por reducir el número de argumentos que aparecen antes del verbo. Esta reducción la

consigue tanto mediante la omisión de argumentos como con el movimiento de argumentos a posición postverbal. Aun así, se observa una preferencia por la omisión de argumentos frente al uso de argumentos postverbales.

4.4 Discusión

En este capítulo he llevado a cabo un estudio de corpus en castellano y euskera para examinar la influencia del orden de palabras en el uso de las dos estrategias (omisión de argumentos y uso de oraciones intransitivas) propuestas por Ueno y Polinsky (2009) para reducir el número de argumentos preverbales que se han de mantener en la memoria antes de procesar el verbo. Los resultados del estudio de corpus muestran, tanto en castellano como en euskera, una preferencia de usar con mayor frecuencia argumentos omitidos en oraciones transitivas que en intransitivas. Además, al comparar el uso de argumentos omitidos entre ambas lenguas, el euskera muestra un mayor uso de argumentos omitidos que el castellano. En cuanto al uso de oraciones intransitivas, los resultados no muestran diferencias en ninguna de las lenguas ni entre las lenguas: ambas usan con una frecuencia similar las oraciones intransitivas y transitivas. Sin embargo, sí que se observa una diferencia en el género libros: el castellano usa más oraciones intransitivas que el euskera. Por tanto, los resultados sustentan solo en parte las predicciones planteadas por Ueno y Polinsky, 2009: (a) la omisión de argumentos es una estrategia universal que utilizan las lenguas VO y OV para reducir el número de argumentos en oraciones transitivas; pero (b) el uso el uso de oraciones intransitivas no parece ser una estrategia específica de las lenguas OV para reducir el número de argumentos.

4.4.1 La omisión de argumentos para reducir el área preverbal

Los resultados del estudio de corpus revelan una mayor frecuencia de argumentos omitidos en oraciones transitivas que en oraciones intransitivas tanto en castellano como en euskera. Argumenté que el euskera recurriría con mayor frecuencia que el castellano a la omisión de argumentos en oraciones transitivas para reducir el coste de procesamiento que supone mantenerlos en la memoria. Siguiendo la hipótesis con la que parten Ueno y Polinsky (2009) de que en la posición del verbo se resuelven las relaciones argumentales (Boland, Tanenhaus y Garnsey, 1990; Trueswell et al., 1993; MacDonald et al., 1994; Garnsey et al., 1997; Vosse y Kempen, 2000), el euskera al tener dos argumentos preverbales (el sujeto y el objeto) y tener que mantenerlos en la memoria hasta que procese el verbo le supondría un coste de procesamiento mayor que al castellano, que solo ha de mantener el sujeto en la memoria; por lo que recurriría con mayor frecuencia a omitir uno de ellos. Los datos del corpus muestran que, efectivamente, el euskera recurre con mayor frecuencia que el castellano a la omisión de argumentos en las oraciones transitivas para

reducir el área preverbal. Estos resultados convergen con los del estudio Ueno y Polinsky (2009), que encuentran que las lenguas OV (japonés y turco) tienen una mayor frecuencia de oraciones transitivas con argumentos omitidos que las lenguas VO (inglés y castellano). Sin embargo, no hacen una comparación estadística entre las lenguas OV y VO. Ueno y Polinsky (2009) asumen que el verbo tiene un papel importante a la hora de procesar una oración y es en la posición del verbo donde se resuelven todas las relaciones argumentales de la oración (i.e., si la oración es intransitiva o transitiva, si el verbo requiere uno, dos o tres argumentos…); por lo que deberían de haber predicho que el japonés y el turco (lenguas OV) tendrían más casos de argumentos omitidos en oraciones transitivas que el castellano (VO). En este capítulo, sí que he llevado a cabo dicha comparación y he observado, tal y como predecía, que en las oraciones transitivas el euskera hace más uso de argumentos omitidos que el castellano. Si nos fijamos en los datos sobre el uso de argumentos omitidos en oraciones intransitivas y transitivas que proporcionan Ueno y Polinsky (2009:700, Tabla 709) en su trabajo, podemos observar cómo en las oraciones transitivas el castellano (67%) tiene más casos de argumentos omitidos que el japonés (33%) y el turco (51%). Estas frecuencias son inconsistentes con las de las del estudio de corpus de este capítulo. Esta discrepancia en los datos de castellano puede deberse a que Ueno y Polinsky (2009) hayan etiquetado como argumentos omitidos omisiones que no lo son. Para esclarecer esta discrepancia es necesario un análisis más profundo de su etiquetaje.

Lo que sí que parece claro es que todas las lenguas analizadas hasta el momento, tanto lenguas VO como OV, reducen el número de argumentos preverbales en las oraciones transitivas mediante el uso de argumentos omitidos (castellano: Bentivoglio, 1992; Ueno y Polinsky, 2009, y esta tesis doctoral; japonés: Ueno y Polinsky, 2009; turco: Ueno y Polinsky, 2009; chino: Tao, 1996; maya sacapulteco: Du Bois, 1987; euskera: esta tesis doctoral). Así, parece que la omisión de argumentos es una estrategia universal de economía en el procesamiento de oraciones que se aplica en todas las lenguas, sin importar el tipo de orden básico de palabras que tengan, para reducir el número de argumentos fonológicamente expresados y así facilitar la producción y comprensión. La omisión de un argumento puede requerir un menor número de recursos cognitivos que exprésalo fonológicamente (Yamashita, Chang y Hirose, 2005; Demiral, 2007; Demiral, Schlesewsky y Bornkessel-Schlesewsky, 2008; Ueno y Garnsey, 2008; Wolff, Schlesewsky, Horie y Bornkessel-Schlesewsky, 2008b; Wolff, 2010; Özge, Marinis y Zeyrek, 2013). Demiral (2007); Demiral et al. (2008) y Wolff et al., 2008b; Wolff, 2010 observaron en hablantes de turco y japonés que las oraciones OSV provocaban una positividad P600 en la posición del verbo en comparación con oraciones SOV. Los autores interpretan que tal positividad se debe al coste asociado a la reinterpretación del primer NP de sujeto a objeto, pues los hablantes comenzarían interpretando las oraciones OSV como una oración transitiva con

sujeto omitido ([*pro*]OV). Es más, se ha encontrado que los hablantes de turco procesaban más rápido las oraciones OVS si el primer NP si estaba en acusativo (i.e., objeto) que si estaba en nominativo (i.e., sujeto), pues interpretarían la oración como una oración transitiva sin sujeto ([*pro*]OV) y no como una oración OVS. Del mismo modo, los argumentos nulos son más fáciles de recuperar durante el discurso, ya que sus referentes están mencionados anteriormente en el discurso (Kameyama, 1985, 1988; Walker, Iida y Cote, 1994; Turan, 1998; Prince, 1999); o son referenciados mediante la concordancia verbal (Meyerhoff, 2000).

En cuanto al tipo de argumento que se omite, el sujeto es el que más tiende a ser omitido, tanto en las lenguas analizadas en el estudio de Ueno y Polinsky (2009) como en las de esta tesis doctoral. Bloom (1990, 1993) arguye que la preferencia de omitir el sujeto antes que el objeto se debe a factores pragmáticos, pues el sujeto tiende a ser una información ya dada (Chafe, 1976; Lambrecht, 1994) y el objeto, por el contrario, una información nueva (Arnold, 1998). Los argumentos con información ya dada tienden a aparecer delante de los argumentos con información nueva (Arnold et al., 2000; Ferreira y Yoshita, 2003), lo cual se asocia con las posiciones de sujeto y objeto (en las lenguas VO y OV). La información que conlleva el sujeto viene dada por un contexto previo y este influye en que el sujeto se omita más que el objeto (Wolff, 2010), pues es fácilmente recuperable. Asimismo, se ha observado que la omisión objeto (O-drop) es más costosa que la de sujeto (S-drop): Ueno y Garnsey (2008) reportan que los hablantes de japonés tardan menos tiempo en leer oraciones con sujetos omitidos ([*pro*]OV) que oraciones con objetos omitidos (S[*pro*]V), y además estas muestran una negatividad N400 en la posición del verbo.

4.4.2 El uso de oraciones intransitivas para reducir el área preverbal

Los resultados del estudio de corpus muestran que el uso de oraciones intransitivas frente al de oraciones transitivas es similar en castellano (VO) y euskera (OV). Este resultado va en contra del reportado por Ueno y Polinsky (2009), quienes encuentran una mayor frecuencia de uso de oraciones intransitivas en lenguas OV (japonés y turco) como el euskera. Los resultado del corpus de este capítulo revelan que en castellano y euskera ambos tipos de oraciones tienen una frecuencia de uso similar, tanto en la comparación entre lenguas como dentro de cada lengua; sin embargo, en japonés y turco las oraciones intransitivas son más frecuentes que las oraciones transitivas, e incluso más frecuentes al compararlas con las oraciones intransitivas de inglés y castellano (Ueno y Polinsky, 2009). Ueno y Polinsky (2009) argumentan que en aras de reducir el número de argumentos preverbales, las lenguas OV recurren con mayor frecuencia a oraciones intransitivas que las lenguas VO. De esta forma, las lenguas OV al hacer un mayor uso de oraciones

intransitivas con orden SV estarían teniendo un orden similar a las lenguas VO, en las que antes del verbo solo aparecen el sujeto. Este orden SV, que se da en las oraciones intransitivas de las lenguas OV y en las oraciones intransitivas y transitivas de las lenguas VO, tendría un coste de procesamiento menor que el orden SOV de las oraciones transitivas de las lenguas OV (Lindsley, 1975; Kempen y Hoenkamp, 1987). Lindsley (1975) encuentra que los hablantes de inglés a la hora de describir viñetas tardan el mismo tiempo en empezar a describir una viñeta con una acción intransitiva y con una acción transitiva. En ambos tipos de oración el verbo aparece en la misma posición, i.e., después del sujeto, por lo que en ambas las relaciones argumentales se resuelven igual.
Siguiendo la hipótesis de Ueno y Polinsky (2009), predije que el euskera recurriría con mayor frecuencia al uso de oraciones intransitivas que de oraciones transitivas, tal y como muestran sus resultados de japonés y turco. Sin embargo, en los resultados de euskera no hay diferencias en la frecuencia de uso entre oraciones intransitivas y transitivas, ni tampoco en la frecuencia de uso de oraciones intransitivas entre castellano y euskera. Estos resultados sugieren que el euskera no recurre a la intransitividad para reducir el número de argumentos preverbales, como sí lo hacen el japonés y el turco. Una posible explicación de ello puede ser el hecho de que el euskera sea una lengua ergativa y las que el japonés y el turco sean lenguas acusativas. Nichols, Peterson y Barnes (2004) observan en una muestra de 80 lenguas como las lenguas con ergatividad son lenguas transitivas, i.e., tienen más verbos transitivos que las lenguas no ergativas, y derivan morfológicamente de estos verbos transitivos los verbos intransitivos; las lenguas acusativas, por el contrario, tienen más verbos intransitivos y derivan los verbos transitivos de los verbos intransitivos.

Otra posible explicación que sugiero es que el euskera difiere del japonés en que tiene mayor libertad para mover los argumentos. Ueno y Polinsky (2009), en su estudio, se percatan de que el turco muestra una frecuencia similar al inglés en el uso de oraciones intransitivas en comparación con el japonés, que tiene más oraciones intransitivas que el inglés. Ellas sugieren que la diferencia en el uso de oraciones intransitivas se debe a la diferencia de libertad para mover los argumentos en las lenguas OV: el japonés al tener un orden más rígido recurriría con mayor frecuencia que el turco al uso de oraciones intransitivas, mientras que el turco, con mayor libertad de argumentos, lo haría con menor frecuencia que el japonés. Así, argumenté que el euskera podría mostrar una mayor frecuencia de oraciones transitivas con argumentos postverbales (SVO-OVS) que oraciones transitivas con orden básico de palabras (SOV) para reducir el área preverbal y tener una linearización similar a la de las oraciones intransitivas (SV). Los resultados en euskera confirman esta predicción: existe una preferencia por reducir el número de argumentos que aparecen antes del verbo moviendo argumentos a posición postverbal. Aunque el uso de argumentos postverbales no estaría motivado exclusivamente por reducir el número de argumentos *per se*, sino que otros factores como el peso de los

argumentos en número de palabras que lo componen también pueden afectar (McDonald, Bock y Kelly, 1993; Wasow, 1997b, 1997a; Stallings et al., 1998; Wasow y Arnold, 2003; Stallings y MacDonald, 2011; Ros et al., 2015; Ros, 2018). De hecho, Ros (2018) y Ros et al. (2015) observan cómo los hablantes de euskera tienden a colocar el verbo en posición intermedia (SVO-OVS) cuando uno de los argumentos es largo y otro corto, siendo el argumento largo el que aparece después del verbo ($NP_{corto}VNP_{largo}$). En cuanto al tipo de argumento que aparece en posición postverbal, los resultados del corpus de este capítulo muestran que en euskera los objetos aparecen con mayor frecuencia en posición postverbal (sujeto (39,6%) y objeto (60,4%%). Esto puede deberse a que el objeto suele ser una información nueva (Chafe, 1976; Lambrecht, 1994; Arnold, 1998), ya que se ha encontrado que los argumentos con información nueva tiende a correlacionarse con argumentos más largos (Arnold et al., 2000; Wasow, 2002). En suma, los resultados sugieren que las lenguas OV con un orden no rígido, como el euskera y el turco, tienden a reducir el número de argumentos preverbales moviendo uno de ellos a posición postverbal mientras que las lenguas OV con un orden rígido, como el japonés, tienden más al uso de oraciones intransitivas.

4.5 Conclusiones

El estudio de corpus de este capítulo muestra que las lenguas OV tienden a reducir el número de argumentos que aparecen en el área preverbal. En especial, está reducción se consiguen mediante un mayor uso (a) de argumentos omitidos y (b) de argumentos postverbales. El uso de oraciones intransitivas, por el contrario, no parece ser una estrategia útil en euskera para reducir el área preverbal. Así, la generalización que hacen Ueno y Polinsky (2009) de que las lenguas OV recurren al uso de la intransitividad no queda confirmada. Los resultados de euskera de este capítulo parecen indicar que no solo el orden básico de palabras influye en la frecuencia con la que las lenguas recurren a ciertos fenómenos sintácticos (uso de argumentos omitidos, de oraciones intransitivas, de argumentos postverbales…) sino también el grado de libertad para mover los argumentos que tienen las lenguas, que se correlaciona con el grado de concordancia verbal, muy rica en euskara pero inexistente en lenguas estrictamente de verbo final como el japonés o el coreano. Por tanto, la frecuencia de uso de recursos gramaticales para facilitar el procesamiento no depende de un solo rasgo tipológico (VO-OV), sino que está modulada por la concurrencia de otros rasgos gramaticales, generando diferentes perfiles según la combinación paramétrica de cada lengua.

Capítulo 5

Estrategias para la reducción de la interferencia por animacidad

ABSTRACT

It is assumed that during sentence planning elements of the sentence are held in memory until they are produced. Semantically similar elements (such as, both animates) create an interference overloading memory and producers tend to develop sentence plans that minimize that type of interference (MacDonald, 2013). Based on this idea, some studies have observed that speakers reduce similarity induced interference by placing similar elements farther apart from each other in the linear order of the sentence or dropping one of them. Gennari, Mirkovic & MacDonald (2012) investigated the effect of animacy induced interference on sentence processing using a picture description task in English, Spanish and Serbian, and they showed that participants reduce the interference between two animate NPs omitting one of them (English and Spanish) and increasing the linear distance between them (English, Spanish and Serbian). Hsiao, Gao & MacDonald (2014) conducted a corpus and production study of transitive sentences in Chinese to investigate the frequency of subject omission when both arguments (subject and object) were animate. Their results showed that subjects were omitted more frequently when both subject and object were animate than when only the subject was animate. This chapter tests those two hypothesized strategies (omission and postverbal scrambling) to minimize animacy interference in Basque and Spanish. I present data from a corpus study about animate argument omissions and postverbal arguments in Basque and Spanish –the former has both subject and object before the verb in transitive sentences (SOV), the latter only the subject (SVO)–. I hypothesize that when both arguments are animate (a) Basque will use more null arguments or would move one of the arguments to the postverbal position in order to avoid similarity-triggered interference; and (b) Spanish will not need to omit subjects more frequently. Results showed that both languages resort to reduce the interference by omitting the subject, which converges with results in previous studies. I conclude that argument omission is less costly than displacement, and hence whenever omission is possible, languages will preferably resort to it in order to avoid animacy-triggered interference.

5.1 Introducción

Antes de producir una oración, esta se planifica, i.e., se prepara una cantidad de información antes de producir la oración (Levelt, 1989), y esta ha de mantenerse en la memoria hasta que es producida (Levelt, 1989; Bock y Levelt, 1994; Rosenbaum, Cohen, Jax, Weiss y van der Wel, 2007; MacDonald, 2013). El alcance de la planificación, i.e., la cantidad de elementos (información) que se planifica puede variar (Konopka, 2009, 2012): puede ser una sola palabra (Meyer, Sleiderink y Levelt, 1998; Griffin y Bock, 2000; Griffin, 2001, 2003; Brown-Schmidt y Konopka, 2008; Zhao y Yang, 2016), un sintagma (Smith y Wheeldon, 1999; Allum y Wheeldon, 2007, 2009; Martin, Crowther, Knight, Tamborello II y Yang, 2010; Wheeldon, Ohlson, Ashby y Gator, 2013; Zhao, Alario y Yang, 2015) o una oración (Garrett, 1980; Meyer, 1996; Lee, Brown-Schmidt y Watson, 2013). Esta diferencia de alcance de planificación puede deberse a diferentes factores como presiones temporales (Ferreira y Swets, 2002, 2005), incremento de la carga cognitiva (Wagner, Jescheniak y Schriefers, 2010), saturación de memoria de trabajo (Slevc, 2011) e incluso a las propiedades tipológicas de las lenguas como la posición del objeto en relación al verbo (SVO/SOV) (Sauppe, Norcliffe, Konopka, Van Valin y Levinson, 2013b; Hwang y Kaiser, 2014; Norcliffe, Konopka, Brown y Levinson, 2015; Momma, Slevc y Phillips, 2016; Sauppe, 2017).

Durante la planificación elementos conceptualmente similares cargan la memoria y dificultan la producción, creando interferencia (Meyer, 1996; Gordon, Hendrick y Johnson, 2001; Ferreira y Firato, 2002; Smith y Wheeldon, 2004; Gordon, Hendrick, Johnson y Lee, 2006; Acheson y MacDonald, 2009; Wagner et al., 2010). Smith y Wheeldon (2004), por ejemplo, en un estudio de descripción de imágenes encontraron que los participantes necesitaban más tiempo para planificar las oraciones cuando las imágenes incluían dos nombres conceptualmente similares (e.g., "sierra" y "hacha") en comparación a las imágenes que incluían dos nombres que no eran conceptualmente similares (e.g., "sierra" y "gato"). Gordon et al. (2001); Gordon, Hendrick y Johnson (2004) y Lee, Lee y Gordon (2007) también encontraron interferencias creadas por elementos conceptualmente similares en varios estudios de compresión: los participantes tardaban más tiempo en leer y comprendían peor las oraciones relativas con dos sintagmas nominales (NPs) si eran del mismo tipo (e.g., ambos profesiones: *the banker that the barber praised climbed…* [el banquero que el barbero alabó subió…]) que cuando ambos eran de diferente tipo (e.g., uno profesión y otro nombre propio: *the banker that Sophie praised climbed…* [el banquero que Sofía alabó subió…]). Gordon et al. (2001, 2004) y Lee et al. (2007) arguyen que esta dificultad se debe a la interferencia creada por dos NPs semánticamente similares durante su procesamiento.

Dado que la planificación es esencial para una producción eficiente y la planificación de la oración se mantiene en la memoria hasta que se produce (MacDonald, 2015, 2016), la carga de memoria de dos elementos similares crea interferencia y, por tanto, dificulta la producción. Por ello, MacDonald (2013) propone que los hablantes tienden a utilizar estrategias para reducir el coste de memoria durante la planificación. Una de esas estrategias, relevante en este capítulo, es la estrategia denominada *Reduce la Interferencia*: los hablantes tienden a favorecer estructuras gramaticales que minimizan la interferencia creada por elementos similares durante la planificación.

El capítulo se organiza de la siguiente manera: la sección 5.2 resume los estudios que han mostrado cómo las lenguas recurren a diferentes estrategias para reducir la interferencia creada por dos NPs similares, como que ambos sean animados. Las secciones 5.3 y 5.4 presentan sendos estudios de corpus escrito en euskera y castellano que analizan estadísticamente la frecuencia y la probabilidad con la que ambas lenguas utilizan estrategias para reducir la interferencia de animacidad. Por último, las secciones 5.5 y 5.6 terminan con la discusión de los resultados y las conclusiones del capítulo.

5.2 La animacidad como factor de interferencia

La animacidad crea interferencia en el procesamiento del lenguaje: durante la planificación oracional los argumentos animados compiten entre sí por aparecer en posición inicial o ser sujeto de la oración (McDonald et al., 1993; Ferreira, 1994; Prat-Sala, 1997; Prat-Sala y Branigan, 2000; Prat-Sala, Shillcock y Sorace, 2000; van Nice y Dietrich, 2003; Tanaka, Branigan y Pickering, 2005, entre otros). Varios estudios de comprensión han encontrado que el coste de procesamiento de las oraciones relativas de objeto está modulada por la animacidad de este (Mak, Vonk y Schriefers, 2002; Fedorenko y Gibson, 2008; Betancort, Carreiras y Sturt, 2009; Wu, Kaiser y Andersen, 2012; Lowder y Gordon, 2014). Mak et al. (2002) en un estudio de *self-paced reading* (lectura autoadministrada) investigaron la influencia de la animacidad en el procesamiento de oraciones de relativo en holandés y alemán. Los resultados mostraron que los participantes de ambas lenguas tardaban más tiempo en leer las relativas de objeto cuando ambos NPs eran animados (e.g., *al ocupante que los ladrones robaron…*) que cuando el antecedente era inanimado y el NP interno de la relativa era animado (e.g., *el ordenador que los ladrones robaron…*), y estás últimas mostraban unos tiempos de lectura similar a las relativas de sujeto. Mak et al. (2002) concluyeron que el coste de procesamiento de las relativas de objeto se debe a la proximidad de los dos NPs animados (NP_{Ani} [que NP_{Ani} V…]), frente a las relativas de sujeto donde la proximidad es menor ya que el verbo aparece entre ambos NPs (NP_{Ani} [que V NP_{Ani}…]). Es más, diferentes estudios de corpus han encontrado que las relativas de objeto con dos animados son poco frecuentes en las lenguas en comparación con las

relativas de sujetos con dos animados (holandés y alemán: Mak et al., 2002; inglés: Roland et al., 2007; chino: Wu, Kaiser y Andersen, 2009; taiwanés: Vasishth, Chen, Li y Guo, 2013).

5.2.1 Reduciendo la interferencia por animacidad

En línea con estos estudios de comprensión y corpus, varios estudios de producción arguyen que los hablantes resuelven la interferencia creada por dos argumentos animados en la planificación recurriendo a dos estrategias: (a) omitir uno de los argumentos animados (Christianson y Ferreira, 2005; Gennari et al., 2012; Hsiao, Gao y MacDonald, 2014) e (b) incrementar la distancia lineal entre ambos argumentos animados (Gennari et al., 2012; Montag y MacDonald, 2014; Hsiao y MacDonald, 2016; Humphreys, Mirković y Gennari, 2016; Perera y Srivastava, 2016; Montag, Matsuki, Kim y MacDonald, 2017).

