"*Oat milk Vegan Chocolate Taste Great!*": Monitoring The Food Transition Debate in Reddit

Greta Zella*, Jan Willem Bolderdijk†, Saskia-Peels-Matthey*, Gerry Wakker*, Tommaso Caselli*
*CLCG, Faculty of Arts, University of Groningen
{g.zella, s.peels, g.c.wakker, t.caselli}@rug.nl
†Section Marketing, Faculty of Economics and Business,University of Amsterdam
Marketing Department, Faculty of Economics and Business,University of Groningen
j.w.bolderdijk@uva.nl

Abstract

We present *DRiFT* (Debates on Reddit involving Food Transition), a new large-scale corpus and set of computational methods for using language as an early indicator of social change in the protein transition, i.e., the shift from a diet predominantly based on animal proteins to one based mainly on plant sources. *DRiFT* comprises 17.5M Reddit comments (2010–2022) from 29 subreddits grouped into two speaker communities: SUSTAINABLE (early adopters/innovators) and GENERIC (general public). Building on neologism analysis, lexical semantic change detection, and connotative profiling, we introduce three linguistic measures of innovation awareness, meaning shift, and attitudinal valence. We extract neonyms and retronyms to quantify awareness; apply static and contextual embedding-based Lexical Semantic Change methods (PPMI, SGNS, BERT substitutions) to probe semantic reconceptualization; and adapt an embedding-based connotation hyperplane to measure polarity changes for targeted terms. Results show marked diastratic differences, with SUSTAINABLE users both using innovation-specific lexicon more frequently and having reconceptualized core food terms in ethical/environmental frames, while the GENERIC community exhibits rapid proportional growth in neologism use and emerging positive connotations for some plant-based products. Diachronic denotational shifts over the 12-year window are weak, suggesting shortcoming of embedding-based methods to capture subtle meaning changes. *DRiFT* and our analyses demonstrate that language can function as a sensitive "thermometer" of subtle social change, revealing attitudinal dynamics before observable behavioral shifts.

Keywords: Computational Social Sciences, Lexical Semantic Change, Connotation Shifts

1. Introduction

The process of social change that characterizes the diffusion of innovation is uncertain, difficult to predict, and normally observable in behavior only after completion (Judge et al., 2024). Over the past decade, public awareness of the environmental and ethical impacts of animal-based food production and consumption has grown substantially. Concerns about greenhouse gas emissions from livestock, animal welfare, and public health have increasingly entered mainstream discourse. This heightened awareness has prompted interest in dietary alternatives, particularly plant-based proteins, though actual behavioral change in consumption patterns remains limited and gradual. While sales of plant-based alternatives have increased in some markets, overall animal protein consumption in the Global North has declined only marginally, suggesting a gap between awareness and action. This gap between attitudes and actions suggests that traditional behavioral metrics such as sales figures and consumption surveys may fail to capture the full extent of social change in progress. Awareness and shifting perceptions represent crucial early stages in the adoption of innovations, yet they often remain invisible to conventional measurement approaches focused solely on revealed behavior. Understanding these "beneath the surface" changes requires methods capable of detecting attitudinal shifts before they crystallize into observable actions.

In this contribution, we introduce *DRiFT* (Debates on Reddit involving Food Transition) [1], a Reddit corpus on protein transition, to investigate attitudes and ongoing changes in society on this topic. Beyond being a tool to shape and accelerate social change (Zella et al., 2025), language is also a thermometer that reflects the degree of social change unfolding "beneath the surface". In particular, we investigate linguistic proxies of social change by analyzing: *a*.) the frequency and distribution of neologisms related to food and protein consumption; *b*.) whether food-related lexical items have undergone semantic change; and *c*.) differences in semantic contexts across communities of speakers. Our contributions can be summarized as follows:

- We create and release *DRiFT*, the first large-scale corpus specifically designed for studying language as a proxy for social change in the context of dietary innovation, containing over

[1] Code is publicly available; data are subject to a Data Sharing Agreement https://github.com/gretazella/DRiFT/tree/main.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11209–11220
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

17 million comments from 12 years of Reddit discussions.

- We establish frequency analysis of neologisms, i.e., neonyms and retronyms as a valid measure of innovation awareness, demonstrating relevant diachronic and diastratic patterns in their usage.
- We apply lexical semantic change detection methods to identify perception shifts, revealing measurable semantic differences between innovator and general populations that align with different stages of innovation adoption.
- We introduce connotation analysis as a method for detecting attitudinal valence shifts, showing increasingly positive associations with plant-based alternatives over time and across communities.

The remainder of this paper is structured as follows: Section 2 reviews related work and theoretical background; Section 3 describes the *DRiFT* corpus and its creation; Section 4 presents our analysis of neonyms and retronyms; Section 5 details our experiments to measure lexical semantic change and connotations shifts; lastly, Section 6 discusses the results and their implications, while conclusions and future work are presented in Section 7.

2. Theoretical Background and Related Work

Language is a dynamic system that evolves in response to social and technological change. As societies innovate, speakers adapt existing linguistic resources to describe new concepts, leading to the expansion of word meanings and the creation of neologisms. Such developments often serve to differentiate innovation from tradition. For instance, with the invention of electronic books, one of the constitutive qualities of books, i.e. they are made of paper, has become optional. This had caused the meaning of the term "book" to expand with the subsequent formation of neologisms that refer to the innovation (neonyms) such as "e-book" and those denoting the tradition (retronyms) like "paper book" (Mäkelä, 2022). These linguistic shifts do not occur in isolation: subtle variations in word meaning and usage can mirror changing social attitudes, making language a sensitive indicator of ongoing social transformation (Judge et al., 2024).

