IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

I. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2015eko maiatzaren 13, 14 eta 15
Durango, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

GIZA ZIENTZIAK

Elzen eta Aditzen frekuentziak SVO eta SOV hizkuntzetan

Luis Pastor

116-120 or.
https://dx.doi.org/10.26876/ikergazte.i.15

ANTOLATZAILEA: BABESLEAK: LAGUNTZAILEAK:

Izen eta Aditzen frekuentziak SVO eta SOV hizkuntzetan

L. Pastor*

*Euskal Herriko Unibertsitatea
luis.pastor@ehu.es

Laburpena

Hizkuntza guztiek desberdintzen dute Izen eta Aditz kategoria sintaktikoak eta unibertsaltasun horretan oinarrituz hainbat ikerketek erakutsi dute badagoela korrelazio bat hizkuntzen artean bi kategori hauen eta zenbait ezaugarri linguistikoen artean: komunztadura eta hitz-ordena. Hemen erakusten dugu izen-aditz erratioen datuak hainbat hizkuntzetan. Hori lortzeko, corpus digitalak erabili dira eta hizkunza bakoitzeko corpusa 300.000 hitzez osatuta dago, prentsa artikuluez osatuta. Emaitzek erakusten dute badagoela korrelazio bat izen-aditz erratioaren eta hitz-ordenaren artean: izen-aditz erratioa baxua duten hizkuntzek OV dira eta erratio handia dutenek VO. Halaber, emaitzek ez dute erakusten inolako korrelaziorik izen-aditz erratioaren eta komunztaduraren artean.

Hitz gakoak: izena, aditza, hitz-ordena, komunztadura, gaztelania, euskara

Abstract

All languages distinguish the syntactic categories Noun and Verb and based on this universality some studies have observed correlations between the relative distribution of these two categories cross-linguistically and certain linguistic traits: agreement and word-order. Here we present data from a cross-linguistic study of the noun-verb ratios in some languages. To this end, we used digital corpora and each digital corpus contains about 300,000 words for each language, obtained from newspaper articles. Our results reveal that there is a correlation between the noun-verb ratio and the headedness: languages with low noun-verb ratio are OV languages, and those with high noun-verb ratio VO languages. However, our results do not show a correlation between the noun-verb ratio and agreement.

Keywords: noun, verb, word order, agreement, spanish, basque

1. Sarrera

Munduko hizkuntza guztietan “izen” eta “aditza” kategoria lexikoak/sintaktikoak desberdindu daitezke (Croft, 1991). Tradizionalki, desberdintasun hori hiru ezaugarrien arabera egin ohi da: (a) semantiko, (b) morfologikoa, eta (c) sintaktikoa (Hopper y Thompson, 1984; Schachter, 1985; Evans, 2000; Givón, 2001; Haspelmath, 2001; Beck, 2002; Schachter y Shopen, 2007). Ezaugarri semantikoaren arabera, izenek “gauzak” edota “entitateak” adierazten dute; aditzak, berriz, “ekintzak”, “egoerak” edota “gertaerak”. Morfologikoki, izenak numeroa eta generoa markatzen duten atzizkiak eraman ohi dute, eta aditzak, ordea, aldia, modua, pertsona markak. Eta sintaktikoki, izenak perpausaren subjektua ala objektua izan daiteke, baina aditzak subjektua zein objektuaren egoerari buruz adierazten du. Hala ere, hainbat kasutan ezaugarri hauek huts egin dezakete, adibidez “suntsipen” hitzean, “ekintza” bat adierazten duelako eta ez “gauza” edota “entitate” bat. Beraz, unibertsaltasun honetan oinarrituz, hainbat ikerketak ikusi dute nolabaiteko korrelazioa dagoela izen-aditzen distribuzio erlatiboaren (izen-aditz erratioa, hemendik aurrea) eta zenbait ezaugarri linguistikoen artean: aditz-komunztadura eta hitz-ordena, batik bat.