De acuerdo con la primera estrategia (a), durante la planificación se inhibe uno de los argumentos animados, pero no se recupera más adelante, dando como resultado oraciones con argumentos omitidos. Gennari et al. (2012) observaron que los participantes de castellano, y en menor medida los de serbio, producían con mayor frecuencia oraciones impersonales relativas de objeto (e.g., *el hombre (al) que están abrazando*) cuando tenían que describir escenas con agente y paciente animados que cuando las escenas tenían solo el agente animado. Hsiao et al. (2014), replicaron el estudio de Gennari et al. (2012) con oraciones transitivas simples, partiendo de la hipótesis de que los hablantes omiten el sujeto con mayor frecuencia cuando ambos argumentos (sujeto y objeto) son animados que cuando solo el sujeto lo es. Hsiao et al. (2014) llevaron a cabo dos estudios en chino, uno de corpus y otro de producción para testear dicha hipótesis. En el estudio de corpus analizaron la frecuencia con la que se omite el sujeto en oraciones transitivas cuando este es animado y el objeto animado o inanimado. El corpus constaba de 4035 oraciones transitivas, extraídas del *Chinese Treebank 7.0* (Xue et al., 2010), y fueron etiquetadas manualmente según la animacidad del sujeto y el objeto. Los resultados mostraron que cuando ambos argumentos eran animados el sujeto se omitía con mayor frecuencia (52%) que cuando el sujeto era animado y el objeto inanimado (28%). En el estudio de producción, replicaron los resultados del estudio de corpus. En el experimento, los hablantes veían dos viñetas: una a la izquierda que incluía un texto de unas dos o tres frases que proporcionaba el contexto sobre el sujeto (e.g., ocupación, hábitos…) de la acción; y otra a la derecha que aparecía acompañada de un verbo para forzar a los participantes a utilizarlo para describir la acción de la viñeta. En las viñetas el sujeto siempre era animado, mientras que el objeto era animado o inanimado. Los resultados mostraron que los participantes producían de forma significativa oraciones transitivas con sujetos omitidos (e.g., *saluda al hombre*) cuando describían viñetas con dos animados que cuando las viñetas tenían un animado y un inanimado.

De acuerdo con la segunda estrategia (b), durante la planificación se inhibe uno de los argumentos animados para que no compita con el otro y se recupera más adelante en la planificación, dando como resultado oraciones donde se ha aumentado la distancia lineal entre la producción de ambos animados. Gennari et al. (2012) investigaron, mediante la descripción de imágenes, si los hablantes de inglés, castellano y serbio utilizaban esta estrategia lineal en oraciones relativas de objeto. En el experimento, los participantes veían escenas en las que un agente animado actuaba sobre un paciente animado (e.g., una mujer abrazando un hombre) y otro agente animado actuaba sobre un paciente inanimado (e.g., una mujer abrazando un peluche). Tras ver las escenas, los participantes debían describir las imágenes respondiendo a preguntas como "*¿quién es calvo?*" para que describieran la escena seleccionando el paciente animado (e.g., *el hombre que la mujer abraza*); o "*¿qué es blando?*" para la describieran con el paciente inanimado (e.g., *el peluche que la mujer abraza*). Los participantes de inglés y castellano, pero no los de serbio, produjeron con mayor frecuencia oraciones pasivas relativas de objeto cuando tenían que describir escenas con agente y paciente animados (e.g., *el hombre que es abrazado por la mujer*) en comparación con escenas que tenían un agente animado y un paciente inanimado, en estos casos producían con mayor frecuencia activas relativas de objeto (e.g., *el peluche que la mujer abraza*). Estudios posteriores han replicado los resultados de Gennari et al. (2012) en diferentes lenguas (Montag y MacDonald, 2014; Hsiao y MacDonald, 2016; Humphreys et al., 2016; Perera y Srivastava, 2016; Montag et al., 2017).

En conjunto, la evidencia indica que los hablantes tienden a reducir la interferencia creada por dos argumentos animados durante la planificación de una oración. En algunos casos recurrirán a aumentar la distancia lineal entre los dos, inhibiendo uno de los argumentos animados durante la planificación y recuperándolo más tarde; y en otros casos recurrirán a oraciones con sujetos omitidos, inhibiendo completamente durante la planificación el animado que se selecciona para ser el sujeto de la oración.

Este capítulo tiene como objetivo investigar si esas estrategias se observan en euskera y el castellano. Es decir, estudiaré si estas lenguas muestran evidencia de recurrir a (a) la omisión de uno de los dos argumentos, y (b) al aumento de la distancia lineal entre los argumentos. Así, si la hipótesis es correcta, en el caso del euskera espero encontrar que la frecuencia de la omisión de los argumentos y el incremento de la distancia lineal entre ambos sea mayor cuando ambos argumentos son animados frente a cuando solo uno de ellos lo es. En el caso del castellano, por el contrario, espero no encontrar diferencias en la frecuencia en el uso de omisión de argumentos cuando los dos argumentos son animados frente a cuando solo uno lo es, dado que en castellano los argumentos ya están separados por el verbo y por tanto la interferencia se vería reducida. En cuanto al incremento de la distancia lineal en castellano, no lo voy a explorar dado que en castellano los órdenes con

ambos argumentos en posición preverbal son casi inexistentes (SOV: 0,1% y OSV: 0%, López, 1997). Para investigar esta hipótesis, realizaré dos estudios de corpus escrito en euskera y castellano.

5.3 Estudio de corpus 1: euskera

En el presente estudio de corpus he examinado si en euskera se recurre a reducir la interferencia de animacidad mostrando una mayor frecuencia de argumentos omitidos o postverbales cuando en una oración transitiva los dos argumentos son animados. Tal y como he explicado en la introducción (sección 5.1.2), en otras lenguas se ha observado que los hablantes recurren a dos estrategias principales para reducir dicha interferencia de animacidad: la omisión de un argumento (Gennari et al., 2012; Hsiao et al., 2014) y el aumento lineal entre ambos argumentos (Gennari et al., 2012). En la mayoría de los estudios previos de lenguas VO se ha examinado la interferencia de animacidad en oraciones relativas de objeto, ya que en este tipo de estructuras aparecen dos NPs en posición preverbal. En euskera, al ser una lengua SOV, tanto el sujeto como el objeto aparecen en posición preverbal y por ello es posible examinar el efecto de la interferencia de animacidad en oraciones transitivas simples. Así, dada la hipótesis de trabajo, predigo que el euskera reducirá la interferencia de sujeto y objeto animados mediante la omisión de uno de ellos (cf. 5.1b,c) o mediante el desplazamiento de uno de los argumentos a posición postverbal para separarlos en la linearización (cf. 5.1d,e).

(5.1) a. Irakasleak ikaslea ikusi zuen — SOV
profesor.SG.ERG alumno.SG.ABS ver AUX
"El profesor vio al alumno"

b. *pro* ikaslea ikusi zuen — OV [omisión de sujeto]
"Vio al alumno"

c. Irakasleak *pro* ikusi zuen — SV [omisión de objeto]
"El profesor lo vio"

d. Irakasleak ikusi zuen ikaslea — SVO

e. Ikaslea ikusi zuen irakasleak — OVS

5.3.1 Materiales

El corpus de euskera consta de 3011 oraciones transitivas. 1011 oraciones pertenecen al corpus descrito en el CAPÍTULO 4 y las 2000 oraciones restantes son oraciones nuevas extraídas de corpus *EPG – Ereduzko Prosa Gaur* (Sarasola et al., 2009) que he incluido para tener una muestra más amplia. El criterio de selección de estas nuevas oraciones ha sido el mismo que el utilizado para el corpus del CAPÍTULO 4, con la intención de tener una

muestra heterogénea: 1050 oraciones del periódico *Berria* (150 x 7 secciones), 600 oraciones de libros (150 x 4 géneros), 50 oraciones de la revista científica *Uztaro*, y 300 oraciones de guiones de la serie televisiva *Goenkale*. Del total de oraciones que componen el corpus, solo he tenido en cuenta para el análisis estadístico las 2409 oraciones transitivas declarativas, descartando las oraciones negativas, interrogativas, las ditransitivas, las que tienen como objeto una subordinada (objeto CP) y aquellas en las que aparece solo el verbo[1]:

declarativas	**negativas**	**interrogativas**	**ditransitivas**	**objeto CP**	**solo verbo**	**TOTAL**
2409	80	21	111	236	154	3011

TABLA 5.1. Distribución del tipo de oraciones transitivas en el corpus escrito de euskera.

5.3.2 Procedimiento

Las oraciones de euskera las he etiquetado manualmente y clasificado según el tipo de omisión (sujeto (cf. 5.2a) u objeto (cf. 5.2b), la animacidad de los argumentos y el orden de palabras (SOV, SVO, OSV, OVS, VSO, VOS). Los sujetos y objetos animados incluyen pronombres de primera, segunda y tercera persona, nombre propios, sustantivos que hacen referencia humanos y organizaciones, y animales (cf. 5.3) (Silverstein, 1976; Dixon, 1979; Garreston, O'Connor, Skarabela y Hogan, 2004; Bresnan, Cueni, Nikitina y Baayen, 2007; Bresnan y Hay, 2008; Baker y Brew, 2010). El resto de sujetos y objeto han sido etiquetados como inanimados. En cuanto al orden de palabras, he etiquetado las oraciones según el orden lineal del sujeto, objeto y verbo: SOV, SVO, OVS, OSV, VSO, VOS.[2]

(5.2) a. *pro* Downing Street nazional bat behar dugu — SOV [omisión de sujeto]
"Necesitamos un Downing Street nacional"
b. Erresuma Batuko fisikari batzuek *pro* aurkitu dute… — SV [omisión de objeto]
"Unos físicos de Reino Unido lo han encontrado…"

(5.3) a. Nik albaniarrak defenditu ditut — [animado: pronombre]
"Yo he defendido a los albanos"
b. Karmenek sorbaldak goratzen ditu — [animado: propio]
"Carmen ha levantado los hombros"
c. …mutil batek oso esaldi polita idatzi du — [animado: humano]
"…un chico ha escrito una frase muy bonita"

[1] Las oraciones las he descartado en el siguiente orden, debido a que algunas coincidían en más de un criterio: aparece solo el verbo (154) > interrogativas (21) > negativas (80) > ditransitivas (111) > objeto como oración subordinada (236) = 602 oraciones excluidas.

[2] Los órdenes OSV, VSO y VOS no los he tenido en cuenta a la hora de hacer el análisis estadístico.

d. LABek toki berean egingo du manifestazioa [animado: organización]
"LAB ha hecho la manifestación en el mismo lugar"
e. …behiak aztertu dituzte [animado: animal]
"…han analizado las vacas"

A la hora de analizar los datos del corpus he usado los análisis estadísticos prueba χ^2 (*Pearson chi-square*) y el modelo de regresión logística binomial. La prueba χ^2 (*Pearson chi-square*) la he usado para analizar si existe una asociación entre que los argumentos de una oración transitiva sean animados y la distribución de los argumentos omitidos y los argumentos expresados. El modelo de regresión logística binomial, por el contrario, lo he utilizado para analizar si la omisión de argumentos y el uso de argumentos postverbales están influenciados por la combinación de animacidad de los argumentos en las oraciones transitivas, en especial, cuando ambos argumentos son animados. La finalidad de este análisis es observar si en euskera el hecho de que el sujeto y el objeto sean animados influye en la probabilidad de omitir o de mover a posición postverbal uno de ellos. Los análisis los he computado mediante el programa estadístico R (R Core Team, 2017) y usando el paquete *lme4* (Bates et al., 2015). He tomado como nivel de referencia la media de las medias de las cuatro condiciones (SANI+OANI, SANI+OINA, SINA+OANI, SINA+OINA), dado que el *intercept* es una media no ponderada, es decir, las variables tienen diferentes frecuencias. Así, el coeficiente de cada condición (*estimate*) representa cuánto se desvía de la gran media. Los resultados los he considerado significativos a un nivel $p < .05$._Los gráficos los he realizado con el paquete *ggplot2* (Wickham, 2009).

5.3.3 Resultados

La TABLA 5.2 muestra la clasificación de las 2409 oraciones transitivas que componen el corpus, según la animacidad del sujeto y del objeto, el orden de palabras, y si la oración tiene algún argumento omitido.

	sujeto						objeto					
	animado		inanimado		Total		animado		inanimado		Total	
SOV	67%	(366)	33%	(179)	100%	(545)	6%	(34)	94%	(511)	100%	(545)
SVO	70%	(191)	30%	(80)	100%	(271)	8%	(23)	92%	(248)	100%	(271)
OSV	74%	(17)	26%	(6)	100%	(23)	4%	(1)	96%	(22)	100%	(23)
OVS	72%	(129)	28%	(49)	100%	(178)	3%	(6)	97%	(172)	100%	(178)
VSO	89%	(17)	11%	(2)	100%	(19)	5%	(1)	95%	(18)	100%	(19)
VOS	86%	(13)	14%	(2)	100%	(15)	0%	(0)	100%	(15)	100%	(15)
OV	94%	(987)	6%	(61)	100%	(1048)	12%	(128)	88%	(920)	100%	(1048)
SV	72%	(39)	28%	(15)	100%	(54)	54%	(29)	46%	(25)	100%	(54)
VO	71%	(10)	29%	(4)	100%	(14)	57%	(8)	43%	(6)	100%	(14)
VS	98%	(236)	2%	(6)	100%	(242)	13%	(31)	87%	(211)	100%	(242)
TOTAL	83%	(2005)	17%	(404)	100%	(2409)	11%	(261)	89%	(2148)	100%	(2409)

TABLA 5.2. Distribución en euskera de las oraciones transitivas según el orden de palabras, la animacidad y la omisión del sujeto y el objeto.

5.3.3.1 Reducción del número de argumentos

Para el análisis de la omisión de argumentos he tenido en cuenta los tipos de oraciones SOV, OV y SV (ver TABLA 5.3 para más detalle). De las 1647 oraciones transitivas utilizadas, 1102 oraciones tienen algún argumento omitido (67%) y 545 oraciones tienen ambos argumentos explícitos (SOV: 33%). De las 1102 oraciones con argumento omitido, 1048 oraciones tienen el sujeto omitido (OV: 64%) y 54 el objeto (SV: 3%).

	sujeto animado						sujeto inanimado						TOTAL	
	objeto animado		objeto inanimado		Total		objeto animado		objeto inanimado		Total			
SOV	8%	(30)	92%	(336)	100%	(366)	2%	(4)	98%	(175)	100%	(179)	33%	(545)
OV	13%	(126)	87%	(861)	100%	(987)	12%	(2)	88%	(59)	100%	(61)	64%	(1048)
SV	38%	(15)	62%	(24)	100%	(39)	54%	(14)	46%	(1)	100%	(15)	3%	(54)
TOTAL	12%	(171)	88%	(1221)	100%	(1392)	11%	(20)	89%	(235)	100%	(255)	100%	(1647)

TABLA 5.3. Distribución en euskera de las oraciones transitivas SOV, OV (omisión de sujeto) y SV (omisión de objeto) según la animacidad del sujeto y el objeto.

El uso de omisión de argumentos es mayor en las oraciones transitivas que tienen algún argumento animado: SANI+OANI (82,5%), SANI+OINA (72,5%) y SINA+OANI (80%) (GRÁFICO 5.1). Este uso de omisión de argumentos es significativamente mayor en las transitivas con los dos argumentos animados frente a la condición SANI+OINA [SANI+OANI vs. SANI+OINA: 82,5% vs. 72,5%, χ^2 (1, N = 1391) = 7.1943, p < .007, odds ratio = 1.78] y la condición SANI+OANI vs. SINA+OINA [82,5% vs. 25,5%, χ^2 (1, N = 406) = 126.04, p < .001, odds ratio = 13.59].

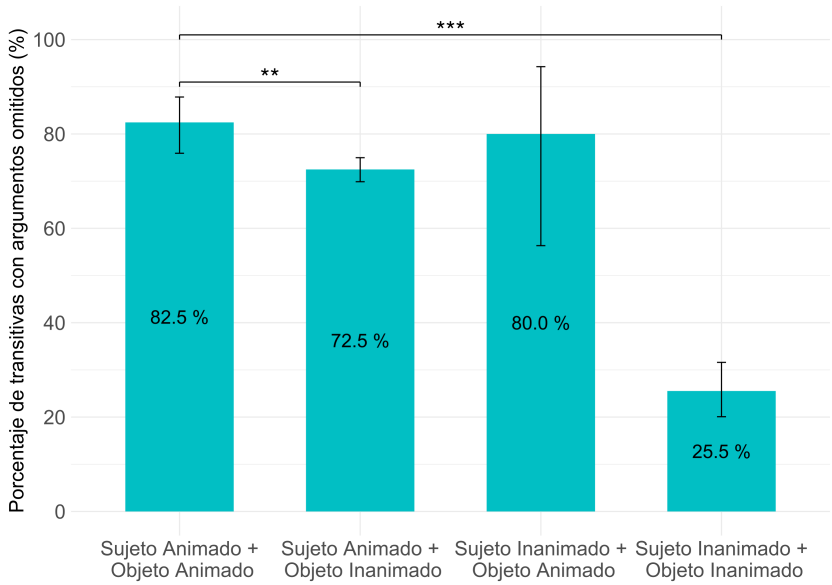


GRÁFICO 5.1. Frecuencia de oraciones transitivas en euskera con argumentos omitidos, agrupadas por la animacidad del sujeto y el objeto. Las barras de error muestran los intervalos de confianza de 95%.

Para determinar si la omisión de los argumentos se debe a que ambos sean animados he llevado a cabo un análisis de regresión logística. La omisión de los argumentos la he analizado en las cuatro posibles condiciones de animacidad: SANI+OANI (sujeto y objeto animados), SANI+OINA (sujeto animado y objeto inanimado), SINA+OANI (sujeto inanimado y objeto animado) y SINA+OINA (sujeto y objeto inanimados). El análisis muestra que existe una correlación significativa entre que ambos argumentos sean animados y la omisión de uno de ellos [SANI+OANI: 82,5%, $\beta = 0.8396$, $z = 4.006$, $< .001$, odds ratio = 2.31, 95% CI = 75.91 – 87.83], comparado con el resto de condiciones [SANI+OINA: 72,5%; SINA+OANI: 80%; SINA+OINA: 25,5%] (TABLA 5.4). La probabilidad de que algún argumento se omita es 2,3 veces mayor cuando el sujeto y el objeto son animados que en el resto de condiciones. Por tanto, el modelo de regresión logística confirma que la animacidad de los argumentos influye significativamente en la probabilidad de omisión de argumentos.

Omisión de argumentos – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	0.7080	0.1540	4.597	0.001 ***
SANI+OANI	0.8396	0.2096	4.006	0.001 ***
SANI+OINA	0.2605	0.1605	1.623	0.105
SINA+OANI	0.6783	0.4242	1.599	0.110
SINA+OINA	-1.7784	0.1868	-9.519	0.001 ***

TABLA 5.4. Resultados del modelo de regresión logística para la animacidad de los argumentos y la omisión de argumentos en euskera.

En cuanto al tipo de argumento que se omite, la frecuencia de omisión del sujeto es significativamente mayor cuando el sujeto y el objeto de la oración son animados, frente a cuando solo el sujeto es animado [SANI+OANI vs. SANI+OINA: 80,8% vs. 71,9%, χ^2 (1, N = 1352) = 5.025, p < .024, odds ratio = 1.63] (GRÁFICO 5.2a). Esta mayor frecuencia de omisión de sujeto también se observa al comparar oraciones transitivas con sujeto y objeto animados frente a transitivas con sujeto y objeto inanimados [SANI+OANI vs. SINA+OINA: 80,8% vs. 25,2%, χ^2 (1, N = 390) = 113.64, p < .001, odds ratio = 12.45], y frente a transitivas con sujeto inanimado y objeto animado [SANI+OANI vs. SINA+OANI: 80,8% vs. 33,3%, p < .018, Fisher's Exact Test, odds ratio = 8.246]. La comparación de transitivas con sujeto inanimado no es significativa: no hay diferencias significativas en la omisión del sujeto inanimado cuando el objeto es animado o inanimado [SINA+OANI vs. SINA+OINA: 33,3% vs. 25,2%, p = 0.64, Fisher's Exact Test, odds ratio = 1.480]. En cuanto a la omisión de objeto, es más frecuente cuando el sujeto es inanimado y el objeto animado, frente a cuando ambos son animados [SINA+OANI vs. SANI+OANI: 77,8% vs. 33,3%, p < .002, Fisher's Exact Test, odds ratio = 0.147] (GRÁFICO 5.2b). Este efecto, sin embargo, puede deberse al pequeño tamaño de la muestra de oraciones transitivas con objetos omitidos. En segundo lugar de frecuencia de omisión de objeto se observa cuando el sujeto y el objeto de la oración son animados frente a cuando el sujeto es animado y el objeto inanimado [SANI+OANI vs. SANI+OINA: 33,3% vs. 6,7%, p < .001, Fisher's Exact Test, odds ratio = 6.942], y frente a cuando ambos son inanimados [SANI+OANI vs. SINA+OINA: 33,3% vs. 0,6%, p < .001, Fisher's Exact Test, odds ratio = 84.969] (GRÁFICO 5.2b).

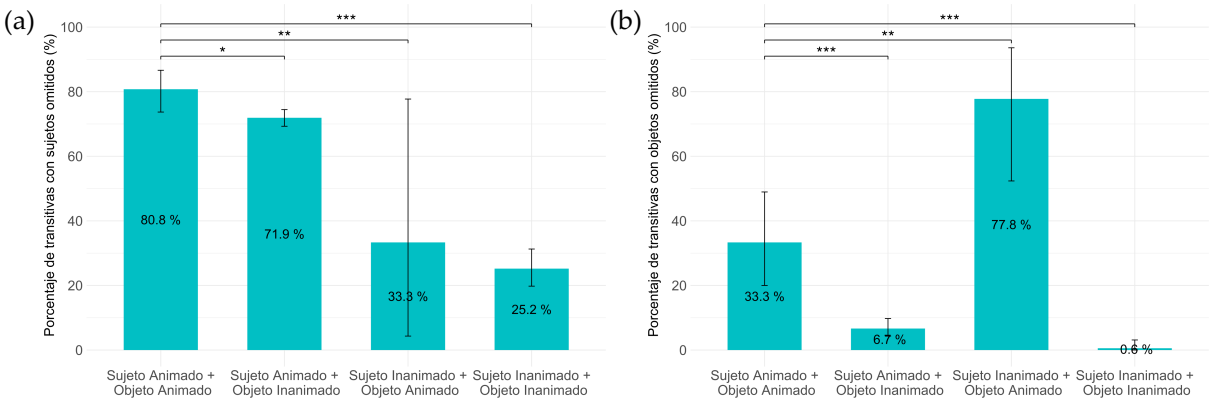


GRÁFICO 5.2. Frecuencia de oraciones transitivas en euskera con (a) sujetos omitidos (OV) y (b) objetos omitidos (SV), agrupadas por la animacidad del sujeto y el objeto. Las barras de error muestran los intervalos de confianza de 95%.

Para determinar si la omisión del sujeto y el objeto se debe a que ambos sean animados he ha llevado a cabo un análisis de regresión logística. El análisis de regresión logística muestra que existe una correlación significativa entre la omisión de sujeto y la animacidad de ambos argumentos: que ambos argumentos (sujeto y objeto) sean animados influye significativamente en la omisión del sujeto [SANI+OANI: 80,8%, β = 1.2862, z = 4.801, < .001,

odds ratio = 3.61, 95% CI = 73.6 – 86.6] en comparación con el resto de condiciones [SANI+OINA: 71,9%; SINA+OANI: 33,3%; SINA+OINA: 25,2%] (TABLA 5.5a). La probabilidad de que el sujeto se omita es 3,6 veces mayor cuando el sujeto y el objeto son animados que en el resto de condiciones. En el caso de la omisión de objeto, que ambos argumentos sean animados no influye tanto en su omisión, ya que la probabilidad de que el objeto se omita es 21,4 veces mayor cuando el sujeto es inanimado y el objeto animado que en el resto de condiciones [SINA+OANI: 77,7%, $\beta = 3.0638$, $z = 6.095$, $p < .001$, odds ratio = 21.40, 95% CI = 52.4 – 93.6] (TABLA 5.5b). Sin embargo, el alto *estimate* (3.0638) indica que el modelo puede estar sobreajustado debido al limitado número de observaciones en dicha condición, por lo que el resultado debe de interpretarse con cautela.

(a) Omisión de sujeto (OV) – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	0.1489	0.2261	0.659	0.510
SANI+OANI	1.2862	0.2679	4.801	0.001 ***
SANI+OINA	0.7921	0.2306	3.434	0.001 ***
SINA+OANI	-0.8421	0.6528	-1.290	0.197
SINA+OINA	-1.2362	0.2499	-4.946	0.001 ***
(b) Omisión de objeto (SV) – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	-1.8111	0.3032	-5.973	0.001 ***
SANI+OANI	1.1179	0.3767	2.967	0.003 **
SANI+OINA	-0.8280	0.3380	-2.450	0.014 *
SINA+OANI	3.0638	0.5026	6.095	0.001 ***
SINA+OINA	-3.3537	0.7710	-4.350	0.001 ***

TABLA 5.5. Resultados del modelo de regresión logística para la animacidad de los argumentos y el tipo de omisión en euskera: (a) de sujeto (SOV) y (b) de objeto (SV).