Previous work on the automatic detection of Lexical Semantic Change (LSC) builds on the distributional hypothesis (Firth, 1957; Harris, 1954), which posits that words appearing in similar contexts tend to have similar meanings. LSC is thus measured as a variation in the similarity of the same target word across embedding representations obtained from corpora covering different time periods. Although contextualized embeddings capture richer contextual nuances, static embeddings have proven more effective for detecting gradual, subtle meaning changes (Tahmasebi et al., 2021; Periti and Montanelli, 2024). These fine-grained changes are especially informative for tracing emergent social dynamics and evolving attitudes. The work by Del Tredici et al. (2019) pioneered the use of embeddings to detect subtle LSC over short periods of time, later extended to social media data by Martinc et al. (2020). More recently, Hoeken et al. (2023) conducted a systematic evaluation of models for subtle semantic variation and shown the potential of contextualized embeddings for capturing short-term meaning shifts.

Beyond denotational meaning, words also convey connotative or affective dimensions. Word embeddings have been successfully employed to model connotative meaning and sentiment (Allaway and McKeown, 2021; Feng et al., 2011; Kang et al., 2014), with sentiment scores often serving as proxies for collective attitudes and evaluative trends (Baccianella et al., 2010; Esuli and Sebastiani, 2006; Hamilton et al., 2016a). While most studies examine sentiment synchronically, only a few have addressed diachronic connotative change. In particular, Basile et al. (2022) proposed a transparent method to detect polarity shifts over time, which we extend in our work.

In the food domain, text mining techniques have been mainly applied to consumer reviews, addressing topics such as food waste reduction, safety, and regional consumption patterns (Islam and Masudul Alam, 2023; Xiong et al., 2024; Siddique et al., 2025; Rong et al., 2019). Additional linguistic resources have been developed from social media data to support text mining and sentiment analysis related to food and sustainability (Brzustewicz and Singh, 2021; Singh and Glińska-Neweś, 2022; Molenaar et al., 2024). Building on this foundation, our study is the first to construct and release a corpus centered on the protein transition. We apply lexical and connotative change detection methods to track how the discourse surrounding food transition evolves over time, using linguistic indicators as proxies for social and attitudinal change.

3. The DRiFT Corpus

In this section we introduce the *DRiFT* corpus (Debates on Reddit involving Food Transition), describing its construction, composition, and quality assurance procedures.

3.1. Data Collection and Corpus Curation

We selected a 12-year period from January 1, 2010 to December 31, 2022, based on two factors: the growing societal interest in sustainable food alternatives during this period, and the increasing availability of relevant language data on social media platforms (McCarthy et al., 2017).

To identify relevant discussions, we compiled a list of 21 keywords from the semantic field of protein transition: beef, burger, cheese, chicken, dairy, eggs, fish, lamb, meat, milk, mutton, pork, poultry, sausage, soy, steak, tofu, turkey, veal, yogurt, and yoghurt. These keywords were manually selected based on their prominence in recent EU and US policy documents concerning food sustainability and protein transition.

We conducted a keyword search on reddit.com, querying for subreddits containing one or more keywords (both singular and plural forms) and having a minimum of 10,000 members. This automated process yielded an initial list of 662 subreddits. To ensure relevance, one of the authors independently reviewed subreddit descriptions and manually filtered the list. Subreddits were included if their discussions regularly addressed topics related to food consumption, sustainability, diet choices, or animal agriculture. Subreddits focused exclusively on recipes, restaurant recommendations, or unrelated topics (despite containing keywords) were excluded. The filtering process resulted in 29 subreddits. The 29 selected subreddits vary in their keyword coverage, with an average of 7 keywords per subreddit, a maximum of 27 for r/vegan, and a minimum of 1 for six subreddits (r/politics, r/science, r/worldnews, r/news, r/technology, r/askreddit).

We aggregated the 29 subreddits into two macrocategories representing distinct communities of speakers, henceforth referred to as GENERIC and SUSTAINABLE (capitalized to distinguish them from general usage):

- **GENERIC** (16 subreddits): Subreddits whose primary topics are not directly related to sustainability or dietary choices (e.g., r/politics, r/science, r/todayilearned). This community serves as a proxy for the general population.
- **SUSTAINABLE** (13 subreddits): Subreddits primarily focused on environmental issues, veganism, and sustainable living (e.g., r/vegan, r/environment, r/vegetarian). This community represents early adopters and innovators in dietary transition.

This categorization allows us to examine whether linguistic signs of social change emerge first among innovators and subsequently diffuse to the general population, consistent with innovation diffusion theory (see Section 2).

Using Reddit data, we cannot control for users' private information such as gender or age, thus we cannot guarantee that our data is representative of the whole population or that there is no overlap between SUSTAINABLE and GENERIC users. However, our main goal is to present a resource of natural language on the topic of protein transition and show that language can detect subtle signs of social change that methods based on behavioral observation (e.g., market research) cannot fully grasp. External measures of validation, such as surveys and questionnaires will be needed to further generalize and support our results.

Data Collection and Preprocessing We used the Pushshift API to collect all comments from the 29 subreddits that replied to posts containing at least one of our keywords. The initial collection yielded 18,266,145 comments.

We applied the following preprocessing steps:

1. Converted emojis to text descriptions
2. Replaced email addresses, usernames, and URLs with standardized placeholders
3. Removed HTML markup and numeric characters
4. Excluded comments with three or fewer words

To identify and remove irrelevant content, two annotators independently reviewed a stratified random sample of 500 comments. They identified unwanted instances of keywords, including sexually explicit references, figurative language unrelated to food (e.g., "*that's the beef between them*"), and spam. All messages containing such instances were removed (full list of noisy keywords reported in the Appendix A). A subsequent manual inspection of randomly selected comments confirmed satisfactory quality.

