Cardiff University Press
Gwasg Prifysgol Caerdydd

JOURNAL OF CORPORA AND DISCOURSE STUDIES 2025 8 : 21-44
E-ISSN 2515-0251
JCaDS

THE ZAPATISTA SEMANTIC STRUGGLE: ANALYSING THE LINGUISTIC INNOVATION OF THE EZLN WITH SEMANTIC DIFFERENCE KEYWORDS (SDKs)

ISABELLE GRIBOMONT
UC LOUVAIN; KBR (ROYAL LIBRARY OF BELGIUM)

CITATION

Gribomont, I. (2025) The Zapatista Semantic Struggle: Analysing the Linguistic Innovation of the EZLN with Semantic Difference Keywords (SDKs) *Journal of Corpora and Discourse Studies*, *8*:21–44

ABSTRACT

This paper extracts Semantic Difference Keywords (SDKs) from the discourse issued by the EZLN (Zapatista Army of National Liberation), a contemporary Mexican social movement, by comparing it to a comparison corpus comprising the discourse produced by other Latin American leftist insurgent movements from the Cuban Revolution onwards. SDKs are keywords selected because of a comparatively high semantic difference between their uses in two or more corpora. These semantic differences point to areas of semantic contestation between the corpora. In this study, SDKs are extracted from Word2Vec word embeddings. The contribution of this paper is twofold. First, it further the understanding of Zapatista discourse and the ways in which it distances itself from the Latin American guerrilla tradition. Results show that the Zapatistas redefine words belonging to the topics of insurgency, oppression and ideology. Notably, they reject the term 'revolution' and create semantic continuums between concepts whose meaning does not appear to be linked in the comparison corpus. For instance, capitalism becomes semantically associated with epistemological violence and the semantic boundaries between thoughts and actions become eroded. Finally, various terms underwent a process of 'subjectivisation', i.e. they acquired a strong correlation with the subjective experience of the communities issuing the discourse. Second, this study demonstrates the potential of SDKs in discourse studies and highlights the way in which semantic contestation can be better understood by observing the shift in semantic relations within word clusters rather than by observing target words in isolation.

KEYWORDS

Keywords, Discourse Analysis, Political Discourse, Word Embeddings, Semantic Change

CONTACT

Centre de traitement automatique du langage, Institut Langage et Communication, Université catholique de Louvain, Place Montesquieu, 3 (bte L2.06.04), 1348 Louvain-la-Neuve, Belgium.
isabelle.gribomont@uclouvain.be

DOI

10.18573/jcads.142

ORCID

0000-0001-7443-5849

ISSUE DOI

10.18573/jcads.v8

Manuscript accepted 2024-10-04

The Zapatista Semantic Struggle: Analysing the Linguistic Innovation of the EZLN with Semantic Difference Keywords (SDKs)

Isabelle Gribomont
UC Louvain; KBR (Royal Library of Belgium)

1. Introduction

On 1st January 1994, the day the NAFTA agreement came into force, a group of Maya peasants of Chiapas declared war on the Mexican government. The armed conflict between the Mexican army and the EZLN (Zapatista Army of National Liberation) took the world by surprise and immediately made headlines. Anti-globalisation groups which had been campaigning against NAFTA sympathised with the plights of the Zapatistas, who denounced the threat of neoliberal policies to their livelihood. Intellectuals, academics and journalists were intrigued by the EZLN and their seemingly anachronistic guerrilla organisation, made media-friendly by their colourful indigenous clothing and charismatic spokesperson, Subcomandante Marcos. The latter soon took centre stage, thanks to his humorous and passionate writing style, exemplified by the sentence he famously addressed to a group of tourists on 1st January 1994: 'Apologies for the inconvenience, this is a revolution'. The international attention garnered by the movement forced the Mexican government to declare a peace agreement and start negotiating with the rebels only ten days after the beginning of the conflict.

Thirty years later, the EZLN is still one of the most influential social movements in leftist circles. Part of the appeal resides in the creativity of Zapatista language (Conant, 2010; Holloway, 2005; Rosset, Martínez-Torres, & Hernandez-Navarro, 2005). This creativity appears in the strong literary dimension of the discourse written by Marcos (known as Galeano since 2016), but also in the renovation of past guerrilla language from Latin America. The EZLN argues that it does not rely on Marxist terminology because the 'old words had become so worn out that they had become harmful for those that used them' (Holloway, 1998, 180). In addition, the encounter with the symbolic and metaphoric language of the indigenous people was a major influence in the shift away from Marxist language (Carr, 1997; Le Bot, 1997, 146, 355).

Gribomont (2019) used corpus-assisted methods to compare Zapatista language to a corpus of Latin American guerrilla discourses. This research identified a continuity between the two corpora through the repetition and appropriation of a series of tropes. However, it showed that the oral style of Zapatista language is novel and provides a rupture with past discourses. This paper further investigates the areas of semantic contestation when comparing these two corpora by extracting Semantic Difference Keywords (SDKs) (Gribomont, 2023). SDKs are words whose meaning differ the most between two or more corpora.

By investigating SDKs extracted from comparing the EZLN discourse with a comparison corpus composed of discourses issued by other past and present Latin American left-wing insurgent groups, this paper also uncovers areas of ideological ruptures between the EZLN and other organisations. Although differences in the ways in which words are used can stem from various reasons (e.g. context-specific terminology, linguistic register, polysemy), they are often symptomatic of differing ideologies or worldviews (see, for instance, Brigadir, Greene, & Cunningham, 2015; Chen, 2019; Garg, Schiebinger, Jurafsky, & Zou, 2018; Rheault & Cochrane, 2020; Würschinger & McGillivray, 2024; Zhao et al., 2019).

The aim of this paper is twofold. On the one hand, it contributes to a better understanding of Zapatista discourse and the ways in which it distances itself from the Latin American guerrilla tradition. In the wake of Zapatismo, various grassroots movements arose, and numerous connections were established transnationally. Their discourse is paramount in their appeal. Khasnabish (2010) suggests that 'the most obvious materialisation of the resonance of Zapatismo amongst North American activists is in the adoption of the rhetoric of Zapatismo, which is most frequently expressed through the communiqués of the Zapatista spokesperson Subcomandante Marcos'. Considering its importance to the movement, the interest it has aroused and its transnational resonance, Zapatista discourse is key to understanding the EZLN's goals and strategies, as well as its influence on the ideology and rhetorical practices of leftist activist circles worldwide.

On the other hand, through this case study, this paper shows how SDKs can be efficiently used in the field of discourse studies. From a computational discourse analysis perspective, statistical keywords, i.e. keywords which occur significantly more often in the corpus under investigation than in a comparison corpus, highlight what is distinctive at the lexical level. These keywords, although revealing of important discursive differences, do not allow for the examination of 'controversial' terms, e.g. terms that are the site of discursive struggle because they are utilised to grant greater or lower relevance to specific realities in sociopolitical debates, or the new meanings a term can acquire when it travels to new domains or new socio-cultural contexts. In other words, they allow for terms which are favoured by different groups to be identified but would not be able to highlight words which are used differently if they occur at a similar rate in both discourses (Dénigot & Burnett, 2021, 299). However, in qualitative discourse studies, keywords usually refer to words which are central to a discourse regardless of patterns of comparative overuse or underuse. Schröter (2008) argues that studying the semantics and use of such expressions is key to understanding these discourses, particularly the ways in which 'the meaning of the word changes relative to the group that uses it'. Since such keywords are often sites of contestation or 'semantic struggle' (Jeffries & Walker, 2017), their rhetorical role is highly context dependent. As will be shown below, by quantifying the semantic variations between context dependent uses of the same word, the automated extraction of SDKs allows researchers to rank words from most to least 'contested'

or 'controversial'.[1] Then, by querying lists of nearest neighbours (i.e. words which are semantically closest to the target keyword in the different corpora), and performing a qualitative analysis of relevant text samples, these semantic variations can be interpreted.