5.3.3.2 Posición postverbal: incremento de la distancia lineal

Para el análisis de la frecuencia de uso de argumentos posverbales he utilizado las condiciones SOV, SVO y OVS (994 oraciones) (ver TABLA 5.6 para más detalle). De las 994 oraciones transitivas que forman el corpus, 545 son SOV (55 %), 271 oraciones son SVO (27%) y 178 son OVS (18%). De las 523 oraciones transitivas SOV, 366 tienen sujeto animado (30 con objeto animado y 336 con objeto inanimado) y 179 sujeto inanimado (4 con objeto animado y 175 con objeto inanimado). En cuanto a las 271 oraciones SVO, 191 oraciones tienen sujeto animado (18 con objeto animado y 173 con objeto inanimado) y 80 sujeto inanimado (5 con objeto animado y 75 con objeto inanimado). Dentro de las 178 oraciones OVS, 7 transitivas tienen objetos animados (6 con sujeto animado y 1 con sujeto

inanimado) y 171 transitivas tienen objetos inanimados (122 con sujetos animados y 49 con sujetos inanimados).

	sujeto animado					**sujeto inanimado**					**TOTAL**
	objeto animado		**objeto inanimado**		**Total**	**objeto animado**		**objeto inanimado**		**Total**	
SOV	8%	(30)	92%	(336)	100% (366)	2%	(4)	98%	(175)	100% (179)	55% (545)
SVO	9%	(18)	91%	(173)	100% (191)	6%	(5)	94%	(75)	100% (80)	27% (271)
OVS	5%	(6)	95%	(122)	100% (128)	2%	(14)	98%	(49)	100% (50)	18% (178)
TOTAL	8%	(54)	92%	(631)	100% (685)	3%	(20)	97%	(299)	100% (309)	100% (994)

TABLA 5.6. Distribución de oraciones transitivas en euskera con todos los argumentos preverbales (SOV) y con algún argumento postverbal (SVO y OVS) según la animacidad del sujeto y el objeto.

Los resultados muestran que no hay diferencias significativas en la frecuencia de uso entre oraciones transitivas con sujeto y objeto preverbales (SOV) y transitivas con argumentos postverbales (SVO+OVS) cuando ambos argumentos son animados que cuando solo el sujeto lo es [S_{ANI}+O_{ANI} vs. S_{ANI}+O_{INA}: 44,4% vs. 46,8%, χ^2 (1, N = 685) = 0.033, p = .854, odds ratio = 0.911] (GRÁFICO 5.3). Tampoco se observan diferencias significativas en la frecuencia de uso de argumentos postverbales al comparar oraciones transitivas con sujeto y objeto animados frente a transitivas con sujeto inanimado y objeto animados [S_{ANI}+O_{ANI} vs. S_{INA}+O_{ANI}: 44,4% vs. 60%, p = 0.494, Fisher's Exact Test, odds ratio = 0.538], y frente a transitivas con sujeto y objeto inanimados [S_{ANI}+O_{ANI} vs. S_{INA}+O_{INA}: 44,4% vs. 41,5%, χ^2 (1, N = 353) = 0.066376, p = .796, odds ratio = 1.128].

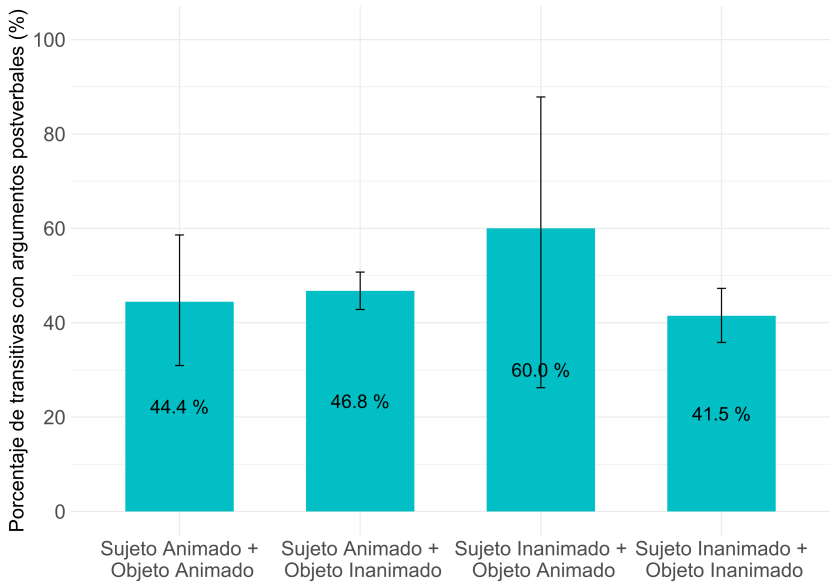


GRÁFICO 5.3. Frecuencia de oraciones transitivas en euskera con postverbales (SVO+OVS), agrupadas por la animacidad del sujeto y el objeto. Las barras de error muestran los intervalos de confianza de 95%.

El análisis de regresión logística muestra que no existe una correlación entre la presencia de argumentos postverbales y la animacidad. El hecho de que ambos argumentos sean animados no influye en la en que alguno de los argumentos aparezca en posición postverbal [SANI+OANI: 44,4%, β = -0.15006, z = -0.569, p = .569, odds ratio = 0.42, 95% CI = 30.91 – 58.59] (TABLA 5.7). La probabilidad de que ambos argumentos (sujeto y objeto) animados aparezcan en posición preverbal (SOV) es 0,86 veces mayor a que aparezca uno en posición preverbal y otro en posición postverbal (SVO+OVS).

Argumentos postverbales – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	-0.07308	0.17885	-0.409	0.683
SANI+OANI	-0.15006	0.26361	-0.569	0.569
SANI+OINA	-0.05706	0.18754	-0.304	0.761
SINA+OANI	0.47854	0.49023	0.976	0.329
SINA+OINA	-0.27142	0.19717	-1.377	0.169

TABLA 5.7. Resultados del modelo de regresión logística para la animacidad de los argumentos y los argumentos postverbales en euskera.

En cuanto al tipo de argumento que aparece en posición postverbal, los resultados tampoco muestran diferencias significativas. En el caso de los objetos postverbales, no se observan diferencias significativas entre las condiciones (GRÁFICO 5.4a): SANI+OANI vs. SANI+OINA [37,5% vs. 34%, X^2 (1, N = 557) = 0.10952, p < .740, odds ratio = 1.16], SANI+OANI vs. SINA+OANI [37,5% vs. 55,5%, p < .461, Fisher's Exact Test, odds ratio = 0.486] y SANI+OANI vs.

SINA+OINA [37,5% vs. 30%, X^2 (1, N = 298) = 0.73465, p < .740, odds ratio = 1.16]. Del mismo modo, en la frecuencia de uso de sujetos postverbales tampoco se observan diferencias significativas entre las condiciones (GRÁFICO 5.4b): SANI+OANI vs. SANI+OINA [16,7% vs. 26,6%, X^2 (1, N = 494) = 1.2481, p = .263, odds ratio = 0.551], SANI+OANI vs. SINA+OANI [16,7% vs. 20%, p < .1, Fisher's Exact Test, odds ratio = 0.804], y SANI+OANI vs. SINA+OINA [16,7% vs. 21,9%, X^2 (1, N = 260) = 0.24049, p < .623, odds ratio = 0.715].

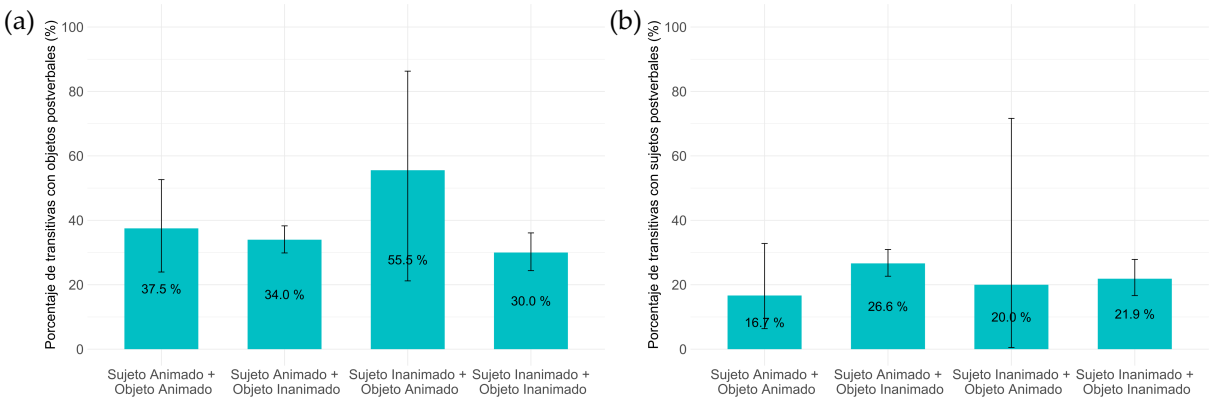


GRÁFICO 5.4. Frecuencia de oraciones transitivas en euskera (a) con objetos postverbales (SVO) y (b) con sujetos postverbales (OVS), agrupadas por la animacidad del sujeto y el objeto. Las barras de error muestran los intervalos de confianza de 95%.

El análisis de regresión logística no muestra una correlación entre el tipo de argumentos postverbales y la animacidad. Es decir, que ambos argumentos sean animados no influye en que el objeto [SANI+OANI: 37,5%, β = -0.06113, z = -0.216, p = .828, odds ratio = 0.94, 95% CI = 23.95 – 52.64] o el sujeto [SANI+OANI: 16,7%, β = -0.28899, z = -0.658, p = .511, odds ratio = 0.74, 95% CI = 06.37 – 32,81] se muevan a posición postverbal (TABLA 5.8). Dicho de otro modo, la probabilidad de que el objeto aparezca en posición postverbal (SVO) cuando ambos argumentos son animados es 0,9 veces menor a que aparezca delante del verbo (SOV), mientras que el sujeto, cuando ambos son animados, tiende a aparecer 0,7 veces menos en posición postverbal (OVS) que en posición preverbal (SVO).

(a) Objeto posverbal (SVO) – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	-0.44970	0.18820	-2.390	0.016 *
SANI+OANI	-0.06113	0.28260	-0.216	0.828
SANI+OINA	-0.21412	0.19949	-1.073	0.283
SINA+OANI	0.67284	0.51031	1.318	0.187
SINA+OINA	-0.39760	0.21200	-1.876	0.060 .
(b) Sujeto posverbal (OVS) – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	-1.32045	0.30489	-4.331	0.001 ***
SANI+OANI	-0.28899	0.43927	-0.658	0.511
SANI+OINA	0.30736	0.31391	0.979	0.328
SINA+OANI	-0.06585	0.84732	-0.078	0.938
SINA+OINA	0.04748	0.32560	0.146	0.884

TABLA 5.8. Resultados del modelo de regresión logística para la animacidad de los argumentos y el tipo de argumentos postverbales en euskera: (a) de objeto (SVO) y (b) de sujeto (OVS).

En esta sección he analizado la omisión de argumentos (sujeto y objeto) y argumentos postverbales en oraciones transitivas en euskera. Los datos señalan que la omisión de argumentos es significativamente mayor en las oraciones transitivas que tienen ambos argumentos animados (sujetos y objetos). En especial, el sujeto es el que se omite con mayor frecuencia en las transitivas cuando ambos argumentos son animados. Por el contrario, los datos de argumentos postverbales señalan que no hay preferencia de uso de oraciones transitivas con argumentos en posición postverbal cuando ambos argumentos son animados en comparación al resto de transitivas. En suma, la animacidad de los argumentos modula su omisión de los argumentos, pero el uso de argumentos en posición postverbal no.

5.4 Estudio de corpus 2: castellano

Este estudio de corpus escrito en castellano busca evidencia de los efectos descritos en los experimentos de Gennari et al. (2012) y de Hsiao et al. (2014). Como he explicado en la introducción (sección 5.2), ambos estudios observan que los hablantes de inglés, castellano, serbio (Gennari et al., 2012) y chino (Hsiao et al., 2014) omiten con mayor frecuencia el sujeto en oraciones transitivas con dos argumentos animados que con uno solo. El castellano presenta un orden de palabras (SVO) donde ambos argumentos aparecen separados por el verbo, de tal forma que no espero encontrar diferencias

significativas en las frecuencias de oraciones transitivas con sujeto omitido y con sujeto no omitido (cf. 5.4a,b), dado que ambos tipos de oraciones mantienen la misma linearización.

(5.4) a. El profesor vio al estudiante [SVO]
b. *pro* vio al estudiante [VO: omisión de sujeto]

5.4.1 Materiales

Las oraciones de castellano las he obtenido del corpus ADESSE [*Base de datos de verbos, alternancias de diátesis y esquemas sintáctico semánticos del español*] (García-Miguel, Vaamonde y González Domínguez, 2010). La obtención de las oraciones las he limitado a oraciones transitivas de textos periodísticos del periódico *La Voz de Galicia*. Del total de oraciones que componen este subcorpus de ADESSE, solo he tenido en cuenta para el análisis estadístico las 2923 oraciones transitivas declarativas (TABLA 5.9), descartando el resto de oraciones (oraciones impersonales, con argumentos expresados mediante clausulas…).

declarativas	**resto de oraciones**	**TOTAL**
2923	266	3189

TABLA 5.9. Distribución del tipo de oraciones transitivas en el corpus escrito de castellano.

5.4.2 Procedimiento

Dado que en el corpus ADESSE ya vienen etiquetadas la omisión y la animacidad de los argumentos, a continuación, detallaré como he llevado a cabo la selección de los diferentes tipos de argumentos (IMAGEN 5.1, interfaz de búsqueda). Los sujetos y los objetos directos los he seleccionado especificando en la opción de búsqueda la función sintáctica de los sintagmas nominales (*frase nominal* en el corpus) y los pronombres personales. El orden de palabras lo he establecido detallando la posición de relativa de los argumentos en cuanto al verbo (*sujeto preverbal + objeto postverbal* = SVO / *sujeto postverbal + objeto preverbal* = OVS). En cuanto la animacidad (*animación* en el corpus), los argumentos animados son todos aquellos etiquetados como "animado" (*animado discontinuo* y *animado colectivo* en el corpus[3]) y los argumentos inanimados los etiquetados como "concreto" y "abstracto"

[3] En ADESSE, los argumentos *animados discontinuos* hacen referencia a entidades animadas y discretas (e.g. *niño, nieto, coronel, funcionario*), y los pronombres personales y los nombres propios de persona (e.g. *yo, tú, Genoveva*). Los argumentos *animados colectivos*, por su parte, hacen referencia a colectivos de entidades animadas (e.g. *gente, muchedumbre*, *sociedad*, *público*, *equipo*).

(*inanimado concreto* e *inanimado abstracto* en el corpus[4]). He considerado como omisión de sujeto seleccionando las categorías sintácticas etiquetadas como "ninguna unidad (vacío)".

IMAGEN 5.1. Ejemplo de búsqueda en la interfaz del corpus ADESSE.

Al igual que en el estudio de corpus en euskera (sección 5.3), he usado la prueba χ^2 (*Pearson chi-square*) para observar si existe una asociación entre que los argumentos de una oración transitiva sean animados y la distribución de los sujetos omitidos. De igual modo, mediante el modelo de regresión logística binomial he analizado si la omisión de sujetos está influenciada por la combinación de animacidad de los argumentos en las oraciones transitivas. La finalidad de este análisis es observar si en castellano el hecho de que el sujeto y el objeto sean animados aumenta la frecuencia de omisión de sujetos. Los análisis los he computado mediante el programa estadístico R (R Core Team, 2017) y usando el paquete *lme4* (Bates et al., 2015). El nivel de referencia del modelo es la media de las medias de las cuatro condiciones (SANI+OANI, SANI+OINA, SINA+OANI, SINA+OINA), dado que las variables de cada condición tienen diferentes frecuencias. De esta forma, el coeficiente de cada condición (*estimate*) representa cuánto se desvía de la gran media. Los resultados los he considerado significativos a un nivel $p < .05$. Los gráficos los he realizado con el paquete *ggplot2* (Wickham, 2009).

[4] En ADESSE, los argumentos *inanimados discontinuos* hacen referencia a objetos físicos contables e incontables (e.g. *camiseta, cigarrillo, libro, mano, verja, cielo, ropa, pólvora, piel, munición*) y los argumentos *inanimados abstractos* hacen referencia a entidades no concretas y abstracciones conceptuales (e.g. *consejo, fragilidad, lenguaje, susto, tiempo*).

5.4.3 Resultados

La TABLA 5.10 muestra la clasificación de las 2923 oraciones transitivas que componen el corpus, según la animacidad del sujeto y del objeto, el orden de palabras, y si la oración tiene algún argumento omitido.

	sujeto						**objeto**					
	animado		**inanimado**		**Total**		**animado**		**inanimado**		**Total**	
SVO	70%	(1063)	30%	(459)	100%	(1522)	8%	(123)	92%	(1399)	100%	(1522)
VO	85%	(1197)	15%	(204)	100%	(1401)	12%	(164)	88%	(1237)	100%	(1401)
TOTAL	77%	(2260)	23%	(663)	100%	(2923)	11%	(287)	89%	(2636)	100%	(2923)

TABLA 5.10. Distribución de oraciones transitivas en castellano según el orden, la animacidad y la omisión del sujeto y el objeto.

Para el análisis de la omisión de sujetos he tenido en cuenta los tipos de oraciones SVO y VO (ver TABLA 5.11 para más detalle). De las 2923 oraciones transitivas utilizadas, 1401 oraciones tienen el sujeto omitido (48%) y 1522 oraciones tienen ambos argumentos explícitos (SOV: 52%) (TABLA 5.11).

	sujeto animado						**sujeto inanimado**						**TOTAL**	
	objeto animado		**objeto inanimado**		**Total**		**objeto animado**		**objeto inanimado**		**Total**			
SVO	10%	(105)	90%	(958)	100%	(1063)	4%	(18)	96%	(441)	100%	(459)	52%	(1522)
VO	13%	(152)	87%	(1045)	100%	(1197)	6%	(12)	94%	(192)	100%	(204)	48%	(1401)
TOTAL	11%	(257)	89%	(2003)	100%	(2260)	5%	(30)	95%	(633)	100%	(663)	100%	(2923)

TABLA 5.11. Distribución de oraciones transitivas en castellano SVO y VO (con omisión del sujeto) según la animacidad del sujeto y el objeto.

La frecuencia de omisión de sujeto (VO) es significativamente mayor cuando el sujeto y el objeto de la oración son animados, frente a cuando solo el sujeto es animado [SANI+OANI vs. SANI+OINA: 59,1% vs. 52,2%, χ^2 (1, N = 2260) = 4.1692, p < .041, odds ratio = 1.32] (GRÁFICO 5.5). Esta misma frecuencia significativa de omisión de sujeto también se observa al comparar oraciones transitivas con sujeto y objeto animados frente a transitivas con sujeto y objeto inanimados [SANI+OANI vs. SINA+OINA: 59,1% vs. 30,3%, χ^2 (1, N = 890) = 62.783, p < .001, odds ratio = 3.32]. La frecuencia de uso de omisión de sujeto entre oraciones transitivas con ambos argumentos animados frente transitivas con sujeto inanimado y objeto animado no es significativa [SANI+OANI vs. SINA+OANI: 59,1% vs. 40%, χ^2 (1, N = 287) = 3.2765, p < .070, odds ratio = 2.17].

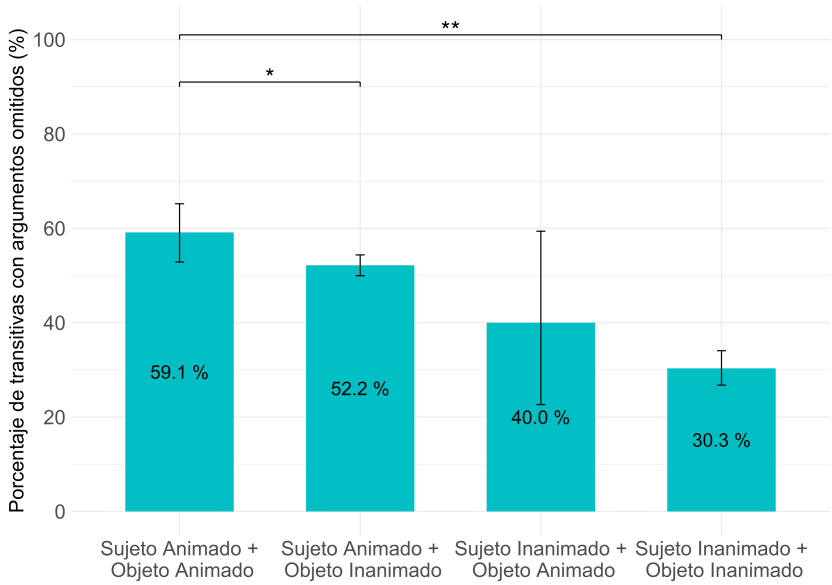


GRÁFICO 5.5. Frecuencia de oraciones transitivas en castellano con sujeto omitido (VO), agrupadas por la animacidad del sujeto y el objeto. Las barras de error muestran los intervalos de confianza de 95%.

El análisis de regresión logística muestra que la omisión del sujeto está correlacionada con que ambos argumentos sean animados [SANI+OANI: 59,1%, $\beta = 0.5650$, $z = 4.173$, < .001, odds ratio = 1.75, 95% CI = 52.82 – 65.21], comparado con el resto de condiciones [SANI+OINA: 52,1%; SINA+OANI: 40%; SINA+OINA: 30,3%] (TABLA 5.12). Dicho de otro modo, la probabilidad de que el sujeto se omita en castellano es 1,75 veces mayor cuando el sujeto y el objeto son animados que en el resto de condiciones. Por tanto, el modelo de regresión logística confirma que la animacidad de los ambos argumentos de una oración transitiva influye significativamente en la probabilidad de la omisión de sujeto.

Omisión de sujeto – Coeficientes:				
	Estimate	**SE**	**z-value**	**p-value**
(Intercept)	-0.1950	0.1014	-1.924	0.054 .
SANI+OANI	0.5650	0.1354	4.173	0.001 ***
SANI+OINA	0.2820	0.1062	2.655	0.007 **
SINA+OANI	-0.2104	0.2824	-0.745	0.456
SINA+OINA	-0.6365	0.1184	-5.376	0.001 ***

TABLA 5.12. Resultados del modelo de regresión logística para la animacidad de los argumentos y la omisión de sujeto en castellano.

En resumen, los análisis revelan diferencias en el uso de sujetos omitidos y sujetos explícitos en oraciones transitivas de castellano en diferentes condiciones. Estas diferencias son significativas: el sujeto tiene más probabilidad de ser omitido en aquellas

oraciones transitivas que tienen tanto el sujeto como el objeto animados que en el resto de transitivas con diferentes combinaciones de animacidad entre el sujeto y el objeto.

5.5 Discusión

En este capítulo he testeado si en euskera y castellano se hacen uso de dos estrategias propuestas en la literatura para reducir la interferencia de animacidad: (a) la omisión de uno de los dos argumentos animados y (b) el incremento de la distancia lineal entre los dos argumentos animados. Los resultados de los estudios de corpus muestran, tanto en euskera como en castellano, una preferencia por la omisión del sujeto en oraciones con dos argumentos animados. Sin embargo, en cuanto al aumento de la distancia lineal, no he encontrado evidencia de que el euskera reduzca la interferencia de animacidad incrementando la distancia lineal entre los dos argumentos animados. Los resultados de ambos estudios de corpus sugieren que la omisión de argumentos, en especial la omisión de sujeto, es una estrategia compartida para reducir la interferencia creada por dos argumentos animados.

5.5.1 Reducir la interferencia omitiendo argumentos

Los resultados de los estudios de corpus revelan que tanto el euskera como el castellano recurren a la omisión de argumentos, particularmente sujetos, cuando ambos argumentos son animados. Este resultado es congruente con la hipótesis de que la omisión de argumentos es una estrategia para minimizar la interferencia creada por dos argumentos animados durante la planificación. En el caso del euskera (ver sección 5.3), argumenté que al tener ambos argumentos adyacentes en posición preverbal (SOV) el uso de argumentos omitidos sería una estrategia que utilizaría para reducir la interferencia creada por dos animados. Mientras que en el caso del castellano (ver sección 5.4), argumenté que no recurriría a la omisión de sujetos al estar ambos argumentos separados por el verbo en la linearización (SVO). Sin embargo, los datos del corpus muestran que el castellano, al igual que el euskera, también recurre a la omisión de argumentos cuando ambos argumentos son animados. Estos resultados convergen con los de estudios previos de Gennari et al. (2012) y Hsiao et al. (2014), quienes encuentran que los hablantes de lenguas VO, que lo permiten, tienden a omitir el sujeto cuando tienen que describir escenas en la que intervienen dos animados. Sin embargo, estos dos estudios previos (Gennari et al., 2012 y Hsiao et al., 2014) difieren tanto en el tipo de oraciones analizadas como en el tipo omisión del sujeto. En Gennari et al. (2012) los hablantes de castellano y serbio producían más frecuentemente oraciones relativas de objeto impersonales cuando describían escenas con dos animados (e.g., *el profesor que saludan…*), en las que el sujeto nulo de la relativa no se pronuncia y el verbo se conjuga en 3ª persona. En Hsiao et al. (2014), por el contrario, los

hablantes de chino preferían omitir el sujeto en oraciones transitivas al describir escenas con dos animados (e.g., *zhāohū le lǎoshī* [lit. *ha saludado al profesor*]), y ese resultado lo replicaban en un estudio de corpus: había mayor frecuencia de omisión de sujeto en transitivas con dos argumentos animados.

