A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)

Fernández, Elena Fernández; Savcisens, Germans

Published in:
Proceedings of Digital Humanities in the Nordic and Baltic Countries 2023. Sustainability, Environment, Community, Data.

Link to article, DOI:
10.5617/dhnbpub.10660

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)

Elena Fernández Fernández[1], Germans Savcisens[2]

[1]*Institute of Computational Linguistics (Univesity of Zurich), Andreas Strasse 15, Zurich, 8050, Switzerland*
[2]*Section for Cognitive Systems (Technical University of Denmark), Building 324, Kongens Lyngby 2800, Denmark*

Abstract
In this article, we analyze the temporal and geographic evolution of sustainability-related discourses over a time frame of twenty years (1999-2018). We use a collection of multilingual newspapers in English, French, German, Spanish, and Italian, as a proxy. We filter documents using four key terms: sustainable development, climate change, environment, and pollution, seeking to explore how different newspapers encode the same message, aiming to detect points of contact (agreement) and rupture (polarity). Our methodology includes Topic Modelling (Pachinko Allocation [1]), word embeddings [2], Ward's hierarchical cluster analysis [3], and network analysis [4]. Our results show a progressive simplification of semantic fields over time, reflecting less polarizing views across countries and, therefore, showing an increasing agreement on sustainability-related discourses in our contemporary societies. Moreover, we also notice little variation of newspapers rhetorics over time. Therefore, this article also contributes with a meta-reflection about newspapers behaviour as information containers.

1. Introduction

The United Nations 2030 Agenda for Sustainable Development (United Nations [5]), organized into seventeen sustainable development goals, is a clear indicator that signals contemporary concerns about the necessity of taking various measures in the present to ensure a well-balanced growth of society. While relatively new, discourses about sustainability have received critical attention across different fields (Drucker [6]). However, we believe that the analysis of the historical and geographical evolution of sustainability-related concepts has not been accomplished yet with enough granularity.

In this article, we address that research gap by using a multilingual dataset of contemporary newspapers in English (*The Times, The New York Times International, Chicago Daily Herald, The Irish Times*), German (*NZZ*), French (*Le Figaro*), Italian (*La Stampa*), and Spanish (*El País*) over a period of twenty years (1999-2018). Our goal is to analyze the temporal evolution of geographic clusters of public opinion in Western societies seeking to detect points of convergence and

DHNB2023: Digital Humanities in the Nordic and Baltic Countries 2023. Sustainability, Environment, Community, Data. March 08-10, 2023, Online Conference
elena.fernandezfernandez@uzh.ch (E. F. Fernández); gersa@dtu.dk (G. Savcisens)
https://elenafernandezfern.wixsite.com/elena-fernandez (E. F. Fernández);
https://carlomarxdk.github.io/portfolio/ (G. Savcisens)
0000-0001-6596-6349 (E. F. Fernández); 0000-0002-5811-3230 (G. Savcisens)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

rupture. To do so, we first filter our dataset by selecting documents that contain three key terms aligning with three of the seventeen United Nations Sustainable Development Goals. Thus, we firstly filter documents using the key terms climate change (which is related to Goal Thirteen, Climate Action), pollution (which is similar to Goal Six, Clean Water, and Sanitation, and to Goal Twelve, Responsible Consumption and Production), and environment (related to Goal Fifteen, Life on Land). We have selected these terms as an initial exploratory approach to analyzing press coverage on sustainability discourses. On top of those three key terms that are semantically related to three of the United Nations Sustainable Development Goals, we also select the concept of sustainable development to inspect the discussion about this topic in a broader context.

To uncover the temporal variations in data, we split the dataset into epochs of five years (1999-2003, 2004-2008, 2009-2013, and 2014-2018). We implement a three-step methodology consisting of Topic Discovery, Topic Evolution Analysis, and Sentiment Analysis. The first two aim to capture the contextual patterns across time and space, while the last one aims to identify the sentimental direction of the discourse.

This article contributes to establishing intellectual bridges between the emerging field of Environmental Humanities, big data and computational methods, and current efforts to create a sustainable world. Environmental Humanities acts as an umbrella term that gathers a variety of interdisciplinary research lines aiming to demystify beliefs that nature crises are a one-sided consequence of technology, therefore ignoring the role of history and culture as actors in that process (Heise et al. [7]): "The environmental humanities, by contrast, envision ecological crises fundamentally as questions of socioeconomic inequality, cultural difference, and divergent stories, values, and ethical frameworks". (2). In the following pages, we will analyze quantitatively public discourses about sustainability in Western societies under that theoretical framework using newspapers as a proxy.

While our work is informed by both Habermas [8] (who considers newspapers as the embodiment of the public sphere where private citizens evaluate government actions and make informed decisions about who to vote in democratic elections) and Herman and Chomsky [9] (who, on the other hand, propose the so-called "propaganda model," that states that during the last decades, private corporations and governments have progressively taken control over mass media worldwide), we do not wish to participate in that debate. However, we are very interested in observing how citizens living in diverse Western societies are exposed to information similarly or differently over time. We believe that an empirical analysis of cultured discourses about sustainability may be beneficial to gain a better understanding of the social effect of the press, laying the ground for further analyses across sectors. In doing so, we dialogue with Hay et al. [10], who state that knowledge is a driver of human action towards sustainability. In this article, we explore the temporal and geographic evolution of those discourses in Western societies to assess the footprint of both culture and zeitgeist in environmental communication. Moreover, we reflect about newspapers as containers of information, arguing that big data and computational methods have the capacity to shed light in a yet unseen manner about the segment of reality they represent.

2. Related Work

State of the art across disciplines has analyzed qualitatively and quantitatively discourses about sustainability under a variety of theoretical perspectives using different corpora. Qualitatively, (Philippon [11]) analyzes sustainable food rhetorics (i.e., the slow food movement), (Lenz [12]) explores points of convergence and divergence in discourses about digital technologies and the future of sustainability internationally and across sectors, and (Beling et al. [13]) discusses tensions in Ecological Economics sustainable development discourses between the Euro-Atlantic cultural region and the Global South north. Moreover, in the field of sustainability communication (defined by (Godemann and Michelsen [14]) as the development of critical awareness about the relationship between humans and their environment in social discourse), Brand [15] observes different frames of sustainability discourses from a sociological perspective in Germany, (Kruse [16]) analyzes human behaviour towards sustainability under the frame of environmental psychology, and (Witt [17]) explores Media Theory and Sustainability Communication. We believe that our work complements these approaches by focusing on the discursive aspects of sustainability (yet using distant reading methodologies). Environmental Humanities is a relatively recent area of scientific interest, and, to our knowledge, there are not many studies focusing on the historical evolution of sustainability rhetorics. In this paper, we seek to fill that research gap.

Quantitatively, Barkemeyer et al. [18] use 115 leading national newspapers from 39 different countries to inspect public opinion of sustainability concepts in the field of finance - their data covers publications over the eighteen-year period, i.e., 1990-2008. They filter the corpus using key terms such as sustainability, sustainable development, and business ethics, as well as a variety of tokens in the semantic family of business responsibility. Their methodology quantifies the number of appearances of key terms historically as a proxy for collective attention of environmental-related discourses, noting a clear increment over time. They continue this line of work in Barkemeyer et al. [19], this time inspecting different media coverage between the Global North and South of sustainability-related terms. Their corpus consists of articles from 115 multilingual newspapers based in 41 countries in eight different languages using the year 2008 as an observational time. They find significant differences in coverage of the selected tokens in both regions, reflecting diverging public policy sustainability agendas in developed and developing countries. Similarly, Kumar and Das [20] analyze sustainability reports of 200 firms internationally between 2008 and 2017. They quantify the number of appearances of 208 key terms related to sustainability in their corpus, noting a general trend of an increasing presence over countries across time. In the area of Media and Communication studies, Fischer et al. [21] analyze the semantic evolution of the terms sustainable and sustainability in German (choosing the key term *nachhaltig* in six German newspapers between 1995 and 2015. Results show an increasing number of mentions of the key terms across newspapers, as well as a progressive semantic technification of the sustainability field as articles approached contemporary times. We are interested in dialoguing with these lines of work, and we further investigate these findings by extending their observational time (we will use two decades, 1999-2018) as well as by specifically inspecting points of contact (similar encoding across multilingual newspapers of a same environmental-related concept) and rupture (different expression of a same message) in our dataset.

