INSTITUTO CARO Y CUERVO

FACULTAD SEMINARIO

ANDRÉS BELLO

MAESTRÍA EN LINGÜÍSTICA

ON THE RELATIONS OF THE DOM IN SPANISH ON TWITTER AND THE WORLD DEVELOPMENT INDICATORS

ANDREA LAUDID GRANADOS DÁVILA

BOGOTÁ

2021

INSTITUTO CARO Y CUERVO

FACULTAD SEMINARIO

ANDRÉS BELLO

MAESTRÍA EN LINGÜÍSTICA

ON THE RELATIONS OF THE DOM IN SPANISH ON TWITTER AND THE WORLD DEVELOPMENT INDICATORS

ANDREA LAUDID GRANADOS DÁVILA

Trabajo de grado para optar por el título de Magister en Lingüística

Directores

SERGIO GONZALO JIMÉNEZ VARGAS

DANIEL EDUARDO CHÁVES PEÑA

BOGOTÁ

2021

BIBLIOTECA JOSÉ MANUEL RIVAS SACCONI

INFORMACION DEL TRABAJO DE GRADO

1. TRABAJO DE GRADO REQUISITO PARA OPTAR AL TÍTULO DE: Magíster en Lingüística

2. TÍTULO DEL TRABAJO DE GRADO: On the Relations of the DOM in Spanish on Twitter and the World Development Indicators (SOBRE LAS RELACIONES ENTRE LA MDO EN ESPAÑOL EN TWITTER Y LOS INDICADORES DE DESARROLLO MUNDIAL)

3. SI AUTORIZO [X] **NO AUTORIZO** [ ]

A la biblioteca José Manuel Rivas Sacconi del Instituto Caro y Cuervo para que con fines académicos:

- Ponga el contenido de este trabajo a disposición de los usuarios en la biblioteca digital Palabra, así como en redes de información del país y del exterior, con las cuales tenga convenio la Facultad Seminario Andrés Bello y el Instituto Caro y Cuervo.
- Permita la consulta a los usuarios interesados en el contenido de este trabajo, para usos de finalidad académica, ya sea formato impreso, CD-ROM o digital desde Internet.
- Socialice la producción intelectual de los egresados de las Maestrías del Instituto Caro y Cuervo con la comunidad académica en general.
- Todos los usos, que tengan finalidad académica; de manera especial la divulgación a través de redes de información académica.

De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, ***"Los derechos morales sobre el trabajo son propiedad de los autores"***, los cuales son irrenunciables, imprescriptibles, inembargables e inalienables. Atendiendo lo anterior, siempre que se consulte la obra, mediante cita bibliográfica se debe dar crédito al trabajo y a su autor.

IDENTIFICACIÓN DEL AUTOR

Nombre completo:	**Documento de Identidad**:
Andrea Laudid Granados Dávila	1014205187

Firma:

DESCRIPCIÓN TRABAJO DE GRADO

AUTOR

Apellidos	Nombres
GRANADOS DÁVILA	ANDREA LAUDID

DIRECTOR (ES)

Apellidos	Nombres
JIMÉNEZ VARGAS CHÁVES PEÑA	SERGIO GONZALO DANIEL EDUARDO

TRABAJO PARA OPTAR POR EL TÍTULO DE: Maestría en Lingüística

TÍTULO DEL TRABAJO DE GRADO: ON THE RELATIONS OF THE DOM IN SPANISH ON TWITTER AND THE WORLD DEVELOPMENT INDICATORS

NOMBRE DEL PROGRAMA ACADÉMICO: Maestría en Lingüística

CIUDAD: Bogotá, 2021 AÑO DE PRESENTACIÓN DEL TRABAJO:

NÚMERO DE PÁGINAS: 61

TIPO DE ILUSTRACIONES: Ilustraciones __ Mapas ___ Retratos ___ Tablas, gráficos y diagramas x Planos ___ Láminas ___ Fotografías ___

MATERIAL ANEXO (Vídeo, audio, multimedia): N/A

Duración del audiovisual: _____ Minutos.

Otro. ¿Cuál? _____

Sistema: Americano NTSC _____ Europeo PAL _____ SECAM _____
Número de archivos dentro del CD, en caso de incluirse un CD-ROM diferente al trabajo de grado: _____

PREMIO O DISTINCIÓN (En caso de ser Laureadas o tener una mención especial): Tesis Laureada

DESCRIPTORES O PALABRAS CLAVES: Son los términos que definen los temas que identifican el contenido. *(En caso de duda para designar estos descriptores, se recomienda consultar a la dirección de biblioteca en el correo electrónico biblioteca@caroycuervo.gov.co):*

ESPAÑOL	**INGLÉS**
Tensión Armónica	Harmonic Tension

Sociolingüística	Sociolinguistics
Indicadores de desarrollo mundiales	World Development Indicators
Migración	Migration

RESUMEN DEL CONTENIDO Español (máximo 250 palabras):

La marcación diferencial de objeto (MDO) en español se realiza a partir del uso de la preposición “a” como marca antes de objetos prototípicamente animados y definidos en oraciones transitivas. Sin embargo, se ha encontrado que existen algunas diferencias en esta realización entre las variedades del español en Hispanoamérica. Estudios presentan como posible explicación a estas diferencias a los niveles de iconicidad de los objetos y el impulso de la economía en el lenguaje: Puentes (2021) tras un estudio cuantitativo basado en un corpus de Twitter demostró que las fuerzas de iconicidad y economía comparten una relación directamente proporcional; este fenómeno fue llamado la Tensión Armónica (T.A.). Se propone como objetivo encontrar posibles relaciones entre la T.A. y factores sociales (Indicadores de desarrollo mundial) con el fin de ampliar el conocimiento de este fenómeno. La metodología propuesta es de tipo observacional y correlacional, por medio de imputación automática de valores faltantes con el fin de permitir una consistencia a través de los años como una forma de encontrar mejores relaciones. Como resultado, los niveles de T.A., basados en un corpus de Twitter, demostraron novedosas e interesantes correlaciones con Producción de cultivos y comida, Inversión en activos no financieros, Escolaridad (índice de paridad de género), Madres Adolescentes, entre otros indicadores. Como principal resultado se afirma una posible conexión entre la producción de comida y cultivos con la T.A. a través de dos factores:1) Migración interna presente en los países de la muestra, así como en 2) una importante participación de la mujer con respecto a la escolaridad y la alfabetización desde la paridad de género.

RESUMEN DEL CONTENIDO Inglés (máximo 250 palabras):

Differential Object Marking (DOM) in Spanish consists of using “a” as a mark before prototypical animated and definite objects in transitive sentences. Nonetheless, it has been found that there exist some differences of this realization among hispano-american Spanish varieties. Studies show that possible reasons for these differences may be the level of iconicity of the objects and the impulse for economy in the language: Puentes (2021) after a quantitative study on Twitter demonstrated that the forces of iconicity and economy shared a direct proportional relation; this phenomenon is referred to as Harmonic Tension (HT). It is set here as objective to find possible relations between HT and social factors (World Development Indicators -WDI) in order to broaden the understanding of this phenomenon. The proposed methodology is observational and correlational, by using automatic imputation of missing values to allow consistency analysis through the years as a way to find better relations. As a result, levels of HT, from a Twitter based corpus, demonstrated novel and interesting relations with *Crop and Food production, Net investment in nonfinancial assets, School Enrollment (GPI), Teenage Mothers*, among other WDIs. As a main finding it was stated a plausible link between crop/food production from two to three decades before Twitter with levels of HT through two factors: 1) internal migration occurring in these countries as well as (2) an important participation of women, literacy rates, and gender parity. Additionally, we report other relationships with HT (not discussed in this study) in the fields of health, tourism, geography, ecology, which open perspectives for interdisciplinary research.

TABLA DE CONTENIDO

LISTA DE TABLAS Y FIGURAS

On the Relations of the DOM in Spanish on Twitter and the World Development Indicators

Abstract

Differential Object Marking (DOM) in Spanish consists of using "a" as a mark before prototypical animated and definite objects in transitive sentences. Nonetheless, it has been found that there exist some differences of this realization among hispano-american Spanish varieties. Studies show that possible reasons for these differences may be the level of iconicity of the objects and the impulse for economy in the language: Puentes (2021) after a quantitative study on Twitter demonstrated that the forces of iconicity and economy shared a direct proportional relation; this phenomenon is referred to as Harmonic Tension (HT). It is set here as objective to find possible relations between HT and social factors (World Development Indicators -WDI) in order to broaden the understanding of this phenomenon. The proposed methodology is observational and correlational, by using automatic imputation of missing values to allow consistency analysis through the years as a way to find better relations. As a result, levels of HT, from a Twitter based corpus, demonstrated novel and interesting relations with *Crop and Food production, Net investment in nonfinancial assets, School Enrollment (GPI), Teenage Mothers*, among other WDIs. As a main finding it was stated a plausible link between crop/food production from two to three decades before Twitter with levels of HT through two factors: 1) internal migration occurring in these countries as well as (2) an important participation of women, literacy rates, and gender parity. Additionally, we report other relationships with HT (not discussed in this study) in the fields of health, tourism, geography, ecology, which open perspectives for interdisciplinary research.

Keywords: Harmonic Tension, Sociolinguistics, World Development Indicators, Differential Object Marking

1. INTRODUCTION

From a sociolinguistics perspective, it is stated that some language features tend to vary more than others and these are strongly connected to social factors; in other words, the way people talk can be shaped according to their social status, age, and even gender (Hickey, 2007). These variations, if accepted by a large number of individuals, tend to generate a follow-up change in language. In the Spanish used in America, it was discovered that there is not a unifying set rule regarding when to use a Differential Object Marking (henceforth DOM). Unfortunately, there exists very little study on the extension of DOM varieties, especially from a geographical perspective (Fábregas, 2013). For example:

1. a. Ve una pareja.

He/She.sees a couple

'He/she sees a couple'

b. Ve a una pareja.

He/She.sees A a couple

'He/she sees a couple'

Even though theory says that the kind of statements found in (1) should carry the mark, empirical results showed a high level of hesitation (Puentes, 2021), then it has been proved that the percentage of cases of omission and use of differential marking fluctuates not only among Spanish-speaking countries but within them as demonstrated in *Quantitative analysis of DOM in Spanish: Iconicity and economy forces driven by a harmonic tension* whose results data will be used in the present study[1]. Another subsequent finding from this work showed that

[1] Regarding the data utilized here, it is important to mention that the findings resulting from Puente's study were based on a corpus made up of 217'928.418 tweets from about 5 million users residing in 333 different cities in 21 Spanish-speaking countries collected by Jimenez, Dueñas, Gelbukh, Rodriguez-Diaz and Mancera (2018) during the period 2009 to 2016.

there is a connection between the variable of omission (economy) and that of the marking (iconicity), that is, that there is a direct proportional relationship that connects them both. This force, that makes both features increase or decrease proportionally, is called “Harmonic Tension " (HT). In that study some of the countries showed a higher level of HT while others demonstrated lower levels; and s, the high correlation between iconicity and economy demonstrates the existence of a pattern.

DOM has been explained through the linguistic features of verbs and objects (Aissen, 2013. Fábregas, 2003, Balasch, 2011, Puentes, 2021, among others), but an explanation for the important variations of the Harmonic Tension among Hispano-American countries is still lacking. For instance, Mexico and Guatemala showed a low intensity of HT, this demonstrates that the speakers in these countries have a clear understanding about which verbs/objects should carry the differential mark, as if the rules were well settled, giving little room for hesitation. On the contrary, countries like Dominican Republic and Spanish speakers from the U.S.A. showed a higher intensity of HT, which means that either speakers hesitate about when to use the mark or there is a greater acceptance towards an optional use in more cases. Considering the complex nature of HT and its recent discovery, there is still no linguistic explanation to these variations and the factors that may be involved in them. Based on this, it is here intended to answer the question about what are those possible social factors that are related to the level of HT in these American countries since it could provide valuable information upon sociolinguistic behavior and the linguistic phenomenon itself. Such a task provides complementary data to the already known influential factors (i.e. phonological influence, frequency, iconicity, markedness, grammaticalization, etc. (Rohdenburg & Mondorf, 2003)), which could tell us more about linguistic variation’s nature and it can even be extrapolated to other linguistic levels.

