Quantitative Methods for: Socialist Indoctrination in Venezuela

Ragan

January 10, 2020

1 Literature

2 Methods

The quantitative text analysis was preformed on the pre-Chavez and post-Chavez textbook sets. The textbooks were all in PDF format when we received them. The post-Chavez textbooks we already in machine readable form and we were able to directly read in those files for our text analysis. The pre-Chavez textbooks were also PDFs, but the text in these PDFs were not machine readable. We attempted to use various methods for converting the PDFs to a machine readable form using optical character recognition (OCR). The results from these attempts were always poor. In order to ensure that all the text from both textbooks sets was available for analysis we used the website Upwork.com to hire several Venezuelans to transcribe the PDFs to plain text for us. Once the pre-Chavez textbooks were transcribed we had five machine readable textbooks (Grade 1 - Grade 5) from both the pre-Chavez and post-Chavez eras. For both sets of textbooks the following steps were taken:[1]

1. Extract all words from the textbooks and store them along with the page number the word was located on.

2. Add the grade level to the data frame for each textbook and combine the five data frames into one data frame.

3. Remove "stop words", short words, underscores, whitespace, numeric digits, and words containing any stray characters.

[1] A detailed description of each step used in calculating the different metrics used in the quantitative analysis is included in the Appendix. Te R code used in the text analysis is available as an R Package that can be installed and run using the free R statistical language. The package can be found on GitHub at `https://github.com/robiRagan/prePostChavezTextbooks`

4. Create Stems and Lemmas for all words so words with the same root are all categorized together as the same word.

5. Calculate the proportions for all lemmas.

   (a) The overall proportions for each lemma are calculated using the following formula:

$$Prop.\ for\ Lemma_i = \frac{Frequency\ of\ Lemma_i}{\sum_{i=1} Frequency\ of\ Lemma_i}$$

   (b) The keyword proportions for each lemma are calculated using the following formula:

$$Prop.\ for\ Keyword\ Lemma_i = \frac{Frequency\ of\ Keyword\ Lemma_i}{\sum_{i=1} Frequency\ of\ Keyword\ Lemma_i}$$

6. Calculate the change in proportion from pre-Chavez to post-Chavez for all lemmas.

   (a) The changes in the proportions are calculated using the following formula:

$$Change\ in\ Prop.\ for\ Lemma_i = Prop.\ for\ Lemma_i\ Post{-}Chavez - Prop.\ for\ Lemma_i\ Pre{-}Chavez$$

7. Calculate the change in proportion from pre-Chavez to post-Chavez for all key lemmas.

   (a) The changes in the proportions are calculated using the following formula:

$$Change\ in\ Prop.\ for\ Lemma_i = Prop.\ for\ Keyword\ Lemma_i\ Post{-}Chavez - Prop.\ for\ Keyword\ Lemma_i\ Pre-$$

3 Findings

We analyze the lemma proportions in two ways. First we will look at all of the lemma proportions across the two textbook sets and then we will limit ourselves to a list of keywords. The list of keywords can be found in the appendix.

Figure 1 displays the ten highest lemma proportions for the two textbook sets. There is a great deal of difference between the most common lemmas used in the two sets. Only a few lemmas appear in both