In the areas of Environmental Management and Public Policy, Sebestyén et al. [22] use the Voluntary National Reviews (mandatory reports that countries worldwide must accomplished after the declaration of the UN Sustainable Development Goals) of 75 countries in English published between 2016 and 2021. Implementing network analysis, they discover clear geographic clusters of data, identifying a Southern European cluster composed of Italy, Spain, Montenegro, Greece, and Portugal; a Northern European one formed by France, Germany, and Sweden; and two less clearly geographically connected ones, yet still in close proximity (a first one containing Samoa, Egypt, and Azerbaijan, while a second one including The Maldives, Georgia, India, Belize, Turkey, and Tajikistan). In a similar line of research (and of particular interest for this paper) is the work of Hallin and Mancini [23], where the authors propose the existence of geographic clusters of media models rooted in different historical, economic, and political traditions. They identify three main regions of media behaviour. Firstly, they point out to a Mediterranean knot, composed of Spain, Greece, and Portugal (sharing very similar contemporary history patterns of military dictatorships during the twentieth century and, therefore, a late-blooming of democracy), Italy, and to a lesser degree France. Secondly, they identify a North-Central European region composed of Scandinavia, the low countries, Germany, Austria, and Switzerland, claiming a similar shared recent history as well as many cultural, economic, and political traits. Thirdly, they identify a third media area, the North-Atlantic Anglo-speaking region, formed by the UK, Ireland, the United States, and Canada, following a similar line of reasoning.

Big data and computational methods, under the analytical frame of environmental humanities, can enrich these academic discussions by providing new findings to inspect those two research dimensions (multilingualism and historical perspective). In this paper, we seek to identify a possible existence of geographic clusters of media discourses in Western societies as well as their plausible temporal evolution during the last twenty years. We believe that doing so will provide innovative perspectives about the dialectical relationship between public opinion and sustainability discourses, shedding light on their reversed influence across time and space. Moreover, we contribute with cutting-edge discoveries to current discussions that investigate the historic multicultural evolution of the press in Western societies. Indeed, our main contribution consists of a novel observation of the textual structure of sustainability-related discourses in newspapers over the last twenty years. We are able to state whether sustainability rhetorics have become more fragmented (showing geographic polarization) or more homogeneous (showing an increasing international agreement) over time, as well as observing patterns of information diversity and emotions in the press of our selected countries.

3. Dataset

Our multilingual dataset is composed of six different newspapers of record (*Le Figaro, El País, La Stampa, NZZ, The Times, Irish Times*, one international newspaper (*The New York Times International*), and one local newspaper (*Chicago Daily Herald*). We believe that it is interesting to compare and contrast discourse dynamics across those three sorts of news outlets. Newspapers of record are considered relatively neutral vehicles of public opinion (i.e., leaning towards center political views) and, therefore, understood as non-partisan units of recorded history. *The New*

Table 1
Number of articles per newspaper (for the period 1999-2018)

The Times	The NYT	Chicago	Irish Times	Le Figaro	El Pais	La Stampa	NZZ
48985	26181	23123	43341	32545	48819	37616	23990

York Times International (previously known as *The International Herald Tribune*), is a newspaper that, from its inception, was conceived as an international news outlet focusing on international, economic, and cultural news (Sterling [24], Greenslade [25]). *Chicago Daily Herald* is a local newspaper yet headquartered in a major city with an international economic weight. We think that, for knowledge discovery purposes, it would be relevant to observe points of contact and rupture in sustainability-related discourses not only geographically and temporally but also across different news outlets publications, and editorial policies.

Our corpus has a total of 261477 documents, consisting of 118507 in English, 32545 in French, 48819 in Spanish, 37616 in Italian, and 23990 in German. We use the database Datalab, powered by LexisNexis, to retrieve the data. Each article includes a title, date of publication, as well as the full text (and other accompanying metadata). Table 1 shows the number of documents per newspaper.

Datalab provides access to a variety of newspapers internationally in more than twenty languages with a time coverage of approximately thirty years (1990-2020, although each newspaper presents a different time availability). Datalab is a news repository linked to a LexisNexis owned Jupyter Notebooks environment where users can execute data analyses using either Python or R programming languages. Datalab provides access to all the articles whose publication rights are owned by the newspapers they host. But, articles written by freelancers are not owned by the newspaper, and thus, they do not appear in the database. Consequently, the number of yearly articles available in the database may not necessarily coincide with the real number of yearly articles published by newspapers. While we are aware that this may represent a bias in our data analysis, we still believe that we can observe general trends in the geographic and temporal evolution of sustainability discourses.

We filter each newspaper using four key terms: Sustainable Development, Climate Change, Pollution, and Environment. That means that a) we create four different datasets per newspaper (each dataset matches one of the four key terms), and b) that we only select documents that contain those key terms. We use each word's native language root form (i.e., stem) to filter the documents (Porter [26]). Table 2 shows each word with its corresponding root. We have followed two different criteria to translate each term into different languages. Firstly, we have consulted how multilingual versions of Wikipedia articles translate our selected key terms in different languages. For example, to infer how climate change is translated into French, we consult the English Wikipedia page of climate change and then select the French version of the same article seeking to observe how Wikipedia users express the same concept in French. Oftentimes, we encounter that there are several synonyms of the same concept. For instance, we note that there are three options to translate climate change into French: *réchauffement climatique, changement climatique, dérèglement climatique.* Aiming to decide which one of them is most semantically relevant in press discourses, we filter our corpus with each one in order to observe the one that produces the higher number of documents. In the case of French,

Table 2
Table of filtering queries for each language and key term

	Queries			
Terms	***Sustainable Development***	***Pollution***	***Climate Change***	***Environment***
EN	sustain! AND develop!	pollut!	climat! AND chang!	environ! AND natur!
FR	développ! AND souten!	pollut!	chang! AND climat!	environ! AND naturel!
IT	svilupp! AND sosten!	inquin!	riscald! AND climat!	ambient! AND natural!
ES	desarroll! AND sosten!	contamin!	cambi! AND climat!	medi! AND ambient! AND natural!
DE	nachhalt! AND entwickl!	verschmutz!	klimawandel!	natur! AND umwelt!}

changement climatique is the most used in *Le Figaro*. In some cases where, after following these two steps, we were still unsure, we consulted with native speakers.

After stemming, we use Datalab procedures to develop a search within their own environment. They facilitate the use of the symbol ! (truncation), that, placed at the end of the stemmed word, includes both its root and inflections. For example, the root of the word pollution is *pollut-*, and followed by the ! sign (that would be, *pollut!*), would return all the inflections of the word such as polluted, pollution, pollutions, etc. Moreover, we use Datalab's operationalization of boolean operators, and we use the word *AND* in their search engine to select documents that contain two words. For the key terms climate change and sustainable development, we need to find articles that include both words. The term environment is a highly polysemous word that appears in very diverse semantic contexts. Therefore, we add the word nature to ensure that we are filtering our corpus efficiently for the kind of research question that we are seeking to explore.

4. Methods

Our pipeline includes five different steps. Firstly, we use the Pachinko Allocation (PA) model (Mimno et al. [1]) to discover inner topics[1] of public opinion. Secondly, we use word embeddings to calculate mean pairwise cosine similarity between words of each topic (Aletras and Stevenson [2]) - this score indicates how similar topics are. Thirdly, we use Hierarchical Cluster Analysis (Ward's linkage function) (Großwendt et al. [3]) to aggregate topics produced by all the newspapers into semantically similar global topics (separately for each epoch). Fourthly, we analyze the temporal and geographical evolution of global topics (Beykikhoshk et al. [4]). Lastly, we analyze the sentiment associated with each global topic (Hutto and Gilbert [27]).

4.1. Inner Topic Modelling

In our work, we split articles into four different epochs: (1) 1999-2003, (2) 2004-2008, (3) 2009-2013, and (4) 2014-2018. Within each epoch, we split the articles based on the newspaper they belong to. As a result, we split data into 24 subsets based on epoch and newspaper. To perform topic modeling, we use PA (Mimno et al. [1]) model on each subset of data separately. For each

[1] inner topics refer to topics that exist within a specific set of articles

subset, the PA model generates a set of topics that exist in that particular subset of articles, and we call these inner topics. These topics represent the dominating context of discourse in a particular set of articles.

We use the Pachinko Allocation model (Mimno et al. [1]) as it has several advantages over other methods, such as LDA-based models (Carlsen and Ralund [28]). It simultaneously models the correlation between words and relations between the topics. Thus, it captures topics that might exist only in a handful of articles (i.e., it is not overwhelmed by the imbalance of articles). To do so, the PA model operates on two levels: sub-topic modeling and super-topic modeling. Each generated topic is a sub-topic that the PA model has discovered. The super-topics are clusters of highly correlated sub-topics (they might share a similar vocabulary but still have distinct distribution over the vocabulary). We are particularly interested in sub-topics. However, we need to specify the number of super- and sub-topics (i.e., k^1 and k^2 parameters, see Mimno et al. [1] for more details) before we fit the PA model. Since we do not know the number of super- and sub-topics for a given dataset, we search for the most optimal set of parameters. We perform a search using Bayesian Optimisation (Head et al. [29]) that goes through many combinations of parameters and converges to the optimal solution.