Seifartek (2010, 2011) izen-aditz erratioa eta aditz-komunztaduraren arteko korrelazioa ikusi du. Erratio hori bost hizkuntzetan ikertzen ditu (Baure, Chintang, Bora, N|uu and Sri Lanka Malay), DoBeS corpusa erabiliz, genero desberdineko narrazioz, elkarrizketaz, epai-testuz, eta abestiz osatuta dagoena; eta izen-aditz erratioa kalkulatzeko hurrengo formula erabiltzen du: izenak / izenak + aditzak. Haren emaitzek eraskuten dute korrelazio bat dagoela izen-aditz erratioaren eta aditz-komunztaduraren artean: zenbat eta komunztadura gehiago eduki hizkuntza batek, orduan eta izen-aditz erratio txikiago edukiko du. Polinskyk (2012), berriz, izen-aditz erratioa eta hitz-ordenaren arteko korrelazioa ikusi du. Izen-aditz erratioa hogeita hamar hizkuntzetan ikertzen ditu WordNet (Miller et al. 1990) datu-base lexikoa erabiliz, eta erratioa kalkulatzen du izenak aditzekin zatikatuz (izenak / aditzak). Haren emaitzek eraskuten dute korrelazio bat dagoela izen-aditz erratioaren eta hitz-ordenaren artean: VO hizkuntzek izen-aditz erratio baxua daukate eta OV hizkuntzek erratio handiagoa.

Ikerketa honen helburua aurreko bi ikerketen erreplika egitea da, hau da, izen-aditzen distribuzioak eakorrelazioa daukan hitz-ordena eta komunztadurarekin ikusi. Izan ere, era berean, korrelazio bat dagoelako hitz-ordena eta komunztaduraren artean: VO hizkuntzek komunztadura gutxi edo eza dute eta OV

hizkuntzek komunztadura gehiago.

2. Ikerketa

Corpus ikerketan honetan sei hizkuntzetan oinarritu da eta hitz-ordenaren arabera bi taldetan banatu dira: VO hizkuntzak (gaztelania, katalana, galegoa, ingelesa eta portugera) eta OV hizkuntzak (euskara, japoniera, koreera, turkiera eta armeniera). Polinskyren ikerketan euskara, gaztelania eta japoniera erabili arren, hizkuntza hauek erabiltzea erabaki da berak erabiltzen dituen corpusak txikiak eta desorekatuak daudelako. Hemen erabilitakoan, ia hizkuntza guztietarako tamaina berako corpusa erabili izan da, 300.000 hitzezkoa alegia (1. taula). Corpus guztiek prentsako artikuluez osatuta daude eta kategoria lexikoen (*Parts-of-Speech – PoS*) arabera etiketatuak.

1. taula. Hizkuntza bakoitzeko corpusaren datuak.

	Corpus	*Hizkuntza*	*hitz kopurua*
VO	**AnCora**	*gaztelania*	302.017
	AnCora	*katalana*	302.927
	CORGA	*galiziera*	300.257
	COCA	*ingelesa*	301.043
	CETEMPúblico	*portugesa*	300.034
OV	**EPEC**	*euskara*	300.000
	METU-Sabanci Turkish Treebank	*turkiera*	*7262 perpaus[1]
	JEITA	*japoniera*	301.789
	HQMSAC	*koreera*	299.532
	EANC	*armeniera*	*653 dokumentu[2]

Izen eta aditzen maiztasunak kontatzeko orduan, kontutan eduki dira izen eta aditz bezala agertzen diren hitz guztien agerpenak. Era berean, izen etiketa dutenen artean bakarrik izen arruntak direnak kontutan hartu dira, hau da, alde batera utzi dira izen bereziak, leku izenak eta siglak. Aditzei dagokienez, aditz laguntzaileak bakarrik geratu dira kontaketatik kanpo. Izenordainen agerpena ere kontutan eduki da, Seifartek (2010, 2011) bere izen-aditz erratioan erabiltzen dituelako, izenak bezala aditzarekin komunztadura ere baitute. Izen-aditz erratioa kalkulatzeko bi formula erabili dira: bata komunztadura erlaziorako (1), eta bestea hitz-ordena erlaziorako (2).

(1) Izen-aditz erratio eta komunztadura korrelazioa

$$\text{izen-aditz erratioa} = \frac{\text{(izenak + izenordainak)}}{\text{aditzak}}$$