2. Data

The EZLN corpus was assembled by scraping the archive of the Zapatista Army of National Liberation (EZLN) from 1994, when the first communiqué was published, up to the end of 2020, when this data was collected.[2] This corpus includes communiques and speeches from a wide variety of individual and collective authors, from Zapatista spokespersons, committees and councils to human rights defence organisations and speakers invited to Zapatista events. It is therefore far from homogeneous. However, this corpus is exhaustive and therefore representative of the Zapatista rhetoric. This corpus totals more than 3.5 million words.

The comparison corpus is extracted from the CEDEMA archive (Centro de Documentación de los movimientos armados), for the most part.[3] This Centre, born in 2005, focuses on archiving source documents linked to the 'revolutionary left and insurgent processes in Latin America'. It collects sources produced by the insurgent organisations themselves, as well as the academic and bibliographic production related to the revolutionary mobilisation in the region. The primary source material, which constitutes the comparison corpus, includes periodical publications, propaganda and internal documentation of the organisations that promoted insurgent projects.

The EZLN is not included in the CEDEMA archive. The decision to exclude the movement was taken by Jorge Lofredo, a specialist of contemporary Mexican guerrilla movements, who deemed that the EZLN is more akin to a social movement than an insurgent group.[4] The Zapatistas have indeed been described as the first 'informational guerrilla movement' (Castells, 1997, 79) and 'postmodern revolutionary movement' (Burbach, 2001, 116). Since the premise of this investigation is precisely that the EZLN offers a renovation of preceding leftist and Marxist traditions, the CEDEMA archive is a suitable starting point for a comparison corpus. For this study, although the archive contains older sources, documents issued by revolutionary movements since 1953 were selected. The year 1953 marks the beginning of the Cuban revolution, the starting point of a new wave of leftist guerrilla organisations in Latin America whose influence can be traced all the way to the Zapatista movement. Documents in languages other than Spanish were excluded. Considering the foundational role of the Cuban revolution, and the fact that it was not well represented in the CEDEMA archive, documents from the 26th of July Movement (the leading organisation of the Cuban revolution) were also harves-

1 See Gribomont (2023) for a contextualisation of SDKs within quantitative and qualitative approaches to discourse studies.

2 This archive can be accessed at: https://enlacezapatista.ezln.org.mx/.

3 This archive can be accessed at: https://cedema.org/digital_items.

4 This information was obtained through an email conversation with the director of the CEDEMA archive, Eudald Cortina Orero in August 2016.

ted.[5] In spite of this addition, the comparison corpus is referred to as the CEDEMA corpus hereafter. This corpus is of course highly heterogeneous and far from exhaustive. Table 1 shows the 10 most represented movements in the CEDEMA corpus, which totals more than 32 million words.[6]

Organisation	Country	Word count
Partido Democrático Popular Revolucionario-Ejército Popular Revolucionario	Mexico	4 309 085
Ejército de Liberación Nacional (ELN)	Colombia	2 366 628
Fuerzas Armadas Revolucionarias de Colombia - Ejército del Pueblo (FARC-EP)	Colombia	2 067 450
Partido Comunista del Perú (PCP) [Sendero Luminoso]	Peru	745 825
Frente Farabundo Martí para la Liberación Nacional (FMLN)	El Salvador	707 652
Movimiento de Izquierda Revolucionaria (MIR)	Chile	634 544
Partido Comunista de Colombia (marxista-leninista) / Ejército Popular de Liberación (EPL)	Colombia	506 265
Movimiento de Izquierda Revolucionaria (MIR)	Chile	390 204
Movimiento Revolucionario Tupac Amaru (MRTA)	Peru	327 303
Frente Sandinista de Liberación Nacional (FSLN)	Nicaragua	258 780

***Table 1**: Ten most represented movements in the CEDEMA corpus.*

3. Methodology

SDKs are the terms for which the semantic difference between their occurrences in several corpora is largest. Different methods to quantify semantic difference can be used to extract SDKs. This case study relies on static word embeddings modelled with the Word2Vec framework (Mikolov et al., 2013). Word2Vec is a neural network-based algorithm designed to generate distributed representations of words in a continuous vector space.

Distributional semantics is a branch of semantics that relies on the idea that the meaning of a word is a function of the words that appear in the same context, i.e. the 'distributional hypothesis' (Harris, 1954). Words are synonyms if they can be replaced by one another in any context without any change in meaning. Conversely, if words are consistently found in different contexts, their meaning is different.

From this premise, static word embeddings model the meaning of each unique word with a numerical vector according to the context in which it appears in a given corpus, taking n words on either side into consideration. The vectors are optimised in such a way that words with similar meanings have similar vector representations. For example, the

5 This archive can be accessed at: http://www.fidelcastro.cu/es/biblioteca/documentos/.

6 The content of the scraped websites/archives is publicly available but cannot be reproduced elsewhere. However, the code, frequency files and link to the resulting Word2Vec embeddings are available on GitHub: https://github.com/isag91/EZLN-SDKs.

vectors for 'king' and 'queen' might be close together in the embedding space, and certain vector arithmetic operations, such as $v_{king} - v_{man} \approx v_{queen}$ are possible, capturing analogical relationships.

Word2Vec takes a large text corpus as input and outputs a vector space which usually comprises several hundred dimensions, i.e. the vectors are composed of several hundred numbers. In the context of this study, Word2Vec learns the vector representations with the skip-gram approach, which means that it learns vector representations of words by predicting the surrounding context words based on a target word.[7] As the model learns to predict context words, it adjusts the word vectors in such a way that words appearing in similar contexts end up close together in the vector space.

Once the model is trained, similarity metrics between vectors, such as the cosine similarity,[8] can effectively be used as a proxy for the semantic and syntactic similarity between words.[9] Because of their ability to map and formalise relationships between words within specific discourses, word embeddings are increasingly used in the field of Critical Discourse Analysis. See Wiedemann and Fedtke (2022) for a relevant survey of the topic.

In this case study, since the goal is to measure the distance between two vectors for one word (one for their occurrences in the EZLN corpus and one for their occurrences in the CEDEMA corpus), the two corpora are modelled together in one semantic space after context-specific strings are added to target words. For instance, all occurrences of the word revolución ('revolution') in the EZLN corpus are replaced by 'revolución_EZLN' so that the model learns two embeddings instead of one. This method has previously been used to compute Lexical Semantic Change. For instance, Dubossarsky, Hengchen, Tahmasebi, and Schlechtweg (2019) add time-specific tokens to target words to model diachronic semantic change. In addition, they demonstrate that this method leads to more accurate models than methods which learn different semantic spaces for each corpus and align them to be able to measure distances across corpora. With this method, the EZLN corpus and CEDEMA corpus are therefore merged and treated as one corpus.

A known issue with static word embeddings is their variability. Various factors can significantly alter the results such as the presence of specific documents, the size of the documents, the size of the corpus, and seeds for random number generators. Consequently, two models learned from the same data and hyperparameters might provide significantly different outcomes when it comes to measuring the cosine similarity

7 According to Mikolov et al. (2013), the skip-gram architecture performs better on semantic tasks, while the bag-of-words architecture performs better on syntactic tasks. In addition, skip-gram works better on less frequent words, but requires more training time.

8 The cosine similarity between two vectors is the dot product of the normalised vectors. It ranges between -1 and 1. 1 suggests that the words have the exact same meaning, 0 that their meaning is entirely unrelated, and -1 that there is an exact opposing semantic relationship between the two words.