Following Del Tredici et al. (2019), we divided the corpus into two non-consecutive time periods, t_0 and t_1, to maximize the likelihood of detecting diachronic changes while maintaining sufficient data in each period. The time periods were determined based on three criteria:

1. Maximum temporal overlap in data availability across SUSTAINABLE and GENERIC communities
2. Balanced token distributions within each community across time periods
3. Sufficient temporal gap between t_0 and t_1 to allow for meaningful semantic change

The final corpus composition is presented in Table 1. The complete *DRiFT* corpus contains

17,574,237 comments comprising 1,169,588,177 tokens. The GENERIC sub-corpus is substantially larger (more than 1 billion tokens vs. $\approx$ 157 million tokens) due to higher activity levels and greater topic diversity in these subreddits.

Subcorpus	Years	Period	Tokens
SUSTAINABLE	2010-2022	–	156,929,310
	2010-2016	t_0	37,511,206
	2021-2022	t_1	39,348,756
GENERIC	2010-2022	–	1,025,958,879
	2010-2014	t_0	211,366,895
	2021-2022	t_1	213,710,353
Total:			**1,169,588,177**

Table 1: Overall data distribution for the *DRiFT* corpus across the communities and time periods.

To focus our analysis on terms most likely to undergo semantic change, we applied a frequency-based validation. Following Del Tredici et al. (2019), we computed the frequency change for each of the 21 keywords from t_0 to t_1 in both communities. Only keywords whose frequency increased by at least 2 standard deviations above the mean increase were retained. This resulted in a final set of 13 keywords: beef, burger, burgers, cheese, chicken, dairy, egg, meat, milk, sausage, sausages, steak, and tofu. This validation ensures that our subsequent analyses focus on terms actively involved in ongoing discussions about dietary change.

4. Neonym and Retronym Analysis

To measure awareness of dietary alternatives, we analyzed the frequency of neonyms (terms identifying innovations, e.g., "plant-based milk") and retronyms (terms retroactively labeling traditional options, e.g., "dairy milk"). We extracted bigrams from each subcorpus using the NLTK pipeline (Elhadad, 2010) with the following parameters:

- Minimum raw frequency: 10 occurrences
- Pointwise Mutual Information (PMI) score $\geq$ 10

PMI measures the strength of association between two words compared to their independent occurrence probabilities. High PMI scores indicate fixed expressions and collocations, making them suitable for identifying stable neonyms and retronyms.

Among the automatically detected bigrams, we manually retained only those that contain one of our 13 keywords preceded by a modifier that specifies the protein source either directly (e.g., "animal milk", "plant-based milk") or indirectly (e.g., "real meat", "fake meat"). This way we made sure to retain all lexical variations (e.g., non-dairy vs. non dairy) present in our dataset. This process yielded 50 neonyms and 33 retronyms. The following exclusion criteria have been applied:

- **Pre-existing distinctions**: Bigrams like "cow milk" or "goat milk", which existed before plant-based alternatives to distinguish among dairy sources, were excluded.
- **Unique innovations**: Terms referring to specific novel technologies (e.g., "lab-grown meat", "cultured meat") were excluded as they represent distinct categories rather than anchored alternatives to traditional products.
- **Ambiguous modifiers**: Potentially ambiguous terms (e.g., "real", "fake") were evaluated by manually reviewing a random sample of comments to make sure that they were used predominantly to clarify the product's source.

The final analysis focused on four keywords: milk, meat, cheese, and burger(s). The remaining keywords did not yield sufficient neonyms and retronyms meeting our criteria. Table 2 presents the top three neonyms and retronyms for each keyword.

Keyword	Neonyms	Retronyms
milk	plant-based milk, non-dairy milk, nondairy milk	dairy milk, animal milk, animals milk
meat	fake meat, non meat, vegetarian meat	real meat, animal meat, actual meat
cheese	vegan cheese, fake cheese, vegetarian cheese	dairy cheese, real cheese, actual cheese
burger(s)	veggie burger(s), vegan burger(s), vegetarian burger(s)	meat burger(s), real burger(s), normal burger(s)

Table 2: Top three neonyms and retronyms for each analyzed keyword

Diachronic Comparison Concerning the SUSTAINABLE Community, both neonyms and retronyms increased significantly from t_0 to t_1 ($p < 0.05$). The average frequency of neonyms rose from 15.24 pmw to 22.52 pmw (48% increase), while retronyms nearly doubled from 5.58 pmw to 10.12 pmw (81% increase). Although retronyms remain less frequent overall, their faster growth rate suggests increasing recognition of the need

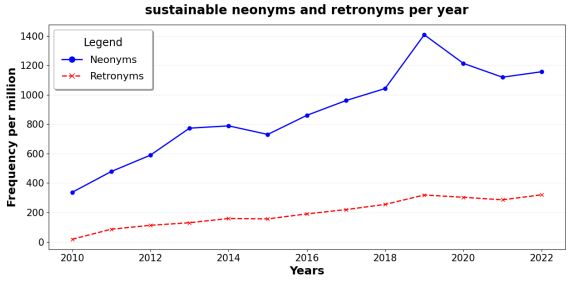

Figure 1: Pmw per year of neonyms (blue line) and retronyms (red line) within the SUSTAINABLE community.