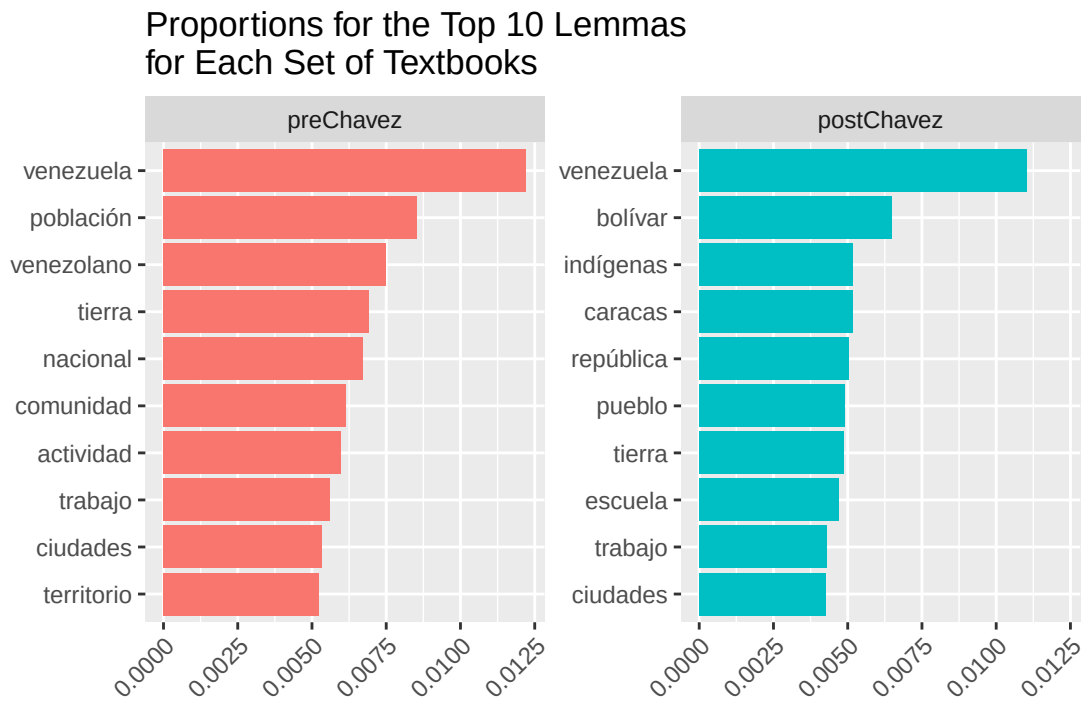

Figure 1: Top 10 Lemmas by Proportion

lists, "venezuela","ciudades", and"trabajo". The rest of the lemmas are unique. Of note is the frequent appearance of "bolívar" in the Post-Chavez textbooks. The differences here could arise for any number of reasons and may not reflect the sort of ideological shift in elementary education we are interested in studying. To better gauge the shift in language form the Pre-Chavez to Post-Chavez era we limit the lemmas in our analysis to only a set of economic and social lemmas.[2]

When we limit the lemmas to only those on our keyword list we now see more similarity across the two sets. But this is to be expected, because we have limited the set of lemmas to only those on our list. However, the differences across the two graphs now represent a much clearer picture of how language about economic and social issues in these textbooks has changed. For example "poblacíon" was the lemma with the highest proportion in the Pre-Chavez textbook set, but it does not appear in the Top 10 of the Post-Chavez textbook set. On the other hand "indígenas" does not appear in the top 10 lemmas Pre-Chavez, but has the second highest keyword lemma proportion in the Post-Chavez textbooks.

[2]A full list of the keywords can be found in the appendix.