9 As part of pre-processing, the corpus was lower-cased, lemmatised and segmented into sentences. The Word2Vec model was built with the Gensim library (Řehůřek & Sojka, 2010) and the hyperparameters were vector size = 300, window = 5, sample = 1e − 5, alpha = 0.03, min_alpha = 0.0007, negative = 20, epoch = 30.

between word vectors. However, Antoniak and Mimno (2018) have shown that this variability can be efficiently mitigated by averaging the results of models trained on different bootstrap samples of the corpus. Bootstrap sampling consists in sampling n documents randomly with replacements, where n is the total number of documents in the corpus. By averaging all results over 50 models trained on 50 different bootstrap samples of our corpus, the impact of the order of the documents, presence or absence of any given document, random initialisations of learned parameters, random negative sampling and randomised subsampling of tokens within documents is mitigated (Antoniak & Mimno, 2018).

This study uses stratified samples, i.e. each bootstrap sample includes specific proportions of documents belonging to predefined groups. Although the outcome should not be skewed by the presence or absence of a single document, it is important that models learn from representative samples of the corpus. Since the distribution of documents per organisation is highly unequal, three different types of stratification groups were implemented. First, the 31 organisations which total at least 50 documents in the corpus were assigned their own stratified group, i.e. each bootstrap sample includes n documents of each of these groups, where n is the number of documents from this group in the original corpus. Second, the remaining organisations were grouped by countries of origin. That is to ensure that each bootstrap sample roughly follows the geographic distribution of the original corpus. Countries which totalled at least 50 documents were assigned their own stratified group.[10] Third, the remaining documents constitute the last stratification group.[11]

To select potential SDKs which are characteristic of the discourse of Latin American leftist insurgency, the full corpus was compared to the Spanish corpus esTenTen18 available in Sketch Engine which contains 16.9 billion words of both European and American Spanish crawled from the Web in 2018. The corpora from the TenTen corpus family aim to offer a picture of 'general language' (Kilgarriff et al., 2014; Kilgarriff & Renau, 2013).[12] The words which obtained a simple maths keyness score higher than one (Kilgarriff, 2009) were selected. In addition, words whose frequency was less than 400 in the EZLN corpus and 1000 in the CEDEMA corpus were discarded, resulting in 107 words. These thresholds were selected empirically so that the words selected are prevalent enough to be discursively significant in the Zapatista discourse, as well as frequent enough in the CEDEMA corpus for the comparison to be robust. The cosine similarity between vector pairs is calculated for all potential SDKs. They are then ranked from smallest to highest similarity.

10 These countries, with the corresponding number of documents between parentheses, are Argentina (227), Chile (333), Colombia (143), Cuba (53), Ecuador (100), El Salvador (84), Guatemala (126), Mexico (223), Peru (56), Puerto Rico (60), Uruguay (75) and Venezuela (117).

11 This last group contains 175 documents from Bolivia, Costa Rica, Haiti, Honduras, Nicaragua, Panama, Paraguay and the Dominican Republic.

12 More information about the esTenTen18 corpus is available here: https://www.sketchengine.eu/estenten-spanish-corpus/#toggle-id-2.

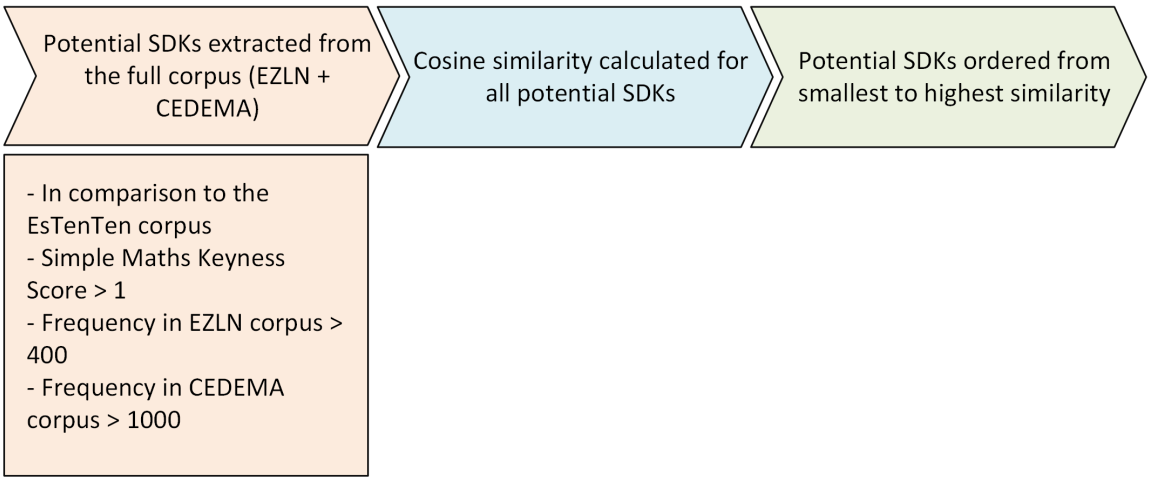

***Figure 1**: Methodology*

4. Results

Table 2 shows the top 25 SDKs ordered by mean rank. The cosine similarity and ranks are averaged over the 50 bootstrap samples of the corpus. Rank is a more robust measure here, since random initialisation imply that distances are not comparable from one model to another. However, since distances should remain proportional, the mean cosine similarity provides a ranking which is very close to the mean rank. The standard deviations of the ranks show that ranks vary quite widely in the 50 models trained on the bootstrap samples. Henceforth, all measures provided are calculated by averaging the measures provided by the 50 models.

	Word	Translation	Cos. sim.	Rank	SD rank
1	rebelde	rebel	0.1995	1.88	1.50
2	clase	class	0.2214	3.71	1.65
3	dirección	direction/address/management	0.2546	5.83	3.16
4	revolución	revolution	0.2562	5.96	3.50
5	desprecio	contempt	0.2626	7.23	4.19
6	destrucción	destruction	0.2766	9.48	6.48
7	rebeldía	rebellion	0.2758	9.66	4.91
8	construcción	construction	0.2856	10.43	5.41
9	pensamiento	thought	0.2858	11.35	6.13
10	neoliberalismo	neoliberalism	0.2867	11.61	6.45
11	indio	Indian	0.2925	12.25	7.57
12	necesidad	necessity	0.2943	12.36	4.89

13	gobernante	leader/ruler	0.2924	12.70	5.79
14	maíz	corn	0.2994	13.91	6.61
15	realidad	reality	0.3028	14.41	5.38
16	unidad	unity	0.3065	15.68	7.66
17	resistencia	resistance	0.3125	16.75	6.49
18	soldado	soldier	0.3197	19.11	8.69
19	capitalismo	capitalism	0.3254	19.85	6.61
20	sociedad	society	0.3273	20.25	5.99
21	humanidad	humanity	0.3312	22.08	9.39
22	despojo	dispossession	0.3402	24.53	8.35
23	mercancía	merchandise	0.3402	26.08	10.67
24	dignidad	dignity	0.3475	26.91	7.22
25	esperanza	hope	0.3483	28.46	11.42

***Table 2**: Top 25 SDKs for the EZLN corpus and the CEDEMA corpus.*

A few words which pertain to Named Entities specific to the Zapatista movements were removed from this list: chiapas, sureste ('southeast'), marcos, base ('base'), apoyo ('support'), clandestino ('clandestine'), comandante ('commandant'), liberación ('liberation'), partido ('party'), acción ('action'), comité ('committee'), montaña ('mountain').[13] The analysis of the results will be divided into three thematic categories of interest: insurgency, oppression and ideology. This analysis is not exhaustive and does not cover all SDKs. However, starting from important semantic categories in the SDK list allows for the main areas of semantic contestation of the EZLN discourse to be identified in comparison with the CEDEMA corpus, before investigating how and why this semantic contestation takes place.