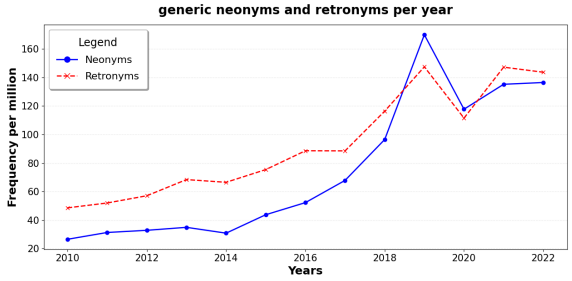

Figure 2: Pmw per year of neonyms (blue line) and retronyms (red line) within the GENERIC community.

to explicitly distinguish traditional products from alternatives. Figure 1 displays yearly frequency trends, showing steady growth with acceleration after 2018.

As for the GENERIC Community, neonyms and retronyms showed similar growth patterns, with the frequency gap between them narrowing over time. Both categories exhibited highly significant increases ($p < 0.001$). From t_0 to t_1, neonyms grew from 1.02 pmw to 3.54 pmw (247% increase), while retronyms increased from 1.68 pmw to 3.87 pmw (130% increase). Despite lower absolute frequencies compared to the SUSTAINABLE community, the proportional growth is substantial. Figure 2 reveals a notable peak in 2019, when neonyms surpassed retronyms for the first time, potentially coinciding with increased mainstream media coverage of plant-based alternatives.

Diastratic Comparison The SUSTAINABLE community uses neonyms and retronyms far more frequently than the GENERIC community, reflecting higher awareness levels among early adopters. Average frequencies across all time periods show neonyms at 18.99 pmw in SUSTAINABLE vs. 1.68 pmw in GENERIC; retronyms at 7.80 pmw in SUSTAINABLE vs. 3.22 pmw in GENERIC. This 10-fold difference in neonym usage confirms that explicit source specification is a linguistic marker of innovation awareness.

Importantly, the diachronic analysis reveals that neonyms and retronyms are growing proportionately faster in the GENERIC community (247% and 130% increases) compared to SUSTAINABLE (48% and 81% increases). This pattern indicates diffusion of awareness from innovators to the general population, consistent with innovation adoption models. The substantial frequency gap between communities also suggests potential semantic differences in how these terms are employed, as frequency shifts are known indicators of semantic change (Del Tredici et al., 2019). We investigate this possibility through lexical semantic change analysis.

5. Language as a Thermometer of Social Change: Experiments

The neonym and retronym analysis (Section 4) has highlighted an increasing awareness of plant-based alternatives in both innovator and general populations. However, awareness is a necessary but insufficient condition for social change. People may know about alternatives without changing their perceptions of them, and they may understand options without developing positive attitudes toward them. The critical question is: *Are people not only aware of plant-based alternatives, but also thinking about them differently and evaluating them more favorably?* We conduct two complementary analyses. First, we examine lexical semantic change (LSC) to detect whether food-related terms have acquired new meanings or shifted their semantic contexts across time periods and communities (§5.1). Changes in meaning indicate that the underlying concepts have expanded: for instance, if "milk" increasingly appears in contexts discussing ethics and sustainability rather than just nutrition and taste, this reflects a conceptual expansion that may precede behavioral change.

Second, we analyze shifts in connotations to measure changes in affective attitudes toward plant-based alternatives (§5.2). While semantic change reveals *what* people associate with these foods, connotation analysis reveals *how* they feel about them. An increase in positive connotations suggests growing acceptance that may not yet be reflected in consumption patterns.

Together, these analyses demonstrate how language can function as an early warning system for social change, detecting shifts in collective thinking that precede shifts in collective action.

5.1. Changes in Meaning as Cues of Changes in Perceptions

Lexical semantic change (LSC) in food-related terms signals that the underlying concepts are being reconceptualized. We investigate whether our

13 validated keywords exhibit LSC across time periods (diachronic) and between communities (diastratic). Following Hoeken et al. (2023), we formulate LSC detection as a classification task: given a set of keyword instances and two corpora (representing different time periods or communities), the goal is to correctly assign instances to their source corpus based on contextual usage.

To ensure fair comparison, we downsampled the GENERIC corpus to match SUSTAINABLE's size (112,791,295 tokens from 1,739,406 comments), maintaining proportional temporal distribution. We then employed four methods shown effective for subtle semantic change detection:

- **PPMI**: Positive Pointwise Mutual Information (context window ±10, minimum frequency 10), a pre-neural network method that registers patterns of co-occurring words in the context of the target keyword, encoded as 1 or 0 respectively (Arora et al., 2016) (Arora et al., 2016).
- **SGNS**: Word2Vec Skipgram with negative sampling (300 dimensions, context window ±5, minimum frequency 10), whereby word vectors are obtained by training a neural network to predict the context of a target keyword. Semantic spaces aligned via Orthogonal Procrustes (Hamilton et al., 2016b), necessary for cross-corpora comparison and to perform the evaluation. The distance between vectors are measure using cosine distance.
- **BERT (target)**: Masked language modeling where target keywords are masked and predicted substitutions are compared across corpora using Jensen-Shannon divergence (JSD). High overlap suggests stable meaning, while low overlap indicates semantic change. BERT-base-uncased (Devlin et al., 2018) is used without fine-tuning.
- **BERT (modifier)**: Same approach but masking the token preceding the keyword to detect modifier-driven semantic shifts.

We pre-process the corpora as follows: messages were split into sentences with NLTK, non-alphabetic characters, including punctuation and trailing whitespace, were removed. All corpora are lemmatized using SpaCy (Honnibal and Montani, 2018) for these experiments.

To distinguish genuine LSC from noise, we create control corpora by randomly splitting reference corpora (t_0 for diachronic, GENERIC for diastratic) into two random halves. This way, we simulate a situation where we expect no meaning change. The two halves are therefore treated as distinct corpora, and the experiments are repeated.

