BOGA: Basque Studies Consortium Journal

Volume 11 | Issue 1 Article 2

May 2024

Word Concreteness and Word Originality: A Basque Case Study

Casey Kennington
Boise State University, caseykennington@boisestate.edu

Recommended Citation

Kennington, Casey (2024) "Word Concreteness and Word Originality: A Basque Case Study," *BOGA: Basque Studies Consortium Journal*: Vol. 11 : Iss. 1 , Article 2.
https://doi.org/10.18122/boga.11.1.2.boisestate
Available at: https://scholarworks.boisestate.edu/boga/vol11/iss1/2

Word Concreteness and Word Originality: A Basque Case Study

Casey Kennington

Introduction

Basque is known to be the oldest spoken language in Western Europe, yet not of the Indo-European language family that makes up all other spoken languages in Western Europe. Basque's remarkable staying power is a testament not only to how those who spoke Basque through the ages continued to speak it in the ancient world even when surrounded by Romans, up to modern times when it was illegal to speak Basque during the tumultuous Franco years (1936-1975).

Post-Franco, the Basque language could be freely spoken, but putting the pieces together of a language full of diverse dialects was a surmountable challenge. Alberdi et al. (2012) explains the history of some of the legal hurdles to give Basque— Euskara—a place of acceptance then prominence within two countries with other official languages. It was the Basques themselves who had to fight for those legal rights, and at the same time implement concrete actions that would help teach adults and the rising generation the Basque language, define a common dialect to teach and speak, motivate settings for

speaking it, and encourage broader use and application of Basque when everyone already could speak another more commonly spoken language.

This begs, I think, an important question: what makes it so a word in Basque remains Basque, where other words adopt neighboring vocabulary? arropak, for example, is the Basque word for clothes borrowed from the Spanish ropa, even though there is a Basque word for clothes tresenak (according to Trask (2008)[1]) Why didn't gorria (red) or thousands of other words suffer the same fate?

In this paper, I explore the hypothesis that concreteness may be a factor in whether or not a word in Basque remains in use today instead of being overpowered by a loan word. Concrete words are words that denote physical entities and properties, such as chair or red, whereas words that are defined by other words and often denote ideas are abstract, such as utopia or democracy. I hypothesize that words that are more concrete (i.e., words that refer to physical objects) are older and therefore less likely to be borrowed, and indeed my experiment shows that it is the case. Knowing whether or not concreteness is a factor in word borrowing can tell us how words that refer to physical things are important to a culture even when abstract ideas from other languages and cultures have a potentially big influence.

I also explore using natural language processing techniques to arrive at an approximation of semantic meaning of words using pre-trained word embeddings where a word's meaning is vectorized (i.e., represented as a list of numbers) and words that have similar "meanings" cluster together. For example, words like cat and dog are closer together to each other than they are to words like car and van. I find that a set of original Basque words remain close together in a vector space compared to commonly spoken contemporary words.

Background

An insight on that process of transitioning to a society where Basque could be freely spoken and more widely used was documented by Urla (2012). The author performed multiple field studies in the Basque Country (starting in Ursibil) during the time that Euskara Batua—the newly proposed unified dialect—was introduced and through some of the challenges of adoption. She met people who were in favor of the change, as well as those who opposed it. The book focuses on the social environment, not the language itself, during the important years during the Batua uptake, including the challenges that speakers of some Basque dialects faced. Some chapters chronicle the efforts to make Basque more mainstream at governmental policy levels, and others focus on grassroots efforts to

[1] Some Basque words came from a time when Sabino Arana de Goiri attempted to make Basque more distinct from outside influence, coining, for example jantziak for clothes.

actually use Basque in everyday interactions and expressions such as in shops and magazines. Of course, the work that it took to bring Basque to life was met with opposition, and the book chronicles some of the struggles that individuals and organizations went through to keep their heritage alive. A reader can appreciate the sociolinguistic approach to the subject, for example it highlights power dynamics from the Basque perspective.

The author puts some focus on an interesting turn of events: the shift from an attempt to make the Basque language the superior language of the region and to cleanse it from the influence of external languages such as French and Spanish, and instead attempt to celebrate the fact that Basque speakers were all bilingual in Basque and either Spanish or French. From my reading of the histories of the Basque people, this seems to be the most Basque way of handling the situation: instead of fighting against outside influences, integrate and even improve upon those influences without sacrificing what it means to be Basque.