(2) Izen-aditz erratio eta hitz-ordena korrelazioa

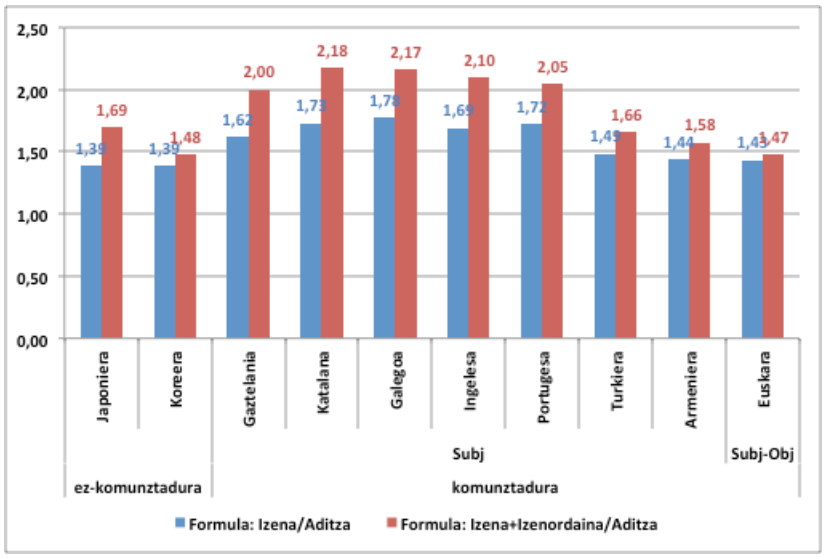
$$\text{izen-aditz erratioa} = \frac{\text{izenak}}{\text{aditzak}}$$

Lortutako izen-aditz erratioak eta komunztadura zein hitz-ordena korrelazioak aztertzeko Wilcoxon testa erabili da, izen-aditzen banaketa VO-OV hizkuntzen artean eta komunztadura eta komunztadura ez duten hizkuntzen artean esanguratsua den ikusteko. Bi azterketarako (p) .05 alfa-maila erabili da.

[1] *METU-Sabanci Turkish Treebank* corpusa *METU Turkish Corpus*etik ateratako perpausez osatuta dago.
[2] Ezin izan da artxiboetara jo, beraz datuak EANCen 2002ko prentsa artikuluetara mugatutako bilaketatik atera dira.
[3] Grafikoetan kolore desberdineko bi zutabe ikus daitezko. Urdinezko zutabeak "izenak / aditzak" formula erabiliz lortutako izen-aditz erratioak erakusten dute, eta gorrizko zutabeak, berriz, "izenak + izenordainak / aditzak"
[2] Ezin izan da artxiboetara jo, beraz datuak EANCen 2002ko prentsa artikuluetara mugatutako bilaketatik atera dira.

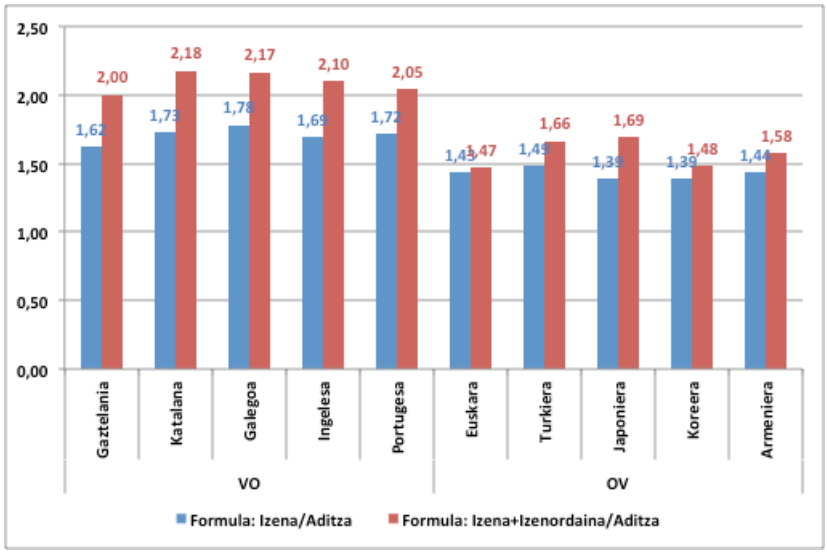
Emaitzek erakusten dute ez dagoela desberdintasun esanguratsurik izen-aditz erratioaren eta komunztaduraren artean (1. irudia). Ez dago inolako desberdintasunik komunztadura (gaztelania, katalana, galegoa, ingelesa, portugesa, turkiera, armeniera eta euskara) eta komunztadura ez duten hizkuntzen artean (japoniera eta koreera), izen-aditzen erabilera erlatiboari dagokionez [$Z = 1.045$, $p = .296$].[3]

1. irudia. Izen-aditz erratioa komunztadura eta komunztadura ez duten hizkuntzetan.



Izen-aditz erratio eta hitz-ordena korrelazioari dagokionez, ordea, bai aurkitzen da desberdintasun esanguratsua (2. irudia). OV hizkuntzek (euskara, turkiera, japoniera, koreera eta armeniera) izen-aditz erratio baxuagoa erakusten dute VO hizkuntzekin konparatuta (gaztelania, katalana, galegoa, ingelesa eta portugesa) [$Z = -2.611$, $p < .009$].