For each of the 13 keywords, we obtain two pairs of vectors: *a.*) test vectors from different corpora (e.g., milk_t_0 vs. milk_t_1); and *b.*) control vectors from the randomly split reference corpus (e.g., milk_t_0_A vs. milk_t_0_B). This results in a total of 26 pairs. Each pair is assigned a manual score, i.e., the gold label: 1 represents "different corpus", 0 represents "control corpus". LSC is confirmed when inter-corpus (different corpus) distance systematically exceeds intra-corpus (control) distance.

The models output a graded value that measures the actual distances between each pair of vectors, which we convert to binary values for comparison with the gold label. The highest value is converted to 1 because it is assumed to represent "different corpus", while the lowest one is annotated as 0, i.e., the "control corpus". If LSC has occurred, the automatic and manual annotations will be aligned, as the distance between vectors from different corpora will have been proved to be systematically higher than those from control corpora. This way, as mentioned earlier, instances of keywords can be ascribed to the corpus they belong to, based on their contexts of use. Figure 3 graphically illustrates the approach and detection of LSC using the diachronic dimension as an example.

The approach of (Hoeken et al., 2023) that we followed for validating the automatic LSC detection has limitations, as the authors themselves highlight. In particular, when converting graded outcomes to binary values, even slight differences between the graded values can flip the prediction. However, we still consider that this is a valuable strategy for comparing automatic and manual LSC. To overcome the limitation, LSC results are assessed not only based on the number of correct predictions, but also measuring Pearson correlation between graded outcomes and gold labels.

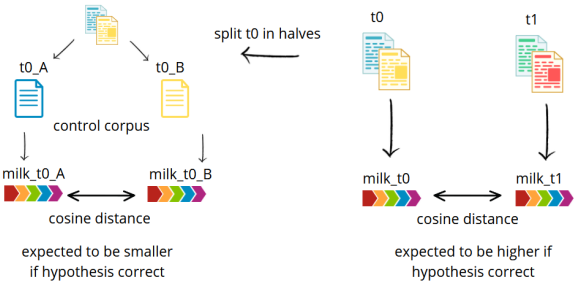

Figure 3: Graphic representation of the method (by Hoeken et al., 2023) used to validate LSC. Graphic created using Canva's Free content. Credits: OpenClipart-Vectors, furqonionii, Macrovector, Fusion Books, Nithish Ramesh.

We compute two metrics: (1) **#correct**: number of keywords where test distance $>$ control distance, and (2) **Pearson's** ρ: correlation between distance scores and binary labels (1 = test comparison, 0 = control comparison). High number of

correct predictions with low correlation suggests unreliable classification, potentially due to chance. The higher the correlation between the gold labels and the distance scores is, the easier it is to discriminate the word's contexts from different corpora compared to the ones from the control corpora. Table 3 presents results across three comparison types: diachronic within SUSTAINABLE (SUS), diachronic within GENERIC (GEN), and diastratic between communities (CROSS).

Diachronic analysis. Despite all correct predictions in both communities, Pearson correlations remain modest, indicating insufficient confidence to conclude diachronic LSC occurred. This suggests that while some keywords may show temporal shifts, the signal is weak and inconsistent across terms. However, BERT substitution analysis reveals subtle patterns. In SUSTAINABLE_t_1, "vegetarian" appears among top substitutes for "meat", reflecting its frequent use as a modifier in retronyms. In GENERIC_t_1, "animals" emerges as a substitute for "meat" suggesting increased salience of the food-source connection. These patterns indicate growing explicitness about protein origins, though not constituting robust LSC.

Diastratic analysis. All models achieve correct predictions with strong correlations, confirming systematic differences between communities. SUSTAINABLE and GENERIC users employ food-related terms in measurably distinct contexts. BERT analysis reveals the nature of this divergence. In SUSTAINABLE, "milk" and "meat" appear as mutual substitutes, suggesting frequent co-occurrence in ethical contexts (e.g., *milk and meat production cause animal suffering*). In GENERIC, these terms remain more semantically distinct, appearing primarily in consumption contexts. This indicates that SUSTAINABLE users have expanded food-related concepts to encompass ethical dimensions, a cognitive shift consistent with innovation adoption.

The absence of robust diachronic LSC contrasts with clear diastratic differences, suggesting two interpretations. First, 12 years may be insufficient for detectable semantic change to consolidate within communities. Second, and more interestingly, the semantic differences between communities may already have existed in 2010, reflecting stable ideological distinctions rather than recent shifts. The diastratic findings are theoretically significant: innovators (SUSTAINABLE) have reconceptualized food-related terms to foreground ethical considerations, while the general population (GENERIC) maintains traditional, consumption-focused semantics. This cognitive divergence represents a crucial stage in innovation diffusion: innovators have not merely adopted new products but developed a fundamentally different conceptual framework for understanding protein sources. The BERT substitution patterns warrant particular attention. The interchangeability of "milk" and "meat" in SUSTAINABLE discourse signals abstraction from specific food types to a generalized ethical category. This conceptual restructuring parallels historical cases where moral reframing preceded social change, such as the abolition movements (Otto et al., 2020).

5.2. Changes in Connotations as Cues of Changes in Attitudes

While LSC reveals shifts in *what* people associate with food terms, connotation analysis reveals *how* they evaluate them. We measure affective shifts along the positive-negative polarity dimension to detect attitudinal changes that may precede behavioral adoption. We employ the connotative hyperplane method (Basile et al., 2022), which positions word embeddings relative to a classifier trained on positive and negative seed words. Each keyword's position relative to this hyperplane indicates its affective valence. We use the same Skipgram embeddings and corpus splits as in Section 5.1 and we focused only on valid neonyms and retronyms from Section 4. To ensure sufficient token frequency, we have used placeholders for their modifiers (e.g., *plant_milk*, *animal_milk*).