The optimization algorithm requires a metric to evaluate the quality of generated topics. We use the coherence score as a proxy for the quality (Mimno et al. [30]) - it considers the words co-occurrence in a set of documents. If the co-occurrence of top-N words[2] associated with the topic is low, the coherence score would also be low (i.e., a topic does not capture any semantic relationships between the words, see Aletras and Stevenson [2]).

Algorithm 1 Inner Topic Modelling

for $t = 1 : \mathrm{T}$ **do** ▷ where T is a number of epochs
  for $n = 1 : \mathrm{N}$ **do** ▷ where N is a number of newspapers
    $\tilde{\mathbf{X}}_{t,n} = \mathrm{PreProcess}\,(\mathbf{X}_{t,n})$
    $k^1_{t,n}, k^2_{t,n} = \mathrm{BayesOptimCV}[\,\mathrm{PA}(\tilde{\mathbf{X}}_{t,n})\,]$
    $\boldsymbol{\phi}_{t,n} = \mathrm{PA}\,(\tilde{\mathbf{X}}_{t,n},\, k^1_{t,n},\, k^2_{t,n})$ ▷ where $\boldsymbol{\phi}_{t,n}$ is a **set** of discovered topics
  end for
end for

Each discovered topic in the set, $\boldsymbol{\phi}_{t,n}$, is represented as a vector, $\boldsymbol{\gamma}_{t,n,d}$ that contains probabilities for each word in that topic (where $d \in D_{t,n}$ is an indicator of a topic in a set $\boldsymbol{\phi}_{t,n}$, and $D_{t,n} = k^2_{t,n}$). It provides an overview of the top words associated with the topic (i.e., to analyze the context) and a probability (i.e., to calculate the cosine similarity between different topics).

4.2. Topic Similarity

To cluster similar topics originating from different newspapers, we calculate the similarity score between each pair of topics. We do so separately per each epoch.

[2] the words that have a high probability of being in a particular topic

First, we start by estimating the similarity between inner topics. To make the multilingual topic vectors $\boldsymbol{\gamma}_{t,n,d}$ comparable, we translate words associated with each topic vector (from Italian, Spanish, German, and French) into English using the Google Translate API.

In some cases, a word translates into a phrase. To add these words to a vocabulary, we split the phrase into separate words and equally redistribute the probability associated with the phrase to these separate words. For example, according to the Google Translate *socialdemócrata* translates into *social democrat*, we pass the phrase through our pre-processing pipeline and get *social* and *democrat*. Thus, now the topic that contained *socialdemócrata* has two separate words instead of the initial term. Since we do not know anything about their probabilities and we cannot assume their independence, we equally redistribute the probabilities as $P(socialdemcrata) = P(social) + P(democrat)$, where $P(social) = P(democrat)$.

Second, as the topic similarity measure is based on word embeddings (Aletras and Stevenson [2]), we need to know all the words across every newspaper and epoch – we create a global vocabulary set, V. Global vocabulary consists of all the unique English words across every epoch and every newspaper. Now we can represent each $\boldsymbol{\gamma}_{t,n,d}$ using the global vocabulary. It might happen that a word that is mentioned in one subset of data (e.g., *cat*) might be missing from the other topic (as it was never mentioned in that particular subset of data). In that case, we update the vocabulary of the $\boldsymbol{\gamma}_{t,n,d}$ and assign a probability of 0 to all newly added words. The aligned vectors can now be used to calculate the similarities between topics.

Further, we stack all $\boldsymbol{\gamma}_{t,n,d}$ into a matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times v}$, where p is the total number of topic vectors (across every epoch and newspaper), and v is the length of the global vocabulary. $\boldsymbol{\Gamma}$ contains the probabilities of each word in every discovered inner topic.

Using matrix $\boldsymbol{\Gamma}$, we can extract word embeddings, i.e., numerical representations of words (Aletras and Stevenson [2]). Each word, V_i is represented as a vector where dimensions correspond to topics and values correspond to the probability of word V_i in each inner topic (i.e., the i-th column of $\boldsymbol{\Gamma}$).

We calculate the topic similarity based on the *average pairwise cosine similarity* of the N-top[3] words in each topic (Aletras and Stevenson [2]).

4.3. Global Topic Aggregation

To analyze whether newspapers share discussion points over a specific epoch, we look at the topic similarity. If multiple inner topics are similar enough, we assume they share similar contexts. The similarity is calculated between each pair of topics (produced by all newspapers over the same epoch). We cluster inner topics with the use of the Hierarchical Cluster Analysis (HCA) with Ward's linkage function (Großwendt et al. [3]).

If the similarity between topics is above a certain threshold, the HCA model clusters topics together. However, we do not have any prior knowledge of the optimal threshold value. Thus, we vary the threshold and look at the quality of the formed clusters. We use the Silhouette method (Rousseeuw [31]) as a proxy for the clustering quality. This method estimates whether topics are closer to the members of their own clusters or vice versa. For each epoch, we find a separate optimal threshold value.

[3] $N = 20$ for Sustainable Development and Climate Change; $N = 15$ for Pollution and Environment

To make clusters comparable, we create cluster representations by averaging the representations of inner topics that end up in the same cluster (i.e., the average of the corresponding rows in $\mathbf{\Gamma}$). We further refer to those as global topics.

The vectors associated with global topics are stacked into another matrix $\tilde{\mathbf{\Gamma}} \in \mathbb{R}^{g \times v}$, where g is the total number of global topics. The representation of a word, V_i, is now based on the matrix of the global topics, $\tilde{\mathbf{\Gamma}}$ (i.e., i-th columns, as in the case with the $\mathbf{\Gamma}$).

Algorithm 2 Global Topic Aggregation

$\mathbf{\Gamma} \in \mathbb{R}^{p \times v}$ is a matrix containing all discovered $\boldsymbol{\gamma}_{t,n,d}$ and p is a total number of topics
$\mathcal{A} = \text{distance}(\mathbf{\Gamma}_i, \mathbf{\Gamma}_j)\ \forall\ i,\ j \in \{1, 2, 3, ..., p\}$ ▷ $\mathcal{A} \in \mathbb{R}^{p \times p}$ is a distance matrix
for $t = 1 : T$ **do**
  $\mathcal{A}^t = \text{subset}(\mathcal{A}, t)$ ▷ $\boldsymbol{A}^t$ contains only between topics of epoch t
  $thr_t = \text{Silhouette}(\text{Ward}(\mathcal{A}^t))$
  $\mathcal{V}_t = \text{Ward}(\mathcal{A}^t, thr_t)$ ▷ $\mathcal{V}_t$ contains sets of similar inner topics
  $\boldsymbol{\nu}_t \leftarrow \text{AverageSimilarTopics}(\mathcal{V}_\text{t}, \mathcal{V})$ ▷ $\boldsymbol{\nu}_t$ contain representations of global topics
end for
$\tilde{\mathcal{V}} = \text{stack}(\boldsymbol{\nu}_t\ \forall\ t \in\ \{1, 2..t\})$

4.4. Temporal evolution of topics

We are interested to see how topics change through time – do they disappear, split into multiple discourses, or stay unchanged, etc. To examine the evolution of global topics, we look at the similarities between global topics between the adjacent epochs, t and $t + 1$ (i.e., between topics of different epochs). We calculate similarities based on the above-mentioned average pairwise cosine similarity – this time, we use word embeddings from $\tilde{\mathbf{\Gamma}}$. We draw connections between the topics of the adjacent epochs if their similarity is above a certain threshold, t. Beykikhoshk et al. [4] suggests setting the threshold based on the n-th quantile of the cumulative distribution of similarity scores. When we estimate similarities between topics of every pair of t (current) and $t + 1$ (next) epochs. We order the scores, find the 90-th quantile of the cumulative distribution, and draw arrows between the pair of topics only if their similarity is higher than the 90-th quantile [4].

Due to the multi-source nature of the data, our algorithm might capture some noise. To substantiate the results, we manually inspect and correct the noisy output of the algorithm. First, we manually inspect the connection between topics if the similarity falls within the 85-th and 90-th quantile. We draw the arrow if the topics share at least one top word. Second, we remove global topics that consist only of one newspaper. We also remove topics if the values of the top ten words (in $\tilde{\mathbf{\Gamma}}$) are below 0.03[5]. We then manually inspect topics if the values of the top twenty words (in $\tilde{\mathbf{\Gamma}}$) are below 0.05. This procedure helps to remove topics lacking consistency (such as this topic with the following set of the top ten words: *leave, world, think, time, know, look, life, people, man, want*).