4.1. The Vocabulary of Insurgency

Vocabulary pertaining to the semantic field of insurgency (rebelde, revolución, rebeldía and resistencia), with three keywords in the top seven, might be the semantic category which is most contested by the Zapatista discourse. When examining the nearest neighbours of these words in both corpora, they appear to be infused with a more subjective

13 Sureste and montaña are used in the phrase which signs off most communiqués: 'Desde las montañas del sureste mexicano' ('From the mountains of the Mexican Southeast'). Base and apoyo are mostly used in the EZLN corpus in the phrase 'bases de apoyo' ('support bases'), which is used to designate the individuals, families or communities who belong to the movement, without being part of the 'Zapatista army'. Clandestino and comité are mostly used in the context of the 'Comité Clandestino Revolucionario Indígena-Comandancia General (CCRI-CG)', the name of the political leadership of the EZLN. Liberación is part of the EZLN acronym ('Ejército Zapatista de Liberación Nacional'). Partido and acción are often used in the context of the 'Partido de Acción Nacional' (PAN), the party of former Mexican presidents Vicente Fox (2000 - 2006) and Felipe Calderón (2006 - 2012).

and emotional component in the EZLN discourse. For instance, the words dolor ('pain') and rabia ('rage') are among the nearest neighbours of rebeldía. When analysing the word in context and looking at its collocates, a major difference is the use of the verb tener ('to have'). Rebellion is consistently described as something that people have within them. Sentences which feature phrases such as 'esa rebeldia que tiene la juventud' ('this rebelliousness that the youth have'), 'esa rebeldía que tenemos' ('that rebelliousness that we have') show that rebellion, like pain and anger is akin to a feeling caused by a negative trigger. However, this emotion pushes people to revolt. The word often simultaneously refers to the emotion (rebelliousness) and its consequence (rebellion). In comparison, in the CEDEMA corpus, rebeldía is presented as something which grows (crecer), sprouts (brotar), comes (venir) or follows (seguir). It is represented more as a concept or an abstraction than an emotion which directly emanates from individuals or collectives. This difference is emphasised by the fact that rebeldía is used much more often with the possessive articles su ('your', 'his', 'her', 'their') and nuestra ('our') in the EZLN corpus. This increased subjectivity in the semantics of the word is mirrored, to a lesser extent, in the words resistencia and rebelde.

To understand the nature of this semantic difference further, looking at the semantic relationships between these terms within the two corpora is relevant. In both cases rebeldia and resistencia are the closest (the cosine similarity is 0.56 in the EZLN corpus and 0.47 in the CEDEMA corpus). These are high similarity scores in the context of this model. For instance, the first negative SDK (i.e. the candidate keyword for which the EZLN vector and the CEDEMA vector are the closest), riqueza ('wealth'), has a cosine similarity score of 0.54. However, while the cosine similarity between revolución and resistencia is nearly as high in the CEDEMA corpus (0.43), it is only 0.23 in the Zapatista corpus. The cosine similarity is always larger in the EZLN corpus for the pairs which do not involve the word revolution and smaller for the pairs which do. In addition, the differences are comparatively small for the pairs which do not involve the word revolution (in terms of percentage difference). We can conclude that revolución is the odd-one-out in the Zapatista corpus.

			Cos. sim. EZLN	**Cos. sim. CEDEMA**
revolución	-	rebelde	0.2011	0.2805
revolución	-	rebeldía	0.1990	0.2720
revolución	-	resistencia	0.2390	0.4375
rebelde	-	resistencia	0.3852	0.3273
rebelde	-	rebeldía	0.3953	0.3043
rebeldía	-	resistencia	0.5649	0.4750

Table 3: *Cosine similarity between pairs of words denoting insurgency in the EZLN corpus and the CEDEMA corpus.*

While resistencia, rebelde and rebeldía are all used differently in the two corpora, as briefly explained above, the proximity between them is maintained from one corpus to

the other. In the case of the word revolución, the word itself is not only a site of contestation and 'distancing' between the Zapatista discourse and the CEDEMA corpus, but the semantic relationships between words are also disrupted.

The ten nearest neighbours of revolución, as used in the EZLN corpus, include partido, 1910, PRD, 1810, and zapata (see Table 4). PRD is the acronym of the Partido de la Revolución Democrática ('Party of the Democratic Revolution'). The PRD originated from the Democratic Current, a political faction formed in 1986 from the Institutional Revolutionary Party (PRI). The PRI (previously called the National Revolutionary Party and the Party of the Mexican Revolution), which also appears further down in the nearest neighbours' list, is a right-wing party which has been in power from 1929 to 2000. This party co-opted the 1910 Mexican revolution and aimed to institutionalise its principles in a new state model. This somewhat oxymoronic process of institutionalising revolution is symbolised by the word revolutionary in the party name, but also involved the use of iconographic and textual propaganda (Plu-Jenvrin, 2019). This institutionalisation process of the Mexican revolution tainted or altered the meaning of the word revolution in Mexico. The existence of a right-wing political party such as the PRI have rendered the word unusable for the EZLN, which defines itself in opposition to institutional politics. This interpretation is supported by the difference in frequency. The word revolución is used 422 times in the EZLN corpus (103.66 times per million words) and 50 668 times in the CEDEMA corpus (1363.3 times per million words).

NN_{EZLN}	Translation	Cos. Sim.	NN_{CEDEMA}	Translation	Cos. Sim.
partido_EZLN	party (EZLN)	0.4093	socialista	socialist	0.7338
1910	1910	0.3380	socialismo	socialism	0.7131
guerra_EZLN	war (EZLN)	0.3280	proletariado	proletariat	0.6811
país_EZLN	country (EZLN)	0.3247	revolucionario	revolutionary	0.6720
movimiento_EZLN	movement (EZLN)	0.3245	comunista	communist	0.6694
lucha_EZLN	fight (EZLN)	0.3233	proletaria	proletarian	0.6478
acción_EZLN	action (EZLN)	0.2961	partido	party	0.6469
prd	PRD	0.2824	triunfo	triumph	0.6438
1810	1810	0.2684	lucha	struggle	0.6400
zapata	Zapata	0.2585	histórico	historical	0.6327

Table 4: *Ten nearest neighbours of the vector for revolución as used in the EZLN corpus and the CEDEMA corpus.*

Close reading of relevant excerpts (see Table 5) confirms that the Zapatistas nearly exclusively use the word revolution to talk about the Mexican revolution of 1910, often mentioned alongside the independence war of 1810 (Excerpt 1), political party such as the PRI and the PRD (Excerpt 2), and changes which are often negatively or neutrally connotated, e.g. 'libertarian revolution', 'industrial revolution', 'technological revolution',

and 'scientific revolution' (Excerpt 3). Excerpt 4 acknowledges a deliberate avoidance of the word revolution, citing the implication of a seizure of power. This implication can of course be linked with the aftermaths of the Mexican revolution. Although rare, there are exceptions to the Zapatista avoidance of the word revolution to describe their own activities (Excerpt 5).