En los estudios de corpus de euskera y castellano presentados y discutidos en este capítulo, todas las oraciones eran transitivas simples y el tipo de omisión que se analizaba era el de sujeto y de objeto (en euskera). Lo que diferencia a estas dos lenguas es la linearización básica o canónica de los argumentos en las oraciones transitivas: el euskera sitúa ambos argumentos juntos en posición preverbal, mientras que el castellano los sitúa separados por el verbo. Por ello, esperaba que los efectos de la interferencia de dos animados fuera reducida mediante la omisión del sujeto en euskera y no en castellano, dado que el verbo situado entre dos animados ya reduciría la interferencia (Traxler, Morris y Seely, 2002; Gordon et al., 2006). Los resultados de castellano, sin embargo, muestran un patrón similar a los de euskera (y los de chino: Hsiao et al., 2014): cuando una oración transitiva tiene dos argumentos animados, es más frecuente que se omita el sujeto. La tendencia a omitir sujetos en transitivas con dos animados en castellano (y chino: Hsiao et al., 2014) puede explicarse si se asume que el alcance de la planificación abarca toda una oración (Meyer, 1996) o al menos llega hasta el verbo principal (Lindsley, 1975; Kempen y Hoenkamp, 1987; Ferreira, 1994; Iwasaki, Vinson, Vigliocco, Watanabe y Arciuli, 2008; Sauppe et al., 2013b; Hwang y Kaiser, 2014; Momma et al., 2016; Sauppe, 2017).

La planificación del verbo puede variar entre las lenguas VO y OV. Hwang y Kaiser (2014) observaron que en inglés la planificación abarca el sujeto y el verbo solo. Mediante una tarea de describir imágenes, los hablantes de inglés fijaban su atención durante más tiempo en la región de la acción que en las regiones del sujeto y el objeto antes de describir las imágenes cuando una palabra distractora (que oían los participantes) estaba relacionada semánticamente con el verbo. Hwang y Kaiser (2014) interpretan dicha demora a que en lenguas VO la planificación del verbo sucede antes de empezar a articular el sujeto. En un estudio similar, Momma et al. (2016) observaron, sin embargo, que en las lenguas OV el verbo se planifica junto al objeto solo. En su estudio, los hablantes de japonés tardaban más en empezar a describir imágenes con oraciones SOV (transitivas) que SV (intransitivas) cuando una palabra distractora (superpuesta en la imagen) estaba relacionada semánticamente con el verbo. Momma et al. (2016) concluyen que el verbo se planifica de antemano solo antes de articular el objeto y no el sujeto porque el objeto y el verbo constituirían una unidad integrada que se planifica a la vez. Esta diferencia de planificación podría estar motivada por el tipo de oraciones que usan (intransitivas vs. transitivas), ya que difieren en el número de argumentos que requiere el verbo. Además, la tardanza en empezar a articular las oraciones transitivas podría deberse a que los

hablantes estarían planificando junto al verbo no solo el objeto sino el sujeto también, aunque luego no lo articulen, es decir, estarían planificando toda la oración transitiva. De hecho, varios estudios sugieren que la planificación puede ser toda una oración transitiva (Griffin y Bock, 2000; Sauppe et al., 2013a; Konopka y Meyer, 2014; van de Velde, Meyer y Konopka, 2014; Kurumada y Jaeger, 2015; Norcliffe et al., 2015). Por ejemplo, Griffin y Bock (2000) encuentran que, en el primer momento de ver una imagen (0-400 ms), los hablantes no mostraban preferencia en fijar su atención en ninguno de los personajes de la imagen antes de empezar a describir la escena; e interpretan dicho resultado como evidencia de que los hablantes planifican toda la escena antes de empezar a articularla. Si la planificación abarca toda una proposición/oración transitiva, esto explica los resultados de omisión de sujeto encontrados en euskera y castellano, y previamente en chino (Hsiao et al., 2014).

5.5.2 Incrementar la distancia lineal: reducir la interferencia posponiendo argumentos

Los resultados de euskera muestran que los hablantes de esta lengua no tienden a aumentar la distancia lineal entre dos argumentos animados. Por tanto, este resultado no es congruente con la hipótesis de que el aumento de la distancia lineal de dos argumentos es una estrategia para minimizar la interferencia creada por dos argumentos animados. Argumenté que en euskera (ver sección 5.3), al tener ambos argumentos en posición preverbal (SOV), el uso de argumentos postverbales (SVO y OVS) sería una estrategia posible para reducir la interferencia creada por dos argumentos animados. Mediante este tipo de oraciones, el euskera separaría ambos argumentos en la linearización poniendo el verbo entre ambos, consiguiendo incrementar la distancia lineal entre estos. Los resultados no convergen con los de estudios previos (Loui y Gennari, 2008; Montag y MacDonald, 2009; Gennari et al., 2012; Montag y MacDonald, 2014; Humphreys et al., 2016; Montag et al., 2017), quienes encuentran que los hablantes tienden a producir oraciones pasivas relativas de objeto (e.g., *el profesor que es saludado por la estudiante…*) que activas relativas de objeto (e.g., *el profesor que la estudiante saluda…*) cuando deben describir escenas en la que intervienen dos animados. El uso de pasivas relativas de objeto ayuda a reducir la interferencia de animacidad ya que separa los dos argumentos animados linealmente, mientras que en las activas relativas de objeto aparecen al lado. Humphreys et al. (2016) sugieren que el uso de pasivas se ve motivado por el hecho de que la planificación de dos animados interfiere en la producción del agente animado y su inhibición lo hace menos accesible favoreciendo la producción de oraciones pasivas.

En euskera, al no haber pasivas, predije que recurriría a una mayor frecuencia de oraciones con argumentos postverbales (SVO y OVS) para incrementar la distancia lineal entre ambos argumentos. Esta predicción parte de la hipótesis de que la interferencia es

reducida por la posición del verbo entre ambos argumentos, tal y como proponen Gordon et al. (2006) y Traxler et al. (2002): lo que reduce la interferencia no es el número de elementos que separan ambos argumentos animados, sino que están separados por el verbo. Pero como ya he mencionado, los resultados de euskera no muestran diferencias significativas entre el uso de oraciones transitivas SOV y SVO/OVS cuando ambos argumentos son animados, por lo que no es congruente con la hipótesis propuesta por Gordon et al. (2006) y Traxler et al. (2002), y tampoco con los resultados de estudios en lenguas VOS (England, 1991; Kubo, Ono, Tanaka, Koizumi y Sakai, 2012; Kiyama, Tamaoka, Kim y Koizumi, 2013; Norcliffe et al., 2015). Estos estudios previos encontraron que en lenguas VOS, que tienen ambos argumentos adyacentes como el euskera, los hablantes producían con mayor frecuencia oraciones en la que los argumentos están separados por el verbo (SVO) cuando ambos eran animados que cuando solo uno de ellos lo era. En la misma línea, estudios con lenguaje de signos han encontrado que los participantes describen con mayor frecuencia escenas que contienen dos animados con el orden SVO que con SOV (Meir, Lifshitz Ben-Basat, Ilkbasaran y Padden, 2010; Hall, Mayberry y Ferreira, 2013; Hall, Ferreira y Mayberry, 2014; Futrell et al., 2015; Meir et al., 2017).

5.5.3 ¿Omitir argumentos o incrementar la distancia lineal? ¿Qué reduce mejor la interferencia de animacidad?

El hecho de que en euskera se recurra solo al uso de la omisión de argumentos (la de sujeto especialmente) para reducir la interferencia de dos argumentos animados, puede deberse a que la omisión sea menos costosa que incrementar la distancia lineal. Numerosos estudios en diferentes lenguas reportan que los órdenes derivados tienen un coste de procesamiento mayor que los órdenes canónicos: los hablantes muestran mayores tiempos de lectura en estudios conductuales y negatividades LAN y N400[5] en estudios de ERPs (Rösler, Pechmann, Streb, Röder y Hennighausen, 1998; Bornkessel et al., 2002; Matzke, Mai, Nager, Rüsseler y Münte, 2002; Mazuka, Itoh y Kondo, 2002; Miyamoto, 2002; Röder, Stock, Neville, Bien y Rösler, 2002; Schlesewsky, Bornkessel y Frisch, 2003; Ben-Shachar, Palti y Grodzinsky, 2004; Casado, Martín-Loeches, Muñoz y Fernández-Frías, 2005; Hagiwara, Soshi, Masami y Imanaka, 2007; Demiral et al., 2008; Kinno, Kawamura, Shioda y Sakai, 2008; Wolff, Schlesewsky, Hirotani y Bornkessel-Schlesewsky, 2008a; Erdocia et al., 2009; Wang, Schlesewsky, Bickel y Bornkessel-Schlesewsky, 2009; Wolff, 2010; Erdocia et al., 2012; Koizumi et al., 2014, entre otros). Por ejemplo, Erdocia et al. (2009) y Erdocia (2006) contrastaron en euskera el coste de procesamiento del orden derivado OSV con el

[5] Dado que parece difícil asociar la negatividad encontrada al procesar órdenes derivados con LAN o N400, Bornkessel, Schlesewsky y Friederici (2002), Bornkessel, Schlesewsky y Friederici (2003) y Schlesewsky y Bornkessel (2003) denominan a dicha negatividad "*scrambling negativity*".

del orden básico SOV en un estudio de *self-paced reading* (lectura autoadministrada) y ERPs. Los resultados revelaron que en las oraciones OSV los participantes tardaban más tiempo en leer y mostraban una negatividad (LAN y N400) en la posición del primer argumento y una positividad (P600) en la posición del verbo en comparación con las oraciones SOV. Los resultados indican un coste de procesamiento asociado con los órdenes derivados. En un estudio similar, Erdocia et al. (2012) compararon el coste de procesamiento de los órdenes derivados SVO y OVS y observaron que los participantes procesaban ambos órdenes de manera similar y que ambos órdenes mostraban un incremento de tiempo de lectura en la posición del segundo argumento. Erdocia et al. (2012) sugieren que tal incremento de tiempo podría deberse al coste de procesamiento de que los participantes interpreten al leer el verbo que la oración ha terminado y que es una oración transitiva con sujeto u objeto omitido.

Esta negatividad asociada al coste de procesar oraciones con órdenes no canónicos desaparece si es posible interpretar la oración con omisión, i.e., que alguno de los argumentos está omitido. Wolff (2010) y Wolff et al. (2008a) compararon en japonés el procesamiento de oraciones con orden básico SOV y con orden no canónico OSV en un experimento de ERPs. En el experimento los participantes escuchaban oraciones transitivas SOV y OSV: la mitad de las oraciones OSV tenían una frontera prosódica después del objeto inicial y la otra mitad no. Sin la frontera prosódica la oración OSV se podría interpretar como una oración transitiva con sujeto omitido (SOV), mientras que con la frontera prosódica tal interpretación desaparece y solo es posible interpretar el orden derivado OSV. Los resultados revelaron que las oraciones OSV con frontera prosódica mostraban un mayor ratio de errores y una negatividad N400 en la posición del objeto; pero en ausencia de la frontera prosódica el ratio de errores disminuían, la negatividad desaparecía y se observaba una positividad P600 en la posición del verbo[6]. Wolff (2010) interpreta dicha positividad P600 como efecto de integrar un argumento del discurso, tal y como se ha observado en otras lenguas (Van Petten, Kutas, Kluender, Mitchiner y McIsaac, 1991; van Berkum, Zwitserlood, Bastiaansen, Brown y Hagoort, 2004; Burkhardt, 2006, 2007; Nieuwland, Petersson y Van Berkum, 2007), y no como un coste de procesamiento generado por la omisión. Resultados similares se encontraron en turco, que también puede tener sujetos omitidos. Demiral (2007) y Demiral et al. (2008) con la intención de observar si el primer NP se interpreta como el sujeto, analizaron el procesamiento de oraciones transitivas: la mitad de ellas con sujeto omitido (OV) y la otra mitad con orden OSV. Las oraciones OV del experimento no tenían la marca de caso acusativo, por lo que eran ambiguas. El NP inicial se podía interpretar como el sujeto o

[6] La positividad P600 está asociada con el coste de integrar el objeto al procesar el verbo (Fiebach, Schlesewsky y Friederici, 2002; Felser, Clahsen y Münte, 2003).

como el objeto de la oración y la ambigüedad se resolvía al llegar al verbo gracias a la concordancia verbal. Los resultados mostraron que ninguno de los dos tipos de oraciones causaba una negatividad N400 y que en la posición del verbo las oraciones SOV generaban una positividad P600. Demiral (2007) y Demiral et al. (2008) interpretaron que tales resultados se debían a la posibilidad de la omisión y que la positividad en la posición del verbo no se debería al coste de procesar un argumento omitido, sino que reflejaría el coste asociado al reanálisis del primer NP de sujeto a objeto. Por último, en otro estudio en turco Özge et al. (2013) compararon los tiempos de escucha de oraciones SVO y OVS y observaron que los participantes procesaban más rápido el primer NP si aparecía en acusativo (i.e., objeto) que si aparecía en nominativo (i.e., sujeto). Özge et al. (2013) concluyen que el objeto en posición inicial no representa un coste de procesamiento extra: los participantes al oír una oración que empieza con un objeto la interpretan como una transitiva con sujeto omitido (SOV) y no como una OVS.

En resumen, los resultados sugieren que la omisión no supone un coste extra de procesamiento como el encontrado en oraciones con orden no canónico. Por tanto, a la hora de reducir el coste de procesamiento creado por la interferencia de dos argumentos animados, los hablantes recurrirán antes a la omisión de los argumentos que al uso de estructuras que los separan en la linearización como las oraciones pasivas o los órdenes no canónicos. Esta preferencia por la omisión es consistente con el principio de "Minimalidad" propuesto por Bornkessel y Schlesewsky (2006), según el cual durante el procesamiento de oraciones se favorecen las estructuras simples (i.e., mínimas). Mediante este principio un NP con caso nominativo al inicio de una oración se interpretará como el sujeto de una oración intransitiva (a pesar de que puede ser el sujeto de una transitiva). Un NP con caso acusativo al inicio de una oración, por el contrario, puede analizarse como el comienzo de un orden derivado o como el único argumento de una oración transitiva con sujeto omitido. Esta segunda interpretación se ve favorecida por el principio de "Minimalidad" porque es la estructura más simple. Así, en las lenguas que permiten omisión (e.g., castellano, euskera, japonés, turco…) las oraciones con objeto inicial pueden ser analizadas siempre como el único argumento de una transitiva con sujeto omitido.

5.6 Conclusiones

Los estudios de corpus de este capítulo tenían como objetivo investigar si en oraciones transitivas el euskera y el castellano reducen la expresión de dos argumentos animados recurriendo a estructuras lingüísticas que inhiban la interferencia de animacidad. En especial, observar si ambas lenguas recurren con mayor frecuencia a (a) la omisión de uno de los argumentos o (b) el incremento de la distancia lineal de ambos argumentos.

Los resultados de este capítulo sustentan la hipótesis de que la interferencia creada por dos argumentos animados durante la planificación de una oración se reduce mediante el uso de estructuras lingüísticas que la inhiben. Sin embargo, observo que no todas las lenguas recurren a las mismas estrategias generales para para reducir la interferencia de animacidad. Hemos visto que el uso de argumentos omitidos es una estrategia general a la que recurren todas las lenguas estudiadas (castellano, euskera, chino). En cuanto al uso de argumentos separados en la linearización, los resultados sugieren que no es una estrategia universal. En estudios previos se encontró una preferencia por el uso de pasivas en oraciones relativas de objeto para separar en la linearización dos argumentos animados adyacentes (el antecedente y el NP de la relativa). Sin embargo, en euskera no he encontrado evidencia de dicha preferencia por separar los dos argumentos preverbales moviendo alguno de ellos a posición postverbal, tal y como había predicho.

En conjunto, estos resultados sugieren que todas las lenguas intentan reducir la interferencia de dos argumentos animados, pero que cada lengua recurre a diferentes estrategias y dependiendo de sus características gramaticales y del tipo de oración en el que surja la interferencia. En lo que se refiere a las dos estrategias tratadas en este capítulo (argumentos omitidos y argumentos postverbales), los resultados de euskera y castellano sugieren que siempre que sea posible la omisión de argumentos las lenguas recurrirán preferiblemente a ella para evitar la interferencia de la animacidad, porque garantiza una estructura mínima y, por tanto, menos costosa.

Capítulo 6

Conclusiones generales

Las principales contribuciones de esta tesis doctoral, en la que he investigado la frecuencia de uso de ciertos elementos gramaticales para aligerar el coste de procesamiento en lenguas OV vs. VO, especialmente, en euskera y castellano, son las siguientes:

1. He querido reivindicar que los corpus son una importante fuente información para tratar de responder preguntas que se plantean sobre el uso del lenguaje y las lenguas. Los datos que se obtienen de ellos son frecuencias de uso y la frecuencia de uso refleja efectos en el procesamiento del lenguaje. Además, aunando los datos de corpus con los de estudios experimentales puede ayudar mucho mejor a responder y comprender el impacto de los costes de procesamiento sobre el uso del lenguaje.

2. He demostrado que el orden básico de palabras de la oración se correlaciona con la frecuencia de uso de ciertas características gramaticales de las lenguas: en la frecuencia de uso de nombres y verbos (véase el CAPÍTULO 3), el uso de argumentos omitidos y argumentos postverbales (véase el CAPÍTULO 4) y el uso de argumentos omitidos para reducir interferencias de animacidad (véase el CAPÍTULO 5). En todos estos capítulos he proporcionado evidencia en favor de que las lenguas OV tienden a reducir el número de argumentos expresados en comparación a las lenguas VO. Esto constituye una fuerte evidencia de que las lenguas tienden a minimizar el coste de procesamiento, recurriendo al uso de ciertas estructuras gramaticales.

Como he mencionado, el principal objetivo de esta tesis doctoral ha sido examinar la correlación entre el orden básico de palabras de una lengua y la frecuencia con la que las lenguas recurren a ciertos fenómenos gramaticales.

Para ello, en el CAPÍTULO 2 he llevado a cabo el estudio de corpus más amplio, más representativo y heterogéneo hasta la fecha sobre la distribución de los órdenes de palabras de la oración en euskera para confirmar si el orden de palabras más frecuente, y

por tanto el canónico, es SOV. Hasta ahora, los corpus anteriores encontraban distribuciones diferentes en el orden de palabras más usado en euskera: de Rijk (1969) y Aldezabal et al. (2003) encuentran que el orden SOV es el más frecuente e Hidalgo (1995a, 1995b) y Aske (1997), por el contrario, encuentran que es SVO. He argumentado que esta diferencia se debe a los errores de etiquetaje llevados a cabo por Hidalgo (1995a, 1995b) y Aske (1997). Los resultados de mi estudio de corpus revelan que el orden básico de palabras en euskera es SOV, y por tanto confirman las distribuciones de observadas por de Rijk (1969) y Aldezabal et al. (2003).

En el CAPÍTULO 3, examino la propuesta de Polinsky (2012) de que existe una correlación entre el orden de palabras y el ratio de nombres y verbos. Polinsky (2012) plantea que las lenguas OV tienden a tener un ratio de nombres-verbos mayor que las lenguas VO. Mediante dos estudios de corpus, he demostrado que existe una correlación inversa: las lenguas OV tienden a tener un ratio menor que las lenguas VO. He argumentado que este menor ratio de nombres y verbos en las lenguas se debe a reducir el coste de procesamiento (Hiranuma, 1999; Ueno y Polinsky, 2009) y de planificación de la oración (Seifart et al., 2018), reduciendo el número de argumentos. La reducción de argumentos facilitaría el coste de procesamiento, ya que en las lenguas OV necesitan mantener más argumentos activos en la memoria hasta procesar el verbo que las lenguas VO (Hawkins, 1994, 2004, 2014; Gibson, 1998, 2000), dado que es en ese momento donde los argumentos son interpretados e integrados (Pickering y Barry, 1991; Gibson y Hickok, 1993; Pickering, 1993; Trueswell et al., 1993; Garnsey et al., 1997).

En el CAPÍTULO 4, he evaluado la hipótesis de Ueno y Polinsky (2009) de que las lenguas OV recurren con mayor frecuencia al empleo argumentos omitidos y oraciones intransitivas como estrategia para facilitar el coste de procesamiento. Para testear esa hipótesis he llevado a cabo un estudio de corpus en castellano (VO) y euskera (OV), y los resultados han revelado que el euskera (OV) recurre con mayor frecuencia al uso de argumentos omitidos en oraciones transitivas que el castellano (VO). Sin embargo, no he observado diferencias significativas entre el uso de oraciones intransitivas entre ambas lenguas. Estos resultados, por tanto, contradicen en parte los de Ueno y Polinsky (2009), quienes reportan un mayor uso de oraciones intransitivas en lenguas OV (japonés y turco). He argumentado que la diferencia entre el euskera y el japonés en el uso de oraciones intransitivas se debe a la mayor libertad de orden de palabras que tiene el euskera. El japonés al tener un orden estricto de verbo final recurre con mayor frecuencia al uso de oraciones intransitivas, mientras que el euskera, con mayor libertad en el orden de palabras, lo hace con menor frecuencia que el japonés y prefiere recurrir al uso de oraciones transitivas con argumentos postverbales (SVO-OVS). En suma, la evidencia de euskera y castellano confirman la propuesta de que las lenguas OV tratan de reducir el

coste de procesamiento mediante la reducción de argumentos preverbales, recurriendo al uso de argumentos omitidos, oraciones intransitivas y argumentos postverbales.

La reducción de argumentos preverbales es compatible con la idea de reducir el coste de planificación oracional. Por ello, en el CAPÍTULO 5, exploro la hipótesis de que la omisión argumentos preverbales y el uso de argumentos postverbales ayudan a reducir la interferencia de dos animados durante la planificación oracional. Gennari et al. (2012) y Hsiao et al. (2014) observan que la omisión de los argumentos y el aumento de la distancia lineal entre los argumentos desplazando uno de ellos a posición postverbal están modulados por la animacidad de los argumentos. Por ello, en este capítulo predecía que en oraciones con dos argumentos animados el euskera recurriría con mayor frecuencia a la omisión de argumentos, pero no el castellano. A su vez, predecía que recurriría también con mayor frecuencia a oraciones con argumentos postverbales, dado que en euskera ambos argumentos está linealmente juntos. Para testear dichas predicciones he llevado a cabo dos estudios de corpus en euskera y en castellano. Los resultados han confirmado en parte estas predicciones: (a) hay mayor omisión de argumentos en oraciones transitivas cuando ambos argumentos son animados, pero está omisión sucede tanto en euskera como en castellano; y (b) el euskera no recurre al uso de argumentos postverbales para aumentar la distancia lineal entre dos argumentos animados. Esperaba que el efecto de interferencia de dos argumentos animados fuera reducido solo mediante la omisión en euskera pero no en castellano, dado que al tener este el verbo entre ambos argumentos animados ya se reducía la interferencia (Traxler et al., 2002; Gordon et al., 2006). He argumentado que la omisión de argumentos oraciones transitivas con dos argumentos animados en castellano se puede explicar si se asume que el alcance de la planificación abarca toda una oración (Meyer, 1996; Griffin y Bock, 2000; Sauppe et al., 2013a; Konopka y Meyer, 2014; van de Velde et al., 2014; Kurumada y Jaeger, 2015; Norcliffe et al., 2015). De esta manera, si la planificación abarca toda una oración, también puede haber interferencia de animacidad en la planificación de oraciones transitivas en las que ambos argumentos no estén linealmente juntos, como en las lenguas VO (castellano y chino). En cuanto al incremento de la distancia lineal, esperaba que el euskera recurriera al uso de argumentos postverbales, situando el verbo entre ambos argumentos animados y así reducir la interferencia de animacidad (Gordon et al., 2006 y Traxler et al., 2002). Sin embargo, no he encontrado diferencias en euskera entre las oraciones SOV y SVO-OVS cuando ambos argumentos son animados. He interpretado los resultados como evidencia de que las lenguas recurren a la omisión de argumentos para reducir la interferencia de animacidad por ser menos costosa (Demiral, 2007; Demiral et al., 2008; Wolff et al., 2008a; Wolff, 2010; Özge et al., 2013) que incrementar la distancia lineal entre ambos argumentos mediante el uso de argumentos postverbales.