[4] we decided on the quantile by manually inspecting graphs
[5] we decided on the threshold by manually inspecting the results

Based on the incoming and outgoing arrows, the life of the global topics can progress in several ways: birth, evolution, split, merge, or death. It signifies how public attention and ideas are refined or transformed (Beykikhoshk et al. [4]). The birth of a topic is characterized by the absence of the incoming arrows - no topic from the previous epoch has a similar context. The death of the topic would be the reverse of this case – no topics in the future share a similar context.

If a topic has only one outgoing arrow – it evolves. The topic does not undergo a drastic change. If a topic has multiple out-coming arrows - it splits. The successors reuse similar words and context, but it also involves new words, e.g., the context surrounding these words changes. If a topic has multiple incoming connections – its ancestors merge. The context of the ancestors significantly overlaps.

4.5. Sentiment Analysis

We also calculate the sentiment score for each global topic by taking the top fifty words and passing them through the Valence Aware Dictionary and Sentiment Reasoner, VADER (Hutto and Gilbert [27]). The score signifies the sentiment associated with the context of a topic. The produced score varies between -1 (strongly negative sentiment) and 1 (strongly positive sentiment), where 0 stands for neutral. We discretize the score into seven categories (the split is based on the five equally distanced quantiles), i.e., each topic belongs to any of the seven categories.

5. Results and Discussion

After filtering our dataset using the four target key terms and the five-step pipeline, we are capable of observing both the geographic and temporal evolution of sustainability-related discourses. We are as well in a good position to observe how newspapers behave as containers of information, and, indeed, we identify consistent patterns across our four selected key terms.

We use three different metrics in the interpretation of our data analysis: diversification, attention, and geography. We are interested in observing whether topics become more diverse or simplified over time, showing incremental rates of polarity or agreement. To measure topic diversification or simplification, we use as a proxy both the labels of the topics (that we manually annotate) as well as broader semantic categories that we implement to create a second classification of topics qualitatively (and we use same colors inside the boxes of the topics to indicate equivalent subject matters). We observe that all the labels of all the newspapers across topics can be consistently grouped into six to eight semantic categories. We also quantify which topics receive the highest and lowest rates of attention by counting the number of represented newspapers as we seek to observe conflict-affinity trends over time. Finally, we are interested in measuring the geographic distribution of newspapers present in each topic, as we wonder whether it is possible to detect consistent trends of cultural diversity influencing information behaviour and, therefore, to engage with state-of-the-art that argues the existence of an Anglo-Atlantic, Southern and Central European different media tradition models (Hallin and Mancini [23]). We also scrutinize sentiment analysis behaviour to detect temporal and geographic trends.

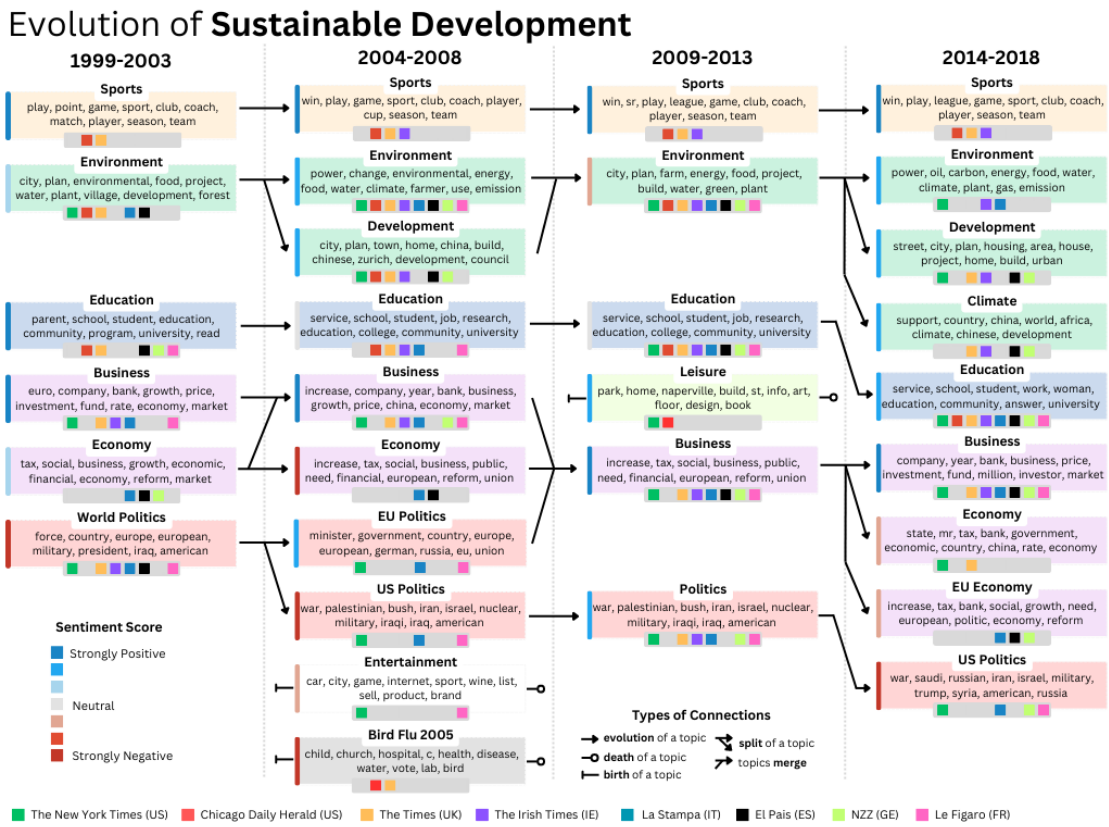

Figure 1: Sustainable Development Discourse Evolution

Finally, we reflect on newspapers as containers of information by inspecting overall discourse behaviour across newspapers, dedicating one subsection to that purpose.

5.1. Sustainable Development

Figure 5.1 shows the historical evolution of discourses about Sustainable Development. We use boxes to represent global topics. Each box includes semantically similar words that we have grouped in the second step of our pipeline using word embeddings on the English translation of the multilingual topics that PAM outputted. We provide information of each newspaper represented in each topic using small coloured squares at the bottom of each box, matching a colour-legend that we include at the bottom of the figure. We also gather topics qualitatively into semantically similar categories (i.e., environment, politics), and we indicate this by colouring the background of each box using the same tone. We add as well sentiment analysis scores by placing a coloured bar at the left of each box, and we include a colour legend in each figure. We depict the connections that the directed graph calculates using arrows, and we explain arrow representation (types of connection) in a legend. We provide similar figures containing the

same information for all our selected four key terms.

Just to provide a reading guide to understand how to interpret our figures using one column as an example, in the epoch 1999-2003, it is possible to observe six different topics (boxes): sports, environment, education, business, economy, and world politics. However, we (subjectively) group those six topics into five different semantic categories that we showcase by colouring the boxes: sports (yellow), environment (green), education (blue), business (lilac), and politics (orange). In this epoch, the topic shared by the biggest number of newspapers is world politics (*The New York Times International, The Times, The Irish Times, La Stampa, El Pais* and *Le Figaro*). The two topics that show the lowest affinity across newspapers are sports (only mentioned in *The Times* and *Chicago Daily Herald*) and economy (only mentioned in *La Stampa, El País* and *NZZ*).

And now, let's analyze the key term sustainable development using our three different criteria: diversification, attention, and geography. In terms of diversification, it is possible to observe little variation of topic labels over time, showing stability in discourse continuity over the last twenty years. That being said, it is interesting to note a progressive specialization of topics such as environment (there is a proliferation of topics in these domains from the first epoch (1999-2003) to the second one (2004-2008), followed by a simplification during the third one (2009-2013), yet increasingly fragmented during the fourth one (2014-2018)). Similarly, the business semantic category fluctuates between two topics (1999-2003, 2004-2008), to one (2009-2013), to eventually get fragmented into three as we approach contemporary times (2014-2018). Some topics are epoch specific, and do not evolve over time (2004-2008 Entertainment and Bird Flu, and 2009-2013 Leisure). The topic of politics remains relatively stable over time.

Overall, it is possible to observe a very clear stability in the evolution of sustainable environment topics over the last twenty years. While there is a clear trend of an initial diversification of topics starting in 2004-2008, it is due to the appearance of two epoch-specific topics (entertainment and Bird Flu 2005). As we have already explained, while there is a proliferation in unique topics over time (in 1999-2003 there are six of them, while in 2014-2018 there are nice), the semantic categories remain stable. So, both in 1999-2003 and in 2014-2018, discourses about sustainable development are framed under the same five semantic fields: sports, environment, education, business, and politics.