	KWIC (sentence)	Translation
(1)	*Ofrecemos esta carga y este problema al resto de los pueblos indios, si no lo hacemos llegaremos a la victoria con las mismas manos vacías con las que llegamos en la **revolución** de 1910 y con las que llegamos en la guerra de independencia de 1810.*	We offer this burden and this problem to the rest of the Indian peoples, if we do not do so we will arrive at victory with the same empty hands with which we arrived at the **revolution** of 1910 and with which we arrived at the war of independence of 1810.
(2)	*Y entonces mientras se mantenga en este terreno el problema de administración, [...] sea la administración de la corrupción [...] que era el partido de la **revolución** institucional, el PRI; o la nueva propuesta de que es posible establecer este equilibrio entre la macroeconomía y la microeconomía, el eje de la propuesta del partido de la **revolución** democrática dice: El problema es de corrupción y honestidad, y entonces se puede mantener el mismo sistema, si los funcionarios ya no son corruptos' [...].*	And then, as long as the problem of administration is maintained in this area, whether it is the administration of corruption [...] which was the party of the institutional **revolution**, the PRI; or the new proposal that it is possible to establish this balance between macroeconomics and microeconomics, the axis of the proposal of the party of the democratic **revolution** says: 'The problem is one of corruption and honesty, and therefore the same system can be maintained, if the officials are no longer corrupt [...].'
(3)	*La nueva **revolución** tecnológica (la informática) y la nueva **revolución** política (las megápolis emergentes sobre las ruinas de los Estados Nacionales) producen una nueva '**revolución**' social.*	The new technological **revolution** (information technology) and the new political **revolution** (the emerging mega-cities on the ruins of the nation states) produce a new social '**revolution**'.
(4)	*Nosotros como zapatistas, [...] siempre hemos sido muy cuidadosos en la diferencia del uso de la palabra de la rebelión y la **revolución**; nosotros siempre hablamos de rebelión, porque el concepto de **revolución** implica la toma del poder, y dentro de los planteamientos del EZLN no está este, no está el de la toma del poder.*	We as Zapatistas [...] have always been very careful about the difference between the use of the word rebellion and **revolution**; we always speak of rebellion, because the concept of **revolution** implies the seizure of power, and the EZLN's approach does not include this, it does not include the seizure of power.
(5)	*El 'mandar obedeciendo' zapatista implica este 'volteo' de la política y es un proceso, no un decreto. Es, para decirlo con la 'modestia' zapatista, una **revolución** que haga posible la **revolución**.*	The Zapatista 'command by obeying' implies this 'overturning' of politics and is a process, not a decree. It is, to put it with Zapatista 'modesty' a **revolution** that makes **revolution** possible.

Table 5: *Examples of the word revolución in context in the EZLN corpus.*

It is worth noting that the semantic reduction of the word revolution in the Zapatista discourse, which mostly limits the use of the word to a few specific contexts distinctly separate from the movement itself, is at odds with several aspects of the way in which the EZLN presented itself in the early days of the movement. First of all, the name of the movement refers to Emiliano Zapata, one of the heroes of the Mexican revolution. By coopting the imagery of the Mexican revolution, much like the PRI a few decades earlier, the Zapatistas place themselves in direct filiation with this event. They appropriate its symbols, presumably to overwrite the institutionalised semantics brought about by decades of PRI governance. If a hero of the Mexican revolution was made one of the flagships of the movement, why was the word revolution not equally re-appropriated? After all, as mentioned in the introduction, Marcos used that word to describe the uprising in January 1994 and the Zapatista leadership is called the 'Revolutionary Indigenous Clandestine Committee'. We can speculate that the 'anti-revolutionary' rhetoric partly arose at the same time as the Zapatistas' initial goal of overthrowing the Mexican government and instigating a national revolution against the rise of neoliberalism died out to give way to a protest platform largely focused on indigenous rights.

This section on the vocabulary of insurgency reveals two kinds of semantic struggles. On the one hand, the word rebeldía undergoes a shift from abstract concept to emotional response. This is not the only example of 'subjectivisation' of the vocabulary of insurgent discourse which will emerge from this study. This semantic shift is a symptom of the EZLN's desire to humanise insurgency. The second semantic shift, undergone by the word revolución, is linked to Mexican political history. In this case, the semantic struggle is more pervasive, rather than contained to Latin American leftist insurgent discourse, and was started many decades ago when the Mexican revolution was institutionalised and co-opted by right-wing politics. These semantic shifts are rooted in different types of contestations and should be interpreted from different contextual foci. In addition, they are different in nature since the former implies a subjectivisation of a term, while the latter consists of a reduction of meaning caused by the rejection of a context-specific ideological connotation of a term.

4.2. The Vocabulary of Oppression

Among the list of top SDKs, destrucción ('destruction'), despojo ('dispossession') and desprecio ('contempt') are the three words which denote the negative experience of the Zapatista communities and indigenous people of Chiapas. Again, their relative distributions in the semantic space differs in the EZLN and the CEDEMA corpus, with much closer semantic proximities in the EZLN corpus (see Table 6). For instance, destrucción_EZLN and desprecio_EZLN are the first and fourth nearest neighbours of despojo_EZLN while such proximity cannot be found in the nearest neighbours of the corresponding CEDEMA vectors.

			Cos. sim. EZLN	Cos. sim. CEDEMA
destrucción	-	despojo	0.4643	0.2674
despojo	-	desprecio	0.3959	0.2083
destrucción	-	desprecio	0.3224	0.1476
destrucción	-	capitalismo	0.3414	0.3133
despojo	-	capitalismo	0.3346	0.2843
desprecio	-	capitalismo	0.2883	0.1486
neoliberalismo	-	capitalismo	0.3462	0.5193

***Table 6**: Cosine similarity between pairs of words denoting oppression in the EZLN corpus and the CEDEMA corpus*

The proximity of these concepts in the EZLN corpus relative to the CEDEMA corpus largely explains the comparatively large semantic distances between their manifestations in the two corpora. In the EZLN corpus, ideological or abstract prejudice (i.e. contempt and humiliation) is closely linked to material prejudice (i.e. poverty and repression). For instance, the nearest neighbours of desprecio include words belonging to the semantic field of contempt and prejudice, such as humilliación ('humilliation'), discriminación ('discrimination'), racismo ('racism'), marginación ('marginalisation') and burla ('moquery'), but also words denoting forms of material discrimination, e.g. despojo, miseria ('misery'/'poverty'), injusticia ('injustice') and represión ('repression'). However, the nearest neighbours of desprecio in the CEDEMA corpus include nearly exclusively words belonging to the semantic field of contempt and prejudice, i.e. odio ('hate'), humiliación, arrogancia ('arrogance'), actitud ('attitude'), discriminación, racismo and cinismo ('cynicism'). The only exception is miseria, in fifth position.

In the context of the Zapatista struggle for the recognition of indigenous cultures and worldviews, as well as the inclusion of indigenous people in the discourses surrounding national identity and history, this increased proximity mirrors a discursive link between epistemic, economic and military violence. Table 7 illustrates the ways in which these concepts are woven together in the Zapatista discourse.

	KWIC (sentence)	Translation
(6)	*Para develar la historia de explotaciones, asesinatos, despojos, **desprecios** y olvidos que se escondía detrás de la historia de arriba. Esa historia de museos, estatuas, libros de texto, monumentos a la mentira. [...] No fueron sólo palabras. La sangre de nuestros caídos y caídas en estos 22 años se sumó a la de años, lustros, décadas, siglos anteriores.*	To unveil the history of exploitations, assassinations, dispossessions, **contempt** and oblivion that was hidden behind the history from above. That history of museums, statues, textbooks, monuments to lies. [...] They were not just words. The blood of our fallen in these 22 years was added to those of previous years, lustrums, decades, centuries.
(7)	*[L]os indígenas aquí en Nayarit, principalmente los huinoricami con los que llevamos una relación de hace años, pueden contar lo que significa en*	[T]he indigenous people here in Nayarit, especially the Huinoricami with whom we have had a relationship for years, can tell you