2. irudia. Izen-aditz erratioa VO eta OV hizkuntzetan.



Emaitza hauek erakusten dute ez datozela bat Seifart (2011) eta Polinskyk (2012) proposatutako hipotesiekin. Alde batetik, Seifarti (2011) dagokionez, komunztadura edukitzeak edo ez edukitzeak ez du inolako eraginik izen-aditz erabilera erlatiboan. Bestetik, Polinskyri (2012) dagokionez, bai ikusten dela

[3] Grafikoetan kolore desberdineko bi zutabe ikus daitezko. Urdinezko zutabeak "izenak / aditzak" formula erabiliz lortutako izen-aditz erratioak erakusten dute, eta gorrizko zutabeak, berriz, "izenak + izenordainak / aditzak" formularen erratioak. Bi formulak irudikatu dira grafikoetan erakusteko ez dagoela desberditasunik formula bat edo beste erabiltzeagatik korrelazioak aztertzerakoan.

desberdintasuna VO-OV hizkuntzen artean izen-aditzen erabilera, baina hemen aurkitzen dena Polinskyk planteatutako hipotesiaren kontrako joeran doa: OV hizkuntzek izen-aditz erratio baxua erakusten dute eta VO hizkuntzek erratio handiagoa.

Hau dena kontutan izanda, bigarren ikerketa bat egin da testu beraz osatutako corpus batean. Honen arrazoia izan da ikustea ea testuaren jatorrizko hizkuntzak bere izen-aditz erratioa beste hizkuntzetan eragina duen edo hizkuntza bakoitzak bere izen-aditz erratioa erakutsiko duen. Lau hizkuntza erabili dira konparaketa honetan: bi VO hizkuntza (ingelesa eta gaztelania) eta bi OV hizkuntza (gaztelania eta euskara). Aukeratu den testua Steven Pinkerren *The Language Instinct*-eko lehenengo kapitulua izan da: jatorrizko testua (ingelesez) eta bere itzulpenak (gaztelaniaz, euskaraz eta koreeraz) (2. taula).

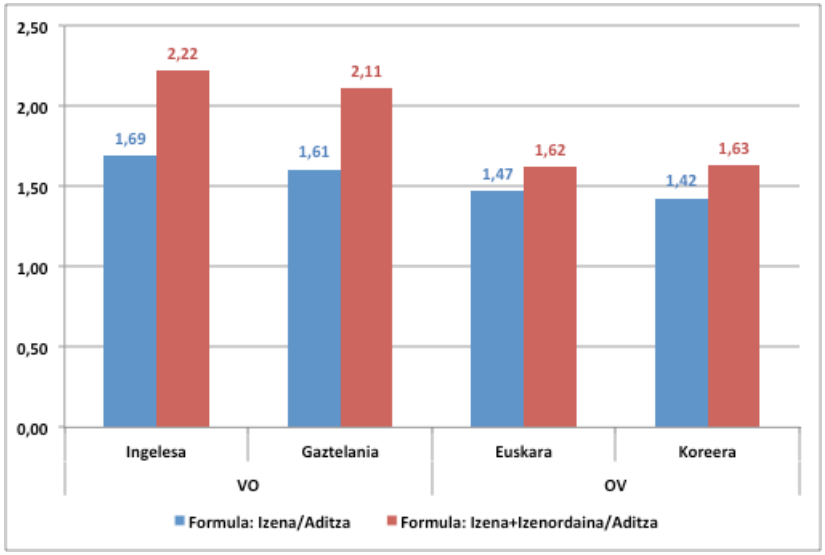
2. taula. Hizkuntza bakoitzeko "Pinker" corpusaren datuak.

	hizkuntza	***hitz kopurua***	***Corpus***
VO	*ingelesa*	3.861	Pinker (1994)
	gaztelania	4.477	Pinker (1995a)
OV	*euskara*	3.158	Pinker (2010)
	koreera	3.046	Pinker (1995b)

Testuak etiketatzerako orduan, bakarrik izen bezala etiketatu dira izen arruntak eta aditz bezala aditz nagusiak, aditz laguntzaileak alde batera utzirik. Era berean, izen-aditz erratioa kalkulatzeko aurreko corpus azterketan erabilitako "izenak / aditzak" formula erabili da (ikus (2)), eta lortutako izen-aditz erratioak eta hitz-ordena korrelazioa aztertzeko Wilcoxon testa erabili da. Ez da izen-aditz erratioa eta komunztaduraren arteko korrelazioa konparatu, aurreko corpus azterketan ez delako desberdintasun esanguratsurik ikusi. Azterketarako (p) .05 alfa-maila erabili da.