En conjunto, la evidencia empírica presentada en los estudios de corpus de esta tesis doctoral muestra que el orden básico de palabras (VO-OV) está estrechamente correlacionado con la frecuencia de uso de determinados recursos gramaticales para facilitar el coste de procesamiento. A su vez, he mostrado que los corpus, estudiados con métodos estadísticos apropiados, proveen de evidencia relevante para dar respuesta a teorías e hipótesis psicolingüísticas y tipológicas, y explicar los patrones tipológicos de las lenguas en términos de procesamiento del lenguaje.

Resumen de resultados

- El orden de palabras más frecuente, y por tanto el orden básico de palabras en euskera es SOV.
- Existe una correlación entre el orden básico de palabras y el ratio de nombres-verbos: las lenguas OV tienden a tener un ratio menor que las lenguas VO.
- Existe una correlación entre el orden básico de palabras y la frecuencia de omisión de argumentos en oraciones transitivas para reducir el número de argumentos preverbales: las lenguas OV tienen a omitir con mayor frecuencia el sujeto que las lenguas VO.
- No existe una correlación entre el orden básico de palabras y la frecuencia de uso de oraciones intransitivas para reducir el número de argumentos preverbales.
- En lenguas OV, la diferencia de frecuencia de uso de oraciones intransitivas puede estar motivada por la libertad de movimiento de argumentos: las lenguas OV con orden rígido recurrirán al uso de oraciones intransitivas y las lenguas OV con mayor libertad al uso de argumentos postverbales.
- No existe una correlación entre el orden básico de palabras y la omisión de argumentos en oraciones transitivas con ambos argumentos animados: ambos tipos de lenguas (VO-OV) recurren de igual modo a la omisión de argumentos animados.
- La mayor frecuencia de uso de argumentos omitidos en lenguas OV frente a oraciones intransitivas y argumentos postverbales para aligerar el coste de procesamiento puede deberse a que la omisión es menos costosa que el uso de los otros recursos gramaticales. Es necesaria más investigación para determinar esta hipótesis.

Referencias

Acheson, D. J., y MacDonald, M. C. (2009). Verbal working memory and language production: common approaches to the serial ordering of verbal information. *Psychological Bulletin, 135*(1), 50-68. doi: 10.1037/a0014411

Aggujaro, S., Crepaldi, D., Pistarini, C., Taricco, M., y Luzzatti, C. (2006). Neuroanatomical correlates of impaired retrieval of verbs and nouns: Interaction of grammatical class, imageability and actionality. *Journal of Neurolinguistics, 19*, 175-194.

Akiyama, N., y Akiyama, C. (2002). *Japanese Grammar*. New York: Barron's Educational Series.

Aldezabal, I., Aranzabe, M. J., Arriola, J. M., y Díaz de Ilarraza, A. (2009). Syntactic annotation in the Reference Corpus for the Processing of Basque (EPEC): Theoretical and practical issues. *Corpus Linguistics and Linguistic Theory, 5*(2), 241-269. doi: 10.1515/CLLT.2009.010

Aldezabal, I., Aranzabe, M. J., Atutxa, A., Gojenola, K., Sarasola, K., y Zabala, I. (2003). *Hitz-hurrenkeraren azterketa masiboa corpusean*.

Allum, P. H., y Wheeldon, L. (2007). Planning scope in spoken sentence production: The role of grammatical units. *Journal of Experimental Psychology, 33*(4), 791-810.

Allum, P. H., y Wheeldon, L. (2009). Scope of Lexical Access in Spoken Sentence Production: Implications for the Conceptual–Syntactic Interface. *Journal of Experimental Psychology, 35*(5), 1240-1255. doi: 10.1037/a0016367

Andor, J. (2004). The master and his performance: an interview with Noam Chomsky. *Intercultural Pragmatics, 1*(1), 93-112.

Arantzeta, M., Bastiaanse, R., Burchert, F., Wieling, M., Martinez-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., Martinez-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.

Aranzabe, M. J. (2008). *Dependentzia-ereduan oinarritutako baliabide sintatikoak: zuhaitz-bankua eta gramatika konputazionala.* (PhD Dissertation), Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU).

Arnold, J. E. (1998). *Reference form and discourse patterns.* (PhD Dissertation), Stanford University.

Arnold, J. E., Wasow, T., Losongco, A., y Ginstrom, R. (2000). Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. *Language, 76*, 28-55.

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.

Azkue, R. M. (1905). *Diccionario Vasco-Español-Frances* (1984 ed.). Bilbao: Euskaltzaindia.

Baker, K., y Brew, C. (2010). Multilingual animacy classification by sparse logistic regression. *Oiho State University Working Papers in Linguistics (OSUWPL), 59*, 52-75.

Baker, M. C. (2003). Lexical categories and the nature of the grammar *Lexical Categories: Verbs, Nouns and Adjectives* (pp. 264-302). Cambridge: Cambridge University Press.

Bassano, D. (2000). Early development of nouns and verbs in French: exploring the interface between lexicon and grammar. *Journal of Child Language, 27*, 521-559.

Batchelor, R. E., y San José, M. Á. (2010). *A Reference Grammar of Spanish*. Cambridge: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., y Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi: 10.18637/jss.v067.i01

Beckwith, R., Fellbaum, C., Gross, D., y Miller, G. A. (1991). WordNet: A Lexical Database Organized on Psycholinguistics Principles. In U. Zernik (Ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. New Jersey: Lawrence Erlbaum Associates.

Ben-Shachar, M., Palti, D., y Grodzinsky, Y. (2004). Neural correlates of syntactic movement: converging evidence from two fMRI experiments. *NeuroImage, 21*(4), 1320-1336. doi: 10.1016/j.neuroimage.2003.11.027

Bentivoglio, P. (1992). Linguistic correlations between subjects of one-argument verbs and subjects of more-than-one-argument verbs in spoken Spanish. In P. Hirschbühler & K. Koerner (Eds.), *Romance languages and modern linguistic theory: 20th Linguistic Symposium on Romance Languages* (Vol. LSRL XX, pp. 11-24). Amsterdam: John Bejamins.

Betancort, M., Carreiras, M., y Sturt, P. (2009). The processing of subject and object relative clauses in Spanish: An eye-tracking study. *The Quarterly Journal of Experimental Psychology, 62*(10), 1915-1929. doi: 10.1080/17470210902866672

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the Development of Language* (pp. 279-362). New York: Wiley & Sons.

Biber, D., Conrad, S., y Cortes, V. (2004). If you look at…: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, D., Conrad, S., y Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., y Jones, J. K. (2009). Corpus Linguistics: An International Handbook. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (Vol. 2, pp. 1287-1304). Berlin: Walter de Gruyter.

Bloom, L. (1970). *Language Development: Form and Function in Emerging Grammars*. Cambridge, MA: MIT Press.

Bloom, P. (1990). Subjectless sentences in child language. *Linguistic Inquiry, 21*, 491-504.

Bloom, P. (1993). Grammatical continuity in language development: The case of subjectless sentences. *Linguistic Inquiry, 24*, 721-734.

Bock, J. K., y Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 945-984). New York: Academic Press.

Boland, J. E., Tanenhaus, M. K., y Garnsey, S. M. (1990). Evidence for the immediate use of verb control information in sentence processing. *Journal of Memory and Language, 29*(4), 413-432. doi: 10.1016/0749-596X(90)90064-7

Bornkessel, I., y Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review, 113*(4), 787-821. doi: 10.1037/0033-295X.113.4.787

Bornkessel, I., Schlesewsky, M., y Friederici, A. D. (2002). Grammar overrides frequency: evidence from the online processing of flexible word order. *Cognition, 85*(2), B21-B30. doi: 10.1016/S0010-0277(02)00076-8

Bornkessel, I., Schlesewsky, M., y Friederici, A. D. (2003). Contextual information modulates initial processes of syntactic integration: The role of inter- versus intrasentential predictions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5), 871-882. doi: 10.1037/0278-7393.29.5.871

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M.-G., Ruel, J., Venuti, P., y Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development, 75*(4), 1115-1139. doi: 10.1111/j.1467-8624.2004.00729.x

Bortolini, U., Tagliavini, C., y Zampolli, A. (1971). *Lessico di frequenza della lingua italiana contemporanea*. Pisa: IBM Italia.

Braine, M. D. S. (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development, 41*, 1-104.

Bresnan, J., Cueni, A., Nikitina, T., y Baayen, H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer & J. Zwarts (Eds.), *Cognitive Foundations of Interpretation* (pp. 69-94). Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Bresnan, J., y Hay, J. (2008). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua, 118*, 245-259. doi: 10.1016/j.lingua.2007.02.007

Brezina, V. (2018). *Statistics in Corpus Linguistics*. Cambridge, UK: Cambridge University Press.

Brown-Schmidt, S., y Konopka, A. E. (2008). Little houses and casas pequeñas: message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition, 109*(2), 274–280. doi: 10.1016/j.cognition.2008.07.011

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language, 98*(2), 159-168. doi: 10.1016/j.bandl.2006.04.005

Burkhardt, P. (2007). The P600 re£ects cost of new information in discourse memory. *Cognitive Neuroscience and neuropsychology, 18*(17), 1851-1854.

Cappa, S. F., y Perani, D. (2003). The neural correlates of noun and verb processing. *Journal of Neurolinguistics, 16*, 183-189.

Caramazza, A., y Hillis, A. E. (1991). Lexical organization of nouns and verbs in the brain. *Nature, 349*(6312), 788-790. doi: 10.1038/349788a0

Casado, P., Martín-Loeches, M., Muñoz, F., y Fernández-Frías, C. (2005). Are semantic and syntactic cues inducing the same processes in the identification of word order? *Cognitive Brain Research, 24*(3), 526-543. doi: 10.1016/j.cogbrainres.2005.03.007

Casart, Y., y Iribarren, C. (2007). Proporción de sustantivos y verbos en el habla del cuidador y en el léxico temprano en español. *Boletín de Lingüística, XIX*, 42-69.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., y Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10*, 155-199.

Comrie, B. (1989). *Language Universals and Linguistic Typology* (2 ed.). Chicago: The University of Chicago Press.

Croft, W. (1991). *Syntactic Categories and Grammatical Relations*. Chicago: The University of Chicago Press.

Chafe, W. (1976). Givenness, contrastiveness, defmiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and Topic* (pp. 25-55). New York: Academic Press.

Chang, S.-J. (1996). *Korean*. Amsterdam: John Benjamins.

Choi, H.-W. (2007). Length and Order: A Corpus Study of Korean Dative-Accusative Construction. *Discourse and Cognition, 14*(3), 207-227. doi: 10.15718/discog.2007.14.3.207

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Christianson, K., y Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition, 98*(2), 105-135. doi: 10.1016/j.cognition.2004.10.006

Chung, S. (2012). Are lexical categories universal? The view of Chamorro. *Theoretical Linguistics, 38*(1-2), 1–56. doi: 10.1515/tl-2012-0001

Daniele, A., Giustolisi, L., Silveri, M. C., Colosimo, C., y Gainotti, G. (1994). Evidence for a possible neuroanatomical basis for lexical processing of nouns and verbs. *Neuropsychologia, 32*(11), 1325-1341.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990-present. Brigham Young University. Disponible en http://corpus.byu.edu/coca/

De Houwer, A., y Gillis, S. (1998). Dutch child language: an overview. In A. De Houwer & S. Gillis (Eds.), *The Acquisition of Dutch* (pp. 1-100). Amsterdam: John Benjamins.

de Rijk, R. P. G. (1969). Is Basque an SOV Language? *Fontes Linguae Vasconum, 1*, 319-351.

Demiral, S. B. (2007). *Incremental argument interpretation in turkish sentence comprehension.* (PhD Dissertation), Universität Leipzig.

Demiral, S. B., Schlesewsky, M., y Bornkessel-Schlesewsky, I. (2008). On the universality of language comprehension strategies: Evidence from Turkish. *Cognition, 106*(1), 484-500. doi: 10.1016/j.cognition.2007.01.008

Dennison, H. Y. (2008). Universal versus language-specific conceptual effects on shifted word-order production in Korean: Evidence from bilinguals. *Working Papers in Linguistics: University of Hawai'i at Mânoa, 39*(2), 1-16.

Desagulier, G. (2017). *Corpus Linguistics and Statistics with R*. Cham: Springer.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology, 25*(2), 108-127. doi: 10.1016/j.newideapsych.2007.02.002

Dixon, R. M. W. (1979). Ergativity. *Language, 55*, 59-138.

Dixon, R. M. W. (2010). *Basic Linguistic Theory: Grammatical Topics* (Vol. 2). New York: Oxford University Press.

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.

Du Bois, J. W. (1987). The discourse basis of ergativity. *Language, 63*, 805-855.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 143-188. doi: 10.1017/S0272263102002024

England, N. C. (1991). Changes in Basic Word Order in Mayan Languages. *International Journal of American Linguistics, 57*(4), 446-486. doi: 10.1086/ijal.57.4.3519735

Erdocia, K. (2006). *Euskal Hitz Hurrenkerak Azterketa Psikolinguistiko eta Neurolinguistikoen Bidez.* (PhD Dissertation), Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU).

Erdocia, K., Laka, I., Mestres-Misse, A., y Rodriguez-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.

Euskaltzaindia. (1991). *Euskal Gramatika. Lehen Urratsak I*. Bilbao: Euskaltzaindia.

Fedorenko, E., y Gibson, E. (2008). *A resource-based account of animacy effects in the processing of relative clauses*. Poster presentado en The 20th Annual CUNY Conference on Human Sentence Processing, San Diego

Felser, C., Clahsen, H., y Münte, T. F. (2003). Storage and integration in the processing of filler-gap dependencies: An ERP study of topicalization and wh-movement in German. *Brain and Language, 87*(3), 345-354. doi: 10.1016/S0093-934X(03)00135-4

Ferreira, F. (1994). Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language, 33*, 715-736.

Ferreira, F., y Swets, B. (2002). How Incremental Is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums. *Journal of Memory and Language, 45*(1), 57-84. doi: 10.1006/jmla.2001.2797

Ferreira, F., y Swets, B. (2005). The production and comprehension of presumptive pronouns in relative clause ''island'' contexts. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 263-278). New Jersey: Lawrence Erlbaum Associates.

Ferreira, V. S., y Firato, C. E. (2002). Proactive interference effects on sentence production. *Psychonomic Bulletin & Review, 9*(4), 795-800. doi: 10.3758/bf03196337

Ferreira, V. S., y Yoshita, H. (2003). Given-New Ordering Effects on the Production of Scrambled Sentences in Japanese. *Journal of Psycholinguistic Research, 32*(6), 669-692. doi: 10.1023/a:1026146332132

Fiebach, C. J., Schlesewsky, M., y Friederici, A. D. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: the processing of German WH-questions. *Journal of Memory and Language, 47*(2), 250-272. doi: 10.1016/S0749-596X(02)00004-9

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., y Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods, 45*(4), 1182-1190. doi: 10.3758/s13428-012-0313-y

Frazier, L. (1979). Parsing and constraints on word order. *University of Massachusetts occasional papers in linguistics, 5*, 177-198.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Kartunnen & A. M. Zwicky (Eds.), *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives* (pp. 129-189). Cambridge: Cambridge University Press.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., y Fedorenko, E. (2018). The Natural Stories Corpus. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk,

S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 76-82). Miyazaki: European Languages Resources Association (ELRA).

Futrell, R., Hickey, T., Lee, A., Lim, E., Luchkina, E., y Gibson, E. (2015). Cross-linguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages. *Cognition, 136*, 215-221. doi: 10.1016/j.cognition.2014.11.022

García-Miguel, J. M., Vaamonde, G., y González Domínguez, F. (2010). ADESSE, a Database with Syntactic and Semantic Annotation of a Corpus of Spanish. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 1903-1910). Valletta: European Language Resources Association (ELRA).

Garnsey, S. M., Pearlmutterb, N. J., Myersa, E., y Lotockyc, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*, 58-93.

Garreston, G., O'Connor, M. C., Skarabela, B., y Hogan, M. (2004). *Coding practices used in the project optimal typology of determiner phrases*.

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language Production: Vol.1. Speech and Talk* (pp. 177-220). London: Academic Press.

Gennari, S. P., Mirkovic, J., y MacDonald, M. C. (2012). Animacy and competition in relative clause production: a cross-linguistic investigation. *Cognitive Psychology, 65*(2), 141-176. doi: 10.1016/j.cogpsych.2012.03.002

Gentner, D. (1982). Why Nouns are Learned Before Verbs: Linguistic Relativity versus Natural Partitioning. In K. S. (Ed.), *Language Development, vol.2: Language, cognition and culture* (Vol. 2).

Gentner, D., y Boroditsky, L. (2009). Early acquisition of nouns and verbs: Evidence from Navajo. In V. Gathercole (Ed.), *Routes to language: Studies in honor of Melissa Bowerman* (pp. 5-36). New York: Taylor & Francis.

Gibson, E. (1991). *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown.* (PhD Dissertation), Carnegie Mellon University.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition, 68*(1), 1-76.

Gibson, E. (2000). The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity. In A. Marantz, Y. Miyashita & W. O'Neil (Eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium* (pp. 95-126). Cambridge: MIT Press.

Gibson, E., y Hickok, G. (1993). Sentence processing with empty categories. *Language and Cognitive Processes, 8*, 147-161.

Gilquin, G., y Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5*(1), 1-26.

Goodglass, H., Klein, B., Carey, P., y Jones, K. (1966). Specific semantic word categories in aphasia. *Cortex, 2*, 74-89.

Gordon, P. C., Hendrick, R., y Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1411-1423. doi: 10.1037/0278-7393.27.6.1411

Gordon, P. C., Hendrick, R., y Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language, 51*(1), 97-114. doi: 10.1016/j.jml.2004.02.003

Gordon, P. C., Hendrick, R., Johnson, M., y Lee, Y. (2006). Similarity-Based Interference During Language Comprehension: Evidence from Eye Tracking During Reading. *Journal of Experimental Psychology, 32*(6), 1304-1321. doi: 10.1037/0278-7393.32.6.1304

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.

Gries, S. T. (2009). What is Corpus Linguistics? *Language and Linguistics Compass, 3*(5), 1225-1241. doi: 10.1111/j.1749-818x.2009.00149.x

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *A mosaic of corpus linguistics: selected approaches* (pp. 269-291). Frankfurt am Main: Peter Lang.

Gries, S. T. (2011). Methodological and interdisciplinary stance in Corpus Linguistics. In V. Viana, S. Zyngier & G. Barnbrook (Eds.), *Perspectives on Corpus Linguistics* (pp. 81-98). Amsterdam: John Benjamins.

Gries, S. T. (2013). *Statistics for Linguistics with R* (2 ed.). Berlin: De Gruyter Mouton.

Gries, S. T. (2016). *Quantitative Corpus Linguistics with R* (2 ed.). New York: Routledge.

Gries, S. T., y Berez, A. L. (2017). Linguistic Annotation in/for Corpus Linguistics. In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 379-409). Dordrecht: Springer.

Gries, S. T., y Newman, J. (2014). Creating and using corpora. In R. J. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 263-293). Cambridge: Cambridge University Press.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition, 82*(1), B1-B14. doi: 10.1016/S0010-0277(01)00138-X

Griffin, Z. M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psychonomic Bulletin & Review, 10*(3), 603-609. doi: 10.3758/bf03196521

Griffin, Z. M., y Bock, J. K. (2000). What the Eyes Say About Speaking. *Psychological Science, 11*(4), 274-279. doi: 10.1111/1467-9280.00255

Hagiwara, H., Soshi, T., Masami, I., y Imanaka, K. (2007). A Topographical Study on the Event-related Potential Correlates of Scrambled Word Order in Japanese Complex Sentences. *Journal of Cognitive Neuroscience, 19*(2), 175-193.

Hagiwara, M. (n.d.). JEITA Public Morphologically Tagged Corpus. Disponible en http://lilyx.net/nltk-japanese-corpus/#jeitac

Hale, J. (2001). *A probabilistic earley parser as a psycholinguistic model*Trabajo presentado en the Conference Name|, Conference Location|.

Hall, M. L., Ferreira, V. S., y Mayberry, R. I. (2014). Investigating Constituent Order Change With Elicited Pantomime: A Functional Account of SVO Emergence. *Cognitive Science, 38*(5), 943-972. doi: 10.1111/cogs.12105

Hall, M. L., Mayberry, R. I., y Ferreira, V. S. (2013). Cognitive constraints on constituent order: Evidence from elicited pantomime. *Cognition, 129*(1), 1-17. doi: 10.1016/j.cognition.2013.05.004

Hawkins, J. A. (1983). *Word Order Universals*. New York: Academic Press.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hawkins, J. A. (2000). The Relative Order of Prepositional Phrases in English: Going beyond Manner–Place–Time. *Language Variation and Change, 11*, 231-266.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.

Helgadóttir, S. (n.d.). Mörkuð íslensk málheild [The Tagged Icelandic Corpus]. The Árni Magnússon Institute for Icelandic Studies. Disponible en http://www.malfong.is

Heylighen, F., y Dewaele, J.-M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science, 7*, 293-340.

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.

Hillis, A. E., y Caramazza, A. (1995). Representation of grammatical categories of words in the brain. *Journal of Cognitive Neuroscience, 7*(3), 396-407. doi: 10.1162/jocn.1995.7.3.396

Hiranuma, S. (1999). Syntactic difficulty in English and Japanese: A textual study. *UCL Working Papers in Linguistics, 11*, 309-322.

Hsiao, Y., Gao, Y., y MacDonald, M. C. (2014). Agent-patient similarity affects sentence structure in language production: evidence from subject omissions in Mandarin. *Frontiers in Psychology, 5*(Article 1015), 1-12. doi: 10.3389/fpsyg.2014.01015

Hsiao, Y., y MacDonald, M. C. (2016). Production predicts comprehension: Animacy effects in Mandarin relative clause processing. *Journal of Memory and Language, 89*, 87-109. doi: 10.1016/j.jml.2015.11.006

Hualde, J. I., Olarrea, A., Escobar, A. M., y Travis, C. E. (2010). *Introducción a la lingüística hispánica*. Cambridge: Cambridge University Press.

Hualde, J. I., y Ortiz de Urbina, J. (2003). *A Grammar of Basque*. Berlin: de Gruyter.

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language, 70*(2), 331-339.

Humphreys, G. F., Mirković, J., y Gennari, S. P. (2016). Similarity-based competition in relative clause production and comprehension. *Journal of Memory and Language, 89*, 200-221. doi: 10.1016/j.jml.2015.12.007

Hwang, H., y Kaiser, E. (2014). The Role of the Verb in Grammatical Function Assignment in English and Korean. *Journal of Experimental Psychology, 40*(5), 1363-1376. doi: 10.1037/a0036797

Imai, M., Haryu, E., Okabe, R., Lianjing, L., y Shigematsu, J. (2006). Revisiting the Noun-Verb Debate: A Cross-Linguistic Comparison of Novel Noun and Verb Learning in English-, Japanese-, and Chinese-Speaking Children. In K. A. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action Meets Word: How children learn verbs* (pp. 450-477): Oxford University Press.

Iwasaki, N., Vinson, D. P., Vigliocco, G., Watanabe, M., y Arciuli, J. (2008). Naming action in Japanese: Effects of semantic similarity and grammatical class. *Language and Cognitive Processes, 23*(6), 889-930. doi: 10.1080/01690960801916196

Jackson–Maldonado, D., Thal, D., Marchman, V., Bates, E., y Gutierrez– Clellen, V. (1993). Early lexical development in Spanish–speaking infants and toddlers. *Journal of Child Language, 20*(3), 523-549.

Jenset, G. B. (2008). *Basic statistics for corpus linguistics*.

Juilland, A., y Traversa, V. (1973). *Frequency Dictionary of Italian Words*. The Hague: de Gruyter.

Kaiser, E., Ichiwaka, Y., Kobayashi, N., y Yamamoto, H. (2010). *Japanese: A Comprehensive Grammar*. Abingdon, Oxon: Routledge.

Kameyama, M. (1985). *Zero anaphora: The case of Japanese.* (PhD Dissertation), Stanford University.

Kameyama, M. (1988). Japanese zero pronominal binding: Where syntax and discourse meet. In W. J. Poser (Ed.), *2nd International Workshop on Japanese Syntax* (pp. 47–73). Stanford: CSLI Publications.

Kempen, G., y Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science, 11*, 201-258.