	*México ser indígena y lo que significa padecer el **desprecio**, la humillación, y la muerte por el color, por la lengua, por la cultura, por el modo, decimos nosotros.*	what it means to be indigenous in Mexico and what it means to suffer **contempt**, humiliation and death because of one's colour, language, culture and way of life, as we say.
(8)	*Somos nosotras la mujeres que sufrimos y enfrentamos más con la presencia militar, aparte de eso desde hace muchos años sufrimos el olvido, el **desprecio** y la marginación por los malos gobernantes [...].*	It is us women who suffer and face more with the military presence, apart from that for many years we suffered the oblivion, **contempt** and marginalization by the bad rulers [...].
(9)	*Pero el EZLN [...] hace una descripción de lo que define a ese sistema social, al capitalismo, y es una relación económica. Esta relación se está basando sobre cuatro ejes fundamentales: es producto del robo o el despojo, producto de la explotación, producto del **desprecio**, del racismo hacia el diferente y producto de la represión; son los cuatro ejes que define.*	But the EZLN [...] makes a description of what defines this social system, capitalism, and it is an economic relationship. This relationship is based on four fundamental axes: it is the product of robbery or dispossession, the product of exploitation, the product of **contempt**, of racism towards the different and the product of repression; these are the four axes it defines.
(10)	*Ya no hay nada ya de que confiar en el capitalismo. Absolutamente nada. Ya lo vivimos cientos de años su sistema, ya las padecimos sus 4 ruedas del carruaje del capitalismo: la explotación, la represión, el despojo y el **desprecio**.*	There is nothing left to rely on in capitalism. Absolutely nothing. We have already lived its system for hundreds of years, we have already suffered its 4 wheels of the capitalist chariot: exploitation, repression, dispossession and **contempt**.

Table 7: *Examples of the word desprecio in context in the EZLN corpus.*

Excerpt 6 relates epistemological violence enacted against indigenous people through the history conveyed in textbooks and museums with (para)military violence. The phrase 'It was not just words' voices the role of cultural and epistemological oppression in the violence suffered by indigenous communities of Chiapas. The link between contempt and material violence does not only apply to the Maya people of Chiapas, but other indigenous peoples of Mexico as well (Excerpt 7). Indigenous women in particular are portrayed as victims of various interconnected layers of oppression (Excerpt 8). Excerpts 9 and 10 show that oppression and contempt are presented as two of the foundations of capitalism. These examples demonstrate that the semantic proximity between these terms mirrors a theoretical premise proposed by the Zapatistas.

This premise is illustrated by the semantic similarity between capitalismo and desprecio, which is twice as large in the EZLN corpus than in the CEDEMA corpus. This trend does not apply to destrucción and despojo, which are roughly equally close to capitalismo in both corpora. In the CEDEMA corpus, capitalism is unsurprisingly closely associated with destruction and dispossession. However, the tie to contempt is specific to the EZLN corpus, suggesting that this relationship contributes to the Zapatistas' resemanticisation

of leftist insurgent language. By producing a semantic continuum between contempt, dispossession, and capitalism, they create a theoretical and semantic co-dependency between epistemological, economic and military violence. The meaning of capitalism, in the Zapatista corpus, is also re-negotiated to expand beyond an economic system and include subjective experiences of oppression.

4.3. The Vocabulary of Ideology

The SDKs clase ('class') and pensamiento ('thought') illustrate an explicit departure from the dominant language and ideology of Marxist Latin American leftist insurgent discourse. Clase is most commonly used in the context of the phrases 'lucha de clase' ('class struggle'), 'conciencia de clase' ('class conscience') and 'clase trabajadora' ('working class') in the CEDEMA corpus but used in the phrase 'clase política' ('political class') in 56% of its occurrences in the EZLN discourse. The activism proposed by the EZLN is intersectional and the redefinition of the word clase to talk nearly exclusively about the ruling class is part of an abandonment of stereotypical Marxist vocabulary, symptomatic of a detachment from past guerrilla movements.

NN_{EZLN}	Translation	Cos. sim.	NN_{CEDEMA}	Translation	Cos. sim.
lucha_EZLN	struggle (EZLN)	0.4186	marxismo	Marxism	0.6687
mundo_EZLN	world (EZLN)	0.4126	ideología	ideology	0.6248
pueblo_EZLN	people/village (EZLN)	0.3971	leninismo	Leninism	0.6191
zapatista	Zapatista	0.3834	gonzalo	Gonzalo	0.6045
palabra	word	0.3814	marxista	Marxist	0.5978
corazón	heart	0.3660	idea	idea	0.5975
indígena_EZLN	indigenous (EZLN)	0.3608	comunista	communist	0.5834
organización_EZLN	organisation (EZLN)	0.3517	partido	party	0.5617
país_EZLN	country (EZLN)	0.3510	maoísmo	Maoism	0.5542
campaña_EZLN	campaign (EZLN)	0.3508	línea	line	0.5510

Table 8: *Ten nearest neighbours of the vector for pensamiento in the EZLN corpus and the CEDEMA corpus.*

The keyword pensamiento ('thought') mirrors both the abandonment of the discourse of Marxist ideology and the alternative mode of action proposed by the Zapatistas. It is strongly associated with political ideology in the CEDEMA corpus. The ten nearest neighbours include the word ideología ('ideology') itself, the word partido ('party'), and words alluding to several branches of communist ideology (see Table 8). In the EZLN corpus, the term appears to have a much more diffuse meaning. First, most occurrences refer to personal or collective thoughts which are not ideological in nature (see Table 9). While pensamiento is often linked to external influence in the CEDEMA corpus (e.g.

Marx, Abimael Guzmán, Lenin, Mao), it is most often emanating from the Zapatistas or the collectives they are dialoguing with in the EZLN corpus. Second, there is an emphasis on sharing one's thoughts and listening to others' (see Excerpt 11). Often, the term is linked to speech as well as insurgency or struggle. In fact, the Zapatistas create a semantic continuum between thoughts, words and weapons. Thoughts are imbued with a power to enact change (see Excerpts 12, 13, 14 and 15). In other words, thoughts are not merely abstractions, but are equated to actions.

	KWIC (sentence)	Translation
(11)	*También, como zapatistas, seguimos abriendo el corazón y el oído para el **pensamiento** de quien con nosotros lucha. [...] Con ese **pensamiento** compañero estamos preparando nuestros siguientes pasos.*	Also, as Zapatistas, we continue to open our hearts and ears to the **thoughts** of those who fight with us. […] With that companion thought, we are preparing our next steps.
(12)	*Hace unas horas ustedes han escuchado las palabras que escribió el comandante Tacho. En ellas les decía que las palabras y **pensamientos** de ustedes son como armas, como buenas armas, que tienen que aprender a usar esas armas para saberlas dirigir adonde se debe y no herir al compañero.*	A few hours ago you heard the words written by commander Tacho. In them he said that your words and **thoughts** are like weapons, like good weapons, that you have to learn how to use these weapons in order to direct them where they should be directed and not to hurt your comrades.
(13)	*Sólo lo que hagamos nosotras, nosotros, cada quien según su calendario y su geografía, según su nombre colectivo, su **pensamiento** y su acción, su origen y su destino.*	Only what we do, we, each according to our calendar and geography, according to our collective name, our **thought** and action, our origin and destiny.
(14)	*Ha llegado ese momento de unirnos, de juntar nuestro **pensamiento**, de juntar nuestro corazón, de juntar nuestras luchas, de juntar nuestras demandas, pero solamente así entre todos vamos a cambiar la realidad.*	The time has come to unite, to unite our **thoughts**, to unite our hearts, to unite our struggles, to unite our demands, but only in this way we will be able to change the reality.
(15)	*El **pensamiento** que no lucha , nada hace más que ruido.*	The **thought** that does not fight, does nothing but noise.
(16)	*El pueblo tiene que volver otra vez a ser pueblo como antes, a gobernarse de acuerdo a sus propios **pensamientos**, a sus propios ideales, de acuerdo a su cultura, de acuerdo a lo que es el **pensamiento** del pueblo indígena.*	The people must once again become a people as before, to govern themselves according to their own **thoughts**, their own ideals, according to their culture, according to what the **thinking** of indigenous people is.