We randomly sampled 300 sentences per neonym/retronym pair (50 per subcorpus), yielding 1,800 sentences. Three annotators assigned polarity scores (1=positive, -1=negative, 0=neutral) based on contextual sentiment. Inter-annotator agreement (Krippendorff's α) ranges from 0.55 to 0.63 (Table 4), indicating moderate agreement and reflecting the task's inherent subjectivity.

Diachronic patterns In the SUSTAINABLE community, attitudes toward plant-based alternatives became more critical over time, particularly for *plant_burgers* and *plant_meat* (Figure 4). Manual analysis reveals this reflects increasingly sophisticated critique: while plant alternatives remain preferred over animal products, users now highlight environmental impacts of soy production, ultra-processing concerns, and additive content. Taste remains central: early positive mentions emphasize flavor and texture, while later criticism cites expense and processing. One t_1 comment exemplifies this shift: *"it doesn't matter that you're ordering soy if your soy burger is paying for the slaughterhouse to be expanded"*. In the GENERIC community, we observe a negative shift for *animal_milk* and a positive shift for *plant_burger(s)*, suggesting the general population is following the innovators' trajectory. Early neutral attitudes toward plant burgers (balanced mentions of taste and availability vs. limited

Models	SUS		GEN		CROSS	
	#Correct	Pearson ρ	#Correct	Pearson ρ	#Correct	Pearson ρ
PPMI	26	0.58	26	0.41	26	0.71
SGNS	22	0.31	22	0.27	26	0.78
BERT - target	26	0.59	26	0.84	26	0.91
BERT - modifier	26	0.65	26	0.75	26	0.86

Table 3: Results of the LSC experiment. SUS (SUSTAINABLE), GEN (GENERIC), CROSS (diastratic).

Term Type	**milk**	**meat**	**burger(s)**
animal_*	0.58	0.57	0.55
plant_*	0.61	0.63	0.62

Table 4: Inter-annotator agreement (Krippendorff's α) for neonyms and retronyms.

options and unfamiliarity) evolved toward more positive evaluation emphasizing improved taste and variety. Conversely, *animal_milk* acquired negative associations, with t_1 comments frequently questioning the "naturalness" of dairy consumption for humans.

Diastratic patterns The SUSTAINABLE community maintains markedly more positive attitudes toward plant alternatives and more negative attitudes toward animal products (Figure 4). The largest difference concerns *animal_meat*: SUSTAINABLE users emphasize animal suffering, resource depletion, and moral concerns, while GENERIC users remain predominantly positive, highlighting nutrition and taste. However, GENERIC users with negative attitudes cite remarkably similar concerns (animal welfare, environmental impact, ethics), suggesting nascent awareness among a minority. Notably, *plant_milk* shows convergence: both communities express largely positive attitudes, with overlapping justifications (taste, health, environmental benefits). This may indicate that *plant_milk* has achieved broader acceptance than other alternatives.

Connotation analysis reveals attitudinal shifts invisible in behavioral data. Overall, we can observe that SUSTAINABLE users' increasingly critical stance toward plant alternatives reflects maturation beyond simple advocacy: they now apply sustainability criteria to evaluate all food production, including plant-based. This suggests genuine integration of environmental values rather than superficial trend-following. Secondly, GENERIC users show early signs of adopting SUSTAINABLE attitudes, particularly regarding plant burgers (increasingly positive) and dairy milk (increasingly negative). However, this diffusion is incomplete and uneven across product categories. Finally, words like "milk" and "burgers" now often reference different products in different communities, necessitating the LSC analysis to disentangle semantic from affective shifts. This indicates that language change is multidimensional, where meaning and valence jointly evolve.

6. Discussion

The linguistic analysis performed on the *DRiFT* corpus reveals subtle attitudinal shifts within the protein transition debate that remain undetectable through traditional market analyses. A clear pattern of innovation diffusion emerges from the SUSTAINABLE community toward the GENERIC one, aligning with established models of innovation spread (Judge et al., 2024).

Within the SUSTAINABLE community, language use reflects a strong integration of environmental and ethical values, accompanied by widespread awareness and positive attitudes toward plant-based alternatives. These attitudes are increasingly visible in the GENERIC community as well, suggesting that early adopters play a key role in shaping broader social acceptance of alternative proteins. However, only SUSTAINABLE users exhibit a genuine reconceptualization of the food lexicon, indicating a deeper perceptual and ethical shift. Terms such as "milk" and "meat" frequently occur in contexts referring to plant-based products, highlighting a semantic broadening rooted in moral and environmental considerations.

These changes are primarily observable at the diastratic level, indicating that while the GENERIC community shows signs of convergence, the sociolinguistic gap between the two groups remains substantial. The absence of significant diachronic LSC indicates that the potential changes in meaning at the denotative level are not captured by approaches grounded in the Distributional Hypothesis: as a matter of fact, the core context of occurrence of the selected terms does not change. Nonetheless, the widespread presence of innovation awareness and favorable attitudes toward the protein transition across both communities provides clear linguistic evidence of emerging social change. Historically, comparable linguistic indicators have preceded major societal transformations, where shifts initially