Kennedy, A., Hill, R., y Pynte, J. (2003). *The Dundee corpus*. presentado en The 12th European Conference on Eye Movement, Dundee

Khurshudian, V. G., Danie, M. A., Levonian, D. V., Plungian, V. A., Polyakov, A. E., y Rubakov, S. V. (2009). EANC: Eastern Armenian National Corpus. Corpus Technologies. Disponible en http://eanc.net

Kim, J.-B., y Yang, J. (2007). On the Syntax and Semantics of the Bound Noun Constructions: With a Computational Implementation *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation* (pp. 223-233). Seoul: The Korean Society for Language and Information (KSLI).

Kinno, R., Kawamura, M., Shioda, S., y Sakai, K. L. (2008). Neural correlates of noncanonical syntactic processing revealed by a picture-sentence matching task. *Human Brain Mapping, 29*(9), 1015-1027. doi: 10.1002/hbm.20441

Kiyama, S., Tamaoka, K., Kim, J., y Koizumi, M. (2013). Effect of Animacy on Word Order Processing in Kaqchikel Maya. *Open Journal of Modern Linguistics, 3*(3), 203-207.

Kizach, J. (2012). Evidence for Weight Effects in Russian. *Russian Linguistics, 36*(3), 251-270. doi: 10.1007/s11185-012-9096-0

Kliegl, R., Nuthmann, A., y Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology, 135*(1), 12-35. doi: 10.1037/0096-3445.135.1.12

Koizumi, M., Yasugi, Y., Tamaoka, K., Kiyama, S., Kim, J., Ajsivinac Sian, J. E., y García Mátzar, L. P. O. (2014). On the (non)universality of the preference for subject-object word order in sentence comprehension: A sentence-processing study in Kaqchikel Maya. *Language, 90*(3), 722-736.

Kondo, T., y Yamashita, H. (2011). Why Speakers Produce Scrambled Sentences: An Analysis of a Spoken Language Corpus in Japanese. In H. Yamashita, Y. Hirose & J. L. Packard (Eds.), *Processing and Producing Head-final Structures* (pp. 195-215). Dordrecht: Springer Netherlands.

Konopka, A. E. (2009). *Planning ahead: how recent experience with structures and words changes the scope of linguistic planning.* (PhD Dissertation), University of Illinois at Urbana-Champaign.

Konopka, A. E. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language, 66*(1), 143-162. doi: 10.1016/j.jml.2011.08.003

Konopka, A. E., y Meyer, A. (2014). Priming sentence planning. *Cognitve Psychology, 73*, 1-40. doi: 10.1016/j.cogpsych.2014.04.001

Kubo, T., Ono, H., Tanaka, M. N., Koizumi, M., y Sakai, H. (2012). *How does animacy affect word order in a VOS language?* Poster presentado en The 25th Annual CUNY Conference on Human Sentence Processing, New York

Kurumada, C., y Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Jorunal of Memory and Language, 83*, 152-178. doi: 10.1016/j.jml.2015.03.003

Laka, I. (1993). Unergatives that assign ergative, unaccusatives that assign accusative. *MIT Working Papers in Linguistics, 18*(Papers on Case and Agreement), 149-172.

Laka, I. (1996). *A Brief Grammar of Euskara, the Basque Language* Disponible en www.ehu.eus/es/web/eins/a-brief-grammar-of-euskara

Lambrecht, K. (1994). *Information structure and sentence form. Topic, focus, and the mental representations of discourse referents.* Cambridge: Cambridge University Press.

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., y Kliegl, R. (2019). Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods, 51*(3), 1161-1178. doi: 10.3758/s13428-018-1051-6

Lee, E.-K., Brown-Schmidt, S., y Watson, D. G. (2013). Ways of looking ahead: Hierarchical planning in language production. *Cognition, 129*(3), 544-562. doi: 10.1016/j.cognition.2013.08.007

Lee, W.-J., y Choi, K.-S. (1999). *모듈화된 형태소 분석기의 구현*. presentado en 한글 및 한국어 정보처리 학술대회-형태소 분석기 및 품사태거 평가 워크숍, 전주. Recuperado de http://semanticweb.kaist.ac.kr/home/index.php/Corpus2

Lee, Y., Lee, H., y Gordon, P. C. (2007). Linguistic complexity and information structure in Korean: Evidence from eye-tracking during reading. *Cognition, 104*(3), 495-534. doi: 10.1016/j.cognition.2006.07.013

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177. doi: 10.1016/j.cognition.2007.05.006

Lindsley, J. R. (1975). Producing simple utterances: How far ahead do we plan? *Cognitive Psychology, 7*, 1-19.

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 Filoloxía, 24*, 45-82.

Loui, S., y Gennari, S. P. (2008). The role of animacy in the production of Greek relative clauses. In A. Botinis (Ed.), *The proceedings of 2nd Tutorial and Research Workshop on Experimental Linguistics* (pp. 145-148). Athens: National and Kapodistrian University of Athens.

Lowder, M. W., y Gordon, P. C. (2014). Effects of animacy and noun-phrase relatedness on the processing of complex sentences. *Memory & Cognition, 42*(5), 794-805. doi: 10.3758/s13421-013-0393-7

Luke, S. G., y Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods, 50*(2), 826-833. doi: 10.3758/s13428-017-0908-4

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology, 4*(Article 226), 1-16.

MacDonald, M. C. (2015). The Emergence of Language Comprehension. In B. MacWhinney & W. O'Grady (Eds.), *The Handbook of Language Emergence* (pp. 81-99). Oxford: Wiley-Blackwell.

MacDonald, M. C. (2016). Speak, Act, Remember. *Current Directions in Psychological Science, 25*(1), 47-53. doi: 10.1177/0963721415620776

MacDonald, M. C., Pearlmutterb, N. J., y Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676-703.

Maital, S. L., Dromi, E., Sagi, A., y Bornstein, M. H. (2000). The Hebrew communicative development inventory: language specific properties and cross–linguistic generalizations. *Journal of Child Language, 27*(1), 43-67.

Mak, W. M., Vonk, W., y Schriefers, H. (2002). The Influence of Animacy on Relative Clause Processing. *Journal of Memory and Language, 47*, 50-68. doi: 10.1006/jmla.2001.2837

Martí, M. A., Taulé, M., Bertran, M., y Màrquez, L. (2008). AnCora: Multilingual and Multilevel Annotated Corpora. CLiC- Centre de Llenguatge i Computació, Universitat de Barcelona. Disponible en http://clic.ub.edu/corpus/es/ancora-descarregues

Martín Herrero, C. (2009). Aproximación a ciertas perspectivas en lingüística de corpus. In P. Cantos Gómez & A. Sánchez Pérez (Eds.), *A Survey on Corpus-based Research / Panorama de investigaciones basadas en corpus* (pp. 1020-1032). Murcia: AELINCO (Asociación Española de Lingüística del Corpus).

Martin, R. C., Crowther, J. E., Knight, M., Tamborello II, F. P., y Yang, C.-L. (2010). Planning in sentence production: Evidence for the phrase as a default planning scope. *Cognition, 116*, 177-192. doi: 10.1016/j.cognition.2010.04.010

Matzke, M., Mai, H., Nager, W., Rüsseler, J., y Münte, T. (2002). The costs of freedom: an ERP – study of non-canonical sentences. *Clinical Neurophysiology, 113*(6), 844-852. doi: 10.1016/S1388-2457(02)00059-7

Mayer, M. (1969). *Frog, where are you?* . New York: Dial.

Mazuka, R., Itoh, K., y Kondo, T. (2002). Cost of scrambling in Japanese sentence processing. In M. Nakayama (Ed.), *Sentence processing in East Asian languages*. Stanford: CSLI.

Mazuka, R., Lust, B., Wakayama, T., y Snyder, W. (1986). Distinguishing effects of parameters in early syntax acquisition: A cross-linguistic study of Japanese and English. *Papers and Reports on Child Language Development, 25*, 73-82.

McDonald, J. L., Bock, J. K., y Kelly, M. H. (1993). Word and World Order: Semantic, Phonological, and Metrical Determinants of Serial Position. *Cognitive Psychology, 25*, 188-230. doi: 10.1006/cogp.1993.1005

McEnery, T., y Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Meir, I., Aronoff, M., Börstell, C., Hwang, S.-O., Ilkbasaran, D., Kastner, I., Lepic, R., Lifshitz Ben-Basat, A., Padden, C., y Sandler, W. (2017). The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. *Cognition, 158*, 189-207. doi: 10.1016/j.cognition.2016.10.011

Meir, I., Lifshitz Ben-Basat, A., Ilkbasaran, D., y Padden, C. (2010). The interaction of animacy and word order in human languages: A study of strategies in a novel communication task. In A. Smith, D. M., M. Schouwstra, B. de Boer & K. Smith (Eds.), *The evolution of language. Proceedings of the 8th international conference (EvoLang8)* (pp. 455-456). Singapore: World Scientific Publishing.

Meurer, P. (2011). The Georgian National Corpus. Uni Research Computing. Disponible en http://gnc.gov.ge/

Meyer, A. S. (1996). Lexical access in sentence production: results from picture-word interference experiments. *Journal of Memory and Language, 35*, 477-496.

Meyer, A. S., Sleiderink, A. M., y Levelt, W. J. M. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition, 66*(2), B25-B33. doi: 10.1016/S0010-0277(98)00009-2

Meyer, C. F., y Tao, H. (2005). Response to Newmeyer's 'Grammar is grammar and usage is usage'. *Language, 81*(1), 226-228.

Meyerhoff, M. (2000). The emergence of creole subject–verb agreement and the licensing of null subjects. *Language Variation and Change, 12*, 203-230.

Miceli, G., Silveri, C., Villa, G., y Caramazza, A. (1984). On the basis for the agrammatic's difficulty in producing main verbs. *Cortex, 20*, 207-220.

Miller, G. A. (1956a). Human memory and the storage of informartion. *Ire Transactions on Information Theory, 2*, 129-137.

Miller, G. A. (1956b). The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information. *Psychological Review, 101*, 343-352.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., y Miller, K. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography, 3*(4), 235-244. doi: 10.1093/ijl/3.4.235

Miller, G. A., y Chomsky, N. (1963). Finitary Models of Language Users. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology. Volume II* (pp. 269-321). New York: Wiley.

Mithun, M. (2007). Linguistics in the face of language endangerment. In W. L. Wetzels (Ed.), *Language Endangerment and Endangered Languages* (pp. 15-35). Leiden University, The Netherlands: ILLA. Publications of the Research School of Asian, African, and Amerindian Studies (CNWS).

Miyamoto, E. T. (2002). Case Markers as Clause Boundary Inducers in Japanese. *Journal of Psycholinguistic Research, 31*, 307-347.

Momma, S., Slevc, L. R., y Phillips, C. (2016). The timing of verb selection in Japanese sentence production. *Journal of Experimental Psychology, 42*(5), 813-824. doi: 10.1037/xlm0000195

Montag, J. L., y MacDonald, M. C. (2009). Word order doesn't matter: Relative clause production in English and Japanese. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Meeting of the Cognitive Science Society* (pp. 2594-2599). Austin: Cognitive Science Society.

Montag, J. L., y MacDonald, M. C. (2014). Visual Salience Modulates Structure Choice in Relative Clause Production. *Language and Speech, 57*(2), 163-180. doi: 10.1177/0023830913495656

Montag, J. L., Matsuki, K., Kim, J. Y., y MacDonald, M. C. (2017). Language Specific and Language General Motivations of Production Choices: A Multi-Clause and Multi-Language Investigation. *Collabra: Psychology, 3*(1), 1-22. doi: 10.1525/collabra.94

Newmeyer, F. J. (2003). Grammar is grammar and usage is usage. *Language, 79*(4), 682-707. doi: 10.1353/lan.2003.0260

Nichols, J., Peterson, D. A., y Barnes, J. (2004). Transitivizing and detransitivizing languages. *Linguistic Typology, 8*, 149-211.

Nieuwland, M. S., Petersson, K. M., y Van Berkum, J. J. A. (2007). On sense and reference: Examining the functional neuroanatomy of referential processing. *NeuroImage, 37*(3), 993-1004. doi: 10.1016/j.neuroimage.2007.05.048

Norcliffe, E., Konopka, A. E., Brown, P., y Levinson, S. C. (2015). Word order affects the time course of sentence formulation in Tzeltal. *Language, Cognition and Neuroscience, 30*(9), 1187-1208. doi: 10.1080/23273798.2015.1006238

Oflazer, K., Say, B., Hakkani-Tür, D. Z., y Tür, G. (2003). Building a Turkish Treebank. In A. Abeille (Ed.), *Building and Exploiting Syntactically-annotated Corpora* (pp. 1-17). Dordrecht, Netherlands: Kluwer Academic Publishers.

Özge, D., Marinis, T., y Zeyrek, D. (2013). Object-first orders in Turkish do not pose a challenge during processing. In U. Özge (Ed.), *Proceedings of the 8th Workshop on Altaic Formal Linguistics* (pp. 269-280): MIT Working Papers in Linguistics.

Parisse, C., y Le Normand, M. T. (2000). How children build their morphosyntax: the case of French. *Journal of Child Language, 27*, 267–292.

Perera, C. K., y Srivastava, A. K. (2016). Animacy-Based Accessibility and Competition in Relative Clause Production in Hindi and Malayalam. *Journal of Psycholinguistic Research, 45*(4), 915-930. doi: 10.1007/s10936-015-9384-0

Pickering, M. J. (1993). Direct association and sentence processing: A reply to Gorrell and to Gibson and Hickok. *Language and Cognitive Processes, 8*, 163-196.

Pickering, M. J., y Barry, G. (1991). Sentence processing without empty categories. *Language and Cognitive Processes, 6*, 229-259.

Pinker, S. (1994). An Instinct to Acquire an Art *The Language Instinct: How the Mind Creates Language* (2007, P.S. ed., pp. 15-24). New York: Harper Collins Publishers.

Pinker, S. (1995a). El instinto para adquirir un arte (J. M. I. González, Trans.) *El instinton del lenguaje* (2012, 2ª ed., pp. 15-24). Madrid: Alianza Editorial.

Pinker, S. (1995b). 기술 습득을 위한 본능 *언어본능*(pp. 15-30). Korea: Greenbee Publishing Co.

Pinker, S. (2010). Arte bat gureganatzeko sena (G. Knörr, Trans.) *Hizkuntza-sena* (pp. 7-18). Zarautz: ehupress.

Polinsky, M. (2012). Headedness, again. In T. Graf, D. Paperno, A. Szabolsci & J. Tellings (Eds.), *Theories of Everything: In Honor of Ed Keenan* (Vol. 17, pp. 348-359). Los Angeles: UCLA.

Poulin–Dubois, D., Graham, S., y Sippola, L. (1995). Early lexical development: the contribution of parental labeling and infants' categorization abilities. *Journal of Child Language, 22*, 325-343.

Prat-Sala, M. (1997). *The production of different word orders: a psycholinguistic and developmental approach.* (PhD Dissertation), University of Edinburgh.

Prat-Sala, M., y Branigan, H. P. (2000). Discourse Constraints on Syntactic Processing in Language Production: A Cross-Linguistic Study in English and Spanish. *Journal of Memory and Language, 42*, 168-182. doi: 10.1006/jmla.1999.2668

Prat-Sala, M., Shillcock, R., y Sorace, A. (2000). *Animacy effect on the production of object-dislocated descriptions by Catalan-speaking children* (Vol. 27).

Prince, E. F. (1999). Subject pro-drop in Yiddish. In P. Bosch & R. van der Sandt (Eds.), *Focus: Linguistic, cognitive, and computational perspectives* (pp. 82-104). Cambridge: Cambridge University Press.

Pullum, G. K. (2007). Ungrammaticality, rarity, and corpus use. *Corpus Linguis- tics and Linguistic Theory, 3*, 33-47.

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

RAE. Banco de datos (CORPES XXI). Corpus del Español del Siglo XXI (CORPES). Disponible en http://web.frl.es/CORPES/

Rocha, P. A., y Santos, D. (2000, 19-22 de noviembre de 2000). *CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa*Trabajo presentado en the Conference Name|, Conference Location|.

Röder, B., Stock, O., Neville, H., Bien, S., y Rösler, F. (2002). Brain Activation Modulated by the Comprehension of Normal and Pseudo-word Sentences of Different Processing Demands: A Functional Magnetic Resonance Imaging Study. *NeuroImage, 15*(4), 1003-1014. doi: 10.1006/nimg.2001.1026

Rojo, G., López, M., Domínguez, E., y Barcala, F. M. (2010). CORGAetq: Corpus de Referencia do Galego Actual etiquetado (versión 2.4). Centro Ramón Piñeiro para a Investigación en Humanidades (Xunta de Galicia). Disponible en http://corpus.cirp.es/corgaetq

Roland, D., Dick, F., y Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language, 57*, 348-379. doi: 10.1016/j.jml.2007.03.002

Ros, I. (2018). *Minimizing dependencies across languages and speakers.* (PhD Dissertation), Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU).

Ros, I., Santesteban, M., Fukumura, K., y Laka, I. (2015). Aiming at shorter dependencies: the role of agreement morphology. *Language, Cognition and Neuroescience, 30*(9), 1156-1174. doi: 10.1080/23273798.2014.994009

Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., y van der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human Movement Science, 26*(4), 525-554. doi: 10.1016/j.humov.2007.04.001

Rösler, F., Pechmann, T., Streb, J., Röder, B., y Hennighausen, E. (1998). Parsing of Sentences in a Language with Varying Word Order: Word-by-Word Variations of Processing Demands Are Revealed by Event-Related Brain Potentials. *Journal of Memory and Language, 38*(2), 150-176. doi: 10.1006/jmla.1997.2551

Sakurai, C. (1998). Why nouns before verbs? A case study of Japanese early lexical acquisition. *CLS, 34*, 511-519.

Sampson, G. (2007). Grammar without Grammaticality. *Corpus Linguistics and Linguistic Theory, 3*(1), 1-32.

Santos, D., y Rocha, P. A. (2001). Evaluating CETEMPúblico, a free resource for Portuguese *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 442-449). Toulouse.

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/

Sauppe, S. (2017). Word Order and Voice Influence the Timing of Verb Planning in German Sentence Production. *Frontiers in Psychology, 8*(1648). doi: 10.3389/fpsyg.2017.01648

Sauppe, S., Norcliffe, E., Konopka, A. E., Brown, P., Van Valin, R. D., y Levinson, S. C. (2013a). *Typology and planning scope in sentence production: eye tracking evidence from Tzeltal and Tagalog*. Poster presentado en The 10th Biennial Conference of the Association for Linguistic Typology, Leipzig

Sauppe, S., Norcliffe, E., Konopka, A. E., Van Valin, R. D., y Levinson, S. C. (2013b). Dependencies First: Eye Tracking Evidence from Sentence Production in Tagalog. *Proceedings of the Annual Meeting of the Cognitive Science Society, 35*, 1265-1270.

Saussure, F. (1916). *Cours de Linguistique Générale*. Paris: Payot.

Schachter, P. (1985). Parts-of-speech systems. In T. Shopen (Ed.), *Language Typology and Syntactic Description* (Vol. 1, pp. 3-61). Cambridge: Cambridge University Press.

Schachter, P., y Shopen, T. (2007). Parts-of-speech systems. In T. Shopen (Ed.), *Language Typology and Syntactic Description* (2 ed., Vol. 1, pp. 1-60). Cambridge: Cambridge University Press.

Schlesewsky, M., y Bornkessel, I. (2003). Ungrammaticality detection and garden path strength: A commentary on Meng and Bader"s (2000) evidence for serial parsing. *Language and Cognitive Processes, 18*(3), 299-311. doi: 10.1080/01690960244000027

Schlesewsky, M., Bornkessel, I., y Frisch, S. (2003). The neurophysiological basis of word order variations in German. *Brain and Language, 86*(1), 116-128. doi: 10.1016/S0093-934X(02)00540-0

Seifart, F. (2011). *Cross-linguistic variation in the noun-to-verb ratio: the role of verb morphology and narrative strategies*Trabajo presentado en the Conference Name|, Conference Location|.

Seifart, F., Strunk, J., Danielsen, S., Hartmann, I., Pakendorf, B., Wichmann, S., Witzlack-Makarevich, A., de Jong, N. H., y Bickel, B. (2018). Nouns slow down speech across structurally and culturally diverse languages. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1800708115

Shapiro, K. A., y Caramazza, A. (2003). Grammatical processing of nouns and verbs in left frontal cortex? *Neuropsychologia, 41*(9), 1189-1198.

Siewierska, A. (1993). Syntactic Weight vs Information Structure and Word Order Variation in Polish. *Journal of Linguistics, 29*(2), 233-265.

Silverstein, M. (1976). Hierarchy of Features and Ergativity. In R. M. W. Dixon (Ed.), *Grammatical Categories in Australian Languages* (pp. 112-171). Canberra: Australian National University.

Slevc, L. R. (2011). Saying what's on your mind: Working memory effects on sentence production. *Journal of Experimental Psychology, 37*(6), 1503-1514. doi: 10.1037/a0024350

Smith, M., y Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition, 73*(3), 205-246.

Smith, M., y Wheeldon, L. (2004). Horizontal information flow in spoken sentence production. *Journal of Experimental Psychology, 30*(3), 675-686. doi: 10.1037/0278-7393.30.3.675

Soto, G., Martínez, R., y Sadowsky, S. (2005). Verbos y sustantivos en textos científicos. Análisis de variación en un corpus de textos de ciencias aplicadas, naturales, sociales y humanidades. *Philologia Hispalensis, 19*, 169-187.

Stallings, L. M., y MacDonald, M. C. (2011). It's not Just the "Heavy NP": Relative Phrase Length Modulates the Production of Heavy-NP Shift. *Journal of Psycholinguistic Research, 40*, 177-187.

Stallings, L. M., MacDonald, M. C., y O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language, 39*, 392-417.

Szekely, A., D'Amico, S., Devescovi, A., Federmeier, K., Herron, D., Iyer, G., Jacobsen, T., y Bates, E. (2002). Timed action and object naming. *Cortex, 41*(7-25).

Tanaka, M. N., Branigan, H. P., y Pickering, M. J. (2005). *The role of animacy in Japanese sentence production*. Trabajo presentado en The 18th Annual CUNY Conference on Human Sentence Processing, Tucson, AZ

Tao, H. (1996). *Units in Mandarin conversation: Prosody, discourse, and grammar*. Philadelphia: John Benjamins.

Tardif, T., Shatz, M., y Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: a comparison of English, Italian, and Mandarin. *Journal of Child Language, 24*, 535-565.

Taulé, M., Martí, M. A., y Recasens, M. (2008). Ancora: Multi level annotated corpora for Catalan and Spanish *Proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakesh.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work* (Vol. 6). Amsterdam: John Benjamins.

Traxler, M. J., Morris, R. K., y Seely, R. E. (2002). Processing Subject and Object Relative Clauses: Evidence from Eye Movements. *Journal of Memory and Language, 47*, 69-90. doi: 10.1006/jmla.2001.2836

Trueswell, J. C., Tanenhaus, M. K., y Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(3), 528-553.

Tummers, J., Heylen, K., y Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics:A technical state of the art. *Corpus Linguistics and LinguisticTheory, 1*(2), 225–261.

Turan, Ü. D. (1998). Ranking forward-looking centers in Turkish: Universal and language specific properties. In M. A. Walker, A. K. Joshi & E. F. Prince (Eds.), *Centering theory in discourse* (pp. 139-161). Oxford: Clarendon Press.

Ueno, M., y Garnsey, S. M. (2008). An ERP study of the processing of subject and object relative clauses in Japanese. *Language and Cognitive Processes, 23*, 646-688.

Ueno, M., y Polinsky, M. (2009). Does headedness affect processing? A new look at the VO-OV contrast. *Journal of Linguistics, 45*, 675-710.

Uit den Boogaert, P. C. (1975). *Woordfrekwenties in Geschreven en Gesproken Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.

van Berkum, J., Zwitserlood, P., Bastiaansen, M., Brown, C. M., y Hagoort, P. (2004). So who's "he" anyway? Differential EEG effects of referential ambiguity and referential failure during spoken language comprehension. *Journal of Cognitive Neuroscience, Supplement*.

van de Velde, M., Meyer, A. S., y Konopka, A. E. (2014). Message formulation and structural assembly: Describing "easy" and "hard" events with preferred and dispreferred syntactic structures. *Journal of Memory and Language, 71*(1), 124-144. doi: 10.1016/j.jml.2013.11.001

van Nice, K. Y., y Dietrich, R. (2003). Task sensitivity of animacy effects: evidence from German picture descriptions. *Linguistics, 41*(5), 825-849.