To fulfill this goal, the methodology was based on an observational, correlational and factorial scope as a way to find more significant relations. Some results showed that HT demonstrated possible relations with Crop and Food production, Net investment in nonfinancial assets, Adolescent Fertility, School Enrollment (GPI), among others. As a main finding it was stated a highly possible incidence of internal migration and women participation in these countries with the levels of HT.

2. BACKGROUND

As aforementioned, for many academic fields, especially humanities, social media has become an important source of information. In the case of linguistics, it is well known these platforms collect a vast amount of data related to language use, change, variety, etc. As an example, Abitbol *et al*. (2018), carried a study based on tweets from people all over France. As their objective was to find correlations between language cues and social factors, they constructed a dataset combining the largest French Twitter corpus and socio-economic maps from the national census. The linguistic variables were, in the first place, the standard usage of the basic form of negation in French: *ne* and *pas*, in a written context. By calculating its rate, the authors proved the *ne* realization was more frequent among high status speakers, while low status speakers are more prone to omit it. Second, since plural marks are mute in French (-s,-x), their omission is a common spelling also associated with a low social status. Finally, a high level of lexical diversity would indicate a prominent socioeconomic status and a higher educational level of speakers. In short, the results show that:

> "(i) people of higher socio-economic status, active to a greater degree during the daytime, use a more standard language; (ii) the southern part of the country is more prone to use more standard language than the northern one, while locally the used variety

or dialect is determined by the spatial distribution of socioeconomic status; and (iii) individuals connected in the social network are closer linguistically than disconnected ones, even after the effects of status homophily have been removed." (Abitbol et. al, 2018, p. 2).

Table 1The R coefficient of determination and the corresponding p-values computed for the pairwise correlations of SES indicators and linguistic variables.

	S^i_{inc}	S^i_{own}	S^i_{den}
$\overline{L}_{\text{cn}}$	$0.19\ (p < 10^{-2})$	$0.59\ (p < 10^{-2})$	$0.74\ (p < 10^{-2})$
$\overline{L}_{\text{cp}}$	$0.59\ (p < 10^{-2})$	$0.66\ (p < 10^{-2})$	$0.76\ (p < 10^{-2})$
$\overline{L}_{\text{vs}}$	$0.70\ (p < 10^{-2})$	$0.32\ (p < 10^{-2})$	$0.41\ (p < 10^{-2})$

Note: inc: Income; own: Ownership; den: density. cn: Common negation; cp: common plural; vs: vocabulary set. (Abitbol et. al, 2018, p. 6).

As shown in fig.1 the correlations between socioeconomic indicators and linguistic variables are statistically significant ($p < 10^{-2}$) suggesting that people with lower socio-economic status are more prone to use non-standard expressions.

Now, the study of large-scale databases can even be extrapolated to fields such as psychology and even medicine. For instance, Eichstaedt *et al*. (2015) and a group of members from the Department of Psychology, Computer and Information Science, School of Education, and the School of medicine approved by the University of Pennsylvania Institutional Review Board developed an investigation based on the language used by Twitter users in the U.S.A. They were aimed to create a model that could predict the risk of having Atherosclerotic Heart Disease (AHD) being hostility and chronic stress the main known risk factors. In order to extract the data, they implemented a high-frequency process for phrases and words; afterward, the types

of language use were divided into a) dictionaries and b) topics. Dictionaries would classify the variables into positive or negative-emotion words (to avoid contextual nuances human raters were in charge of evaluating the tweets). Topic-based variables are explained as semantically word-related clusters. Both variables, dictionary and topic language, were correlated to AHD mortality rates taking education and income as controls. As a result, it was shown that anger, negative-relationship, negative-emotion, and disengagement words were significantly correlated with AHD age-adjusted mortality as well as themes related to hostility and aggression, hate and interpersonal tension, boredom and fatigue. On the other hand, as well as topics regarding positive experiences, skilled occupations, and optimism (Eichstaedt, *et al*., 2015). This model was compared with previous ones (such as phone surveys and household visits) based on demographic, socioeconomic, and health risk factors and it demonstrated a more accurate performance at predicting AHD mortality.

On a more linguistic level, similar to Abitbol's *et al*. (2018) investigation, Nigel Armstrong and Alan Smith (2001) from the University of Leeds developed a study in order to record the progressive decline of the use of *ne* in French and, most importantly, the linguistic and social factors related to this tendency. To carry out this task, two types of corpora were analyzed: the archival corpora and a contemporary corpus. The first included two recorded radio programs broadcasted during 1960 to 1961 (referred by the authors as Agren), in which topics related to politics, economy, sports, etc. were discussed as well as topics addressed to a young audience. The second, the contemporary corpus, is built from a public radio program which also draws discussion over current affairs with little audience participation (over 20 hours of speech) called TS after the name of the program (*Le Téléphone sonne*). In both cases non-spontaneous speech was not taken into account.

After applying the chi-square test and obtaining a significant p value, the authors stated that the percentage of *ne* realisation in Agren was 92.6% in contrast to a 72.5% in TS corpus (for

more detail on statistics data https://eprints.whiterose.ac.uk/1006/1/armstrongn1.pdf). With regard to related social factors, Smith (1996 cited in Armstrong & Smith, 2001) explained that during late 60's France was experiencing a significant symbolic social rupture switching from a hierarchical to a more collaborative social system and even today the country has changed to a more open attitude towards middle and low status communities making the division lines to blur. Additionally, during and after post-war reconstruction, young people have been widely recognized as an important group of consumers. According to the authors, this attitude is reflected in language when agents known to use a standard and prestigious register as radio hosts accept and use structures otherwise considered as informal or colloquial (Armstrong & Smith, 2001). This phenomenon has more recently been studied as a characteristic feature of European French called *levelling* consisting of a series of linguistic attitudes addressed towards the acceptance of change and the decline of deference as a way to welcome individual worth and young people value (Armstrong & Pooley, 2010)

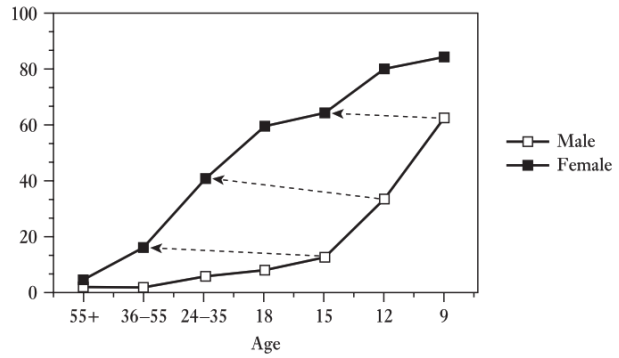
Now, taking into account that one of the main topics studied in Sociolinguistics is related to linguistic variation, migration has been a social factor with a high incidence in this phenomenon. In this sense, Cécile B, Vigouroux's work on globalization, migration and language vitality (2008) is an example of how migration has shed light upon language spread and linguistic change. After researching upon the population of Black African Francophones migrating towards Cape Town, South Africa she recognized an adaptive tendency from speakers in search of engaging into new social dynamics; more specifically, speakers "switched to Cape Town's dominant language [...] suggesting some social stratification among the migrants based on the time of settlement" (Vigouroux, 2008, p.248).

Finally, being women's participation also a related factor with the results here presented, Wolf and Jiménez (1979, cited in Labov, 2010) concluded that, after having registered the devoicing use of /dz/ in the Spanish of Buenos Aires (as in *calle* or *llama*), women showed to be ahead in

the process by a full generation (Fig. 1)

Figure 1

Gender differentiation in 2006 by age of a simulated sound change beginning in 1942, at four-year intervals



Note. "El ensordecimienteo del yeismo porteño, un cambio fonológico en marcha" by Wolf and Jiménez (1979, cited in Labov, 2010). Dashed arrows in the figure indicate agreement between male values and the last maternal generation which demonstrates how women lead the process having an almost linear increment through different ages while men are behind by a complete generation. In terms of Labov (2010), this is a clear example of the predominance of women in changes from below which most of the times would end up retarding or even eliminating male change since future generations would acquire minimized versions from mothers.

THEORETICAL FRAMEWORK

a. DOM in Spanish on Twitter

Linguistic systems have different ways to distinguish who or what is performing the action in a statement and whom or what it is acted upon, in other words, there exists different types of coding events in which a voluntary action, usually human-performed targets an affected

patient", well known as transitivity (Kittilä, 2011, p.346). The definition of transitivity depends on the approach one may follow: on the one hand, from a formal perspective, it can be said that transitivity is a dual system in which, depending on the features of verbs, these can be classified into transitive or intransitive, in other words, verbs are divided by the number of arguments they need to construct a complete statement (Kittilä, 2011), also known as verb valency (Liu, 2011)

Under this approach, transitive verbs refer to those in which two arguments take place in the statement (e.g. eat, give keeping in mind there is not always a syntactic object (2.a) but it can be inferred that something has been eaten), disregarding the semantic differences that appear in some constructions when using verbs like "meet" in which the relation "agent- patient" is not quite accurate, as shown in (2.b) On the other hand, from a semantic approach, transitivity is understood as a continuum from intransitivity to transitivity that includes a range of different levels of connection among arguments at syntactical level (Kittilä, 2011).

(2) a. Ian eats at 8:00 am.

b. Lina met her new boss yesterday.

From a typological view, some languages may use the order of the syntactic elements to express transitivity where the agent usually comes first (English, for instance). Other languages even have a noun hierarchy that tells that, no matter in which order you may present the participants, the agent or the patient is given by this hierarchical system (e.g. Sikuani and Algonquian languages) (Kittilä, 2011); and others, such as Spanish, Hebrew and Romanian, have a Differential Object Marking system in which some objects are marked according to their inner features.

In simple terms, DOM in Spanish divides direct objects in two groups, one receives a mark, the other does not (as in 3). Some features of marked objects are related to their subject-like

properties (3.a.), such as, potential agency, animacy, active role in the statement, and so on (Fábregas, 2013). Whereas, recent studies have shown a high level of variation among Spanish-speaking countries given that what seems prescriptively correct in some regions is considered unacceptable in others (Fábregas, 2013). Another alternative feature that influences the probability of objects being differentially marked is related to the level of iconicity attributed to the Direct Object and the principle of economy as presented below.

(3) a. Juan ama a su mamá

Juan loves A his mother

'Juan loves his mother'

b. Juan odia la playa

Juan hates the beach

'Juan hates the beach'

b. Iconicity and Economy

As a general rule (not yet absolute), languages tend to mark those elements that are not prototypical or to differentiate them from others. In phonetics, for instance, nasality and voicing are marking mechanisms that may influence meaning. Similarly, in structural terms, it was found that in most languages negative forms are always marked in contrast to the affirmative forms; the same happens with singular in contrast to plural forms, present and past tense, etc. All in all, *grosso modo*, frequent and most common elements do not need to be marked, but those out of the prototypical structure do (Bybee, 2011).

Prototypicality should be understood, not only in terms of the inner characteristics of the element, but in respect to what is expected of a syntactic construction. For instance, in the case of transitive sentences, an agent (commonly animated and definite) is expected to perform an

action over an object (commonly, inanimate, etc.) (Bybee, 2011); however, there is a wide range of relations between agents and objects.