Besides word concepts, pronunciation is another aspect of language that has been influenced by neighboring languages. Michelena (1995) looked at the ancient Basque consonants, their similarities and differences from contemporary Basque, a difficult feat, as Basque has only recently–within the past few hundred years– developed a writing system (Salaburu and Alberdi, 2012). Furthermore, in Hualde (1995), the authors attempt to reconstruct the ancient Basque accentual system showing that contemporary pronunciation differs dramatically from ancient Basque and that is due in part to influence of neighboring languages such as Spanish and French (for example, syllable stress; see Hualde 2022). This shouldn't be a surprise as many neighboring languages influence each other, and in this case even the converse is true: a recent study has shown that Castilian spoken around the Basque Country has been influenced by Basque pronunciation (Elordieta and Romera, 2021).

Pronunciation, however, is only one aspect of language. Another linguistic aspect that makes Basque stand out from other Indo-European languages is its unique syntax. Basque is ergative (identifies the subject of transitive verbs; i.e., verbs that have an object) and agglutinative (affixes are added directly to the word instead of as separate words), both important features that make it very different from neighboring languages. It could be argued that Basque would be just as unique of a language if it adopted all of its vocabulary from Spanish but retained the unique grammar, but that doesn't answer the question as to how some words were borrowed whereas other words were not. This question is challenging because the answer relies to some degree on being able to approximate and compare the semantic meanings of words.

In the following section, I explain the set of old words and contemporary words I used and how I arrived at concreteness scores for those words. I then explain the analysis using the word embeddings. I then provide discussion and conclude.

Concreteness of Old Basque Words

Methodology

I used the Etymological Dictionary of Basque (Trask, 2008)[2] which identifies etymological origins for several hundred Basque words, including their dialect of origin or whether they were borrowed from other languages such as Spanish. The author used nine dialects identified by Michelena, but further identified 4 old dialects: old Bizkaian (B), Gipuzkoan (G), Low Navarrese (LN) and Zuberoan (Z). I identified 478 words that belong to at least one of the old dialects, not borrowed from any known language.

As this list of words is known to be very old, as a point of comparison we used Basque Wikipedia (taken from 2019-09-14; 3,676,125 tokens) and identified the 400 (roughly the same number of words) most commonly used words. The average number of times those words appeared in Wikipedia was over 48,000 times each (the most common word was eta), whereas the average count of the words in the original 429 old words list was 1,438 times–much lower token count, but still fairly representative. It is important to note that we tokenized (i.e., identified individual words within a text) words based on space and removed punctuation; we did not identify the root lemma (i.e., the most basic form of the word) the words, so all possible conjugated forms of a root word is considered a different word (e.g., katu, katua, katuak three ways to use the word cat) are separate words Moreover, dictionary entries are root forms and are uncommonly seen in written or spoken text without some kind of suffix. That means that the list of 429 old words likely has greater representation in the text if we applied correct suffixes, but we leave better identification of root words for future work. This led to our comparison set of words having an overlap of only one word with the old list: baino. We don't check to see if the words in the contemporary list are of Spanish origin; whatever the distribution of original and loan Basque words should be distributed properly in the word list, with preference to commonly spoken and written words. Therefore, these lists should represent well a sample of old words and contemporary words and can be used as a proper comparison of concreteness.

Unfortunately, there is no direct way to compute concreteness for Basque words because there is no known dataset that links Basque words to any measure of concreteness. However, Brysbaert et al. (2014) published a list of concreteness scores for over 40,000 English words. The authors collected concreteness ratings by presenting participants with English words and asked them to rate the words for their concreteness (after explaining what was meant by concreteness; see below) one a scale of 1-5 and averaging over all participant ratings for each word. For example, traindriver has a concreteness score of

[2] I was unable to find other options for Basque word origins. Trask sites well-known authors in this area while constructing the dictionary, and, more importantly, the dictionary is in a format that is accessible for research using computational tools.

4.54 (very concrete) and essentialness has a score of 1.04 (very abstract). The authors explained concreteness to participants as follows:

Some words refer to things or actions in reality, which you can experience directly through one of the five senses. We call these words concrete words. Other words refer to meanings that cannot be experienced directly but which we know because the meanings can be defined by other words. These are abstract words. Still other words fall in-between the two extremes, because we can experience them to some extent and in addition we rely on language to understand them. We want you to indicate how concrete the meaning of each word is for you by using a 5-point rating scale going from abstract to concrete.

A concrete word comes with a higher rating and refers to something that exists in reality; you can have immediate experience of it through your senses (smelling, tasting, touching, hearing, seeing) and the actions you do. The easiest way to explain a word is by pointing to it or by demonstrating it (e.g. To explain 'sweet' you could have someone eat sugar; To explain 'jump' you could simply jump up and down or show people a movie clip about someone jumping up and down; To explain 'couch', you could point to a couch or show a picture of a couch).