Topics that receive highest rates of attention (shared by all selected newspapers) are environment (2004-2008, 2009-2013), and education (2009-201, 2014-2018). The ones that receives the least attention are 1999-2003 sports (*The Times, Chicago Daily Herald*), 2004-2008 economy (*La Stampa, El Pais*), 2004-2008 entertainment (*The New York Times International, Le Figaro*), and Bird Flu 2005 (*Chicago Daily Herald, The Times*, 2004-2008); 2009-2013 leisure (*The New York Times International, Chicago Daily Herald*), and 2014-2018 economy (*The New York Times International, The Times*). We believe that attention could be as well be considered as a metric of polarity. The epoch that shows the highest number of low-attention topics is 2004-2008, which is precisely when discourses about sustainability fragment the most. However, we do not observe significant variations if we compare the epochs 1999-2003 and 2014-2018, as they both show similar attention rates, therefore reinforcing our observation regarding the stability of this discourse over time.

It is not possible to observe clear geographic clusters of topic affinity, with the exception of sports, that is dominated by English speaking newspapers in 1999-2003 (*The Times, Chicago*

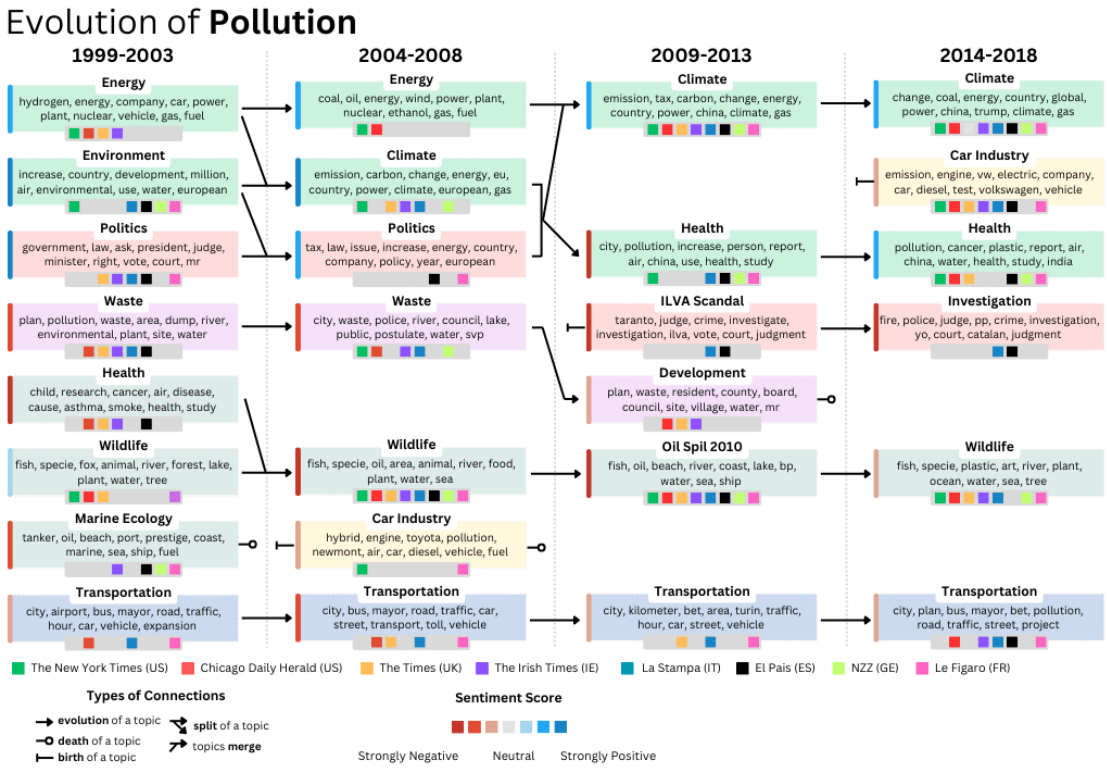

Figure 2: Pollution Discourse Evolution

Daily Herald), 2004-2008/2009-2013/2014-2018 (*The Times, Chicago Daily Herald,* and *Irish Times*), Bird Flu 2005 in 2004-2008 (*The Times, Chicago Daily Herald*, leisure in 2009-2009 (*Chicago Daily Herald, The New York Times International*, and 2014-2014 economy (*The New York Times International, The Times.* There is one Southern European cluster concerned about the Economy in 2004-2008 (*La Stampa* and *El País*), showing low Sentiment Scores.

In terms of sentiments, there is a clear trend of increasing negative views in the semantic category of the economy topics as we approach contemporary times. While in the first epoch (1999-2003), scores are quite high (blue colour), in the last one (2014-2018), their evolution is negative (lower scores in the red range). The rest of the semantic categories do not show significant variations over time, highlighting once again the historic and geographic stability of sustainable development related rhetorics.

5.2. Pollution

In terms of diversification of topics, pollution shows a very clear trend of historic simplification, and, similarly to sustainable development, very little variation over time in overall discourses. In

the 1999-2003 cluster, we have manually annotated eight different topics: energy, environment, politics, waste, health, wildlife, marine ecology, and transportation. In the 2014-2018, only six: climate, car industry, health, investigation, wildlife, and transportation. Furthermore, we observe how thematic categories (that, again, we showcase by colouring the topic boxes) remain very stable in the whole observational time. For example, the 1999-2003 thematic category of climate (green) remains relatively stable over time with topics such as energy and environment evolving in 2004-2008 into energy and climate, and ultimately merging into climate (2009-2013, 2014-2018). Similarly, the semantic field of politics (orange), while fragmented in two different family of topics (1999-2003 to 2004-2008 in the first place, talking about policy related issues, and 2009-2013 to 2014-2018, talking about political controversies (i.e., Ilva Scandal)), however, endures continuity in terms of representation. Waste (lilac) shows no major fluctuations, dying in 2009-2013. The semantic field of wildlife (green), appears initially highly fragmented into three different topics (health, wildlife, and marine ecology), and gets eventually simplified into just one (wildlife) in 2014-2018. Transportation (blue) appears consistently equally represented over time. Finally, the thematic category of car industry appears isolated in two epochs (2004-2008, 2014-2018).

Topics that receive the highest rates of attention are wildlife (2004-2008), climate (2009-2013), and oil spill 2010 (2009-2013). Topics that receive the lowest rates of attention are 2004-2008 energy (*The New York Times International, Chicago Daily Herald*), 2004-2008 politics (*El País, Le Figaro*), 2004-2008 car industry (*The New York Times International, Le Figaro*), 2009-2013 Ilva Scandal (*La Stampa, El País*), 2014-2018 Investigation (*La Stampa, El País*). It is possible to observe a progressive transition from polarity (topics showing low numbers of newspaper representation), to agreement (more newspapers included), as we approach contemporary times, displaying higher rates of information homogeneity in Western societies.

In terms of the geographic distribution of clusters, we observe similar patterns as in sustainable development. There seems to be some English dominated clusters, such as 1999-2003 energy (*The Times, The New York Times International, Chicago Daily Herald, Irish Times*), 2004-2008 energy (*The New York Times International, Chicago Daily Herald*); and 2009-2013 development (*Chicago Daily Herald, The Times, The Irish Times*). There is as well a clear Southern European cluster related to politics, that appears in 2004-2008 politics (*El País, Le Figaro*), with high sentiment scores; and 2009-2013 Ilva Scandal (*La Stampa, El País*), and 2014-2018 Investigation (*La Stampa, El País*). local politics (*El País, La Stampa*, and in the 2009-2013 Ilva Scandal (*El País, La Stampa*), both with low sentiment scores.

Overall, and as compared with Sustainable development, discourses appear a bit more fragmented in terms of geographic representation (there are more regional clusters), and diversity of information (there are more individual topics). However, there is a tendency towards a historic simplification as we approach contemporary times that we interpret as an increasing homogenization of views about pollution. In terms of thematic categories, and as we have just explained, there is not much variation over time and it is possible to observe how all discourses appear gathered in just six different semantic fields.