Mayerthaler (1987, cited in Bybee, 2011) stated that "the more accessible an entity is to the speaker and the more it resembles non-biological properties of the speaker, the less marked it is." (p.141). This is complemented by the thesis presented by Aissen (2003) who affirms that, in transitive constructions, the element in the position of "object" has more chances of being marked by having more prototypical characteristics of a subject (more iconic) as a means to differentiate it from the actual subject. The principle of iconicity operates in these situations in order to disambiguate the grammatical functions of subject and object. Needless to say, that this phenomenon varies among languages, but a general idea relies on the level of prominence, closely related to the level of animacy and definiteness as shown in the following scales:

a. Animacy scale: Human > Animate > Inanimate

b. Definiteness scale: Personal pronoun > Proper name > Definite NP > Indefinite specific NP > Non-specific NP (Aissen, 2003, p. 437)

In contrast, the principle of economy of language aims to maintain a balance between the characteristics that ensure efficient and direct communication on the one hand, and the natural need for the least effort, on the other (Vicentini, 2003). This principle leads to avoid differential marking but, as formulated below, these two forces, iconicity and economy, exert influence on the phenomenon in Spanish varieties at different levels.

c. Harmonic Tension

Puentes (2021), in her study *Quantitative analysis of DOM in Spanish: Iconicity and economy forces driven by a harmonic tension,* explained how DOM takes place in different Spanish varieties and how the forces of *iconicity* and *economy* influence the phenomenon. Through a

statistical analysis of a corpus of 218 million tweets, it was found that iconicity and economy are not independent but, on contrary, they are highly correlated. After having identified iconicity with the tendency to mark direct objects and economy with its absence, the statistical analysis presented by the author showed a pattern of relationship between these two measures across the spanish speaking countries. This close relationship is called *The Harmonic Tension (HT)*. Specifically, it was found that the correlation between the values for economy and iconicity was r=0.92, significantly above the critical value with a statistical significance of p<0,01. This phenomenon indicates that in those varieties in which a higher Harmonic Tension is placed, the greater the possibility of having an intermediate percentage between economy and iconicity; in other words, there would be fewer cases of prominent omission or markedness of DOM. Accordingly, in those varieties in which a weak Harmonic Tension takes over, the probability of having prevalence over either markedness or omission will be greater as shown in Table 2 (based on Puentes, 2021).

Table 2

Transcription of the Harmonic Tension values for Spanish Speaking countries

Country	Code	Iconicity α	Economy β	Harmonic Tension α + β
Peru	PER	0.0227	0.1119	0.1346
Mexico	MEX	0.0232	0.1132	0.1364
Costa Rica	CRI	0.0217	0.1167	0.1384
Guatemala	GTM	0.0240	0.1155	0.1395
Spain	ESP	0.0239	0.1205	0.1444
Argentina	ARG	0.0228	0.1250	0.1478
Chile	CHL	0.0235	0.1255	0.1490
Panama	PAN	0.0255	0.1273	0.1528
Paraguay	PRY	0.0260	0.1282	0.1542
Uruguay	URY	0.0261	0.1440	0.1701
Colombia	COL	0.0259	0.1452	0.1711
Venezuela	VEN	0.0275	0.1448	0.1723
Ecuador	ECU	0.0292	0.1450	0.1742
Nicaragua/El Salvador/Honduras	NSH	0.0295	0.1494	0.1789
Puerto Rico	PRI	0.0281	0.1528	0.1809
Dominican Republic	DOM	0.0293	0.1655	0.1948
United States	USA	0.0325	0.1702	0.2027

Note. Based on Análisis cuantitativo de la MDO en el español: tensión armónica entre iconicidad y economía (Puentes, 2021).

As it was stated before, the sample used included different Spanish varieties, specifically from 19 countries. The Spanish speakers from the USA and the Dominican Republic, for instance, showed a stronger correlation between iconicity and economy, that is to say, a stronger Harmonic Tension. On the other hand, Peru, Mexico, and Guatemala resulted in weaker tension. The countries of Cuba and Bolivia were not included in the Puentes' study because of the scarcity of their Twitter data. Figure 2 shows a visualization of the information from Table 1 in a map.

Figure 2

Map visualization of the levels of harmonic tension in Latin America (dark purple means high HT, light purple low HT)

Low HT High HT

Having in mind that social factors have great influence on language variety, the aim in this paper is to search for plausible correlations between social indicators and HT that could help to better understand this phenomenon.

d. Sociolinguistics

In general terms, sociolinguistics is well known for its diametrical opposition to the notion of "speaker's intuition" as a source of data and to the structuralism perspective which handles languages isolated from their social context. Conversely, sociolinguistics places itself in the

performance dimension as it considers languages to have a communicative and social purpose (Silva-Corvalán, 2001). Bearing this in mind, this discipline is aimed at solving issues related to "correctness", language and dialect opposition, language change, identity and, most importantly, linguistic variation.

This kind of scope brings with it a series of features to keep in mind taking into account that social aspects are not fixed, but dynamic. One of the features sociolinguists have to deal with is related to the level of agency speakers have. The agency allows speakers to diverge from conventional language patterns that, if accepted by an important number of speakers, would lead to language change expressed by lexical choice, phonological variants, grammatical patterns, etc. Additionally to agency, other aspects like demographic variables such as age and gender would determine, for example, in which context one variable is more likely to be used (Nguyen, Doğruöz, Rosé, de Jong, 2016).

On this regard, sociolinguistics has used the Theory of Variation to fulfill the purpose of proving that language variation is not by all means, a matter of luck but, it is rather conditioned by internal features of the language itself as well as by external social factors (Silva-Corvalán, 2001). Even though social factors relevance and influence differs from one community from another, generally speaking sex, age, education level, social status and ethnicity are the ones which have shown higher influence in linguistic variation (Moreno, 2009). This type of insight may provide this study of important elements to analyze observable data and its relations.

3. METHODOLOGY

This study will be of observational, correlational and factorial kind. It is observational taking into account that the variables to be related (HT and demographics) occurred in the past, naturally and without intervention from the researchers. Second, it is a correlational study because its objective relies on finding possible relations between HT and demographic factors

in order to broaden the understanding of HT. Finally, it is factorial in the sense of taking HT as a base for comparison against a large group of demographic indicators. This factorial aspect contrasts with traditional correlational studies where the variable to be studied (that is, HT) is compared with one or a small group of demographic variables whose selection has a clear predetermined motivation. In this scenario, the objective is to test the binary hypothesis of whether there is a relationship between the variables. In our case, there is no predetermined motivation to select demographic variables given the novelty of the HT results. Thus, this factorial study seeks to first discover the possible relationships of HT with a large group of demographic variables, then contextualize them, and analyze the plausibility of the relationships discovered. Additionally, , for some factors we interpret their influence and possible direction of causality with HT. Finally, data (Corpus:Harmonic Tension:World Demographic Indexes) will be triangulated in order to confirm results are not based on co-founding factors (*cf. Data Triangulation*). Information regarding the nature of the data, analysis, methodological challenges and data interpretation will be extended next (*cf. Results*).

a. The World Bank's World Development Indicators (WDI)

World Development Indicators is an open access database collection that includes over 312 time-series indicators for almost 220 economies and more than 45 country groups for a period of time of 61 years (from 1960 to 2020). This information is compiled from official international sources and includes national, regional and global estimates. Some themes found in this database are related to poverty, inequality, health, environment, education, economy, global links, among others. In methodological terms, it is possible to access all the data and can be downloaded[2] in MS-Excel format (The World Bank, 2021).

b. Sample description

[2] https://datatopics.worldbank.org/world-development-indicators/

Due to the diversity of indicators registered in the World Bank database, some researchers have opted to narrow down the number of indicators used in their studies. If a solid criterion is followed, it may facilitate the correlation discovery process; however, it can also be an important limitation since it obstructs the emergence of possible new findings or it may result in few or none significant results. As an example, Rahman *et al*. (2019) used a very similar sample size to the one used here, that is 17 countries, to find correlations with the outcomes of congenital heart surgery but only 8 of the WDI indicators available were considered. As a consequence, through the whole study only one correlation was found barely surpassing the critical value of statistical significance $p<0.05$ (*i.e*. Under-five mortality rate). Bearing this in mind, all WDI consistent in time (i.e. from 1960 to 2010) were considered in the present study as a way to lessen any possible bias of manual selection that may limit the study.

On the other hand, as mentioned before, Harmonic Tension (second variable) was measured over 218-million tweets from 19 Spanish-speaking countries (Puentes,2021). However, in order to have a more homogeneous sample, the information from USA and Spain may not be included keeping in mind their, socially speaking, differences with the rest of Latin American countries. Also, HT data reported a single value for Honduras, Nicaragua and El Salvador due to the small size of their corpora (labeled as NSH); this group will be kept in this study.

c. Statistical Analysis

The analysis consists of a correlational approach in search of the tendency between two variables, in which it is possible to identify direct and indirect relations among the data. There exist different coefficients to measure correlations, Pearson, Spearman and Kendall's coefficients being the most common ones. Pearson's, on the one hand, is more suitable for lineal tendencies, in cases with almost none atypical values, highly recommended for numeral data and when a big size sample is at disposal, but also it is assumed the normality of the variables.

All these conditions are difficult to fulfill with the World Bank data since there is no way a normal distribution of almost 312 indicators could be guaranteed, outliers are common, and n=17 countries is a rather small sample. Consequently, Spearman's coefficient is more suitable for our purpose since the relation of the variables is not necessary to be linear, and it is more robust with the presence of outliers (Mangiafico, 2016). Therefore, it is here considered that Spearman's correlation (denoted as ρ or "rho" in this study) would provide a more precise result since it is able to identify any monotonic tendency between rankings instead of using the actual data. Another consideration related to n is that we decided to strengthen the level of statistical significance to p<0.01 (instead of the traditional p<0.05) to avoid the analysis of correlations produced by effects of chance. Thus, we will only consider correlations above the critical value of $\rho = 0.57$[3] ($p < 0.01$, $n = 17$) plus a safety factor of 0.04 (which is justified later).

As stated before, all indicators were considered in order to be open to any possible relation that may have come out; both correlational and significance values were examined as a way to strengthen statistical reliability. Besides, even though data includes most part of Spanish speaking territory, 17 countries can be considered a relatively small sample from a statistical point of view (Elementary Statistics and Computer Application, 2012), accordingly Spearman's coefficient will be implemented as it adjusts better to this type of data.

Finally, although the World Bank initiative is probably the most complete demographic data in

[3] See https://www.york.ac.uk/depts/maths/tables/spearman.pdf

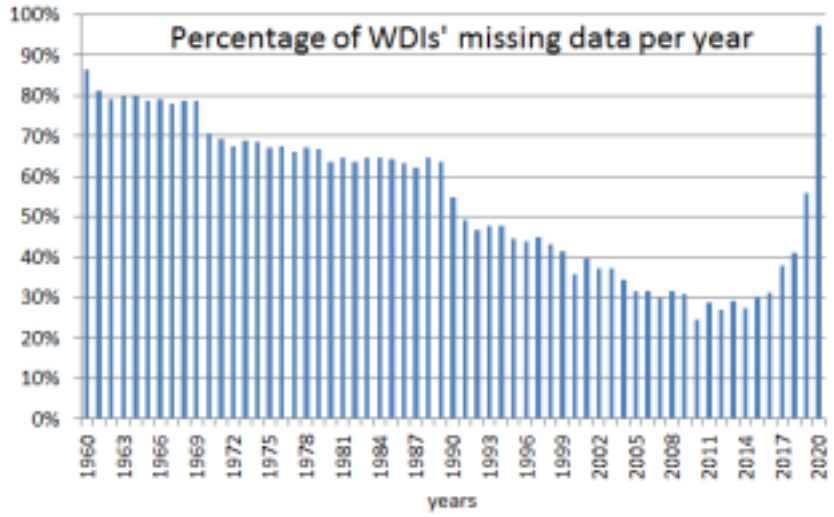
coverage of time and number of countries, it has a considerable amount of missing data[4][5], for what it was necessary to use methods for data imputation to complete the fields with reliable predictions as, otherwise, it would not be possible to find consistent correlations over time. Figure 3 shows the distribution of missing data over the years and how they have been decreasing in recent years. Figure 4 shows the same distribution for the countries indicating a relatively uniform missing data rate of around 50% with the exception of Puerto Rico. In this sense, *k*-Nearest Neighbors (KNN) (Fix & Hodge, 1989), Random Forest (Breiman, 2001) and Bayesian Ridge (Bishop, 2006) methods have demonstrated being robust in cases of missing values as proved by Taylor *et al*. (2017), who after imputing up to 50% of the entries in different biological datasets it resulted that the effect on the observed correlations were systematically lower in contrast with the ones found with the complete datasets. This result means that there is a low possibility (0.01) that after using these methods, false high correlations would come out by chance.