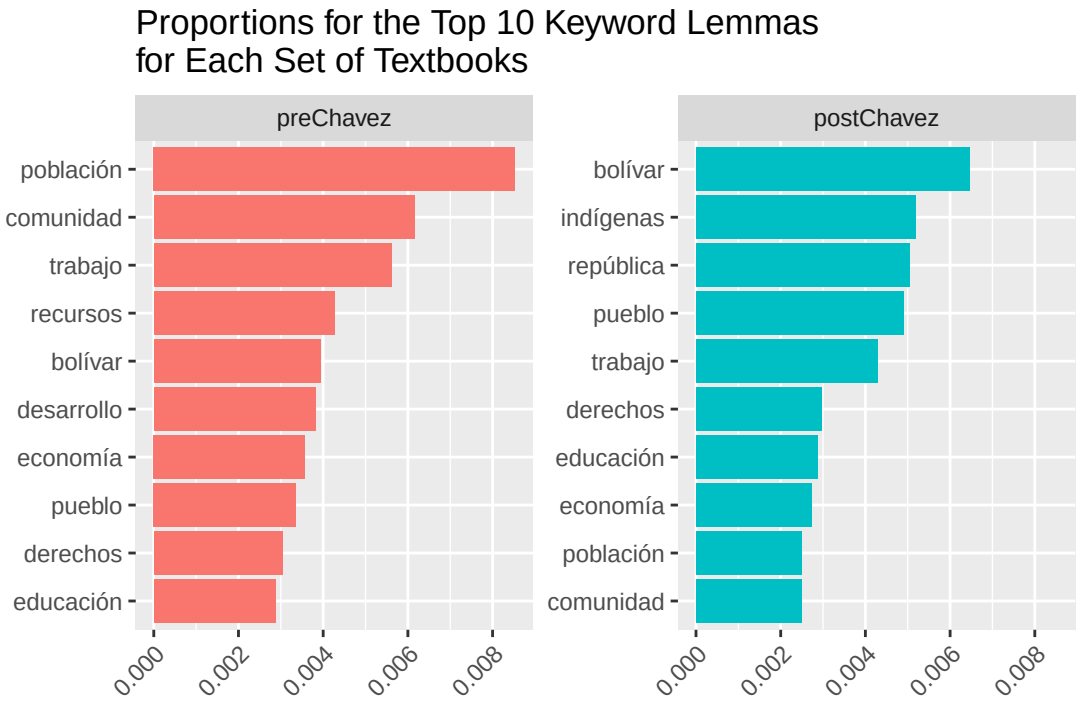

Figure 2: Top 10 Keyword Lemmas by Proportion

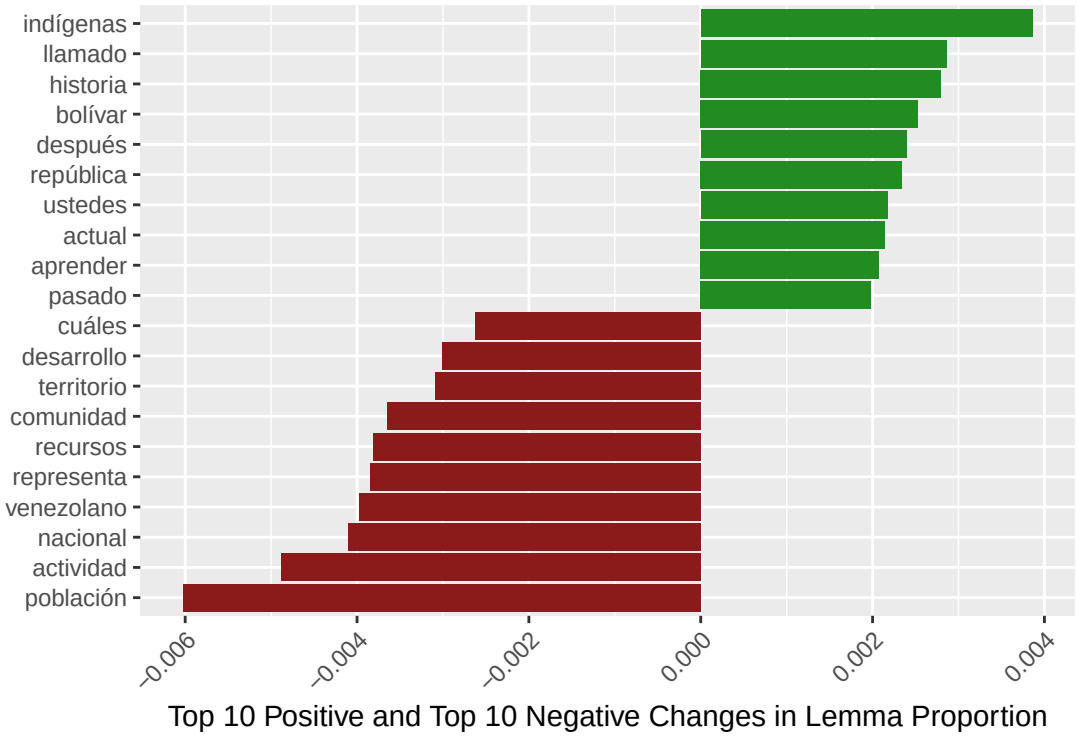

Figure 3: Change in All Lemmas Across the Two Textbook Sets

In Figure 3 we return to the set of all lemmas, not just the keyword lemmas. This graph looks at the lemmas whose proportion changed the most from the Pre-Chavez textbooks to the Post-Chavez Textbooks displays the ten highest lemma proportions for the two textbook sets. Here we see large increases in lemmas like, "indígenas","bolívar", in line with the overall proportion graph. Lemmas like "territorio", "comunidad" and "representa" see a large drop in their proportions. As before the changes in the proportions when looking at all lemmas may not hone in on changes in ideology as well as we would like them to. So we again limit the lemmas in our analysis to only a set of economic and social lemmas.