Our analysis of sentiments shows a subtle feeling of increasing optimism as we approach contemporary times. The climate semantic field (green), remains positively stable over time showing similar high scores in 1999-2003 and in 2004-2008. Politics (orange) transitions from positive (1999-2003) to negative (2014-2018). The wildlife thematic category (forest green)

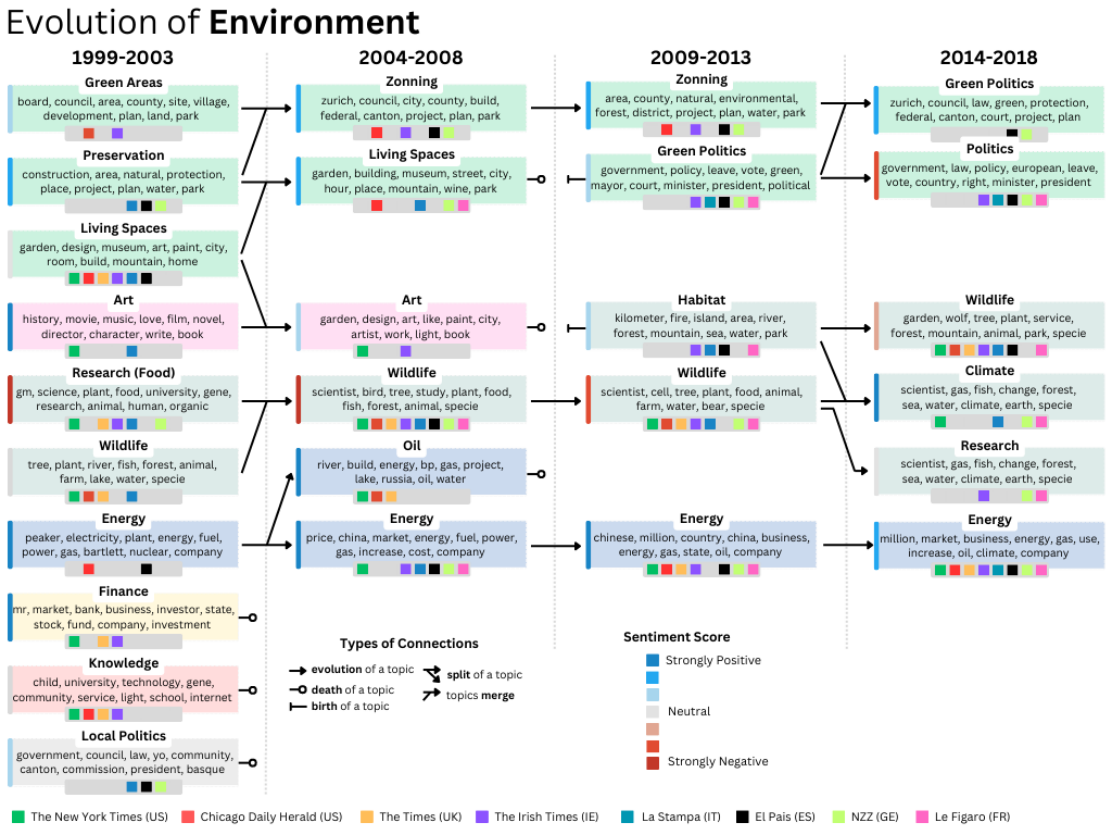

Figure 3: Environment Discourse Evolution

evolves from overall highly negative feelings in 1999-2003 to lesser ones in 2014-2018. Other isolated topics such as car industry remain negatively similar in their two respective epochs.

5.3. Environment

Environment shows very similar trends as pollution does in terms of diversification, attention, and geography. It is possible to observe a clear trend of simplification of topics over time. The first epoch (1999-2003) contains ten different topics (green areas, preservation, living spaces, art, research (food), wildlife, energy, finance, knowledge, local politics), while the last epoch (2014-2018), gets reduced to just six. In terms of semantic fields, there is a simplification ranging from six in 1999-2004 (green areas (green), art (lilac), wildlife (forest green), energy (blue), finance (yellow), knowledge (orange), local politics (grey)), to just three during 2014-2018 (green areas (green), wildlife (forest green), energy (blue)). And, overall, there are only eight different thematic categories during the whole observational time.

Topics related to environmental spaces in 1999-2003 (green areas, preservation, and living spaces), get progressively simplified into two different categories in all the rest of the epochs. The

thematic field of wildlife (forest green) firstly gets simplified, and then eventually fragmented as we reach contemporary times. There is a trend of topics epoch-bounded, such as the art family (lilac), and the three 1999-2003 isolated topics of finance, knowledge, and local politics. We interpret this fragmentation as an initial high polarization across countries about environmentally related discourses.

In terms of attention, topics that receive the highest rates (all the newspapers appear represented) are wildlife (2004-2008), and energy (2014-2018). Topics that receive the lowest rates (only two newspapers represented) are 1999-2003 green areas, art, and energy; 2004-2008 art, and 2014-2018 green politics. The metric of attention therefore shows as well an initial polarization with many two-newspapers topics in the 1999-2003 epoch, progressively reaching higher rates of press agreement as we approach contemporary times.

Geographic distribution shows yet again the presence of an Anglo-speaking cluster of topics. 1999-2003 green areas (*Irish Times and Chicago Daily Herald*), finance (*The Times, The New York Times International, Irish Times*), and knowledge (*The New York Times International, Chicago Daily Herald, The Times, Irish Times*); 2004-2008 oil (*The Times, The New York Times, The Irish Times*) and art (*The New York Times, The Irish Times*). But, there is no Southern European one.

Sentiment Analysis scores show a light trend of increasing pessimism as we approach contemporary times. The green areas thematic category (green) shifts from neutral-positive sentiments in the 1999-2003 cluster to negative ones in politics related topics. The wildlife family of topics (forest green), on the other hand, becomes slightly more positive in the 2014-2018 epoch.

5.4. Climate Change

Climate change shows a simplification of topics until 2009-2013, followed by a diversification in 2014-2018. The 1999-2003 and 2004-2008 topics in the semantic field of economy (orange) disappear in 2009-2013, to re-appear in 2014-2018. Politics related topics (green) simplify in 2009-2013, to later on fragment in 2014-2018. Similarly, the thematic category of global warming (forest green), gets dramatically fragmented in 2014-2018. The leisure family of topics (lilac) follows as well this pattern. Car industry (yellow) and education (dark blue), on the other hand, remain relatively stable. Some topics are time-specific, such as the semantic category of city planning (blue). We interpret these results as a clear trend of Western societies reaching a point of international agreement in climate related topics in 2009-2013, followed by a very strong polarization in 2014-2018.

In terms of attention, topics showing the highest rates of newspapers presence include energy (2004-2008), leisure (2004-2008, 2009-2013), global warming (2009-2013), and politics (2009-2013). Topics with lowest rates of attention include 1999-2003 wine industry and car Industry, and 2014-2018 global warming, car industry, and US economy. Attention aligns with diversity, showing a higher number of low attention topics in 1999-2003 and 2014-2018 (therefore displaying polarity), and greater newspapers representation in 2009-2013 showing an agreement peak.

We note yet again an Anglo-speaking cluster of topics. 1999-2003 economy (*Irish Times, New York Times International, The Times, Chicago Daily Herald*), and sports (*Chicago Daily Herald, The New York Times International, The Times, Irish Times*), 2004-2008 education (*The Times, Chicago Daily Herald, Irish Times*), 2004-2008 Education (*The Times, Chicago Daily Herald, Irish Times*), and 2014-2018 Us Economy (*New York Times, Chicago Daily Herald*), are solely English