Figure 3

Distribution of the missing values in WDIs per year

[4] The World Bank explains the missing data as a result of: "Some indicators are derived from sporadic surveys and are only available every few years. Some data sets or indicators are only available from the year they were initiated [...] Some countries do not regularly report data due to conflict, lack of statistical capacity, or other reasons [...] And some countries do not have data for earlier years simply because they did not exist. (The World Bank, 2021).

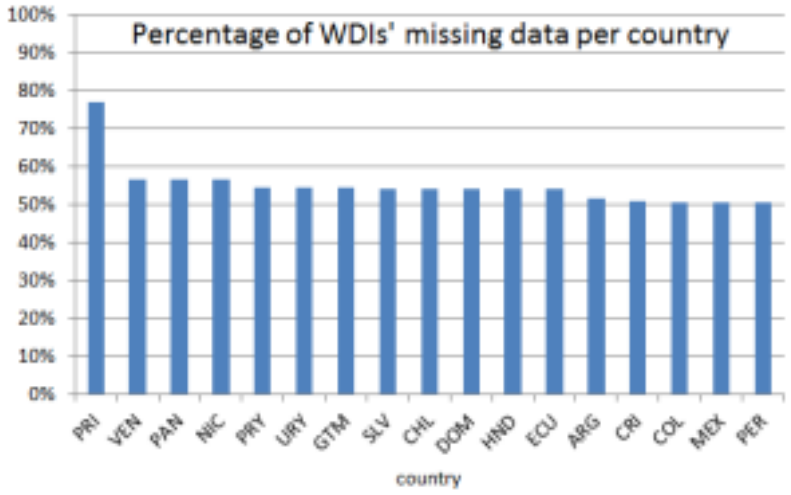
[5] Our data matrix has 312 columns for WDIs by 1,037 rows (17 countries in 61 years from 1960 to 2020), of which 54.9% are missing values. The countries with the larger number of missing values are ARG and CHL



.

Figure 4

Distribution of the missing values in WDIs per country



Note: PRI: Puerto Rico; VEN: Venezuela; PAN: Panama; NIC: Nicaragua; PRY: Paraguay; URY: Uruguay; GTM: Guatemala; SLV: Salvador; CHL: Chile; DOM: Dominican Republic; HND: Honduras; ECU: Ecuador; ARG: Argentina; CRI: Costa Rica; COL: Colombia; MEX: Mexico and PER: Perú.

In order to test this approach, a random 5% of the non-missing entries (7,419) in the data matrix were marked as missing and their true values were saved for error assessment. Several imputation methods were applied on that matrix and the imputed values on the 7,419 were

compared with their true values using the Symmetric Mean Absolute Percentage Error[6] (SMAPE). SMAPE is convenient in our scenario because it measures error as percentages dealing adequately with the variety of numeric ranges of the WDIs. The table 2 shows the results for several company used methods for data imputation, highlighting the Random Forest regression method with 1 iteration as the best option (Tang & Ishwaran, 2017). This method has shown to be more sophisticated and effective in contrast to other strategies like replacing missing data with 0 or with the mean or the median, which could lead to inaccurate results. By employing imputation methods, it will be possible to identify the correlations, not only in terms of high level of correlation and statistical significance, but also by their consistency through time, that is, during the 61 years registered. These experiments were carried out using the imputer implementations of the Python programming language package Sklearn (Pedregosa *et al*., 2011). The value of sMAPE=0.145 can be interpreted as an average error rate of 7.25% for the imputed values (i.e. the half of sMAPE expressed as percentage), which we considered acceptable for this study.

As a means to confirm this acceptability, 10,000 data series sized n= 17 were randomly generated obtaining their correlations with HT. Concurrently, an error of 7.25% was randomly introduced in 54.9% of the data (the same percentage of missing values in WDI data) resulting in a second set 10,000 of data series. Again, we measured correlations with HT. Accordingly, both sets of correlation results were compared, showing a difference in Mean Absolute Error (MAE) of 0.018. To avoid the effects of this error in our analysis, we add a safety factor of 0.040, which is greater than this error, to the threshold of statistical significance 0.57. Thus, in this study, we only consider correlations greater than 0.61. Additionally, each of the 10,000 correlation pairs were compared to verify the significance of their differences. Using Zou's test

[6] See https://en.wikipedia.org/wiki/Symmetric_mean_absolute_percentage_error

(2017), it was observed that 20.1% of them did not show any significant difference (p<0.05)[7]. We can conclude that the effect on the remaining 79.1% of significant differences in correlations is reasonably controlled with the increased
correlation threshold.

Then we consider that the threshold of 0.61 is safe to observe highly significant correlations, and without affectation due to the imputation of missing data. As a result, the estimated error of the missing data imputation (7.25%), the low MAE obtained (0.018) in the test, ant the safety margin (0.040) above of the critical value (0.57), importantly support that the data imputation won't result in false correlations.

However, it is important to mention that some WDIs with more that 50% of missing data could be affected, given that the correlations would be obtained mainly from the predictions of the Random Forest method instead of real data. Consequently, it is expected that, in the case of indicators with a big amount or even total missing data, Random Forest method will generate similar values to the ones shown in other years for the same indicator. For instance, HIV related indicators during the 60's and 70's imputed data show similar correlations to the ones from the 80's and 90's.

[7] The test is available in the following Google Colab Notebook
https://colab.research.google.com/drive/1bNHYReI5GW51LgyfROuTCETbpoE5CKcZ?usp=sharing

Table 3 Estimates of error rates of different imputing methods for WDIs' missing values.

Imputing method	sMAPE	Imputing method	sMAPE
constant value (zero)	2.000	KNN *k*=40	0.491
constant value (1)	1.705	Bayesian Ridge 1 iterations	0.426
mean	0.648	Bayesian Ridge 5 iterations	0.383
median	0.573	Bayesian Ridge 10 iterations	0.394
KNN *k*=1	0.602	**Random Forest 1 iteration***	**0.145**
KNN *k*=5	0.503	Random Forest 2 iterations	0.148
KNN *k*=20	0.489	Random Forest 3 iterations	0.147

In short, this methodology proposed a correlational study that compensates the small sample restriction (*n*=17) by means of four different strategies: 1) To reduce researcher's bias by covering all demographic indicators available; 2) To decrease the classical *p* value threshold from 0.05 to 0.01 to reduce the risk of analyzing correlations resulting by chance; 3) by imputing missing data and adding a safety margin in the threshold to cope with the induced error; and 4) regarding time consistency, it is expected to analyze only those WDIs correlated with HT over a relatively long period of years and ignoring those that are sporadic. The combination of these strategies contributed to more consistent results by preventing the analysis of false, random or spurious relations.

5. RESULTS

After applying the method to measure correlations presented in the previous section, it was found that of the 19,032 series of data that were compared against HT (312 WDIs multiplied by 61 years), 249 correlations revealed to be greater than 0.61 Spearman's value. These correlations corresponded to 60 WDIs of which 26 were positively correlated (directly proportional) and 34 negatively (indirectly proportional). For all these WDIs obtained that were correlated in more than one year, all the correlations were consistently positive or negative.

Once these indicators were arranged by the number of years in which they were correlated with HT, it was here considered as necessary to set a minimum threshold of 5 years as a sign of an adequate consistency over time of the correlation. With this in mind, Tables 3 and 4 show the WDIs with correlations that were present in more than 5 years (*N*=number of years in which the correlation was found) and values above 0.61. Table 3 includes indicators that show a direct correlation against HT while Table 4 indicates inverse or negative correlations with the linguistic phenomenon. In these tables the column labeled "years-span" contains the farthest and closest year in which the WDI was correlated. Similarly, "rho-span" indicates the maximum and minimum correlation observed in the *N* years.

Table 4 Most frequent WDIs correlated directly against HT (p<0.01)

Indicator DIRECT w.r.t the Harmonic Tension	*N**	years span	rho span
Crop production index (2004-2006 = 100)	25	1961:1986	0.71:0.61
Net investment in nonfinancial assets (% of GDP)	12	1988:2007	0.74:0.65
Vulnerable employment, female (% of female employment)	9	1968:1985	0.70:0.62
Land area where elevation is below 5 meters (% of total land area)	9	1965:1989	0.73:0.61
Intentional homicides (per 100,000 people)	9	1969:1984	0.68:0.62
School enrollment, primary (gross), gender parity index (GPI)	5	1980:1985	0.71:0.61
School enrollment, primary and secondary (gross), gender parity index (GPI)**	3	1998:2002	0.67:0.61
Domestic credit provided by financial sector (% of GDP)	7	1967:1977	0.71:0.61
Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)	7	1969:2018	0.63:0.61
Food production index (2004-2006 = 100)	6	1971:1979	0.65:0.61
Prevalence of HIV, male (% ages 15-24)	5	1962:1989	0.81:0.61

* N is the number of years the WDI is correlated significantly against the HT.

**This indicator does not exceed the threshold of N>5 but is included due to its connection with the *School enrollment primary, Gender Parity index.*

Table 5 Most frequent WDIs correlated inversely against the HT (p<0.01)

Indicator INVERSE w.r.t the Harmonic Tension	*N**	years span	rho span
Depth of credit information index (0=low to 8=high)	16	1961:1986	0.80:0.61
Literacy rate, youth male (% of males ages 15-24)	15	1977:2018	0.72:0.61
Literacy rate, adult male (% of males ages 15 and above)	8	1977:2018	0.70:0.61
School enrollment, primary (% net)	8	2004:2018	0.72:0.61
International tourism, expenditures (% of total imports)	8	1971:1984	0.69:0.61
Mammal species, threatened	6	1964:2018	0.67:0.61
External debt stocks, private nonguaranteed (PNG)	6	2014:2019	0.67:0.61
Contributing family workers, male (% of male employment)	6	1972:1979	0.68:0.61
Investment in water and sanitation with private participation	5	1965:1973	0.69:0.61

* *N* is the number of years the WDI is correlated significantly against the HT.

Note. Some names of the indicators in Tables 3 and 4 are self-explanatory, but others are less so. Therefore, in the Annex we include the expanded descriptions of these WDIs provided by the World Bank.

As observed, indicators show a great variety of topics which can be classified mainly as economy related (such as Crop Production, Net Investment, Domestic Credit, Depth of credit, External debt stocks, etc.), social issues (like Vulnerable Employment for Women, Intentional Homicides, Teenage Mothers, Contributing Family Workers, etc.) and educational matters (School enrollment and Literacy rate), among others. Needless to say that some of these correlations may seem far-fetched due to, on one hand, some indicators were correlated over an important percentage of imputed data as it could be seen in *Prevalence of HIV* where results were obtained from decades in which the disease had not been discovered due to the missing data replaced with the one from following years; or, on the other hand, if there is a relation, it is not yet evident from our field of expertise or may be produced by an unknown confounding factor (i.e. Land area where elevation is below 5 meters, Prevalence of HIV, Intentional homicides, and Mammal Species Threatened, which also show few years of consistency). One important example of a possible confounding factor involved is present in the indicator *External debt stocks, private nonguaranteed* which, despite having time consistency, the reasons for which private subjects acquire credits at national and international level are hard to spot on;

additionally, some countries do not report data on this matter for having outstanding debt with the World Bank, other financial entities or private creditors (WB, 2021).

Another aspect to consider is related to the time span in which the correlations are distributed. For example, Crop Production Index's results range from 1961 and 1986, that is, one correlation for each year in the time span demonstrating consistency and continuity. In contrast, *Mammal species threatened* Index's results were distributed in a longer period of time, that is, 6 years from 1964 to 2018 which may be obstructive for a possible linguistic association. Under these criteria, from the economic factors *Crop production*, *Food production* and *Net investment* (this one in lower levels) show to have a better time consistency (See Figure 5) as well as *School enrollment* (both primary and primary and secondary GPI) (Figure 6) along with *Literacy rate* (both in adults and young male in contrast to female) (Figure 7), and *Contributing family workers*.