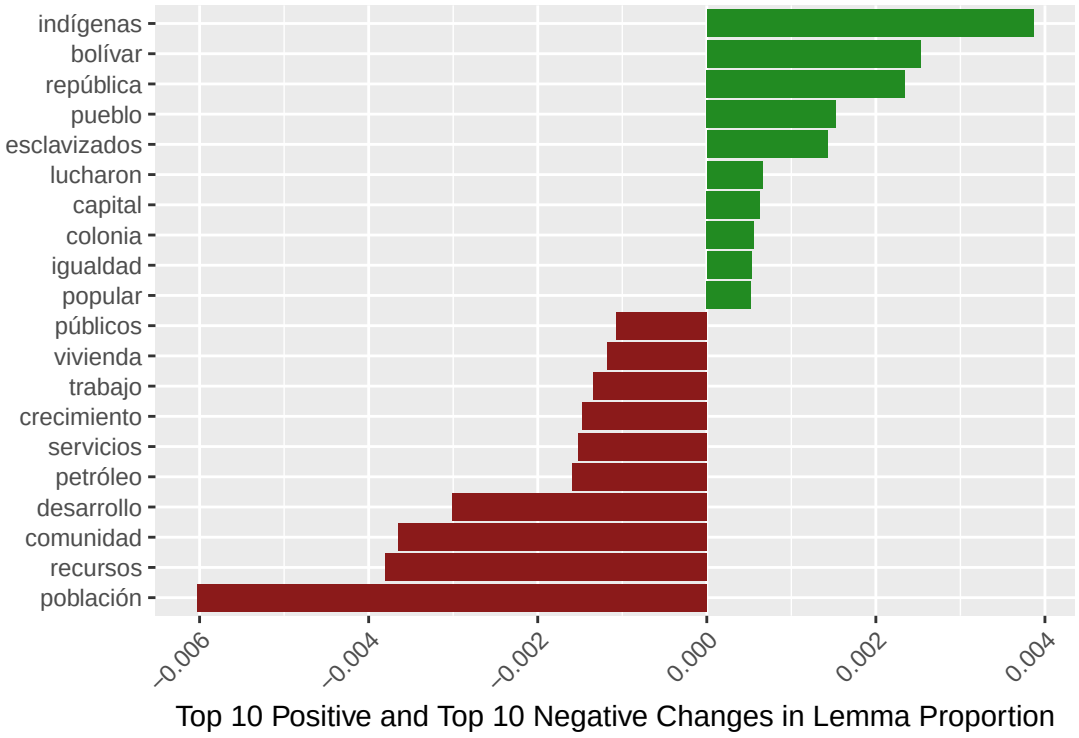

Figure 4: Change in Keyword Lemmas Across the Two Textbook Sets

In Figure 4 we again limit the lemmas to only those on our keyword list we now see more similarity across the two sets. But this is to be expected, because we have limited the set of lemmas to only those on our list. The growth in the proportion for "indígenas" is still present here, as is the growth in the proportion for "bolívar" and "pueblo". In contrast terms like "recursos", and "vivienda" say their proportions fall.

Appendix 1: Data Cleaning Details

1. Extract all words from the textbook and store them along with the page number the word was located on.

   (a) The raw text from the PDF or Text file for a single textbook is read in and stored as an object. Here we will refer to this as **rawTextbookObject**.

   (b) **rawTextbookObject** is then split with each page in the original text becoming its own data frame. Here we will refer to these as **rawPageDataFrames**.

   (c) The **rawPageDataFrames** are stored as a single R list. With each page being a data frame that is one element of the R list. Here we will refer to this list as **rawPagesList**.

   (d) For each page the raw text is tokenized. With each word receiving its own row in the first column of the data frame. The data frames will now be referred to as **tokenizedPageDataFrames** and the list storing all the data frames will be called **tokenizedPagesList**.

   (e) The second column in each of the **tokenizedPageDataFrames** contains the page number that each word appeared on. At this stage, each page has been converted to a two column data frame. The first column contains all the words on the page and the second column contains the page number the word appeared on. All of the **tokenizedPageDataFrames** are stored in an R list, **tokenizedPagesList**.

   (f) All of the **tokenizedPageDataFrames** in **tokenizedPagesList** are merged into one data frame that contains all of the words in the textbook as the first column, and the page the word was found on as the second column. Here we will refer to this as **oneTokenizedTextbookDataFrame**.