An abstract word comes with a lower rating and refers to something you cannot experience directly through your senses or actions. Its meaning depends on language. The easiest way to explain it is by using other words (e.g. There is no simple way to demonstrate 'justice'; but we can explain the meaning of the word by using other words that capture parts of its meaning). (Brysbaert et al. 2014)

Procedure

We therefore first translated each word to English using the Google Translate API by treating each word as an individual word without lexical context. This introduces a problem that the translation is biased towards newer words, and would likely translate incorrectly due to ambiguity in that some words have multiple senses, introducing noise into the data. However, the lists of words are long enough to give us the impression that the noise will not affect the results in either direction (i.e., more or less concrete overall).

For each word in each list, we translated the word into English then found its concreteness score using the ratings from Brysbaert et al. (2014). We report the average concreteness score for both word lists. For words that were not in the concreteness dataset, we simply discarded them. This resulted in 122 words from the original old list of 429 words and 215 words in the list of words derived from Wikipedia.[3]

[3] We used the 2019 Wikipedia dump for Basque: https://archive.org/details/incr-euwiki-20190609.

Results

The mean concreteness score for the old words list was 3.52 and the mean concreteness score for the contemporary word list was 3.02. We conducted a statistical significance test and found that the difference was indeed not due to random chance (ttest=-4.13, p=0.000049).

Discussion

It may not be surprising that, on average, words that are more concrete are the words that are original to Basque. We conjecture that older words are more commonly used especially in day-to-day life where work consisted of farming, fishing, factory work, and household necessities which required reference to concrete objects and actions. More abstract terms and neologisms were later introduced through education when Spanish happened to be the language of instruction. Moreover, learners of Basque (children as well as adults) generally first learn concrete terms before abstract terms, making concrete terms a more foundational part of the language. Though opinions vary, it is generally believed that 20-25% of spoken Basque comes from loan words, though some argue that up to half of spoken Basque vocabulary are loan words (Trask 1998).

Further Analysis: Word-level Embeddings

Concreteness and abstractness is only one dimension of a word's meaning. To further estimate a word's meaning, the distributional hypothesis posits that a word's meaning can be computed based on words that commonly surround it. Word embeddings are "trained" using this idea: as the model observes words in their lexical context, it shifts their location in a high dimensional vector such that, when the model has observed all of the data words that have similar meanings are in similar areas of the embedding space; for example the animals cat and dog have similar vectors because the vector model has seen many examples of how those two words are used in text. Similarly, the model has seen many examples of the vehicles and car and van, which likewise are close to each other but distant from the animal groups.

Such vectors exist for Basque words. The vectors provided by Agerri et al., (2020) were trained using Basque Wikipedia (similar to the one we used), Berria newspaper corpus, EiTB transcriptions, Argia magazine and text from local news sites, totalling 224.6 million word tokens. The resulting model yields a high-dimensional vector (N=300) for each word in the corpus that we used for comparison.

Conclusion

Basque is unique, not just because of the history of the Basque people, but also because the language itself has uncommon linguistic properties that distinguish it from Indo-European languages. I explored comparing concreteness scores of old Basque words with common contemporary Basque words and found a significant difference between them. I also showed that old Basque words cluster in n-dimensional vector space more tightly than contemporary words. More can be done to explore the question of how it is that specific, original Basque words remain spoken today, whereas other words are loaned from foreign languages. Exploring this further may uncover some deeper meanings of old Basque words that could give us hints as to what the history of Basque culture and language has to offer the world.

References

Rodrigo Agerri, Inaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. *Give your text representation models some love: the case for basque*.

Xabier Alberdi, Miren A Villar, Kepa Erdocia, Jon Kortazar, Itziar Laka, Alberto L Basaguren, Julian Maia, Jesus M Makazaga, Ludger Mees, Pello Salaburu, et al. 2012. The challenge of a bilingual society in the Basque country.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. Behav. Res. Methods, 46(3):904–911.

Gorka Elordieta and Magdalena Romera. 2021. The influence of social factors on the prosody of spanish in contact with basque. International Journal of Bilingualism, 25(1):286–317.

Jose Ignacio Hualde. 1995.´ Reconstructing the ancient basque accentual system: Hypotheses and evidence. AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4, pages 171–188.

Hualde, J. I. (2022). The reconstruction of Old Common Basque accentuation: Closing open issues. Anuario del Seminario de Filología Vasca" Julio de Urquijo", 56(2), 77-106.

Luis Michelena. 1995. The ancient basque consonants. AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4, pages 101–136.

Pello Salaburu and Xabier Alberdi. 2012. The search for a common code. The Challenge of a Bilingual Society in the Basque Country, pages 93–112.

Larry Trask. 2008. Etymological dictionary of Basque. University of Sussex.

Larry Trask. 1998. “The Typological Position of Basque: Then and Now.” Language Sciences 20 (3): 313–24.

Jacqueline Urla. 2012. Reclaiming Basque: Language, nation, and cultural activism. University of Nevada Press.