Van Petten, C., Kutas, M., Kluender, R., Mitchiner, M., y McIsaac, H. (1991). Fractionating the Word Repetition Effect with Event-Related Potentials. *Journal of Cognitive Neuroscience, 3*(2), 131-150. doi: 10.1162/jocn.1991.3.2.131

Vasishth, S., Chen, Z., Li, Q., y Guo, G. (2013). Processing Chinese Relative Clauses: Evidence for the Subject-Relative Advantage. *PLOS ONE, 8*(10), e77006. doi: 10.1371/journal.pone.0077006

Verlinden, A., y Gillis, S. (1988). Nouns and verbs in the input: Gentner (1982) reconsidered. In F. Van Besien (Ed.), *First Language Acquisition* (pp. 163–187). Belgium: ABLA.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., y Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and Biobehavioral Reviews, 35*(3), 407-426. doi: 10.1016/j.neubiorev.2010.04.007

Vosse, T., y Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition, 75*(2), 105-143.

Wagner, V., Jescheniak, J. D., y Schriefers, H. (2010). On the flexibility of grammatical advance planning during sentence production: Effects of cognitive load on multiple lexical access. *Journal of Experimental Psychology, 36*(2), 423-440. doi: 10.1037/a0018619

Walker, M. A., Iida, M., y Cote, S. (1994). Japanese discourse and the process of centering. *Computacional Linguistics, 20*, 193-231.

Wang, L., Schlesewsky, M., Bickel, B., y Bornkessel-Schlesewsky, I. (2009). Exploring the nature of the 'subject'-preference: Evidence from the online comprehension of simple sentences in Mandarin Chinese. *Language and Cognitive Processes, 24*(7-8), 1180-1226. doi: 10.1080/01690960802159937

Wasow, T. (1997a). End-Weight from the Speaker's Perspective. *Journal of Psycholinguistic Research, 26*, 347-361.

Wasow, T. (1997b). Remarks on grammatical weight. *Language Variation and Change, 9*, 81-105.

Wasow, T. (2002). *Postverbal behavior*. Stanford: CSLI Publications.

Wasow, T., y Arnold, J. E. (2003). Post-verbal constituent ordering in English. In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of Grammatical Variation in English* (pp. 119-154). Berlin: de Gruyter.

Whaley, L. J. (1997). *Introduction to typology: The unity and diversity of language*. Thousand Oaks, CA: SAGE.

Wheeldon, L., Ohlson, N., Ashby, A., y Gator, S. (2013). Lexical availability and grammatical encoding scope during spoken sentence production. *Quarterly Journal of Experimental Psychology, 66*(8), 1653-1673. doi: 10.1080/17470218.2012.754913

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wolff, S. (2010). *The interplay of free word order and pro-drop in incremental sentence processing: Neurophysiological evidence from Japanese.* (PhD Dissertation), Max Planck Institute for Human Cognitive and Brain Sciences.

Wolff, S., Schlesewsky, M., Hirotani, M., y Bornkessel-Schlesewsky, I. (2008a). The neural mechanisms of word order processing revisited: Electrophysiological evidence from Japanese. *Brain & Language, 107*, 133-157. doi: 10.1016/j.bandl.2008.06.003

Wolff, S., Schlesewsky, M., Horie, K., y Bornkessel-Schlesewsky, I. (2008b). Understanding "missing" arguments: An electrophysiological investigation of subject drop in Japanese. *Journal of Cognitive Neuroscience, Supplement*(113).

Wu, F., Kaiser, E., y Andersen, E. (2009). Animacy effects in Chinese relative clause processing. In M. Grosvald & D. Soares (Eds.), *Proceedings of the Western Conference on Linguistics (WECOL)* (pp. 318-329). Davis: University of California.

Wu, F., Kaiser, E., y Andersen, E. (2012). Animacy effects in Chinese relative clause processing. *Language and Cognitive Processes, 27*(10), 1489-1524. doi: 10.1080/01690965.2011.614423

Xue, N., Jiang, Z., Zhong, X., Palmer, M., Xia, F., Choiu, F.-D., y Chang, M. (2010). *Chinese Treebank 7.0*.

Yamashita, H. (2002). Scrambled sentences in Japanese: Linguistic properties and motivations for production. *Text, 22*(4), 597-633.

Yamashita, H., y Chang, F. (2001). 'Long before short' preference in the production of a head-final language. *Cognition, 81*, B45-B55.

Yamashita, H., Chang, F., y Hirose, Y. (2005). *Producers build structures only with overt arguments*. Poster presentado en The 8th Annual CUNY Conference on Human Sentence Processing, Tucson

Yamashita, Y. (1999). The acquisition of nouns and verbs in young Japanese children: why do verbal nouns emerge early? *Proceedings of the 23rd annual Boston University Conference on Language Development* (pp. 741-752): Cascadilla Press.

Yeon, J., y Brown, L. (2011). *Korean: A Comprehensive Grammar*. Abingdon, Oxon: Routledge.

Yngve, V. H. (1960). A model and an hypothesis for language structure *Proceedings of the American Philosophical Society* (Vol. 140, pp. 444-466).

Zagona, K. (2002). *The Syntax of Spanish*. Cambridge: Cambridge University Press.

Zampolli, A. (1977). Statistique linguistique et dépouillements automatiques. In P. J. G. Van Sterkenburgh (Ed.), *Lexicologie* (pp. 325-358). Groningen: Wolters-Noordhoff.

Zhao, L.-M., Alario, F. X., y Yang, Y.-F. (2015). Grammatical planning scope in sentence production: Further evidence for the functional phrase hypothesis. *Applied Psycholinguistics*, *36*(5), 1059-1075. doi: 10.1017/S0142716414000046

Zhao, L.-M., y Yang, Y.-F. (2016). Lexical Planning in Sentence Production Is Highly Incremental: Evidence from ERPs. *PLOS ONE, 11*(1), e0146359. doi: 10.1371/journal.pone.0146359

Apéndices

A.1. Oraciones usadas en el corpus del CAPÍTULO 2

Breve muestra de las oraciones transitivas de euskera etiquetadas para el estudio de corpus del CAPÍTULO 2. Para poder obtener todas las oraciones hay que acceder al siguiente link (www.ehu.eus/HEB/Corpus_Tesis/Cap2.xlsx).

	Oración	**orden**	**género**
1.	Alemaniak eskuhartze nuklearraren akordioa urratu du.	SOV	periódico
2.	Algoritmoak, gainera, badu beste abantaila bat:	SVO	revista
3.	ANE.— Umeek jolas-ordua dute...	SOV	guiones
4.	artikulu "irakurrerraza" egin du egileak.	OVS	revista
5.	Aukera honek abantaila batzuk eskaintzen ditu:	SOV	revista
6.	Begiak distiratsu zekartzan neskatoak,	OVS	libros
7.	Beraz, 500.000 pertsona inguruk dute eritasuna.	SVO	periódico
8.	Berehala irentsi zuen amua katxaloteak:	VOS	libros
9.	Honek CSI Planaren onarpenaren atzerapena ekarri zuen.	SOV	revista
10.	Bi dorrek gotortzen dute Orthezeko zubia,	SVO	libros
11.	Burges miliziakoek karrikaburuak zaintzen dituzte.	SOV	libros
12.	eta hezur bakoitzak istorio bat kontatzen du.	SOV	revista
13.	eta softwareak berehala ematen du emaitza",	SVO	revista
14.	Euskaltzainburuak eskutitz bana egin du zentzu horretan.	SOV	libros
15.	GAIZKAk irekitzen du atea.	SVO	guiones
16.	Gehienek xanpainbotila itxiak dituzte eskuan,	SOV	libros
17.	GREGORIO.— Zerbaiten akabera ematen du honek.	OVS	guiones
18.	Hegoaldean, 200 metroko sakonera du arroilak,	OVS	revista
19.	hitzekin bukatuko du aitak gosaria,	VSO	libros
20.	Honek CSI Planaren onarpenaren atzerapena ekarri zuen.	SOV	revista
21.	Honek istilu handiak sortuko ditu.	SOV	revista
22.	Horiek lehen postua ziurtatua dute,	SOV	periódico
23.	HORTENTSI.— Guk bestelako kontuak ditugu.	SOV	guiones
24.	Idazle on askok erabili dute izen bikoitza.	SVO	libros
25.	IÑIGOk, isil-isilik, argia pizten du.	SOV	guiones
26.	italiar kulturaren astea antolatu zuen udalak.	OVS	libros
27.	KARMENek sorbaldak goratzen ditu.	SOV	guiones
28.	Kasu horietarako epaitegi bat du elkarte horrek.	OVS	periódico
29.	katu batek irentsi behar ditu kisteak.	SVO	revista
30.	LABek toki berean egingo du manifestazioa,	SVO	periódico
31.	Lana sei ikerketa-taldek egin dute,	OSV	revista
32.	lau txanda aurreikusi ditu batzordeak.	OVS	periódico
33.	Lehen azterketa filologikoa Lakarrak egin du,	OSV	periódico

34. Logia nagusi bakoitzak badu konstituzio bat.	SVO	periódico
35. MAIDERrek luntxera gonbidatu du IKER.	SVO	guiones
36. MERTXE.— Karlosek goizez ditu klaseak,	SVO	guiones
37. MERTXE.— nik zainduko dut Itxaro,	SVO	guiones
38. MERTXEk KARLOS ikusten du,	SOV	guiones
39. mugak beti dakar kontrabandoa,	SVO	libros
40. Neuk ordaintzen dut gosaria.	SVO	libros
41. Nik albaniarrak defenditu ditut,	SOV	periódico
42. Nik literatura ona sortu nahi dut,	SOV	libros
43. Nobelak bere dinamika propioa du:	SOV	periódico
44. ohiko telebista dohanekoak bateratu zuen alderdia,	SVO	revista
45. PAGALDAY.— Bazkari legea egingo dugu denok elkarrekin…	OVS	guiones
46. Perezek oraingo lan hau hortxe argitaratu du,	SOV	libros
47. Teknologia berriek bizimoduaren erritmoa aldatu dute,	SOV	periódico
48. Txinako Gobernuak adopzioak gestionatzeko sail bat du.	SOV	periódico
49. Txinatarrek ezagutzen dute grabitatearen legea,	SVO	periódico
50. XABIER.— Horrek erremedio erraza du,	SOV	guiones

A.2. Oraciones usadas en el corpus paralelo del CAPÍTULO 3

Breve muestra de las oraciones etiquetadas del primer capítulo del libro de Steven Pinker *The Language Instinct* en sus cuatro ediciones (inglés, castellano, euskera y coreano) para el estudio de corpus paralelo del CAPÍTULO 3. Para poder obtener todas las oraciones hay que acceder al siguiente link (www.ehu.eus/HEB/Corpus_Tesis/Cap3.xlsx).

	Oraciones en inglés
1.	In any natural history[N] of the human species[N], language[N] would(VMD) stand out[V] as the preeminent trait[N].
2.	To be[V] sure, a solitary human[N] is[V] an impressive problem-solver[N] and engineer[N].
3.	But a race[N] of Robinson Crusoes would(VMD) not give[V] an extraterrestrial observer[N] all(PI) that much(PQ) to remark[V] on.
4.	What(PR) is(AUX) truly arresting(G) about our kind[N] is[V] better captured(PART) in the story[N] of the Tower[N] of Babel, in which(PR) humanity[N], speaking(G) a single language[N], came[V] so close to reaching(G) heaven[N] that(PR) God himself(PP) felt[V] threatened(PART).
5.	A common language[N] connects[V] the members[N] of a community[N] into an information-sharing[N] network[N] with formidable collective powers[N].
6.	Anyone(PI) can(VMD) benefit[V] from the strokes[N] of genius[N], lucky accidents[N], and trial-and-error[N] wisdom[N] accumulated[V] by anyone(PI) else, present or past.
7.	And people[N] can(VMD) work[V] in teams[N], their efforts[N] coordinated[V] by negotiated(PART) agreements[N].
8.	As a result[N], Homo sapiens is[V] a species[N], like blue-green algae[N] and earthworms[N], that(PR) has(AUX) wrought[V] far-reaching(G) changes[N] on the planet[N].
9.	Archeologists[N] have(AUX) discovered[V] the bones[N] of ten thousand wild horses[N] at the bottom[N] of a cliff[N] in France, the remains[N] of herds[N] stampeded(PART) over the clifftop[N] by groups[N] of paleolithic hunters[N] seventeen thousand years[N] ago.
10.	These fossils[N] of ancient cooperation[N] and shared(PART) ingenuity[N] may(VMD) shed[V] light[N] on why(PQ) saber-tooth[N] tigers[N], mastodons[N], giant woolly rhinoceroses[N], and dozens (PN) of other large mammals[N] went[V] extinct around the time[N] that(PR) modern humans[N] arrived[V] in their habitats[N].
11.	Our ancestors[N], apparently, killed[V] them(PP) off.
12.	Language[N] is[V] so tightly woven(PART) into human experience[N] that it(PP) is[V] scarcely possible to imagine[V] life[N] without it(PP).
13.	Chances[N] are[V] that if you(PP) find two or more people[N] together anywhere on Earth, they(PP) will(VMD) soon be(AUX) exchanging(G) words[N].
14.	When there is[V] no one(PI) to talk[V] with, people[N] talk[V] to themselves(PP), to their dogs[N], even to their plants[N].

15. In our social relations[N], the race[N] is[V] not to the swift[N] but to the verbal[N]—the spellbinding orator[N], the silver-tongued seducer[N], the persuasive child[N] who(PR) wins[V] the battle[N] of wills[N] against a brawnier parent[N].
16. Aphasia[N], the loss[N] of language[N] following brain[N] injury[N], is(AUX) devastating(G), and in severe cases[N] family[N] members[N] may(VMD) feel[V] that the whole person[N] is[V] lost(PART) forever.
17. This book[N] is[V] about human language[N].
18. Unlike most books[N] with "language[N]" in the title[N], it(PP) will(VMD) not chide[V] you(PP) about proper usage[N], trace[V] the origins[N] of idioms[N] and slang[N], or divert[V] you(PP) with palindromes[N], anagrams[N], eponyms[N], or those precious names[N] for groups[N] of animals[N] like "exaltation[N] of larks[N].
19. For I(PP) will(VMD) be(AUX) writing(G) not about the English language[N] or any other language[N], but about something[N] much more basic: the instinct[N] to learn[V], speak[V], and understand[V] language[N].
20. For the first time[N] in history[N], there is[V] something[N] to write[V] about it(PP).
21. Some thirty-five years[N] ago a new science[N] was(AUX) born[V].
22. Now called[V] "cognitive science[N]," it(PP) combines[V] tools[N] from psychology[N], computer[N] science[N], linguistics[N], philosophy[N], and neurobiology[N] to explain[V] the workings[N] of human intelligence[N].
23. The science[N] of language[N], in particular[N], has(AUX) seen[V] spectacular advances[N] in the years[N] since.
24. There are[V] many phenomena[N] of language[N] that(PR) we(PP) are(AUX) coming(G) to understand[V] nearly as well as we(PP) understand[V] how a camera[N] works[V] or what the spleen[N] is[V] for.
25. I(PP) hope[V] to communicate[V] these exciting discoveries[N], some(PI) of them(PP) as elegant as anything(PI) in modern science[N], but I(PP) have[V] another agenda[N] as well.

Oraciones en castellano

1. En cualquier historia[N] natural de la especie[N] humana, el lenguaje[N] se(PP) destaca[V] como rasgo[N] prominente.
2. Con toda seguridad[N], un ser[N] humano aislado(PART) es[V] una impresionante obra[N] de ingeniería[N] y una máquina[N] de resolver[V] problemas[N].
3. No obstante, una raza[N] de Robinson Crusoes no llamaría[V] demasiado la atención[N] a un observador[N] extraterrestre.
4. Lo(PP) verdaderamente notable de la condición[N] humana se(PP) refleja[V] mejor en la historia[N] de la Torre[N] de Babel, en la que(PR) la humanidad[N], con el don[N] de una única lengua[N], se(PP) aproximó[V] tanto a los poderes[N] divinos que Dios se(PP) sintió[V] amenazado(PART).

5. Una lengua[N] común conecta[V] a los miembros[N] de una comunidad[N] con una red[N] de información[N] compartida(PART) con unos formidables poderes[N] colectivos.
6. Cualquiera(PI) se(PP) puede(VMD) beneficiar[V] de los toques[N] de genialidad[N], los golpes[N] de fortuna[N] o el saber[V] espontáneo de cualquier otra persona[N], viva(PART) o muerta(PART); Además, las personas[N] pueden(VMD) trabajar[V] en equipo[N], coordinando(G) sus esfuerzos[N] mediante acuerdos[N] negociados(PART).
7. Como consecuencia[N] de ello(PP), el Homo sapiens es[V] una especie[N] que(PR), sin ser[V] muy distinta de las algas[N] marinas o las lombrices[N] de tierra[N], ha(AUX) originado[V] cambios[N] perdurables en este planeta[N].
8. Los arqueólogos[N] han(AUX) descubierto[V] los esqueletos[N] de diez mil caballos[N] salvajes al pie[N] de un acantilado[N] en Francia, restos[N] de una manada[N] que(PR) se(PP) despeñó[V] empujada[V] por varios grupos[N] de cazadores[N] del paleolítico[N] hace[V] unos diecisiete mil años[N].
9. Estos fósiles[N] de la colaboración[N] y del ingenio[N] compartido(PART) nos(PP) puede(VMD) aclarar[V] el motivo[N] por el que(PR) los tigres[N] de colmillos[N] de sable[N], los mastodontes[N], los rinocerontes[N] lanudos gigantes y otras muchas especies[N] de mamíferos[N] se(PP) extinguieron[V] en los tiempos[N] en que los modernos humanos[N] arribaron[V] a sus hábitats[N].
10. Todo(PI) parece[V] indicar[V] que nuestros antecesores[N] los(PP) aniquilaron[V].
11. El lenguaje[N] se(PP) halla[V] tan íntimamente entrelazado(PART) con la experiencia[N] humana que apenas es[V] posible imaginar[V] la vida[N] sin él(PP).
12. Si uno se(PP) encuentra[V] a dos o más personas[N] juntas en cualquier rincón[N] de la tierra[N], lo(PP) más probable es[V] que estén(AUX) conversando(G).
13. Cuando uno(PN) no tiene[V] con quién(PQ) hablar[V], se(PP) pone[V] a hablar[V] consigo(PP) mismo, con su perro[N] o incluso con sus plantas[N].
14. En nuestras relaciones[N] sociales no se(PP) admira[V] la rapidez[N], sino la labia[N]: el orador[N] que(PR) nos(PP) hechiza[V] con sus palabras[N], el seductor[N] que(PR) nos(PP) conquista[V] con su verbo[N], o el niño[N] persuasivo que convence[V] a su testarudo padre[N].
15. La afasia[N], la pérdida[N] del lenguaje[N] a consecuencia[N] de un daño[N] cerebral, es[V] un mal[N] devastador, y en casos[N] muy severos de esta enfermedad[N], la familia[N] del afectado[N] llega[V] a sentir[V] que lo(PP) han(AUX) perdido[V] para siempre.
16. Este libro[N] trata[V] del lenguaje[N] humano.
17. A diferencia[N] de la mayoría[N] de los libros[N] que(PR) llevan[V] la palabra[N] «lenguaje»[N] en su título[N], no tiene[V] la intención[N] de reñir[V] a nadie(PI) por usarlo[V](PP) incorrectamente, de informar[V] sobre el origen[N] de los giros[N] idiomáticos o las expresiones[N] coloquiales, o de entretener[V] a base[N] de palíndromos[N], anagramas[N], epónimos[N] o con curiosos nombres[N] de animales[N] como «piara[N] de cerdos[N]».
18. Mi propósito[N] no es[V] hablar[V] del inglés[N], el español[N] u otra lengua[N] en particular, sino de algo(PI) mucho más elemental: el instinto[N] de aprender[V], hablar[V] y entender[V] el lenguaje[N].

19. Por primera vez[N] en la historia[N], ya hay[V] algo(PI) de lo(PP) que hablar[V] al respecto[N].
20. Hace unos treinta y cinco años[N] nacía[V] una nueva ciencia[N].
21. Lo(PP) que ahora se(PP) conoce[V] como «ciencia[N] cognitiva» combina[V] procedimientos[N] tomados(PART) de la psicología[N], las ciencias[N] de la computación[N], la lingüística[N], la filosofía[N] y la neurobiología[N] para explicar[V] el funcionamiento[V] de la inteligencia[N] humana.
22. La ciencia[N] del lenguaje[N], en particular, ha(AUX) sido[V] testigo[N] de espectaculares avances[N] desde entonces.
23. Hay[V] muchos fenómenos[N] del lenguaje[N] que(PR) estamos(AUX) empezando(G) a entender[V] casi tan bien como el funcionamiento[N] de una cámara[N] o para qué(PQ) sirve[V] el bazo[N].
24. Mi intención[N] es[V] transmitir[V] estos apasionantes descubrimientos[N], algunos(PI) de los cuales(PR) fisuran[V] entre los más sugerentes de la ciencia[N] moderna.
25. Sin embargo, también tengo[V] otros planes[N].

Oraciones en euskera

1. Giza espeziearen[N] edozein historia[N] naturaletan, hizkuntza[N] ageri[V] da(AUX) ezaugarri[N] nagusi.
2. Zalantzarik[N] gabe, gizabanakoa[N], bakarka harturik(PART), problema-ebazle[N] eta ingeniari[N] ikaragarria da[V].
3. Robinson Crusoez osaturiko(PART) arraza[N] bati, baina, ez lioke(AUX) aparteko zer (PQ) aipagarririk ikusiko[V] behatzaile[N] estralurtar batek.
4. Gizakiongan[N] benetan txundigarria dena[V](Rel) hobeki jasotzen[V] da(AUX) Babelgo Dorrearen[N] istorioan[N], non gizateria[N], hizkuntza[N] bakarra eginaz[V], hainbeste gerturatu[V] baitzen(AUX) zerura[N], ezen Jainkoa bera(PP) mehatxaturik(PART) sentitu[V] baitzen(AUX).
5. Hizkuntza[N] batek, hain zuzen, informazioa[N] partekatzeko[V] sare[N] erkide izugarri ahaltsu batean konektatzen[V] ditu(AUX) taldeko[N] kideak[N].
6. Edonork(PI) erabil[V] dezake(AUX) beste batek(PI) –garaikide zein iraganeko– sormen-txinpartez[N], zorioneko ustekabeez[N] edo ahaleginez[N] lortutako(PART) ezagutza[N].
7. Eta jendeak[N] taldean[N] jardun[V] dezake(AUX), lanak[N] hitzarmenen[N] bitartez uztartuz[V].
8. Horren ondorioz[N], Homo sapiens espezieak[N] –espezie[N] bat baita[V], alga[N] urdinak edo zizareak[N] bezala– aldaketa[N] sakon eta iraunkorrak eragin[V] ditu(AUX) planetan[N].
9. Arkeologoek[N] hamar mila zaldiren[N] hezurrak[N] aurkitu[V] dituzte(AUX) amildegi[N] baten hondoan[N], Frantzian: duela[V] hamazazpi mila urte[N] paleolitikoko[N] ehiztari[N] taldeek[N] bultzaturik(PART), hara amildutako[V] zaldi[N] saldoen[N] hondarrak[N] omen dira[V].
10. Antzinako elkarlanaren[N] eta gaitasun[N] partekatuen[V] fosil[N] horiek argi[V] dezakete(AUX) nola desagertu[V] ziren(AUX) sable-hortz[N] tigreak[N], mastodonteak[N], errinozero[N] iletsu erraldoia eta beste hainbat ugaztun[N] handi, gizaki[N] modernoak haien bizilekuetara[N] iritzi[V] berritan.