Additionally, it was taken into account the fact that some results were placed on a different time interval from the one in which HT corpus was registered (2009-2016) while others coincide. This can provide information regarding the characteristics of the population involved and of the development of the social processes going on.

Figure 4 Correlations over the years of HT with the most correlated and consistent direct factors

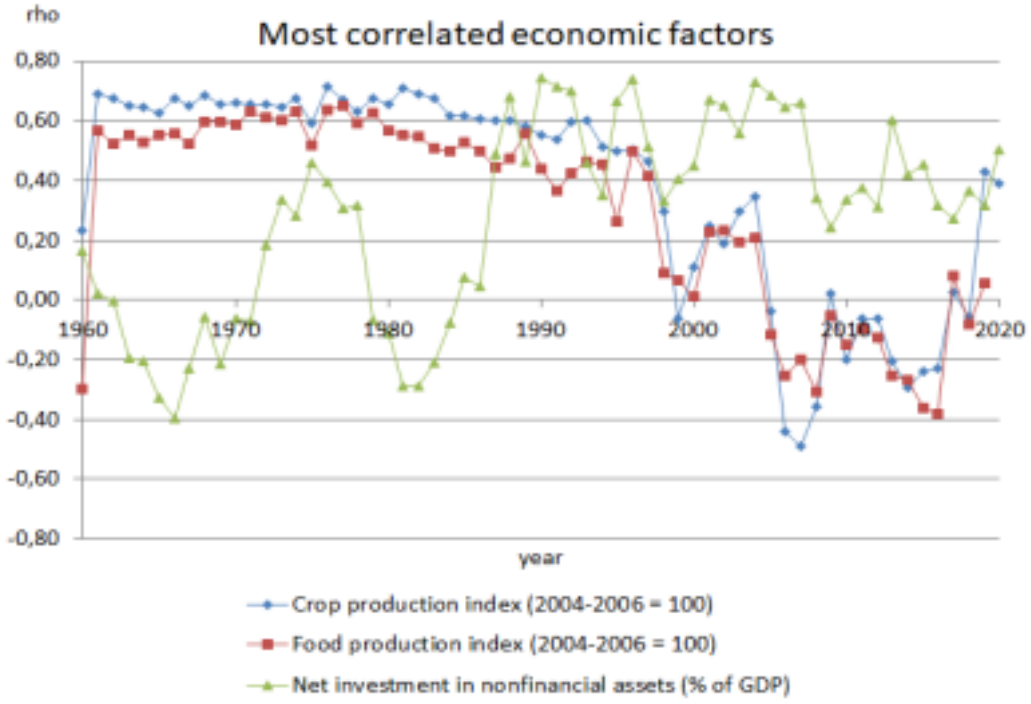
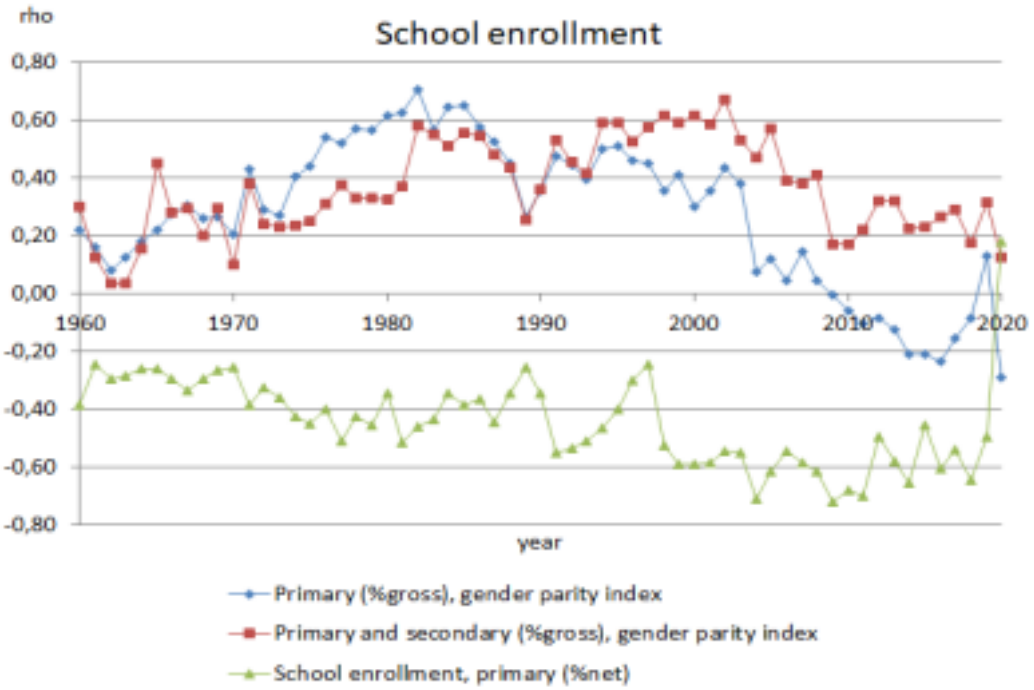


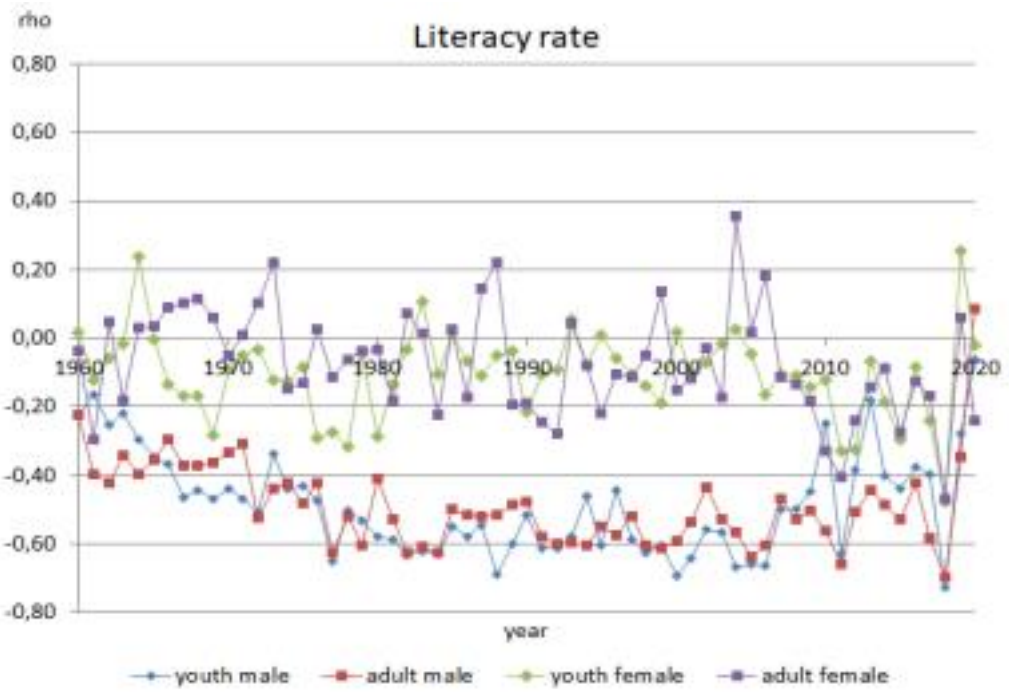
Figure 5 shows *Crop production, Food Production* and *Net investment* as the indexes with higher correlations with HT as well as the ones with better time consistency. Even though crop and food production show a decline during 2000's they are still considered as important economic activities in the region (Graesser *et al.*, 2015). Net investment shows a significant uprising tendency from the late 80's.

Figure 5 Correlations across the years of HT with school enrollment associated with the gender parity index and for both sexes



In figure 6 it is possible to observe how gender parity indexes demonstrate higher correlations with HT than the ones found with School enrollment index in which gender of students enrolled is not taken into account.

Figure 6 Correlations across the years of HT with literacy rate for both sexes



In agreement with the results in figure 6, figure 7 indicates low correlations between male

(young and adult) with HT, while female indexes show higher connections.

All in all, level of significance, field relation and time continuity of correlations were used as filters to narrow down the relations to discuss in the next section as the most relevant findings. Additionally, given the similarities among some indicators, they will be explained in the same section. Finally, due to time and space limitations discarded indicators are invited to be studied in further studies.

a. Data triangulation

Data triangulation has as main objective to challenge or confirm if results are reliable as not based on co-founding factors. Having this in mind, it is here necessary then to confront 1) Corpus data against Harmonic Tension; 2) Harmonic Tension against WDI; and 3) Corpus Data against WDI. As 2) has been largely explained before (*cf. Results*), results for 1) and 3) will be following presented:

Table 5

Corpus and Harmonic Tension

Country	**ISO**	**# words**	**# tweets**	**V.O*. pairs**
Argentina	ARG	254,982,258	26,933,107	4065
Bolivia	BOL	3,136,167	289,683	n.a.
Chile	CHL	155,791,513	15,291,490	4115
Colombia	COL	209,085,865	19,875,419	4967
Costa_Rica	CRI	43,905,034	4,272,517	3763
Cuba	CUB	122,595	13,246	n.a.

Ecuador	ECU	49,016,999	4,483,875	4552
El_Salvador	SLV	19,898,193	1,835,850	4492
Guatemala	GTM	31,753,056	3,131,936	4360
Honduras	HND	18,282,159	1,710,399	4492
Mexico	MEX	453,724,537	43,544,549	4387
Nicaragua	NIC	10,982,904	1,222,135	4492
Panama	PAN	33,237,123	3,078,389	4374
Paraguay	PRY	39,753,880	3,968,928	4287
Peru	PER	35,355,182	3,329,937	4280
Puerto_Rico	PRI	35,230,113	3,863,552	4145
Dominican_Rep.	DOM	86,657,210	8,608,484	4380
Spain	ESP	499,630,471	45,276,446	4091
USA	USA	59,974,018	6,172,521	4520
Uruguay	URY	37,121,241	4,252,022	3772
Venezuela	VEN	194,073,318	16,773,933	4648
Correlation with HT		-0.28	-0.26	0.30
p-value		0.24	0.28	0.21

Note:* Verb-Object Pairs

Table 5 gives key information regarding the corpus (gathered by Jimenez *et al*. (2018)) used in Puentes (2021) in terms of the number of words it included, the amount of tweets and verb-object pairs they used in order to calculate DOM's marking and omission in each country from 2009 to 2016. In total, 4947 verb-object pairs were analyzed starting with a subcorpus from Colombia and later contrasted with other countries' data as shown in the last column. By observing the percentage distribution for object marking in each country the authors were able

to indicate the levels of iconicity and economy in each one of them while realizing the intrinsic connection already mentioned as Harmonic Tension. The last two rows in the table include the Spearman's value for the correlations of each column against HT along with significance *p* value; both of them demonstrate that there is no relation between HT and descriptive corpus variables, size or V.O. pairs occurrences used in Puentes (2021) confirming HT as an independent variable.

Table 6 Corpus and World Demographic Indicators

Description	**# words per country**	**# tweets per county**	**# occuring v.o. pairs per country**
First p<0.01 correlated WDI during most years	CO2 emissions (kt)	CO2 emissions (kt)	Population in the largest city (% of urban population)
Number of years with correlation p>0.01 (first)	61	61	27
Second p<0.01 correlated WDI during most years	GDP (current US$)	Labor force, total	Investment in water and sanitation with private participation (current US$)
Number of years with correlation p>0.01 (second)	60	61	11
Crop production index (2004-2006 = 100)	1	1	3
Net investment in nonfinancial assets (% of GDP)	5	6	0
Food production index (2004-2006 = 100)	0	1	0
Vulnerable employment, female (% of female employment) (modeled ILO estimate)	14	10	0
School enrollment, primary (gross), gender parity index (GPI)	0	0	**11**
School enrollment, primary and secondary (gross), gender parity index (GPI)	0	0	1
Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)	28	14	0

In the last side of the triangulation (Table 6), that is Corpus against WDI, it resulted in

correlations between Corpus and some WDI (*i.e.Vulnerable employment female*, *School enrollment GPI* and *Teenage mothers*); however, these results could not be considered as significant keeping in mind that no relation between HT and Corpus was found which means that this side of the triangle does not exist for what a triangular relation among HT:Corpus:WDI is not possible. Nonetheless, indicators such as *Crop production*, *Food Production* and *Net Investment* did not show correlations, which helps to affirm their strong relations with HT as referred before (c*f. Results*).