***Table 9:** Examples of the word pensamiento in context in the EZLN corpus.*

This drastic semantic variation in the word pensamiento is correlated to the Zapatistas' partial rejection of theoretical thinking, which they associate with a lack of attention to lived realities. Subcomandante Marcos compares the 'pensamiento de arriba' ('thinking

from above') to a stone which would be thrown in a pond and whose ripple would not reach the shore, i.e. reality. He suggests that the irrelevance of reality in theoretical thinking can be linked to the primacy of idea over matter proposed by Descartes. For him, 'I think therefore I am' places the self at the centre and the others in an irrelevant periphery unaffected by the perception of that self (Marcos, 2007). Marcos' distrust of Descartes and theories which privilege ideas over the lived circumstances of the other echoes the writings of several decolonial philosophers. Nelson Maldonado-Torres, for example, also associates Cartesian philosophy with a form of individualism which prevents dialogue (Maldonado-Torres, 2006, 187). The link between the Cartesian self, centred on individual perception, and epistemological violence has been discussed extensively. Ramón Grosfoguel argues that the gap between philosophy and lived realities, which Marcos denounces, is linked to Western man's propensity to consider his truth universal. He contends that, within Western philosophy, the location of a claim is often dismissed: the ego is taken away from its body and materiality, that is, from its ethnic, gendered and sexual epistemological location, which results in the production of 'Truthful Universal knowledge' (Grosfoguel, 2009, 15). Grosfoguel's diagnosis mirrors Marcos' denunciation of 'the primacy of idea over matter'.

The semantic contestation at play around the word pensamiento exposes the Zapatistas' distrust of ideology and mirrors a decolonial approach to knowledge. Pensamiento is not only more concrete and able to intervene within reality, but it is also more situated and context-dependant. In addition, thinking, like speech and actions, comes from within rather than being external to the (collective) self. Once again, the renegotiation of the term pensamiento happens in relation to other concepts by entering a conceptual continuum with the word lucha ('struggle') and palabra ('word').

5. Discussion and Future Works

5.1. Sense disambiguation

SDKs point to areas of semantic contestation when comparing two or more corpora. However, differing worldviews or ideologies are only some of the factors which can account for semantic variations. As mentioned above, the list of SDKs was filtered to exclude words whose semantic variation could be explained by their use within group specific phrases.[14] To address this issue, the recognition of Named Entities will be a useful pre-processing step. As a consequence, it will be possible to differentiate the semantic representation of the term as used within the Named Entity vs. elsewhere in the corpus. In the context of this analysis, for instance, it would be convenient to represent the word acción ('action') without including the occurrences of 'Partido de Acción Nacional' ('National Action Party'), and the word liberación ('liberation') separately from the occurrences of 'Ejército Zapatista de Liberación Nacional' ('Zapatista Army of National Liberation').

14 Distinct language varieties (including dialects, registers and styles) can also account for high semantic differences.

Likewise, identifying relevant bigrams and trigrams would allow for the semantic representations of phrases such as 'sociedad civil' ('civil society') or 'lucha de clase' ('class struggle'), which would lead to more granular results, since the semantics of one word would be divided into several representations.

Differentiating between word senses would also be productive, since polysemy often leads to high semantic variation based on the sense favoured by different groups (e.g. the second SDK is the word dirección which can be translated as 'address', 'management' and 'direction'). However, favoured word-senses can be insightful in themselves. For instance, principio is used most often in the sense of 'values' or 'norms' in the CEDEMA corpus and in the sense of 'beginning' in the EZLN corpus. This difference is symptomatic of the Zapatistas' rhetoric, which is based on revolutionary practices more than revolutionary principles, but it also reflects the more narrative and oral writing styles adopted by Zapatista representatives.

Finally, although this method requires that both corpora are learned together for the SDKs to be computed, the nearest neighbour calculations would be more robust if they were extracted from separate models. In the current architecture, the nearest neighbours of an EZLN vectors could technically be a word which is never used in the EZLN corpus or used widely differently in the two corpora. If they are not among the list of candidate SDKs (and, therefore, not contextually referenced), the lack of granularity obscures the interpretability of these lists. Extracting nearest neighbours from separate models, corpus-size permitting, would 'disambiguate' these context words and therefore improve their interpretability.

5.2. Negative SDKs

Although this case study focuses on areas of semantic contestation, negative SDKs, i.e. words which are semantically the closest across corpora, also provide valuable insights on the areas of semantic continuities and similarities. In this case, mano ('hand'), sangre ('blood'), tierra ('land'/'earth'), dinero ('money') and riqueza ('wealth') are the top five negative keywords. For each of these, the nearest neighbour for the EZLN vector is the corresponding vector in the CEDEMA corpus. Conversely, the EZLN vector is in the top five nearest neighbours of the CEDEMA vector.

5.3. Beyond Binary Comparisons

This study focuses on the EZLN and its alleged renovation of the discourse of Latin American leftist insurgency. However, by limiting the contextual referencing to the EZLN corpus, the potential of the methodology is limited. In future works, potential SDKs will be referenced for all movements (frequency permitting) and divided into three periods informed by historical research on Latin American leftist guerrilla movements (Wickham-Crowley, 2014). Some movements have been active for several decades and significantly evolved over time. This more granular referencing will be used to identify ideological clusters as well as patterns of continuity and rupture in the discourse of insurgency in Latin America (Chasteen, 1993). From a methodological viewpoint, by calculat-

ing all pairwise semantic similarities for potential SDKs, we will be able to extract keywords which are most susceptible to semantic variability across the board, rather than focusing on one movement. In addition, when focusing on one movement, it will be interesting to look at words whose pairwise distances are abnormally large in comparison to the pairwise distances involving the other movements. For instance, we would be able to ask whether the rejection of the word revolution extends to other Mexican insurgent organisations.

Moreover, relying on contextual embeddings, i.e. models for which an embedding is created for each occurrence of a term (Periti & Montanelli, 2024; Wiedemann & Fedtke, 2022), would be productive for this area of research, since it would allow for the assessment of the stability of word meaning within one (sub-)corpus as well as across different corpora. It would be possible to examine whether the semantic difference observed leads to a 'shrinking' or 'expansion' of the meaning. For instance, with contextual embeddings, we could verify the hypothesis that the high semantic difference between the word revolución in the EZLN and the CEDEMA corpus is due to a strong reduction of the semantics of the word.

6. Conclusions

Due to their well-documented appeal, Zapatista writings are relevant beyond the boundaries of Chiapas. As noted by Popke (2004), Marcos' communiqués 'reflect and contribute to, through their broader engagement with global civil society, the development of a new conception of social and cultural agency, within which a different form of ethics and politics is at stake'. In this article, SDKs were extracted and interpreted to identify the areas in which these new conceptions are semantically negotiated. By analysing the vocabulary of insurgency among the SDK list, the radical rejection of the word revolución was uncovered, alongside a subjectivisation of the word rebelión. SDKs related to oppression revealed an area of semantic struggle surrounding the meaning of capitalism, which is imbued with a strong connotation of epistemological violence in the EZLN corpus. Finally, the semantic contestation of the word pensamiento in a leftist insurgent context was found to mirror the Zapatista rejection of theoretical knowledge rooted in a decolonial understanding of the primacy of ideas over lived reality in Western philosophy. Moreover, the semantic shift undergone by this word highlights the Zapatista engagement with a plurality of voices, as well as the creation of a semantic continuum between thoughts and subversive processes. A common denominator throughout this analysis is the 'materialisation', as well as the 'subjectivisation' of the vocabulary surrounding insurgency. In comparison with the CEDEMA corpus, words move away from the theoretical plane to enter the lived reality and subjectivity of the people involved in the movement. This subjectivity, in opposition to the Cartesian subject (as understood by Marcos and Grosfoguel), erodes the boundaries between thoughts and actions, as well as between ideas and reality.