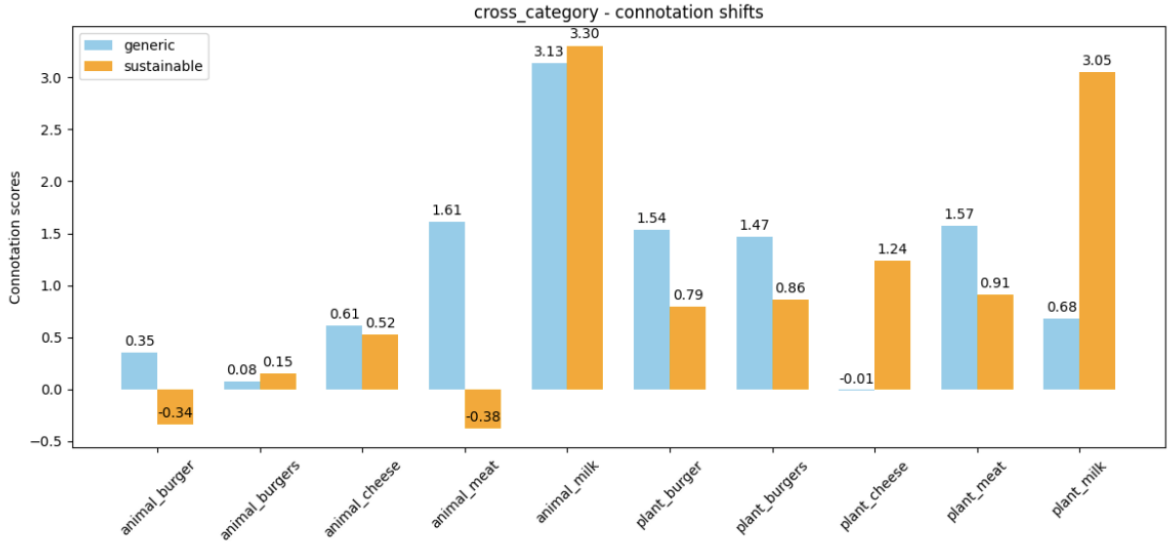

Figure 4: Diastratic connotative shifts for the neonyms and retronyms of “milk”, “meat”, “cheese” and “burger(s)”.

confined to small groups of innovators eventually reshaped mainstream discourse and collective values (Otto et al., 2020).

Overall, these findings confirm that language functions as a sensitive barometer of social change, capable of detecting emerging trends “beneath the surface”. Linguistic analyses such as ours offer a valuable complement to behavioral and market-based approaches, which typically register societal change only once it has already occurred.

7. Conclusion

We introduce *DRiFT* (Debates on Reddit involving Food Transition), a large, curated corpus (2010–2022) and a set of language-based measures designed to detect early signs of social change in the ongoing protein transition. Using complementary linguistic probes, such as neonym/retronym frequency, lexical semantic change (LSC) detection, and embedding-based connotation analysis, we document a consistent pattern of innovation diffusion: linguistic innovation and ethical reframing originate in SUSTAINABLE (innovator) communities and are spreading into GENERIC (general-public) communities. Crucially, our results show strong diastratic differences (systematic semantic and connotative divergence across communities) but weak evidence of consolidated diachronic denotational shifts within the 12-year window.

Taken together, the findings validate the claim that language functions as a sensitive “thermometer” of subtle social change: attitudinal shifts (positive valence toward plant-based products; moralization of animal-based products) are already detectable in discourse even where behavioral metrics remain inconclusive. At the same time, the diastratic–diachronic contrast highlights important methodological caveats: Reddit subcommunities are valuable but partial proxies for public opinion, and distributional LSC methods may miss subtle, slow-moving semantic consolidation without longer or broader corpora.

Future work will target three main venues: *a.*) extending temporal coverage and data sources (e.g., news, reviews, other social platforms) to test whether diachronic reconceptualization emerges over longer spans; *b.*) applying multilingual and cross-platform analyses to assess generalizability beyond English Reddit; *c.*) integrating behavioral data to link linguistic indicators with downstream adoption.

8. Bibliographical References

Emily Allaway and Kathleen McKeown. 2021. A unified feature representation for lexical connotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2145–2163, Online. Association for Computational Linguistics.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Iuri Y.F. Baptista and Hendrik N.J. Schifferstein. 2023. Milk, mylk or drink: Do packaging cues affect consumers' understanding of plant-based products? *Food Quality and Preference*, 108:104885.

Valerio Basile, Tommaso Caselli, Anna Koufakou, and Viviana Patti. 2022. *Automatically Computing Connotative Shifts of Lexical Items*, pages 425–436.

Paweł Brzustewicz and Anupam Singh. 2021. Sustainable consumption in consumer behavior in the time of covid-19: Topic modeling on twitter data using lda. *Energies*, 14(18).

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Eugenio Demartini, Daniel Vecchiato, Livio Finos, Simone Mattavelli, and Anna Gaviglio. 2022. Would you buy vegan meatballs? the policy issues around vegan and meat-sounding labelling of plant-based meat alternatives. *Food Policy*, 111:102310.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Michael Elhadad. 2010. Natural language processing with python steven bird, ewan klein, and edward loper. *Computational Linguistics*, 36:767–771.

Boelie Elzen, Barbara van Mierlo, and Cees Leeuwis. 2012. Anchoring of innovations: Assessing dutch efforts to harvest energy from glasshouses. *Environmental Innovation and Societal Transitions*, 5:1–18.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Song Feng, Ritwik Bose, and Yejin Choi. 2011. Learning general connotation of words using graph-based algorithms. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1103, Edinburgh, Scotland, UK. Association for Computational Linguistics.

J. R. Firth. 1957. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, pages 10–32.

Dirk Geeraerts. 2009. *Theories of Lexical Semantics*. Oxford University Press.

Gábor Győri. 2002. Semantic change and cognition. *Cognitive Linguistics*, 13:123–166.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2–3):146–162.

Narine Harutyunyan. 2024. Lexicon leap: Unveiling the neologisms of modern world. *Foreign Languages in Higher Education*, 28:20–40.