6. DISCUSSION

First, it will be explained how HT is probably related mainly to Crop and Food Production (25 and 6 years respectively). *Crop Production Index* is understood as the amount of primary crops (e.g. cereals, grains, fruit, treenuts, etc.) calculated in terms of area harvested, production quantity and yield, excluding animal feeding (Food and Agriculture Organization of the United Nation [FAO], 2021); while *Food Production Index* refers to the group of edible crops that contain nutrients (e.g. coffee and tea are excluded for not having nutritive value) (The World Bank [WB], 2021). Additionally, *Net Investment in Nonfinancial assets*[8] *(cf. Crop and food production and Net Investment in nonfinancial assets)* and *Contributing family workers, male (% of male employment)* will be here included as a subsequent relation, as it will be explained next (*cf. School enrollment and Male Literacy and School enrollment Gender Parity Index*).

In second place, it will be intended to suggest an answer to the possible relation between School enrollment, primary (Gender Parity Index) and School enrollment, primary and secondary (Gender Parity Index) with the linguistic phenomenon. They will be explained together given

[8]"Nonfinancial assets are stores of value and provide benefits either through their use in the production of goods and services or in the form of property income and holding gains. Net investment in nonfinancial assets also includes consumption of fixed capital." (WB, 2021)

that both indicate: “the ratio of girls to boys enrolled at primary and secondary levels in public and private schools” (WB, 2021). These last results will be contrasted with the inverse relations shown in Table 4, specifically with the index *Literacy rate, adult male*[9] and *School enrollment, primary (% net)* since they are related to the same phenomenon (in this case, participation of women and men in social systems).

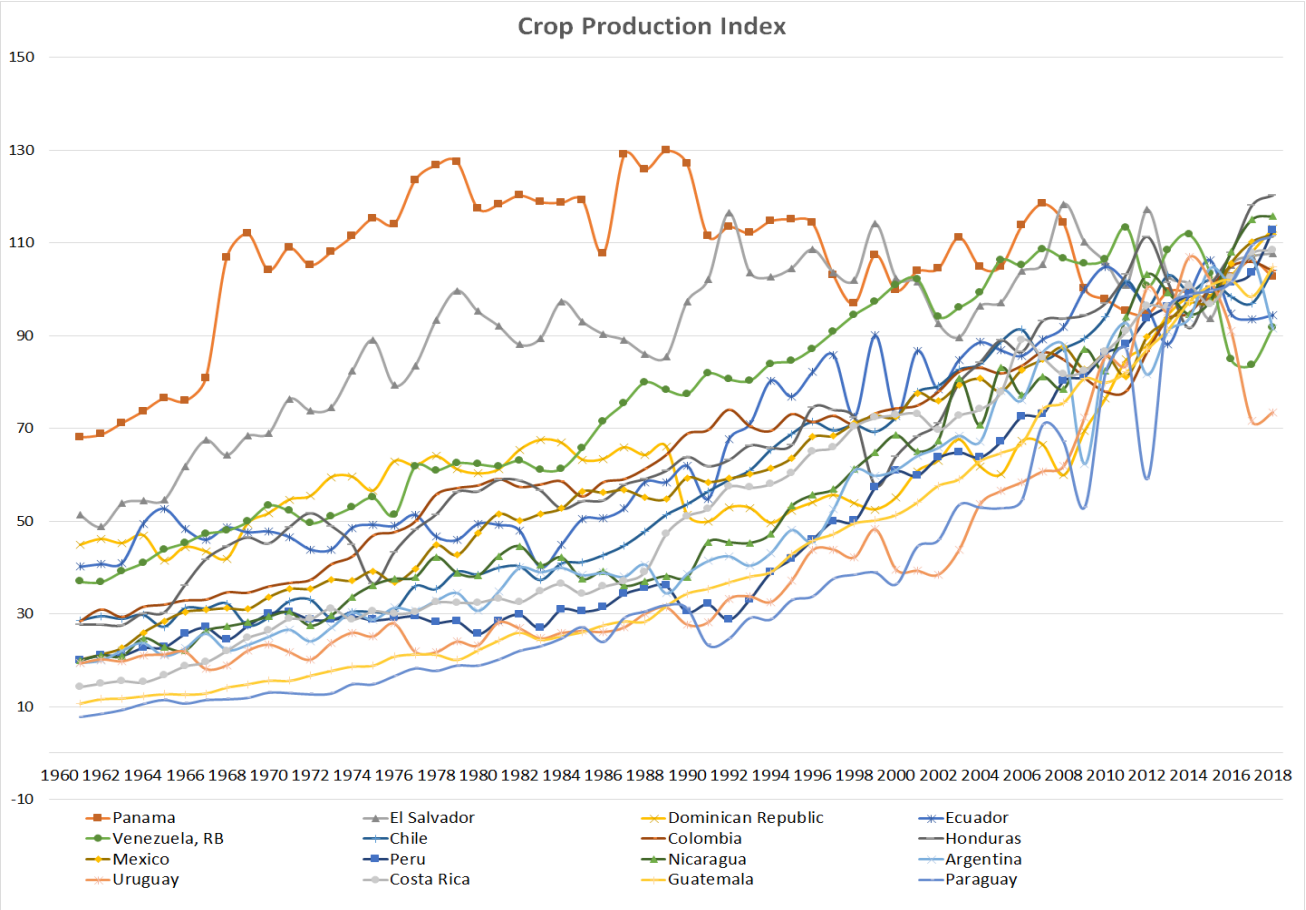
Crop and food production and *Net Investment in nonfinancial assets*

During the last five decades variationist sociolinguistics research had been focused on phenomena taking place in urban areas due to, among many other reasons, the idea around cities as centers of congregation of people from different parts of the country in a constant dynamic interaction: a perfect breeding ground for linguistic change, variation, etc. (Gordon, 2019). This tendency makes sense if we consider that for years cities experienced great expansion and a high level of migration, a social phenomenon known as *urbanization*. This movement is attributed, mostly, to the new global economy (Cohen, 2004) and, as a result, in South America around 84% of the population lives in urban areas (United Nation data, 2018; FAO, 2018).

Despite this scenario, during the last years it has been recorded that minor cities and rural areas in Latin America are now growing bigger in terms of built-up expansion, net investment (correlated index), population, among others (Andreade-Nuñez & Aide, 2018). One of the main reasons for this development is associated with the rising production of agricultural products and raw materials in the region (Graesser *et al.*, 2015) as shown in Figure 8.

Figure 7 Crop production indicator in Hispanoamerican countries (WDI)

[9] [it] is the percentage of people ages 15 and above who can both read and write with understanding a short simple statement about their everyday life. (WB, 2021)



Tacoli (2003) explains that in order to make agricultural activities an asset it is necessary to invest in infrastructure and mobility from and towards urban areas where consumers, markets, exportation means, among others, take place (Net Investment in non-financial assets). This coincides with the built-up expansion experienced in around 56% of municipalities in South America, suggesting that this kind of economic development is happening far from urban agglomerations (Andrade & Aide, 2018) "supporting the argument that growth in many countries is shifting to small and medium cities" (ONU HABITAT, 2012).

As expected, this important shift in the economy has brought with it a series of changes at local level. In the first place, due to the expansion and a more constant interaction between rural/urban communities "many aspects of the traditional distinction between urban and rural are becoming redundant" (Cohen, 2004, p. 23) for what it is now difficult to draw a limit (for

example, peri-urban communities or near-by villages) (Tacoli, 2003). Also, small cities and villages' development is now recognized as "New Ruralism", in other words, as regions no longer considered as underdeveloped and segregated but as centers of economic growth.

This dynamic brings a continuous flow of people "moving between rural and urban settlements, either commuting on a regular basis, for occasional visits to urban-based services and administrative centres, or migrating temporarily or permanently"[10] (Tacoli, 2003, p.4) also known as *floating population* or *circular migration*[11](International Organization for Migration, [IOM], 2010).

Another related issue with circular and inverse migration is observed in the negative correlation existing with *Contributing family workers, male (% of male employment) index* understood as the percentage of "own-account workers in a market-oriented establishment operated by a related person living in the same household" also known as unpaid family workers usually founded in rural areas (WB, 2021). The connection is given by the fact that, even though crop and food production usually indicate better economic situation, the truth is that there is an important amount of inequality in land distribution in Latin America. This can be translated in, as mentioned before, an important rural development but mainly coming from great industries displacing rural families and, as consequence, reducing the number of self-employed family workers in rural areas as stated in the results (Moloney, 2016)

Migrants' characteristics

In general terms, internal migrants are characterized as being mostly young driven by reasons

[10] More than 1738 municipalities in South America had an increasing densification of 15.1% (Andrade & Aide, 2018).

[11] [...] fluid movement of people between areas, often linked to labour needs in areas of origin and destination (IOM, 2011).

such as university studies, introduction to the labor field, partnership and reproduction along with a psychological disposition to take risks and experimentation. In terms of gender, women in the region have demonstrated a higher predisposition to migrate given the activities mostly attributed to women such as services field and domestic work (Rodríguez, 2004).

Regarding education level, it was previously thought that poor under qualified peasants were the main migrant population under the idea of seeking better opportunities given the bad conditions of rural life (FAO, 2018). Nonetheless, it is now discussed that in many cases high level of education facilitates displacement since it provides information, allows prior hiring, better economic solvency and, as consequence, permits to cover moving expenses (Rodríguez, 2004).

Language change and social networks

Recalling the social nature of language, it's likely to expect that the dynamic around this "new ruralism" leads us to new social interactions well known for being "the driving force both of language change and of the maintenance of conventions" (Baxter, 2016, p.257). From this, it could be affirmed that Harmonic Tension may be revealing a linguistic change in process.

One important aspect in human social life are the bonds we create with the surrounding others: family, neighbors, co-workers, etc. These links get stronger with frequent and consistent interactions among the members of a community which usually has little contact to other groups. Linguistically speaking, this helps to keep common realizations among members as they tend to be closed to varieties coming from other communities or social groups as a form to construct their identity. Nonetheless, as explained above, the region is showing great internal movement which makes these networks go weaker (Milroy, 2000). Moreno (2009) explains that weak social networks welcome new ways of speaking since social groups (commonly

middle class) get in contact with more variants making "speakers [to] learn and adapt to the usage patterns of those around them" (Baxter, 2016, p.258). In a more stable scenario these nuances would be kept as isolated, but what the data suggests is that a great amount of people is moving which may provoke or accelerate linguistic change, in our case the use of DOM in Spanish, but it may be happening on many other levels.

This clash of distinct communities provokes new agreements about language form. Nonetheless, in the meanwhile it is good to expect hesitation and fluctuation between different variables as seen in Dominican Republic, Puerto Rico, Colombia, etc. in respect to use/omission of DOM. On the other hand, in the countries in which internal migration have low numbers, those connections or social networks are built stronger as people seem to be stable in their interactions and, as a consequence, there is also a sense of stability in the variety they use (as seen in Peru, Mexico, Guatemala, etc.) (Milroy, 2000).

As a side effect to this phenomenon, it is important to point out that "prestigious" or "high social status" groups are considered to be the ones leading the linguistic norms. However, this paradigm has to be re-evaluated as well since in rural areas the concept of "prestigious" or "powerful" does not entail the same as in urban areas; for instance, economic growth does not necessarily come along with a high level of education, but on the extension of land or the profitability of crops or raw materials as well as social recognition in the community (Tacoli, 2003) which may contribute to the growing uncertainty upon the linguistic norm.