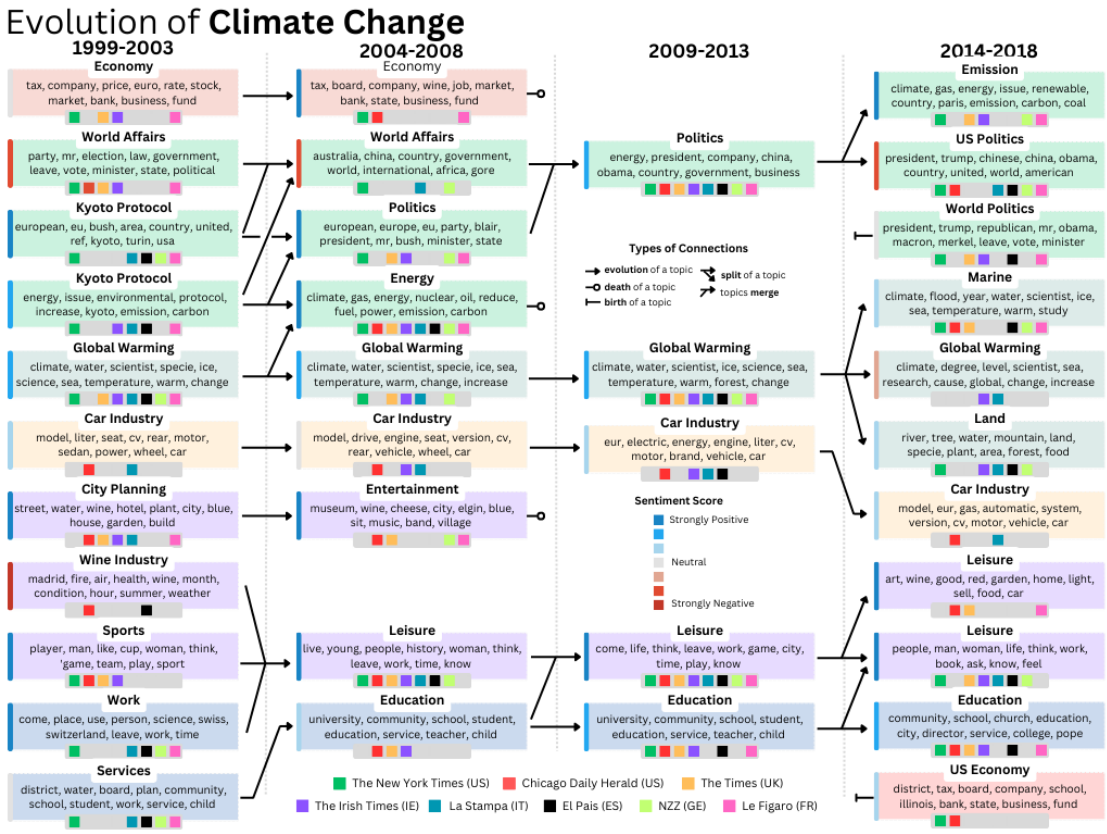

Figure 4: Climate Change Discourse Evolution

speaking newspapers. There is no Southern European cluster.

In terms of sentiment, we observe a very clear trend of positive emotions in 2009-2013, followed by a trend of pessimism in 2014-2018 (with special emphasis in US Politics and global warming). Therefore, in this case, sentiments line up with the three other metrics (diversification, attention, and geography), showcasing a historic break in 2009-2013.

5.5. General Trends

Overall, we observe very little variation of newspapers sustainability discourses across time of our selected four terms. While there is an array of topics that we label differently, and that, as we have explained, fluctuates in each epoch, broadly speaking it is possible to detect a very small number of thematic categories across all newspapers that could be grouped in a range of six to eight. Again, we subjectively decide how to classify topics into those broad categories. Yet we believe that our choice of semantically similar topics should not be controversial, as we think that our grouping criteria is quite self-explanatory.

For example, Sustainable development only shows eight different categories in the four

epochs: 1. sports (yellow boxes), 2. environment (green boxes), 3. education (blue boxes), 4. business (lilac boxes), 5. politics (orange boxes), 6. entertainment (white box), 7. Bird Flu 2005 (grey box), 8. Leisure (lime green box).

Pollution appears framed under 6 areas: 1. energy (green boxes), 2. politics (orange boxes), 3. waste (lilac boxes), 4. wildlife (forest green boxes), 5. transportation (blue boxes), 6. car industry (yellow boxes.)

Documents containing the key terms environment and nature could be group in seven major thematic areas: 1. green areas (green boxes), 2. art (lilac boxes), 3. wildlife (forest green boxes), 4. energy (blue boxes), 5. finance (yellow box), 6. knowledge (orange box), 7. local politics (grey box).

Climate change could be classified under six major areas: 1. economy (orange boxes), 2. environmental politics (green boxes), 3. global warming (forest green boxes), 4. car industry (yellow boxes), 5. leisure (lilac boxes), 6. education (blue boxes).

Our data analysis therefore shows how newspapers content appears relatively stable over time, not showing dramatic shifts in themes represented. So, while it may be possible to observe how some topics fragment over time into an increasing number of semantically similar topics, we do not consider these as expanding information fractures but as a semantic evolution of a same thematic category. And, that is why we state that there is an overall trend of simplification of discourses as we approach contemporary times: while there may be some topic proliferation, broad semantic categories tend to get reduced, showcasing higher rates of international agreement, possibly reflecting processes of media globalization. That being said, our analysis of climate change discloses an opposite effect, highlighting a trend of agreement in 2009-2013 followed by an international public discourse crack in 2014-2018.

As we mentioned in the "Related Work" section, we seek to contribute with data-driven findings to discussions that have analyzed the agency of history, economics, and politics, in shaping different media traditions ([23]). While we do not wish to oversimplify the complexity of factors determining those different media cultures and their prevalence in today's journalistic practices, we do believe that, with the exception of climate change, our data analysis reveals an increasing homogenization of Western views as we become approach present times, although we are aware that our sample of newspapers is not big enough to make confident assertions about all Western media. That being said, these findings resonate with [23]'s views about the evolution of media systems:

> Media systems have historically been rooted in the institutions of the nation state, in part because of their close relationship to the political world. National differentiation of media systems is clearly diminishing; whether that process of convergence will stop at a certain point or continue until national differentiation becomes irrelevant we cannot yet know. (13).

At the same time, our findings reveal how sustainability related discourses have not significantly changed during the last twenty years. Even in the case of climate change, where we show a polarity breach in 2014-2018, the semantic categories are the same as in 1999-2003 (green politics, global warming, car industry, leisure, education, economy). It is important to note that our data analysis only focuses on finding semantic clusters shared by at least two newspapers.

Therefore, what we are intentionally selecting is overlapping multilingual discourses that talk about sustainability similarly: our findings are reporting points of contact in the international press. And, what we have found is that, in the majority of our selected terms, there is a solid demonstrated history of shared views among Western countries that can be traced back in time to the last twenty years. While we observe some polarity in most of the 1999-2003 epochs across terms, there is a tendency of reaching a global agreement as we become more contemporary. Furthermore, we show that the majority of the semantic categories across newspapers can be grouped in six-to-eight topics (and therefore, showing low rates of information diversity) once again reinforcing the idea that public discourses about sustainability related discourses showcase high numbers of shared views in Eurocentric societies.

Our analysis of the geographic distribution of topics empirically shows that there are no clear trends with the exception of an English speaking cluster that appears consistently over time, and a marginal southern European one related to the economy that only appears in 2/4 key terms (sustainable development and pollution). We are aware that, as our dataset lacks data from central Europe (we don´t have any newspaper representation from Germany, Austria, The Netherlands, or Belgium), as well as the Scandinavian region; it is difficult to make conclusive statements. That being said, we do use a highly diverse geographic sample that includes seven countries (Spain, France, Italy, Switzerland, UK, USA, and Ireland), and still, our clustering shows inconclusive geographic patterns.

Similarly, our study of sentiments highlights that it is not possible to notice any consistent trends across time and space. So, some terms become slightly more positive, and others a bit more negative. Consequently, we can´t detect any significant identifiable data behaviour patterns.

Finally, we do not observe significant data behaviour differences across our news outlets. While *Chicago Daily Herald* appears marginalized in discussions about politics in the key terms sustainable development, environment, and pollution (probably showing different semantics in these two areas as the rest of the newspapers), it does appear in those discussions in climate change. We do not detect any noticeable trends in *The New York Times International* signaling any differences with the rest of the newspapers.

6. Conclusion

Our article contributes to current discussions across fields about sustainability-related discourses with two major findings: a) it is possible to note a progressive homogenization of discourses across countries as we approach contemporary times showing increasing shared views across Western countries about sustainability-related topics (with the exception of climate change), and b) newspapers content in our selected key terms shows very little variation over the last twenty years, empirically demonstrating high rates of agreement in sustainability discourses historically across our selection of Western countries. We believe these are highly interesting results and that two main conclusions can be reached. Firstly, we show that even if "the West" is a multicultural space composed of a diversity of countries with different histories, economic models, religions, or languages, in terms of sustainability press discourses, it is possible to observe a rather monolithic behaviour where a common voice can be identified. And furthermore,

there is a tendency to reach higher rates of agreement as we approach contemporary times, which we identify as a possible side effect of globalization processes. Secondly, we show that, in spite of recent efforts by governments worldwide, international organizations, industry, policy hubs, research funding institutions, and social activism (just to name a few), to raise awareness among civil society about the necessity of creating more sustainable societies, press discourses remain relatively stable, and therefore, not necessarily capturing recent shifts in environmental social morality. Yet, when there is strong polarity, such as in the case of climate change, newspapers do reflect discourse antagonism. Therefore, we state that newspapers discourses do not necessarily capture subtle changes in sustainability-related rhetorics, but they do reflect abrupt disagreements. Yet, as we have shown, even when there are clear trends of polarity (as it happens with climate change), newspaper content does not necessarily change: as already explained, the 1999-2003 and 2014-2018 epochs show the same semantic fields.