   (g) Steps 1a through 1f are run for each of the five textbooks in a set.

2. Add the grade level to the data frame for each textbook and combine the five data frames into one data frame.

   (a) A third column is added to each **oneTokenizedTextbookDataFrame** that contains the grade level for the textbook.

   (b) The five **oneTokenizedTextbookDataFrame**s are all combined into one data frame.

3. Steps 1 and 2 above are run for the second set of five textbooks. The result is one data frame that contains all of the words from all five of the pre-Chavez textbooks, and one data frame that contains

all of the words from all five of the post-Chavez textbooks.

4. Remove "stop words", short words, underscores, whitespace, numeric digits, and words containing any stray characters.

   (a) Stop words are common short words that contain little information on their own. They are frequently discarded when doing text analysis on single words. The complete list of stop words is listed in an appendix. These words were all removed from the two data frames.

   (b) Any word under five characters was removed form the analysis. Such words rarely provide any useful information on their own.

   (c) Underscores and whitespace characters are sometimes generated when PDF files are read in for text analysis. These characters are removed.

   (d) Numeric digits without any context contain little information, so such numbers are removed.

   (e) When reading in text from a PDF file some words can contain stray characters like periods or numbers. This is usually due to an encoding issue with the way PDFs store some characters. Words with these stray characters are removed. In a related issue, the apostrophe symbol used in PDFs is encoded in a different way that the apostrophe symbol in text files. So all apostrophes were corrected to be the same across the two data frames.

5. Create Stems and Lemmas for all words.

   (a) Many words with the same meaning are not identical due to prefixes, suffixes or conjugation. For example in English the words"run" and "running" would show up as different words in a word count, but in many cases we would want to count them in the same way. This is done through a process called Stemming and Lemmaing.

   (b) Stemming basically takes words and reduces them to a shortened version of the word. For example the words: "distribución", "distribuye", "distribuida", "distribuyen", "distribuido", "distribuidos", "distribuirlas", and "distribuyes" are all stemmed to become "distribu".

   (c) To create lemmas, we take the simple approach of assigning the most common word in a stem group as the lemma for all of the words in the stem group. For example, in the pre-Chavez data frame the stem family for "boliv" contains the words "bolívar" 327 times, "bolivariana" 82 times, and bolivariano 1 time. So for this stem group the lemma is the most common word 'bolívar".

(d) All of the counts for the two textbook sets are based on the number of times a lemma appears in each textbook set. In the appendix we list some of the most common words across the data sets along with their stems and lemmas.

6. Calculate the proportions for all lemmas.

   (a) The proportions for each lemma are calculated using the following formula:

$$Prop.\ for\ Lemma_i = \frac{Frequency\ of\ Lemma_i}{\sum_{i=1} Frequency\ of\ Lemma_i}$$

7. Calculate the change in proportion from pre-Chavez to post-Chavez for all lemmas.

   (a) The changes in the proportions are calculated using the following formula:

$$Change\ in\ Prop.\ for\ Lemma_i = Prop.\ for\ Lemma_i\ Post{-}Chavez - Prop.\ for\ Lemma_i\ Pre{-}Chavez$$

Appendix 2: Stop Words

(NOTE: Accent marks and other non-english characters are giving LaTeX problems make sure to double check this list)

The stop words used in Step 4 of the data cleaning:

de, la, que, el, en, y, a, los, del, se, las, por, un, para, con, no, una, su, al, lo, como, ms, pero, sus, le, ya, o, este, sí, porque, esta, entre, cuando, muy, sin, sobre, también, me, hasta, hay, donde, quien, desde, todo, nos, durante, todos, uno, les, ni, contra, otros, ese, eso, ante, ellos, e, esto, mí, antes, algunos, qué, unos, yo, otro, otras, otra, él, tanto, esa, estos, mucho, quienes, nada, muchos, cual, poco, ella, estar, estas, algunas, algo, nosotros, mi, mis, tú, te, ti, tu, tus, ellas, nosotras, vosotros, vosotras, os, mío, mía, míos, mías, tuyo, tuya, tuyos, tuyas, suyo, suya, suyos, suyas, nuestro, nuestra, nuestros, nuestras, vuestro, vuestra, vuestros, vuestras, esos, esas, estoy, ests, est, estamos, estis, estn, esté, estés, estemos, estéis, estén, estaré, estars, estar, estaremos, estaréis, estarn, estaría, estarías, estaríamos, estaríais, estarían, estaba, estabas, estbamos, estabais, estaban, estuve, estuviste, estuvo, estuvimos, estuvisteis, estuvieron, estuviera, estuvieras, estuviéramos ,estuvierais, estuvieran, estuviese, estuvieses, estuviésemos, estuvieseis, estuviesen, estando, estado, estada, estados, estadas, estad, he, has, ha, hemos, habéis, han, haya, hayas, hayamos, hayis, hayan,