Emaitzek erakusten dute ez dagoela desberdintasun esanguratsurik izen-aditz erratio eta hitz-ordenaren artean (3. irudia) [Z = -1.549, p = .121]. Hala ere, bai ikusten da OV hizkuntzek (euskara eta koreera) izen-aditz erratio baxuago edukitzeko joera dutela VO hizkuntzekin (ingelesa eta gaztelania) konparatuta. Baliteke hizkuntza gehiago erabiliko balira konparaketan, desberdintasun esanguratsuak aurkituko liratekeela.

3. irudia. Izen-aditz erratioa VO eta OV hizkuntzetan.



3. Ondorioak

Orokorrean, eta kontutan izanda emaitza guztiak, ondoriozta daiteke badagoela korrelazioa bat izen-aditz erratio eta hitz-ordenarean artean. Halaber, ez dago korrelaziorik izen-aditz erratio eta komunztaduraren artean. Polinskyren lanak (2012) erakusten du OV hizkuntzek izen gehiago erabiltzen dituztela aditzekin konparatuta, eta VO hizkuntzek, ordea, izen gutxiago aditzekin konparatuta. Beraz, badago korrelazio bat izen-aditz erratio eta hitz-ordenaren artean. Hamen ere ikusi eta erakutsi egin da badagoela korrelazio hori,

baina ez Polinskyk (2012) planteatzen duen moduan. Corpus ikerketa honetako emaitzek kontrako joera erakusten dute: OV hizkuntzek izen gutxiago erabiltzen dituztela VO hizkuntzekin konparatuta. Honen arrazoietako bat izan daiteke *pro*-droparen (argumentuen murrizketa) erabilera, OV hizkuntzek pro-drop kasu gehiago dituztelako VO hizkuntzen aldean (Ueno & Polinsky, 2009; Pastor & Laka, 2013).

4. Etorkizunerako planteatzen den norabidea

Hurrengo pausoak izango dira, alde batetik, Pinkerren testua beste hizkuntza gehiagotan aztertzea ikusteko ea izen-aditz erratioen desberdintasunak esanguratsuak diren OV-VO hizkuntzen artean; eta bestetik, ikusi *pro*-droparen eragina izen-aditz erratioetan.

5. Erreferentziak

Beck, D. (2002). *The Typology of Parts of Speech Systems: The Markedness of Adjectives*. New York: Routledge.

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

Evans, N. (2000). Word classes in the world's languages. In G. Booij, C. Lehmann, J. Mudgan & S. Skopeteas (Eds.), *Morphology: An International Handbook on Inflection and Word-Formation* (Vol. 1, pp. 708-732). Berlin: de Gruyter.

Givón, T. (2001). *Syntax* (Vol. 1). Amsterdam: John Benjamins Publishing.

Haspelmath, M. (2001). Word Classes and Parts of Speech. In P. B. Baltes & N. J. Smelser (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 16538-16545). Amsterdam: Elsevier.

Hopper, P. J., y Thompson, S. A. (1984). The Discourse Basis for Lexical Categories in Universal Grammar. *Language, 60*(4), 703-752.

Pastor, L., Laka, I. (2013) Processing facilitation strategies in OV and VO languages: a corpus study. *Open Journal of Modern Linguistics, 3,* 252-258.

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.

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* (pp. 1-60). Cambridge: Cambridge University Press.

Seifart, F. (2011). *Cross-linguistic variation in the noun-to-verb ratio: the role of verb morphology and narrative strategies*. Paper presented at the Association for Linguistic Typology 9th Biennial Conference, University of Hong Kong, Japan. Poster retrieved from

Seifart, F., Meyer, R., Zakharko, T., Bickel, B., Danielsen, S., Nordhoff, S., y Witzlack-Makarevich, A. (2010). *Cross-linguistic variation in the noun-to-verb ratio: Exploring automatic tagging and quantitative corpus analysis*. Paper presented at the DobeS Workshop "Advances in Documentary Linguistics", Nijmegen.

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

7. Eskerrak eta oharrak

Ikertzaileek BES-2010-030196 (MICINN), FFI2012-31360 (MICINN) eta IT665-13 (Eusko Jaurlaritza) proiektuen babesa jaso dute. Masha Polinsky (Harvard) eta Maxux Aranzabe (EHU) irakasleak beraien denbora eta laguntza eskertzekoa da ere.