Sanne Hoeken, Özge Alacam, Antske Fokkens, and Pia Sommerauer. 2023. Methodological insights in detecting subtle semantic shifts with contextualized and static language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3662–3675, Singapore. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2018. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Md. Shymon Islam and Kazi Masudul Alam. 2023. Sentiment analysis on bangla food reviews using machine learning and explainable nlp. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Madeline Judge, Thijs Bouman, Linda Steg, and Jan Bolderdijk. 2024. Accelerating social tipping points in sustainable behaviors: Insights from a dynamic model of moralized social change. *One Earth*, 7.

Jun Seok Kang, Song Feng, Leman Akoglu, and Yejin Choi. 2014. ConnotationWordNet: Learning connotation over the Word+Sense network. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1544–1554, Baltimore, Maryland. Association for Computational Linguistics.

Danae Marshall, Faiza Bano, and Kasia Banas. 2022. A meaty issue: The effect of meat-related label terminology on the willingness to eat vegetarian foods. *Food Quality and Preference*, 96:104413.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

K.S. McCarthy, M. Parker, A. Ameerally, and S.L. Drake. 2017. Drivers of choice for fluid milk versus plant-based alternatives: What are consumer perceptions of fluid milk? *Journal of Dairy Science*, 100.

Zan Mlakar, Jan Bolderdijk, Hans Risselada, Bob Fennis, Mengbin Ye, Lorenzo Zino, and Ming Cao. 2024. Social tipping games: Experimental paradigms for studying consumer movements. *Journal of the Association for Consumer Research*, 9.

Annika Molenaar, Eva L Jenkins, Linda Brennan, Dickson Lukose, and Tracy A McCaffrey. 2024. The use of sentiment and emotion analysis and data science to assess the language of nutrition-, food- and cooking-related content on social media: a systematic scoping review. *Nutrition Research Reviews*, 37(1):43–78.

Eetu Mäkelä. 2022. *Retronyms and neonyms: A corpus-based study*. Master's thesis, Tampere University, Tampere, Finaland.

Ilona M. Otto, Jonathan F. Donges, Roger Cremades, Avit Bhowmik, Richard J. Hewitt, Wolfgang Lucht, Johan Rockström, Franziska Allerberger, Mark McCaffrey, Sylvanus S. P. Doe, Alex Lenferna, Nerea Morán, Detlef P. van Vuuren, and Hans Joachim Schellnhuber. 2020. Social tipping dynamics for stabilizing earth's climate by 2050. *Proceedings of the National Academy of Sciences*, 117(5):2354–2365.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*, 56(11):1–38.

Chuitian Rong, Zhaopei Liu, Na Huo, and Huabo Sun. 2019. Exploring chinese dietary habits using recipes extracted from websites. *IEEE Access*, 7:24354–24361.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233.

Aftab Siddique, Ashish Gupta, Jason Sawyer, T. Huang, and Amit Morey. 2025. Big data analytics in food industry: a state-of-the-art literature review. *npj Science of Food*, 9.

Anupam Singh and Aldona Glińska-Neweś. 2022. Modeling the public attitude towards organic foods: a big data and text mining approach. *Journal of Big Data*, 9.

Ineke Sluiter. 2016. Anchoring innovation: A classical research agenda. *European Review*, 25:1–19.

John Suler. 2004. The online disinhibition effect. *Cyberpsychology and behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 7:321–6.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors. 2021. *Computational approaches to semantic change*. Number 6 in Language Variation. Language Science Press, Berlin.

S. P. Wagner. 2021. Consumer attitudes toward milk and plant-based dairy alternative beverages.

Shufeng Xiong, Wenjie Tian, Haiping Si, Guipei Zhang, and Lei Shi. 2024. A survey of the applications of text mining for the food domain. *Algorithms*, 17(5).

Greta Zella, Jan Bolderdijk, Tommaso Caselli, and Saskia Peels-Matthey. 2025. The role of neologisms in the climate change debate: Can new words help to speed up social change? *Wiley Interdisciplinary Reviews: Climate Change*, 16.

9. Appendices

9.1. Appendix A - noisy keywords

The manually isolated keywords (and phrases), based on a sample of 500 comments are reported

in Table 5. While some of the listed keywords may not completely eliminate unwanted content, or on the contrary, eliminate some comments that are not entirely unrelated to the topic of protein transition, we aimed at a finding a good balance between the two.

Keyword/phrase	Reason
remember the milk	computer app
cock	sexual, explicit
beat my meat	sexual, explicit
porn	sexual, explicit
have beef with	figurative
turkey calls	hunting device, unrelated
leader of turkey	Turkey (country)
kindler egg	typo, confectionary product
reel big fish	music band
the milk man	figurative, joke
your man meat	sexual, explicit
chicken pox	disease
left to get milk	figurative
string cheese incident	music band
lay eggs	biological, unrelated
laid a egg	biological, unrelated
pregnancies	biological, unrelated
egg them	figurative
like dead fish	figurative
ages like milk	figurative
i'm a bot	automation
nipples	sexual, explicit
titty	sexual, explicit
chicken and egg	figurative
catching fish	social, recreational
caught a fish	social, recreational

Table 5: Data distribution across the time periods and the communities

10. Acknowledgments

This study was partly funded by the University of Groningen. This research was partly funded by Anchoring Innovation. Anchoring Innovation is the Gravitation Grant research agenda of the Dutch National Research School in Classical Studies, OIKOS. It is financially supported by the Dutch ministry of Education, Culture, and Science (NWO project number 024.003.012). For more information about the research program and its results, see the website www.anchoringinnovation.nl. This article was written as part of a research project carried out in the context of a Dutch research grant (NWO-016.Veni.185.103, title: Polytheism as language. A linguistic approach to divine plurality in the religious experience of Greek worshippers).