habré, habrs, habr, habremos, habréis, habrn, habría, habrías, habríamos, habríais, habrían, había, habías, habíamos, habíais, habían, hube, hubiste, hubo, hubimos, hubisteis, hubieron, hubiera, hubieras, hubiéramos, hubierais, hubieran, hubiese, hubieses, hubiésemos, hubieseis, hubiesen, habiendo, habido, habida, habidos, habidas, soy, eres, es, somos, sois, son, sea, seas, seamos, seis, sean, seré, sers, ser, seremos, seréis, sern, sería, serías, seríamos, seríais, serían, era, eras, éramos, erais, eran, fui, fuiste, fue, fuimos, fuisteis, fueron, fuera, fueras, fuéramos, fuerais, fueran, fuese, fueses, fuésemos, fueseis, fuesen, siendo, sido, tengo, tienes, tiene, tenemos, tenéis, tienen, tenga, tengas, tengamos, tengis, tengan, tendré, tendrs, tendr, tendremos, tendréis, tendrn, tendría, tendrías, tendríamos, tendríais, tendrían, tenía, tenías, teníamos, teníais, tenían, tuve, tuviste, tuvo, tuvimos, tuvisteis, tuvieron, tuviera, tuvieras, tuviéramos, tuvierais, tuvieran, tuviese, tuvieses, tuviésemos, tuvieseis, tuviesen, teniendo, tenido, tenida, tenidos, tenidas, tened.

Appendix 3: Examples of Stems and Lemmas

(NOTE: Accent marks and other non-english characters are giving LaTeX problems make sure to double check this list)

word	stem	lemma
venezuela	venezuel	venezuela
bolívar	boliv	bolívar
caracas	carac	caracas
escuela	escuel	escuela
después	despues	después
república	republ	república
historia	histori	historia
tiempo	tiemp	tiempo
indígenas	indigen	indígenas
tierra	tierr	tierra
pueblo	puebl	pueblo
américa	amer	américa
españa	españ	españa
personas	person	personas
población	poblacion	población
ciudad	ciud	ciudades
nacional	nacional	nacional
comunidad	comun	comunidad
familia	famili	familia
ustedes	usted	ustedes
trabajo	trabaj	trabajo
educación	educ	educación
indígena	indigen	indígenas
independencia	independent	independencia
territorio	territori	territorio

Appendix 4: Keywords

(NOTE: Accent marks and other non-english characters are giving LaTeX problems make sure to double check this list)

Category	Keyword	Category	Keyword
surplus value	abuso	public goods	bienes
surplus value	apropiación	public goods	públicos
surplus value	aprovechar	public goods	bienestar
surplus value	capital	public goods	educación
surplus value	despojo	public goods	explotación
surplus value	obreros	public goods	gratis
objective theory of value	costo	public goods	gratuita
objective theory of value	injusto	public goods	salud
objective theory of value	justo	public goods	servicios
objective theory of value	trabajo	public goods	vivienda
objective theory of value	esclavo	nationalism	bolívar
subjective theory of value	oferta	nationalism	patria
subjective theory of value	demanda	nationalism	Latinoamrica
subjective theory of value	precio	nationalism	república
subjective theory of value	mercado	nationalism	Venezuela
income/wealth redistribution	comunidad	nationalism	colonia
income/wealth redistribution	derecho	production	crecimiento
income/wealth redistribution	desigualdad	production	desarrollo
income/wealth redistribution	distribución	production	recursos
income/wealth redistribution	igualdad	production	población
income/wealth redistribution	injusticia	production	economía
income/wealth redistribution	justicia	production	petróleo
income/wealth redistribution	repartición		
income/wealth redistribution	indígena		
income/wealth redistribution	pueblo		
income/wealth redistribution	lucha		
income/wealth redistribution	popular		