That being said, we wonder whether our method poses some biases in these findings. We are purposely selecting only semantic regions across countries that show the highest scores of affinity. As mentioned, in the second step of our pipeline, we use word embeddings to calculate pairwise cosine similarities of the English-translated words outputted by PAM. And in the third step, we use Ward Hierarchical Clustering to only select the highest ranked clusters. Therefore, we are intentionally eliminating topics that show low rates of agreement and that indicate polarity. Moreover, in our topic evolution graph, we are discriminating our selection of the temporal evolution of topics by only picking the ones that have the highest connection weights across epochs. So, again, we are only choosing clusters that indicate high-affinity rates across countries, leaving behind those that score low and that therefore indicate geographic disagreement on public discourses.

Yet, when there is a lot of polarity, our method is capable of uncovering it. It is possible to observe a variety of epoch-specific isolated topics such as the Ilva Scandal, the 2005 Bird Flu, the Kyoto Protocol, or the Volkswagen Car Scandal. Therefore, when newspapers content is diverse enough, our method successfully detects information variance. Moreover, our four different metrics (diversity, attention, geography, and sentiment) are capable of gradating our analysis of discourses to a rather close level. Therefore, while we may be only selecting topics that show high rates of semantic affinity over time, we have empirically demonstrated how, nevertheless, in terms of newspaper content, sustainability-related discourses have not significantly changed over the last twenty years. And, by extension, as already mentioned, we show that Western societies show relatively similar views in quite a stable way over time (even though, within that stability, there is space for disagreement, such as in climate change). Consequently, we believe that our method is capable of successfully analyzing numerically the geographic and temporal evolution of newspapers rhetorics.

Accordingly, and following this line of reasoning, we suspect that newspapers materiality may influence the composition of discourses. Newspapers are commodified information designed to be sold in a market for profit. There are certain editorial policies that most newspapers follow, such as organizing information in relatively stable sections across countries during the last twenty years (i.e., politics, economy, sports, or entertainment), number of pages, article length, font size, or publication periodicity. Therefore, we are intrigued by the thought of how the form determines the content in this specific information scenario.

Hence, future lines of work include extending our four selected words to other sustainability-

related terms and expanding our current newspaper corpus to other under-represented Western regions (central Europe, the low countries, and the Scandinavian region), as well as using other computational methods to further assess the relationship between newspapers materiality and information behaviour. It would be highly interesting as well to implement a mix-methods qualitative-quantitative research approach on selected articles. Doing close readings on documents belonging to specific topics outputted by our computational analysis (i.e., comparing articles that receive low and high attention rates in terms of newspapers representation) could shed light on media environmental rhetorics and to what extent they do capture recent shifts in sustainability social morality. Finally, we are also interested to use social media (i.e., Twitter or Reddit) to test cross-platform data trends. In this article, we show how monolingual readers of their respective national press across our selected Western countries are exposed to very similar information in sustainability-related discourses. We are curious whether they may (or not) express similar views as well in social media, therefore being able to have a more well-informed judgment about the effect of the content on the form in the specific case of multilingual newspapers.

7. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie (MSC) grant agreement No 101024996. The authors would like to thank Prof. Julia Flanders for her guidance and useful advice.

References

[1] D. Mimno, W. Li, A. McCallum, Mixtures of hierarchical topics with pachinko allocation, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 633–640.

[2] N. Aletras, M. Stevenson, Measuring the similarity between automatically generated topics, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, 2014, pp. 22–27.

[3] A. Großwendt, H. Röglin, M. Schmidt, Analysis of ward's method, in: Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2019, pp. 2939–2957.

[4] A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, Discovering topic structures of a temporally evolving document corpus, Knowledge and Information Systems 55 (2018) 599–632.

[5] United Nations, Transforming our world: The 2030 agenda for sustainable development, 2015. URL: https://wedocs.unep.org/20.500.11822/9814.

[6] J. Drucker, Sustainability and complexity: Knowledge and authority in the digital humanities, Digital Scholarship in the Humanities 36 (2021) ii86–ii94.

[7] U. K. Heise, J. Christensen, M. Niemann (Eds.), The Routledge Companion to the Environmental Humanities, Routledge, New York, NY, 2017.

[8] J. Habermas, The Structural Transformation of the Public Sphere, Cambridge, Massachusetts, 1993.

[9] E. S. Herman, N. Chomsky, Manufacturing Consent: The Political Economy of the Mass Media, Pantheon Books, New York, 2002.
[10] L. Hay, A. Duffy, R. Whitfield, Sustainability and complexity: Knowledge and authority in digital humanities, Digital Scholarship in the Humanities 32 (2021) ii86–ii94. doi:https://doi.org/10.1093/llc/fqab025.
[11] D. J. Philippon, Sustainability and the humanities: An extensive pleasure, Journal of Environmental Management 133 (2014) 232–257.
[12] S. Lenz, Is digitalization a problem solver or a fire accelerator? situating digital technologies in sustainability discourses, Social Science Information 60 (2021) 188–208. doi:doi.org/10.1177/05390184211012179.
[13] A. E. Beling, J. Vanhulst, F. Demaria, V. Rabi, A. E. Carballo, J. Pelenc, Discursive synergies for a 'great transformation' towards sustainability: Pragmatic contributions to a necessary dialogue between human development, degrowth, and buen vivir, Ecological Economics 144 (2018) 304–313.
[14] J. Godemann, G. Michelsen, Sustainability Communication: An Introduction, Springer, 2011. doi:10.1007/978-94-007-1697-1.
[15] K.-W. Brand, Sociological Perspectives on Sustainability Communication, Springer, 2011. doi:10.1007/978-94-007-1697-1.
[16] L. Kruse, Psychological Aspects of Sustainability Communication, Springer, 2011. doi:10.1007/978-94-007-1697-1.
[17] C. d. Witt, Media Theory and Sustainability Communication, Springer, 2011. doi:10.1007/978-94-007-1697-1.
[18] R. Barkemeyer, F. Figge, D. Holt, T. Hahn, What the papers say: Trends in sustainability: A comparative analysis of 115 leading national newspapers worldwide, The Journal of Corporate Citizenship 33 (2009) 69–86.
[19] R. Barkemeyer, F. Figge, D. Holt, Sustainability-related media coverage and socioeconomic development: a regional and north–south perspective, Environment and Planning C: Government and Policy 31 (2013) 716–740.
[20] A. Kumar, N. Das, A text-mining approach to the evaluation of sustainability reporting practices: Evidence from a cross-country study, Problems of Sustainable Development 16 (2021). doi:10.35784/pe.2021.1.06.
[21] D. Fischer, F. Haucke, A. Sundermann, What does the media mean by 'sustainability' or 'sustainable development'? an empirical analysis of sustainability terminology in german newspapers over two decades, Sustainable Development 25 (2017). doi:10.1002/sd.1681.
[22] V. Sebestyén, E. Domokos, J. Abonyi, Focal points for sustainable development strategies. text mining-based comparative analysis of voluntary national reviews, Journal of Environmental Management 263 (2020). doi:https://doi.org/10.1016/j.jenvman.2020.110414.
[23] D. C. Hallin, P. M. Mancini, Comparing Media Systems: Three Models of Media and Politics, Cambridge Univesity Press, Cambridge, UK, 2004.
[24] C. Sterling, Encyclopedia of Journalism, SAGE Publications, Inc., 2009. URL: https://doi.org/10.4135/9781412972048. doi:10.4135/9781412972048.
[25] R. Greenslade, The new york times introduces its new 'international edition',

The Guardian (2016). URL: https://www.theguardian.com/media/greenslade/2016/oct/11/the-new-york-times-introduces-its-new-international-edition.
[26] M. F. Porter, Snowball: A language for stemming algorithms, 2001.
[27] C. J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text., in: ICWSM, The AAAI Press, 2014.
[28] H. B. Carlsen, S. Ralund, Computational grounded theory revisited: From computer-led to computer-assisted text analysis, Big Data & Society 9 (2022) 20539517221080146.
[29] T. Head, MechCoder, G. Louppe, Iaroslav Shcherbatyi, Fcharras, Zé Vinícius, Cmmalone, C. Schröder, Nel215, N. Campos, T. Young, S. Cereda, T. Fan, Rene-Rex, Kejia (KJ) Shi, J. Schwabedal, Carlosdanielcsantos, Hvass-Labs, M. Pak, SoManyUsernamesTaken, F. Callaway, L. Estève, L. Besson, M. Cherti, Karlson Pfannschmidt, F. Linzberger, C. Cauet, A. Gut, A. Mueller, A. Fabisch, Scikit-optimize/scikit-optimize: V0.5.2, 2018. URL: https://zenodo.org/record/1207017. doi:10.5281/ZENODO.1207017.
[30] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 262–272.
[31] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.