In addition to furthering the understanding of the Zapatista rhetoric, this study demonstrates that the automated extraction of SDKs allows for the efficient identification of sites of semantic contestation between two or more corpora. To assess whether the resulting keyword list significantly overlaps with the terms which would be deemed most semantically contested by subject experts would require the development of annotated datasets. Considering the subjectivity of the matter, such datasets would be complex to develop. Nevertheless, a comparatively large cosine distance between vectors pairs undoubtedly signifies a form of contextually dependant semantic difference which can then be interpreted by close-reading relevant text excerpts. This method also highlights the fact that word meanings are not negotiated in isolation, but in relation to other words. Although this is self-evident, this study demonstrates the ricochet effect of semantic shifts, which could be compared to a ripple effect within the semantic space. This means that once a semantically contested word is identified, a network of renegotiated semantic relationships can be identified by 'pulling on the thread'.

Acknowledgements.

This work was supported by a FED-tWIN grant (Prf-2020- 026 KBR-DLL) funded by BELSPO (Belgian Science Policy).

Competing interests

The authors have no competing interests to declare.

References

Antoniak, M., & Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107–119. https://doi.org/10.1162/tacl_a_00008

Brigadir, I., Greene, D., Cunningham, P. (2015). Analyzing discourse communities with distributional semantic models. *Proceedings of the ACM Web Science Conference*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2786451.2786470

Burbach, R. (2001). *Globalization and postmodern politics: From zapatistas to high tech robber barons*. Pluto Press.

Carr, B. (1997). From the mountains of the southeast: A review of recent writings on the zapatistas of chiapas. *Journal of Iberian and Latin American Studies*, *3* (2), 109–123.

Castells, M. (1997). *The information age: Economy, society and culture. Volume ii: The power of identity*. Blackwell Publishers.

Chasteen, J.C. (1993). Fighting words: The discourse of insurgency in Latin American history. *Latin American Research Review, 28* (3), 83–111. https://doi.org/10.1017/S002387910001696

Chen, W. (2019). Towards a discourse approach to critical lexicography. *International Journal of Lexicography, 32*(3), 362-388. https://doi.org/10.1093/ijl/ecz003

Conant, J. (2010). *A poetics of resistance: The revolutionary public relations of the zapatista insurgency*. AK Press.

Dénigot, Q., & Burnett, H. (2021). Using word embeddings to uncover discourses. *Proceedings of the society for computation in linguistics* (pp. 298–312).

Dubossarsky, H., Hengchen, S., Tahmasebi, N., Schlechtweg, D. (2019). Time out: Temporal referencing for robust modeling of lexical semantic change. A. Korhonen, D. Traum, & L. Márquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 457–470), https://doi.org/10.18653/v1/P19-1044

Garg, N., Schiebinger, L., Jurafsky, D., Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115 PMid:29615513 PMCid:PMC5910851

Gribomont, I. (2019). The zapatista linguistic revolution: A corpus-assisted analysis. *Discourses from Latin America and the Caribbean: Current Concepts and Challenges*, 139–171. https://doi.org/10.1007/978-3-319-93623-9_5

Gribomont, I. (2023). From diachronic to contextual lexical semantic change: Introducing semantic difference keywords (SDKs) for discourse studies. *Proceedings of the 4th workshop on computational approaches to historical language change* (pp. 153–160). https://doi.org/10.18653/v1/2023.lchange-1.16

Grosfoguel, R. (2009). A decolonial approach to political-economy: Transmodernity, border thinking and global coloniality. *Kult*, *6*(1), 10–38.

Harris, Z.S. (1954). Distributional structure. *WORD, 10* (2-3), 146-162. https://doi.org/10.1080/00437956.1954.11659520

Holloway, J. (1998). Dignity's revolt. J. Holloway & E. Peláez (Eds.), *Zapatista!: Reinventing revolution in Mexico* (pp. 159–198). Pluto Press.

Holloway, J. (2005). Zapatismo urbano. *Humboldt Journal of Social Relations*, *29* (1), 168–178,

Jeffries, L., & Walker, B. (2017). *Keywords in the press: The new labour years*. Bloomsbury Publishing.

Khasnabish, A. (2010). The international order of hope: Zapatismo and the fourth world war. M. Blaser, R. De Costa, D. McGregor, & W.D. Coleman (Eds.), *Indigenous peoples and autonomy: Insights for a global age* (pp. 229–240). University of British Columbia Press. https://doi.org/10.59962/9780774817943-012

Kilgarriff, A. (2009). Simple maths for keywords. *Proceedings of Corpus Linguistics Conference (CL 2009).*

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Su chomel, V. (2014). The sketch engine: ten years on. *Lexicography*, *1*, 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., & Renau, I. (2013). esTenTen, a vast web corpus of peninsular and American Spanish. *Procedia-Social and Behavioral Sciences, 95*, 12–19. https://doi.org/10.1016/j.sbspro.2013.10.617

Le Bot, Y. (1997). *El sueño zapatista por subcomandante marcos*. Anagrama.

Maldonado-Torres, N. (2006). Aimé césaire y la crisis del hombre europeo. A. Césaire (Ed.), *Discurso sobre el colonialismo* (pp. 173–203). Akal.

Marcos, S. (2007). *Coloquio aubry. Parte I. Pensar el blanco.* Retrieved from https://enlacezapatista.ezln.org.mx/2007/12/13/conferencia-del dia-13-de-diciembre-a-las-900-am/

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. *International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26). https://doi.org/10.48550/arXiv.1310.4546

Periti, F., & Montanelli, S. (2024). Lexical Semantic Change through Large Language Models: a Survey. ACM Computing Surveys. https://doi.org/10.1145/3672393

Plu-Jenvrin, R. (2019). Los años de la 'revolución institucionalizada' en México. Políticas de imagen y contenidos visuales de una construcción institucional (1940- 1960). *Revista de El Colegio de San Luis, 9*(20), 367–406, https://doi.org/10.21696/rcsl92020191001

Popke, E.J. (2004). The face of the other: Zapatismo, responsibility and the ethics of deconstruction. *Social & Cultural Geography, 5*(2), 301-317. https://doi.org/10.1080/1464936040010690277

Řehůřek R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis, 28*(1), 112–133, https://doi.org/10.1017/pan.2019.26

Rosset, P., Martínez-Torres, M.E., Hernandez-Navarro, L. (2005). Zapatismo in the movement of movements. *Development*, *48*(2), 35–41, https://doi.org/10.1057/palgrave.development.1100139

Schröter, M. (2008). Discourse in a nutshell: Key words in public discourse and lexicography. *German as a foreign language* (2), 42–57.

Wickham-Crowley, T. (2014). Two 'waves' of guerrilla-movement organizing in Latin America, 1956–1990. *Comparative Studies in Society and History*, *56*(1), 215–242, https://doi.org/10.1017/S0010417513000674

Wiedemann, G., & Fedtke, C. (2022). From frequency counts to contextualized word embeddings: The saussurean turn in automatic content analysis. *Handbook of computational social science, volume 2.* Taylor & Francis. https://doi.org/10.4324/9781003025245-25

Würschinger, Q., & McGillivray, B. (2024). Semantic change and socio-semantic variation: the case of covid-related neologisms on reddit. *Linguistics Vanguard*, https://doi.org/10.1515/lingvan-2023-0106

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.-W. (2019). Gender bias in contextualized word embeddings. J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 629–634). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1064