Apart from what has been stated before, it is also important to mention that the use of social media has provided linguistics with the advantage of being able to collect and analyze big-sized data as it was here intended with the Twitter corpus time placed from 2009 and 2016. This kind of corpora provides scholars a way to record the linguistic phenomena not only related to ongoing social events, but also those connected to previous ones.

As the correlations here presented took place years before the corpus sample (i.e. *Crop Production* 1961-1986, *Food production* 1961-1986, *Net Investment* 1988-2007 among others) it could be added that HT is a resulting phenomenon from the economic and social changes Spanish speaking countries faced during that time but it is only observable afterwards because of, first, the access to great corpora thanks to social media, and, second, to the nature of the linguistic measure itself which is usually identified once it has been transmitted (mainly through social networks such family members, one generation to another, friends, etc.) to an important part of the speaking community.

Something similar is expected with more recent results (i.e. *School enrollment, primary and secondary (gross), gender parity index (GPI)* 1988-2007, *School enrollment primary* 2004-2018 and *Teenage mothers* 1969-2018) except by the fact that, according to Twitter users' characterization, data gathered by the World Bank and the one coming from the Twitter corpus may have been provided from the same population; in other words, being Twitter users mostly young adults and adults (Salzman, 2015) , they probably also belong to the group reflected in *School enrollment primary* during the years from 2004-2018, in *School enrollment, primary and secondary* in 1988 and 2007 and, in a lesser extent, also *Teenage mothers* 1969-2018.

School enrollment and Male Literacy and School Enrollment Gender Parity Index

It is not surprising to find here an inverse correlation between linguistic change represented by HT, and level of education (specifically with "Literacy rate, Adult Male" and "School enrollment" indexes) when highly educated groups tend to make a more standard use of language. In spite of this "rule" it was also observed that there exists a relation between Harmonic Tension and school enrollment but specifically in the *Gender Parity* Index; in other words, the communities in which there is a more equal percentage of female and male students are more prone to have a higher Harmonic Tension, it means, they are more flexible upon the

omission or use of DOM; meanwhile, the countries in which this parity index is lower, that is to say, there are more male students than female, are characterized by a more restricted use of DOM, that is, lower harmonic tension. This phenomenon can be seen in
Figure 6 where the school enrollment for men is compared with the same indicator adjusted by gender parity, illustrating the effect of the inclusion of women in education with HT. Similarly, Figure 7 shows how the male literacy rate has an inverse effect on HT, while the female does not.

Jennifer Coates (2013) argues that nowadays it is no longer possible to attribute the process of language change exclusively to women or men since evidence has demonstrated important agency from both, although "it is true to say that male/female differences in language seem to be intimately involved in the mechanism of linguistic change" (Coates, 2013. page, 187). One important distinction between changes led by women and the ones led by men, it is that the former is related to conscious change and the latter is defined as unconscious (in Labov's terms: change from above and change from below (Labov, 2010)); it is to say that women tend to initiate changes directed to prestige forms above of social awareness, while changes led by men tend to get away from norms.

This situation in which women seem to have a major participation in the phenomenon could be supported, in lower extent though, by the correlations found with *Teenage Mothers* and *Vulnerable employment, female* indexes. This could be explained by recalling that women are, in most cases and for a long time, in charge of raising children who acquire their mother's variety (Furrow, Nelson & Benedict, 2008). If we add that adolescent pregnancy provokes a faster introduction to the work field (probably in vulnerable work, (Azevedo, Favara, Haddock, Lopez-Calva, Müller & Perova, 2021)), this situation can be somehow related with the migration phenomenon explained before, having in mind that migration would mean job opportunities, and it agrees with the high number of migrant women in the region (Rodríguez,

2004).

In linguistic terms, women and specifically adolescents are considered as one of the main agents of change. It is well known that women tend to make use of a more standard and close to the norm way of speaking which on first sight may look incongruent with the idea of women as language change agents (Moreno, 2009). Chi Luu (2015) shed light on this phenomenon showing how, for example, the use of "uptalk" (rising terminal, like in questions) and "vocal fry" (creaky voice, produced by vibrations in the larynx) is front upon when used by women but not when used by men; surprisingly, this negative attitude acts as catalyzer for change since it makes the variation to spread bigger and faster than those accepted (often not perceived by speakers) (Luu, 2015). In other languages, like Arabic and Hebrew women tend to use *male talk* in order to blend into society and to receive approval from their peers (Sa'ar, 2007); in japanese, female speakers use particles (*wa* and *on*) to sound more polite when making requests as a remain of old traditional women's language (Ide, 2003).

Then, it could be affirmed that the linguistic change on going is likely to be conscious for speakers given that the last correlations found (*School Enrollment Gender Parity*, *Teenage Mothers* and *Vulnerable employment, female*) demonstrate a major participation of women in this phenomenon. The first implies better access to education for women and the second introduces young women to the labor field. This is reflected in a more equal participation in society and therefore, in language change as well.

7. CONCLUSIONS

Social and linguistic factors were here studied as two complementary sides as one is reflected on the other. To start with, this paper had as a starting point the concept developed in Puentes (2021) called *Harmonic Tension* as the connection between the levels of iconicity and economy, known for their incidence upon language change. Among the results presented by the author it

was stated that even though some previous studies have based the problem of use/omission of a mark in transitive sentences upon verbal (telicity, volition, etc.) and object (animacy, definiteness, etc.) features, these do not explain the variation existing in the different Spanish speaking countries which demonstrated distinct levels of HT, *i.e.* a more or less uniform use of DOM in each territory. Having this in mind, it was here intended to provide an explanation for this variation from a sociolinguistic point of view by looking for statistical correlations between HT levels and World Development Indicators. From this study it was concluded that:

To start with, in terms of linguistic attitudes high levels of HT, that is a more fluctuating use of DOM, reflects a dubitative use of the mark showing that speakers are more flexible towards optional use of the element, in other words, it could be affirmed that there exists an open attitude regarding language variation. In contrast, low levels of HT indicate that speakers of this population demonstrate a more restrictive system of marking as if the rules were well settled and beyond discussion (*e.g*. Peru and Mexico).

If we reflect this attitude in terms of social phenomena, recalling how crop and food production (indexes that show higher correlations against HT) brought out a new migration scheme to these territories, it is possible to say that a community in which many peoples converge tends to be correspondingly more open and flexible towards social/linguistic change. Being in contact with members of communities, culturally and socially different, makes it harder to identify the "correct" form of speaking and as a result, in search for economy and efficiency in communication, speakers are more flexible and show acceptance towards new variations (as in Dominican Republic and Puerto Rico).

Same way, recalling that migrants are mostly women (*cf. Migrants' characteristics*) a more important female participation is also expected in this plural scenario. This fact confirms the idea of women as main linguistic change agents and helps to classify this phenomenon as a

Change from below, known for being led by women, for resulting from social factors and for being usually introduced by other dialects (Labov, 2010) as it is expected in a migration process. Additionally, being school enrollment equality index also related to HT it supports that equal education access results into important women's incidence in social and linguistic phenomena.

All in all, Harmonic Tension reflects linguistic processes when it portrays fluctuating linguistic behavior as a way to document a possible coming linguistic change, and also social processes such as migration and women's role incidence. Even though HT was analyzed upon DOM, it is here believed it can be also studied in more diverse linguistic structures in which iconicity and economy forces also play a key role.

Bibliography

Abitbol, J. L., Karsai, M., Magué, J. P., Chevrot, J. P., & Fleury, E. (2018). Socioeconomic dependencies of linguistic patterns in twitter: A multivariate analysis. *In Proceedings of the 2018 World Wide Web Conference*, 1125-1134. https://arxiv.org/pdf/1804.01155.pdf

Aissen, J. (2003). Differential object marking: iconicity vs. economy. *Natural Language and Linguistic Theory*, 21 (3), 435-483. http://dx.doi.org/10.1023/A:1024109008573

Armstrong, N. & Pooley, T. (2010). *Social and Linguistic Change in European French*. Palgrave Macmillan, New York

Azevedo, J.P., Favara, M., Haddock, S.E., Lopez-Calva, L. F., Müller, M. & Perova E. (2021). *Teenage Pregnancy and Opportunities in Latin America and the Caribbean: On Teenage Fertility Decisions, Poverty and Economic Achievement*. 2 International Bank

for Reconstruction and Development / The World Bank, Washington DC https://openknowledge.worldbank.org/bitstream/handle/10986/16978/831670v20REVIS00Bo x385190B00PUBLIC0.pdf?sequence=5&isAllowed=y

Armstrong, N. & Smith, A. (2002) The influence of linguistic and social factors on the recent decline of French ne. *Journal of French Language Studies*, 12 (1). 23-41. ISSN 1474-0079

Andrade-Nuñez, M. & Aide, M. (2018). Built-up expansion between 2001 and 2011 in South America continues well beyond the cities. *Environmental Research Letters*. 13, 1-6, https://doi.org/10.1088/1748-9326/aad2e3

Balasch, S. (2011). Factors determining Spanish differential object marking within its domain of variation, in J. Michnowicz & R. Dodsworth (eds.), *Selected proceedings of the 5th Workshop on Spanish Sociolinguistics*. Somerville, Cascadilla Press, 113-124.

Bárány, A. (2018). DOM and dative case. *Glossa: a journal of general linguistics*, 3 (1), 971–40, DOI: https://doi.org/10.5334/gjgl.639

Baxter, G. (2016). Social Networks and Beyond in Language Change. In Mehler, A., Lücking, A., Banisch, S., Blanchard, P., & Frank-Job, B. (Eds.). *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, 257-277. DOI 10.1007/978-3-662-47238-5

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And %20Machine%20Learning%20-%20Springer%20%202006.pdf

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

https://math.usu.edu/adele/Forests/567.ps.Z

Bybee, J. (2011). Markedness: Iconicity, Economy, and Frequency. En Jung Song, J. (Ed.) *The Oxford handbook of Linguistic Typology* (pp. 131-147). Nueva York: Oxford University Press.

Coates, J. (2013). *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. New York, USA, Routledge

Cohen, B. (2004). Urban Growth in Developing Countries: A Review of Current Trends and a Caution Regarding Existing Forecasts. *World Development*, Elsevier, vol. 32(1), 23-51.

Eichstaedt, J.C., Schwartz, H.A., Kern, M.L., Park, G., Labarthe, D.R., Merchant, R.M., Jha, S., Agrawal, M., Dziurzynski, L.A., Sap, M., Weeg, C., Larson, E.E., Ungar, L.H., & Seligman, M.E.P. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science*, 159-169.

Elementary Statistics and Computer Application (2012). http://ecoursesonline.iasri.res.in/mod/page/view.php?id=15455

Food and Agriculture Organization of the United Nations (2021). *Production Indices.* http://www.fao.org/faostat/en/#data/QI

Food and Agriculture Organization of the United Nations, the International Fund for Agricultural Development & the International Organization for Migration and the World Food Programme. (2018). *The Linkages between Migration, Agriculture, Food Security and Rural Development. Rome*. http://www.fao.org/3/CA0922EN/CA0922EN.pdf. Licence: CC BY-NC-SA 3.0 IGO

Fábregas, A. (2013). Differential Object Marking in Spanish: state of the art. Borealis. *An*

International Journal of Hispanic Linguistics, 2(2), 1-80. https://doi.org/10.7557/1.2.2.2603

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3), 238-247.