11. Itxura[N] guztien arabera, gure arbasoek[N] suntsitu[V] zituzten(AUX).
12. Hizkuntza[N] hain dago[V] giza esperientziaren[N] enborrean[N] txertatua(PART), non zaila baita[V] hura(PD) gabeko bizitzarik[N] irudikatzea[V].
13. Hartu[V] munduko[N] edozein tokitan[N] elkarrekin dauden[V] bi lagun[N] edo gehiago: baietz berehala beren(PP) artean hizketan[V] hasi[V].
14. Norekin(PQ) hitz(NV) egin[V] ez dugunean(AUX), geure buruari[N] hitz(NV) egiten[V] diogu(AUX), edo txakurrari[N], baita landareei[N] ere.
15. Gure gizarte-harremanetan[N], ez du(AUX) azkarrenak[N] irabazten[V] lehia[N], etorririk[N] handiena duenak[N] baizik: mintzoaz[N] xarmatzen[V] gaituen(AUX)(Rel) hizlariak[N], urrezko[N] mihia[N] duen[V](Rel) limurtzaileak[N], guraso[N] egoskorra bairatzea[V] lortzen[V] duen(AUX)(Rel) haurrak[N]… Afasia[N], burmuineko[N] asalduen[N] ondoriozko hizkuntza-galera[N], gaitz[N] suntsigarria da[V] eta, kasurik[N] larrienetan, pertsona[N] osoa betiko galdu[V] dela(AUX) uste izaten[V] dute(AUX) senideek[N].
16. Giza hizkuntzari[N] buruzkoa da[V] liburu[N] hau.
17. Izenburuan[N] «hizkuntza»[N] duten[V](Rel) liburu[N] gehienek ez bezala, honek(PD) ez dizu(AUX) errieta(NV) egingo[V] erabilera[N] zuzena dela[V] eta ez dela[V], ez da(AUX) esapide[N] eta argot-hitzen[N] jatorriaren[N] bila arituko[V], edo ez zaitu(AUX) dibertiaraziko[V] palindromo[N], anagrama[N] eta eponimoekin[N] edo «gintzaizkiokezun»[V] bezalako aditz-formekin[N].
18. Izan ere, ez dut(AUX) ingeles hizkuntza[N] edo beste edozein hizkuntza[N] hartuko[V] hizpide[N], baizik eta askoz oinarrizkoagoa den[V](Rel) zerbait(PI): hizkuntza[N] ikasi[V], erabili[V] eta ulertzeko[V] sena[N].
19. Aurrenekoz historian[N], bada(AUX) horren(PD) gainean zer(PQ) idatzi[V].
20. Duela[V] hogeita hamar bat urte[N] zientzia[N] berri bat sortu[V] zen(AUX), orain «zientzia[N] kognitiboa» deritzona[V](Rel).
21. Zientzia[N] horrek psikologiako[N], informatikako[N], hizkuntzalaritzako[N], filosofiako[N] eta neurobiologiako[N] ekaiak[N] uztartzen[V] ditu(AUX) giza adimenaren[N] funtzionamendua[N] azaltzeko[V].
22. Hizkuntzaren[N] zientziak[N], bereziki, aurrerapen[N] ikusgarriak egin[V] ditu(AUX) ordutik.
23. Hizkuntzaren[N] fenomeno[N] asko ongi ulertzen[V] hasi[V] gara(AUX), kamera[N] bat nola dabilen[V] edo barea[N] zertarako(PQ) dugun[V] ulertzen[V] dugun(AUX) bezain ongi ia.
24. Aurkikuntza[N] zirraragarri horiek ezagutarazi[V] nahi(VMD) nituzke(AUX) hemen, horietako(PD) batzuk(PI) zientzia[N] modernoko beste edozer(PI) bezain ederrak eta argiak baitira[V], baina badut[V] beste asmorik[N] ere.

Oraciones en coreano

1. 어떤 박물학[N]이든 인간[N]을 대상[N]으로 한다[V]면, 언어[N]는 두드러진 특성[N]으로 드러날[V]것(ND)이다.
2. 홀로 고립된[V] 인간[N]이라도 문제[N]를 해결하[V]고 도구[N]를 만들[V] 수(ND) 있[V]는 능력[N]을 지니[V]고 있다[V].

3. 그러나 로빈슨 크루소류[N]의 사람들[N]에게는 외계[N]의 관찰자[N]들에게 언급할[V] 만한 주제[N]가 그리 많지 않(VN)을 것(ND)이다.
4. 바벨탑의 이야기[N]는 우리(Pp) 종[N]의 경이로운 능력[N]을 단적으로 보여주[V]고 있다[V].
5. 여기에서 인간[N]은 단 하나(PS)의 언어[N]를 사용함[V]으로써 신[N]이 위협[N]을 느낄[V] 정도[N]로 하늘[N]에 근접한다[V].
6. 이렇듯 공통[N]의 언어[N]는 대단한 결집력[N]을 가지[V]고 한 사회[N]의 구성원[N]들을 정보[N] 공유[N]의 그물[N] 속[N]에 결합시킨다[V].
7. 누구(PI)든 현재[N]나 과거[N]의 타인[N]들이 축적해[V] 놓[V]은 모든 천재적인 업적[N]과 우연한 행운[N] 또는 시행착오[N]를 거친[V] 지혜[N]의 혜택[N]을 누릴[V] 수(ND) 있다[V].
8. 또한 사람[N]들은 팀[N]을 짜서[V] 일[N]을 할[V] 수(ND) 있[V]고, 그 노력[N]은 협의된[V] 결과[N]를 통해[V] 조정될[V] 수(ND) 있다[V].
9. 그 결과[N], 호모 사피엔스는 남조식물[N]이나 지렁이[N]처럼 하나(PN)의 종이[N]면서도 지구상[N]에 광범위한 변화[N]를 가져왔다[V].
10. 고고학자[N]들은 프랑스의 한 절벽[N] 밑[N]에서 1만 마리(ND)나 되[V]는 야생마[N]의 뼈[N]를 발견했다[V].
11. 이 잔해[N]는 1만 7000년(ND) 전[N]인 구석기시대[N]에 사냥꾼[N]들에게 쫓겨[V] 절벽[N] 아래[N]로 떨어진[V] 야생마[N]들의 것(ND)이다.
12. 고대[N]의 협동[N]과 단합[V]된 재주[N]를 보여주[V]는 이 화석[N]들을 통해[V] 검치호랑이[N]와 마스토돈[N] (신생대[N] 제3기(ND)의 대형 코끼리[N]), 거대하고 털[N]이 많은 무소[N] 그리고 다른 수십 종[N]의 커다란 포유동물[N]들이 어째서 오늘날[N]의 인간[N]들이 그들의(PS) 거주지[N]에 도착했[V]을 무렵[N]에 멸종되[V]어 갔[V]는지 밝힐[V] 수(ND) 있다[V].
13. 분명히 우리의(PS) 조상[N]이 그들(Pp)을 죽여[V] 없앴[V]을 것(ND)이다.
14. 언어[N]는 인간[N]의 경험[N] 속[N]에 아주 단단히 짜여져[V] 있[V]어서 언어[N] 없는 생활[N]이란 상상하기[V]조차 어렵다.
15. 아마도 지구[N] 상[N]의 어느 곳[N]에서든 두 명(ND) 이상[N]의 인간[N]이 모이면[V] 그들(Pp)은 곧 말[N]을 주고받[V]을 것(ND)이다.
16. 사람[N]들은 대화[V]할 상대[N]가 없으면 자기(Pd) 자신(Pd)에게, 자신(Pd)이 기르[V]는 개[N]에게, 심지어는 자신(Pd)이 기르[V]는 식물[N]에게까지 말[N]을 건다[V].
17. 우리의(PS) 사회관계[N] 속[N]에서 승리[N]는 재빠른 자[N]의 것(ND)이 아니(VN)라 언어적인 자[N]의 것(ND)이다.
18. 즉, 승리[N]는 매혹적인 변사[N]와 달변[N]의 바람둥이[N] 그리고 완력[N]으로는 상대[N]가 안되는(VN) 부모[N]와의 싸움[N]에서 어떻게든 자신의(PS) 의지[N]를 관철하고[V] 마[V]는 설득력[N] 있[V]는 아이[N]에게 돌아간다[V].
19. 때문에 뇌[N] 손상[N]의 결과[N]로 발생하[V]는 실어증[N]은 어떤 병[N]보다도 참혹하게 느껴지[V]는데, 심한 경우[N] 가족[N]들은 그 사람[N]이 완전히 그리고 영원히 사라졌다[V]고 여기기[V]도 한다[V].

20. 이 책[N]은 인간[N]의 언어[N]에 관한[V] 것(ND)이다.
21. 제목[N]에 '언어'[N]라는 단어[N]가 들어간[V] 대부분[N]의 책[N]들과는 달리 여기[N]에서는 용법[N]을 문제삼[V]거나 숙어[N]와 은어[N]의 기원[N]을 추적하[V]지 않(VN)을 것(ND)이고, 회문[N] (palindromes, eye나 madam처럼 거꾸로 읽[V]어도 같은 말[N]이 되[V]는 말[N])이나 글자 수수께끼[N] (anagrams, 예[N]를 들[V]어 emit－time, item－mite 등), 이름[N]의 시조[N] (eponyms, 예[N]를 들[V]어 Rome의 시조[N]는 Romulus) 혹은 '종다리 떼 (exaltation of larks)'[N]와 같은 고상한 이름[N]을 들[V]면서 기분전환[N]을 제공하[V]지도 않(VN)을 것(ND)이다.
22. 내(Pp)가 쓰[V]고자 하[V]는 것(ND)은 영어[N]나 그(Pd) 밖[N]의 구체적인 언어[N]에 관[V]해서가 아니다(VN).
23. 그(Pd)보다 훨씬 더 기본적인 것(ND)이다.
24. 그것(Pd)은 바로 언어[N]를 학습하[V]고 말하[V]고 이해하[V]는 본능[N]이다.
25. 역사상[N] 처음[N]으로 그 본능[N]에 관해[V] 쓸[V] 내용[N]이 생겼다[V].
26. 약 35[N]년 전[N]에 새로운 과학[N]이 탄생[V]했다.
27. 현재 '인지과학'[N]이라 불리[V]는 그것(Pd)은 심리학[N], 컴퓨터과학[N], 언어학[N], 철학[N], 신경생물학[N]에서 도구[N]들을 모아[V] 인간 지능[N]의 활동[N]을 설명한다[V].
28. 특히 언어[N]에 관한[V] 과학[N]은 그 후[N] 몇 년(ND)간 눈부신 발전[N]을 거듭했다[V].
29. 카메라[N]가 어떻게 작동하[V]고 비장[N]의 기능[N]이 무엇(PQ)인지 이해하[V]는 것(ND)만큼 우리(PP)는 수많은 언어[N]적 현상[N] 또한 잘 이해해[V] 나가[V]고 있다[V].
30. 나(Pp)는 이 흥미로운 발견[N]들을 전달하고자[V] 한다[V].
31. 그 일부[N]는 현대과학[N]의 어떤 것(ND) 못지않게 훌륭한 것(ND)들이다.
32. 그러나 나(PP)에게는 또 하나(PN)의 목적[N]이 있다.

A.3. Oraciones usadas en el corpus del CAPÍTULO 4

Breve muestra de las oraciones de castellano y de euskera etiquetadas para el estudio de corpus del Capítulo 4. Para poder obtener todas las oraciones hay que acceder al siguiente link (www.ehu.eus/HEB/Corpus_Tesis/Cap4.xlsx).

castellano	oración	omisión	género
1. 'Doña Manolita' tienta a la suerte.	transitiva	no	periódico
2. «Tarde o temprano conseguiré que se le borre».	transitiva	sujeto	libros
3. atendió a su hermana hasta el final,	transitiva	sujeto	periódico
4. Baleares se sitúa en los 12,2.	intransitiva	no	periódico
5. Bernat aceptó la oferta.	transitiva	no	libros
6. Bernat sonrió al cielo otoñal	intransitiva	no	libros
7. Científicos españoles diseñan androides de salvamento	transitiva	no	revista
8. Comenzó a escribir con cinco años	transitiva	sujeto	periódico
9. Con el título, unió sus apellidos	transitiva	sujeto	libros
10. Dos de las naves orbitarán en paralelo,	intransitiva	no	revista
11. El balón llegó hasta mí,	intransitiva	no	libros
12. Era la hija del guarda mayor,	intransitiva	sujeto	libros
13. Es legal.	intransitiva	sujeto	periódico
14. Giró el volante hacia la izquierda	transitiva	sujeto	periódico
15. Han ganado.	intransitiva	sujeto	periódico
16. Hoy por la mañana visitarán la ciudad	transitiva	sujeto	periódico
17. Joseba Egibar no ha querido el pacto en Guipúzcoa	transitiva	no	periódico
18. La caravana atravesará el corredor del Txorierri,	transitiva	no	periódico
19. La comitiva se movía lentamente.	intransitiva	no	libros
20. La genética es esencial en la naturaleza,	intransitiva	no	revista
21. La última se construyó en 1982.	intransitiva	no	periódico
22. La última se construyó en 1982.	intransitiva	no	periódico
23. los criados tenían prohibido jugar al yoyó.	transitiva	no	libros
24. los payeses celebraban las fiestas de septiembre.	transitiva	no	libros
25. Me aburrió muchísimo	transitiva	sujeto	periódico
26. Mi fortuna no tiene límites.	transitiva	no	libros
27. Parpadeaban despacio,	intransitiva	sujeto	libros
28. Por dentro era como inmensas burbujas rojas,	intransitiva	sujeto	libros
29. Se trataron todo tipo de cuestiones.	intransitiva	no	periódico
30. solo podremos disfrutar del evento en una ciudad,	transitiva	sujeto	revista
31. Suena en cierto modo al tránsito débil,	intransitiva	sujeto	periódico
32. Tomás insiste.	intransitiva	no	libros

33. Viaja en bicicleta por Europa con Google Maps	intransitiva	sujeto	revista
34. Vivir en grandes grupos aumenta la inteligencia	transitiva	no	revista
35. y el coche ha ganado las últimas carreras»	transitiva	no	periódico
36. y fecundé a cinco larvas.	transitiva	sujeto	libros
37. Y finalmente eliminaron de manera selectiva dichos potenciadores	transitiva	sujeto	revista
38. Y nunca se entretenía delante de los espejos.	intransitiva	sujeto	libros
39. y se ha muerto en la cama,	intransitiva	sujeto	periódico
40. y tardan más en escapar que en otros.	intransitiva	sujeto	revista

euskera	**oración**	**omisión**	**género**
1. «Berriro ere injustizi handi baten aurrean gaude,	intransitiva	sujeto	periódico
2. «Ekimen judizial honek bultzada politikoa dauka»	transitiva	no	periódico
3. «Oso arrazoi soil batek bultzatu nau babes talde honetan parte hartzera,	transitiva	objeto	periódico
4. Aldaketarako aldarrikapen handia dago	intransitiva	no	periódico
5. Alderdi honek sortzen ditu eskuratzearen tirabira gehienak.	transitiva	no	periódico
6. Beldur naiz itxi ezineko atea zabalduko duten»,	intransitiva	sujeto	periódico
7. Berandu iritsi da euskara batua,	intransitiva	no	libros
8. Bertzeak Sorbonako liburutegian aurkitu ditu.	transitiva	sujeto	libros
9. Betiko neska-lagunak popatik hartzera bidali zuenetik noraezean zebilen	intransitiva	sujeto	libros
10. Bordeleko bidea hartu du berriz jaun De Lancrek.	transitiva	no	libros
11. Downing Street nazional bat behar dugu,	transitiva	sujeto	periódico
12. Ederki oroitzen dudana ospitaleko amesgaiztoa da,	intransitiva	no	libros
13. eguzki-panel edo sentsore gisa ere erabil daiteke.	intransitiva	sujeto	revista
14. elkarrekin egiten genuen lo haren gelako ohe zabalean,	intransitiva	sujeto	libros
15. elkarrekin sartzen ginen bainerako ur epeletan,	intransitiva	sujeto	libros
16. Erresuma Batuko fisikari batzuek aurkitu dute grafenoarekin,	transitiva	objeto	revista
17. eta hezur bakoitzak istorio bat kontatzen du.	transitiva	no	revista
18. eta orduan ere berdin harritu zuen karrika eta plazetako zalapartak.	transitiva	objeto	libros
19. euskaldunok ez dugu duintasunik»,	transitiva	no	periódico

20. Euskarak zeukan bizitasuna adierazten du horrek ere.	transitiva	no	libros
21. Ezohiko segizio batek utzi du kaka arrastoa.	transitiva	no	libros
22. Gai honi buruz zuzenean ez du hitz egin,	intransitiva	sujeto	periódico
23. gaixotasun hau populazioren %6k pairatzen du.	transitiva	no	periódico
24. Gakoa pantailan dago.	intransitiva	no	revista
25. Herri hau beste agertoki batera doa	intransitiva	no	periódico
26. Herri honetan gauza asko daude konpontzeko.	intransitiva	no	periódico
27. Ikerketa hau Hego Korean egin dute	transitiva	sujeto	periódico
28. Jokalari alemanak erdiko peoi bat aurreratu zuen.	transitiva	no	libros
29. Lana sei ikerketa-taldek egin dute,	transitiva	no	revista
30. Lau kategoria saritu dira,	intransitiva	no	revista
31. Liburua txikia zen,	intransitiva	no	libros
32. Max Planck Astrofisika Institutuan dabil ikertzen,	intransitiva	sujeto	revista
33. Momentu honetan bakea prestatzen ari gara,	intransitiva	sujeto	periódico
34. nire porrotek ez ninduten gehiegi arduratzen,	transitiva	objeto	libros
35. Paristarrak ahapeka mintzo dira.	intransitiva	no	libros
36. Patal samar ibiltzen zen erdaraz egiterakoan	intransitiva	sujeto	libros
37. PNAS zientzia-aldizkarian argitaratu dituzte emaitzak.	transitiva	sujeto	revista
38. Sekulako bainera daukat,	transitiva	sujeto	libros
39. Une honetan Ertzaintzak ez du aplikatzen.	transitiva	objeto	periódico
40. Xenotransplanteen inguruan aurrerakuntzak egon dira	intransitiva	no	periódico

A.4. Oraciones de euskera y castellano en el corpus del CAPÍTULO 5

Breve muestra de las oraciones de euskera etiquetadas para el estudio de corpus del CAPÍTULO 5. Para poder obtener todas las oraciones hay que acceder al siguiente link (www.ehu.eus/HEB/Corpus_Tesis/Cap5.xlsx). Como los datos de castellano los he obtenido directamente del corpus ADESSE no proporciono los ejemplos de las oraciones de castellano.

euskera	**orden**	**omisión**	**sujeto**	**objeto**	
1. AINHOA.— nik itsasoako aingira bat nahi dut...	SOV	no	animado	animado	AA
2. aitaordea begira daukat.	OV	sujeto	animado	animado	AA
3. AMAIA.— Gauza guztiek zaukatek hasiera bat,	SVO	no	inanimado	inanimado	II
4. AMAIA.— jendea erakarriko luke bertara...	OV	sujeto	inanimado	animado	IA
5. Asobaleko bigarren itzulia hasiko duzue gaur.	OV	sujeto	animado	inanimado	AI
6. Atal honetan AHVren bideragarritasuna aztertuko dugu.	OV	sujeto	animado	inanimado	AI
7. Aukera honek abantaila batzuk eskaintzen ditu:	SOV	no	inanimado	inanimado	II
8. Aurrerantzean gai aldetik gauza normalagoak bilatuko ditut,	OV	sujeto	animado	inanimado	AI
9. baina gaur gonbidatu berezia dugu gurekin, Jon Arretxe,	OV	sujeto	animado	animado	AA
10. Bakeak fidelak saritzeko tenorea ekartzen ohi du berekin.	SOV	no	inanimado	animado	IA
11. beste euskarri batean grabatzeko aukera eskaintzen du sistemak.	OVS	no	inanimado	inanimado	II
12. biak kutsatzeko gai den birus bat sor dezake.	OV	sujeto	inanimado	animado	IA
13. bidaiariak zerbitzaria besarkatu du ostatuko zoko batean.	SOV	no	animado	animado	AA
14. Bigarren itzuliko lehen aurkaria Bidasoa duzue, Irunen.	OV	sujeto	animado	animado	AA
15. Dena dela, aurrean Titin eta Goñi izango dituzue.	OV	sujeto	animado	animado	AA
16. EBko kideek koordinazio handiagoa hitzartu dute.	SOV	no	animado	inanimado	AI

17. Enplegua arautzeko espedientea aurkeztu dute Landabenen .	OV	sujeto	animado	inanimado	AI
18. ERCk garaipen handia lortu du hegoaldean.	SOV	no	animado	inanimado	AI
19. eta balio juridikoa izango du medikuaren aurrean.	OV	sujeto	inanimado	inanimado	II
20. eta nire bizimoduak beste norabide bat hartu zuen.	SOV	no	inanimado	inanimado	II
21. eta punta lodia daukate.	OV	sujeto	inanimado	inanimado	II
22. GREGORIO.— lan ona topatu omen du:	OV	sujeto	animado	inanimado	AI
23. Historikoki, garai hartako gizarteari buruzko zantzuak eman ditzake.	OV	sujeto	inanimado	inanimado	II
24. Honek istilu handiak sortuko ditu.	SOV	no	inanimado	inanimado	II
25. HORTENTSI.— neuk zabalduko dut kioskoa…	SVO	no	animado	inanimado	AI
26. italiar kulturaren astea antolatu zuen udalak.	OVS	no	animado	inanimado	AI
27. jendeak inguruetako ostatuetan bazkaltzeko erabili zuen,	SV	objeto	animado	inanimado	AI
28. King gitarristaren ibilbidea ekartzen du gogora.	OV	sujeto	inanimado	inanimado	II
29. laguna dute Emiliak eta Mikelek;	OVS	no	animado	animado	AA
30. MARIA LUISA.— Oso jende ona ezagutu dugu,	OV	sujeto	animado	animado	AA
31. Medikuntza alorrean aurkikuntza garrantzitsuak egin zituen,	OV	sujeto	animado	inanimado	AI
32. Mutazio horietako askok hil egingo dute birusa,	SVO	no	inanimado	animado	IA
33. Neska besarkatzen du,	OV	sujeto	animado	animado	AA
34. Nik mutil asko zarrantzatu ditut,	SOV	no	animado	animado	AA
35. Osagarri guztiak dauzka sekulako arrakasta izateko:	OV	sujeto	inanimado	inanimado	II
36. Teknologia berriek gizarte osoa aldatu dute oso denbora gutxian.	SOV	no	inanimado	animado	IA
37. Txinatarrek ezagutzen dute grabitatearen legea,	SVO	no	animado	inanimado	AI
38. TXOMIN.— Estanpa asko dauzka.	OV	sujeto	inanimado	inanimado	II
39. zuk lagunduko diguzu horretan.	SV	objeto	animado	animado	AA
40. Zuk mendekatuko duzu gure Catherine.	SVO	no	animado	animado	AA

A.5. Textos utilizados por Hidalgo (1995) en su estudio de corpus

a) *Textos del siglo XVI*

Urgell, B. (1985). *Refranes y sentencias-eko hitzor denaz, zenbait ohar*. Manuscrito.

Sarasola, I. (1983). Fray Juan de Zumarragaren Gutuna. *ASJU*, XVII, 97-102.

b) *Textos del siglo XVII*

Axular, P. (1643). *Gero*. Aranzazu: Jakin (Ed. 1976).

Tartas, J. (1666). Onsa hilceco bidia. *Orthez* (RIEV, I-III).

Micoleta, R. (1653). *Modo breve para aprender la lengua vizcayna*. Barcelona (Ed. 1880) / Sevilla (Ed. 1897).

Capanaga, O. (1656). *Exposición breve de la doctrina christiana*. Viso: Dodgsonen, (Ed. 1893).

Mitxelena, L. (1954) [ed.]. Viva Jesus. BAP, X.

Arejita, A. (1983). Domingo egiaren kanta. *Euskeraren Iker Atalak*, 2, 139-181.

Lakarra, J. (1984) [ed.]. Bertso bizkaitarrak. *ASJU*, XVIII, 89-197.

c) *Textos del siglo XIX*

Moguel, J. A. (1802). *Peru Abarca*. Durango (Ed. 1881; facsímil, 1981).

Aguirre, J. B. (1808). *Eracusaldiac*. Tolosa (Ed. 1850) / Donostia: Hordago (Ed. 1978).

Duvoisin, J. P. (1858). *Laborantzako liburua*. Donostia: EEE (Ed. 1986).

Apaolaza, A. (1890). *Pachico Cherren*. Bergara (Ed. Erein, 1992).

Azkue, R. M. (1893) Behin da betiko. Lenengo irakurgaia. Bilbao / Donostia: EEE (Ed. 1986).

Webster, W. (1993). *Euskal ipuinak*. Donostia: EEE.

Cerquand, J. F. (1874-1885). Légendes et récits populaires du Pays Basque. *Bulletin de la Societé des Sciences, Lettres et Arts de Pau*. [Ed. euskera (1985-6): IparEuskalherriko legenda eta ipuinak, Donostia: Txertoa].

Urruzuno, P. M. (1988). *Ipuinak*. Donostia: EEE.

d) *Textos del siglo XX*

Anztia, M. (1934). Laburditar ipuñiak. *Anuario de Eusko-Folklore*, XIV, 93-129.

Etxebarria, J. M. (1991). *Zeberio haraneko euskararen azterketa etno-linguistikoa*. Deustu: Ibaizabal.