Food and Agriculture Organization of the United Nations (2021). *Production Indices.* http://www.fao.org/faostat/en/#data/QI

Furrow, D., Nelson, K., & Benedict, H. (2008). Mothers' speech to children and syntactic development: Some simple relationships. *Journal of Child Language, 6*(3), 423-442. doi:10.1017/S0305000900002464

Gordon, J. (2019). Language Variation and Change in Rural Communities. *Annual Review of Linguistics. 5*, 435–53. https://doi.org/10.1146/annurev-linguistics-011817-045545

Ide, S. (2003). Women's language as a group identity marker in Japanese. *Gender Across Languages. The linguistic representation of women and men*, 3, pp. 227-238, https://doi.org/10.1075/impact.11

Milroy, L. (2000). Social Network Analysis and Language Change: Introduction, European Journal of English Studies, 4:3, 217-223, DOI: 10.1076/1382-5577(200012

Hickey, R. (2007). Language and society. Unidue. Essen

International Organization for Migration. (2010). Migration, Environment and Climate Change: Assessing the Evidence. Geneva, International Organization for Migration.

Jimenez S., Dueñas G., Gelbukh A., Rodriguez-Diaz C.A., Mancera S. (2018). Automatic Detection of Regional Words for the Pan-Hispanic Spanish on Twitter. In: Simari G.,

Fermé E., Gutiérrez Segura F., Rodríguez Melquiades J. (eds) *Advances in Artificial Intelligence - IBERAMIA 2018. Lecture Notes in Computer Science*, vol 11238. Springer, Cham. https://doi.org/10.1007/978-3-030-03928-8_33.

Kittilä, S. (2011). Transitivity Typology. En Jung Song, J. (Ed.) *The Oxford Handbook of Linguistic Typology* (p. 346). Nueva York: Oxford University Press.

Labov, W. (2010). Principles of Linguistic Change, Volume 3: *Cognitive and Cultural Factors*. Oxford, Wiley-Blackwell

Liu, H. (2011). Quantitative Properties of English Verb Valency. *Journal of Quantitative Linguistics*, 18(3), 207–233. doi:10.1080/09296174.2011.581849

Luu, C. (2015). Lingua Obscura: Young Women's Language Patterns at the Forefront of Linguistic Change. *Jstor Daily*, https://daily.jstor.org/young-womens-language-patterns-at-the-forefront-of-linguistic-change/

Mangiafico, S. (2016). Summary and Analysis of Extension Program Evaluation in R, version 1.18.8, Rutgers Cooperative Extension, New Brunswick. https://rcompanion.org/handbook/G_09.html

Moloney, A. (2016). Latin America has most unequal land distribution, Colombia fares worst: charity. *Thomson Reuters Foundation*. https://www.reuters.com/article/us-latam-landrights-idUSKBN13P2NX

Moreno, F. (2009). *Principios de la Sociolingüística y Sociología del Lenguaje*. Barcelona, España: Ariel

Nguyen, D., Doğruöz A.S., Rosé C.P, de Jong, F. (2016). Computational Sociolinguistics: A Survey. *Computational Linguistics* 42 (3), 537–593.

https://doi.org/10.1162/COLI_a_00258

ONU-HABITAT (2012). Estado de las ciudades de América Latina y el Caribe 2012 Rumbo a Una Nueva Transición Urbana Technical Report. (Nairobi: Programa de las Naciones Unidas para los Asentamientos Humanos, ONU-Habitat)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, 12, 2825-2830. https://ronw.net/pubs/jmlr2011-scikit-learn.pdf

Puentes, S. (2021). Análisis cuantitativo de la MDO en el español: tensión armónica entre iconicidad y economía, [Tesis de maestría, Instituto Caro y Cuervo], Bogotá, Colombia https://preprints.scielo.org/index.php/scielo/preprint/view/2740

Rahman, S., Zheleva, B., Cherian, K. M., Christenson, J. T., Doherty, K. E., De Ferranti, D., ... & Jenkins, K. J. (2019). Linking world bank development indicators and outcomes of congenital heart surgery in low-income and middle-income countries: retrospective analysis of quality improvement data. *BMJ open*, 9(6), 1-9, https://bmjopen.bmj.com/content/bmjopen/9/6/e028307.full.pdf

Rodríguez, J. (2004). Migración Interna en América Latina y el Caribe: estudio regional del periodo 1980-2000. *Centro Latinoamericano y Caribeño de demografía. Naciones Unidas,* Santiago de Chile

Rohdenburg, G., & Mondorf, B. (2003). Determinants of Grammatical Variation in English. *Topics in English Linguistics.* https://doi.org/10.1515/9783110900019

Sa'ar, A. (2007). Masculine Talk: On the Subconscious Use of Masculine Linguistic Forms among Hebrew- and Arabic-Speaking Women in Israel. *Chicago Journals*, 32(2), pp.

405-429, http://www.jstor.org/stable/10.1086/508501 .

Saladini, F., Betti, G., Ferragina, E., Bouraoui, F., Cupertino, S., Canitano, G., ... & Bastianoni, S. (2018). Linking the water-energy-food nexus and sustainable development indicators for the Mediterranean region. *Ecological Indicators*, 91, 689-697.

Salzman R. (2015). Understanding social media use in Latin America. *Palabra Clave,* 18(3), 842-858. DOI: 10.5294/pacla.2015.18.3.9

Silva-Corvalán, C. (2001). *Sociolingüística y pragmática del español*. Washington, D.C, United States of America: Georgetown University Press

Tacoli, C. (2003). The links between urban and rural development. *Environment & Urbanization*, 5 (1), 3-12. https://journals.sagepub.com/doi/pdf/10.1177/095624780301500111

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6), 363-377.

Taylor, S. L., Ruhaak, L. R., Kelly, K., Weiss, R. H., & Kim, K. (2017). Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices. *Briefings in bioinformatics*, 18(2), 312-320. https://academic.oup.com/bib/article/18/2/312/2562741

The World Bank. (2021). World Development Indicators. https://datatopics.worldbank.org/world-development-indicators/

United Nations, Department of Economic and Social Affairs, Population Division (2018). *World Urbanization Prospects: The 2018 Revision, custom data acquired via website*. World Urbanization Prospects - Population Division - United Nations

Vicentini, A. (2003). The Economy Principle in Language: *Notes and Observations from Early Modern English Grammars*. Università di Milano

Vigouroux, C. (2008). From Africa to Africa: Globalization, Migration and Language Vitality. En Vigouroux, C & Mufwene, S. (Ed.s), *Globalization and Language Vitality: Perspectives from Africa.* Continuum International Publishing Group

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, 12, 399-413

Annex

Additional descriptions for WDIs from:

https://databank.worldbank.org/metadataglossary/all/series

Contributing family workers, male (% of male employment) (modeled ILO estimate)

Contributing family workers are those workers who hold "self-employment jobs" as own-account workers in a market-oriented establishment operated by a related person living in the same household.

Crop production index (2004-2006 = 100)

Crop production index shows agricultural production for each year relative to the base period 2004-2006. It includes all crops except fodder crops. Regional and income group aggregates for the FAO\'s production indexes are calculated from the underlying values in international dollars, normalized to the base period 2004-2006.

Depth of credit information index (0=low to 8=high)

Depth of credit information index measures rules affecting the scope, accessibility, and quality

of credit information available through public or private credit registries. The index ranges from 0 to 8, with higher values indicating the availability of more credit information, from either a public registry or a private bureau, to facilitate lending decisions.

Domestic credit provided by financial sector (% of GDP)

Domestic credit provided by the financial sector includes all credit to various sectors on a gross basis, with the exception of credit to the central government, which is net. The financial sector includes monetary authorities and deposit money banks, as well as other financial corporations where data are available (including corporations that do not accept transferable deposits but do incur such liabilities as time and savings deposits). Examples of other financial corporations are finance and leasing companies, money lenders, insurance corporations, pension funds, and foreign exchange companies.

External debt stocks, private nonguaranteed (PNG) (DOD, current US$) Private nonguaranteed external debt comprises long-term external obligations of private debtors that are not guaranteed for repayment by a public entity. Data are in current U.S. dollars.

Food production index (2004-2006 = 100)

Food production index covers food crops that are considered edible and that contain nutrients. Coffee and tea are excluded because, although edible, they have no nutritive value.

Intentional homicides (per 100,000 people)

Intentional homicides are estimates of unlawful homicides purposely inflicted as a result of domestic disputes, interpersonal violence, violent conflicts over land resources, intergang violence over turf or control, and predatory violence and killing by armed groups. Intentional homicide does not include all intentional killing; the difference is usually in the organization of the killing. Individuals or small groups usually commit homicide, whereas killing in armed conflict is usually committed by fairly cohesive groups of up to several hundred members and

is thus usually excluded.

International tourism, expenditures (% of total imports)

International tourism expenditures are expenditures of international outbound visitors in other countries, including payments to foreign carriers for international transport. These expenditures may include those by residents traveling abroad as same-day visitors, except in cases where these are important enough to justify separate classification. For some countries they do not include expenditures for passenger transport items. Their share in imports is calculated as a ratio to imports of goods and services, which comprise all transactions between residents of a country and the rest of the world involving a change of ownership from nonresidents to residents of general merchandise, goods sent for processing and repairs, nonmonetary gold, and services.

Investment in water and sanitation with private participation (current US$) Investment in water and sanitation projects with private participation refers to commitments to infrastructure projects in water and sanitation that have reached financial closure and directly or indirectly serve the public. Movable assets, incinerators, standalone solid waste projects, and small projects are excluded. The types of projects included are management and lease contracts, operations and management contracts with major capital expenditure, greenfield projects (in which a private entity or a public-private joint venture builds and operates a new facility), and divestitures. Investment commitments are the sum of investments in facilities and investments in government assets. Investments in facilities are the resources the project company commits to invest during the contract period either in new facilities or in expansion and modernization of existing facilities. Investments in government assets are the resources the project company spends on acquiring government assets such as state-owned enterprises, rights to provide services in a specific area, or the use of specific radio spectrums. Data are in current U.S. dollars.

Land area where elevation is below 5 meters (% of total land area) Land area below 5m is the percentage of total land where the elevation is 5 meters or less.

Literacy rate, adult male (% of males ages 15 and above)

Adult literacy rate is the percentage of people ages 15 and above who can both read and write with understanding a short simple statement about their everyday life.

Literacy rate, youth male (% of males ages 15-24)

Youth literacy rate is the percentage of people ages 15-24 who can both read and write with understanding a short simple statement about their everyday life.

Mammal species, threatened

Mammal species are mammals excluding whales and porpoises. Threatened species are the number of species classified by the IUCN as endangered, vulnerable, rare, indeterminate, out of danger, or insufficiently known.

Net investment in nonfinancial assets (% of GDP)

Net investment in government nonfinancial assets includes fixed assets, inventories, valuables, and nonproduced assets. Nonfinancial assets are stores of value and provide benefits either through their use in the production of goods and services or in the form of property income and holding gains. Net investment in nonfinancial assets also includes consumption of fixed capital.

Prevalence of HIV, male (% ages 15-24)

Prevalence of HIV, male is the percentage of males who are infected with HIV. Youth rates are as a percentage of the relevant age group.

School enrollment, primary (% net)

Net enrollment rate is the ratio of children of official school age who are enrolled in school to the population of the corresponding official school age. Primary education provides children

with basic reading, writing, and mathematics skills along with an elementary understanding of such subjects as history, geography, natural science, social science, art, and music.

School enrollment, primary (gross), gender parity index (GPI)

Gender parity index for gross enrollment ratio in primary education is the ratio of girls to boys enrolled at primary level in public and private schools.

School enrollment, primary and secondary (gross), gender parity index (GPI) Gender parity index for gross enrollment ratio in primary and secondary education is the ratio of girls to boys enrolled at primary and secondary levels in public and private schools.

Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)

Having a child during the teenage years limits girls' opportunities for better education, jobs, and income. Pregnancy is more likely to be unintended during the teenage years, and births are more likely to be premature and are associated with greater risks of complications during delivery and of death. In many countries maternal mortality is a leading cause of death among women of reproductive age, although most of those deaths are preventable. Infants of adolescent mothers are also more likely to have low birth weight, which can have a long-term impact on their health and development. Complications from pregnancy and childbirth are the leading cause of death among girls aged 15-19 years in many low- and middle-income countries.

Vulnerable employment, female (% of female employment) (modeled ILO estimate)

Vulnerable employment is contributing family workers and own-account workers as a percentage of total employment.