Learning Through News: Bridging the Gap Between Algorithmic Recommendation and Human Curation

Florian Debaene*, **Loic De Langhe***, **Orphée De Clercq, Véronique Hoste**
LT^3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

News recommendation systems play a central role in how readers access and process current events. Most recommenders' underlying algorithmic strategies, however, prioritize user engagement over comprehension, amplifying risks of misinformation and filter bubbles. This study investigates whether fine-grained content-based recommendation strategies favor human knowledge retention and explores how such a content-based recommendation can be operationalized using event coreference–based document modeling. To this purpose, we first measure the effect of manually curated content-based news recommendation on knowledge retention across five news topics with 126 Dutch speaking participants. Next, we investigate document retrieval by comparing a state-of-the-art event coreference resolution system for Dutch which recommends news articles based on event chains with a document similarity retrieval baseline using state-of-the-art embedding models in three increasingly more complex test settings. The results demonstrate that human-curated content-based recommendation can positively and significantly impact readers' knowledge retention. Moreover, we show that a fine-grained coreference system can approach said level of human curation better than state-of-the-art document retrieval methods. In general, this holds potential for scalable, comprehension-oriented news recommendation.

Keywords: Content-based Recommendation, Event Coreference Resolution, Document Retrieval

1. Introduction

Digital news consumption is increasingly shaped by recommendation algorithms deciding which news consumers encounter. As such, news recommenders not only influence which information consumers access, but possibly also their understanding of current events. This coincides with wider changes in news consumption patterns, such as the rising role of social media and video platforms as the primary news sources for a significant part of the consumer base (Newman, 2025).

Two main recommendation paradigms can be distinguished: collaborative filtering and content-based filtering (Tursunov et al., 2025). Collaborative filtering recommends articles to readers based on their individual preferences and the behaviors of other readers with similar reading profiles. While this approach efficiently boosts user engagement, it also reinforces homogeneity and can create so-called filter bubbles (Pariser, 2011; Nguyen et al., 2014), which limit the readers' exposure to other types of news or viewpoints. In contrast, content-based filtering recommends new articles by comparing textual or topical features of previously read items. Typically, such algorithms are based on keyword overlap, topic classification, or metadata similarity. Although content-based approaches aim to recommend articles of interest, their reliance on superficial lexical similarity makes them ill-suited to identify semantically related or complementary news items which differ in wording, framing, or narrative focus. As a result, these systems often cluster articles with overlapping keywords together while overlooking those that report on the same event from different perspectives or that provide additional, contextually relevant information.

These limitations have raised concerns about the civic and epistemic consequences of algorithmic news personalization. As Joris et al. (2019) and Colruyt et al. (2023) argue, news recommenders are not neutral intermediaries. They operationalize a commercial logic of attention and engagement, which not necessarily corresponds to the primary goal of journalism to foster an informed and critically aware readership and citizenship. A growing body of work therefore calls for the rethinking of personalization in recommendation systems, away from engagement maximization and towards algorithms which enhance contextual understanding and informed news consumption (Shivaram et al., 2022; Ludwig et al., 2023; Joris et al., 2024).

Within this context, content-based recommendation is the preferred starting point for developing citizen-oriented news recommenders. However, the semantic fine-grained modeling required to meaningfully relate news articles across outlets and time remains underexplored. A promising direction is event coreference resolution, an NLP task that detects and links references to the same real-world events across documents. By automat-

* Equal contribution

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 778–794
11–16 May 2026.

ically constructing event chains across news articles describing related or identical events, event coreference models hold the potential to enable more coherent, context-rich news recommendations. Moreover, when the same events can be linked across documents this provides readers access to different points of view.

In this paper we first compare the effect of manually curated content-based news recommendation on knowledge retention across five news topics with 126 Dutch speaking participants. Next, we investigate the feasibility of applying event coreference resolution to automate such recommendations. Our results reveal that fine-grained content-based recommendation impacts readers' knowledge retention when relying on a human-curated set. Moreover, we demonstrate the added value of event coreference resolution to approach this level of human curation when incorporated into fine-grained content-based recommendation, outperforming state-of-the-art document retrieval methods.

The remainder of this paper is structured as follows. Section 2 reviews prior work on recommendation algorithms and event coreference resolution in the context of end-to-end recommendation systems. Section 3 examines the effect of manually curated content-based news on readers' knowledge retention, while Section 4 introduces state-of-the-art Dutch event coreference models to approximate such an ideal recommendation scenario through automated document retrieval. Section 5 presents the results and Section 6 concludes with implications for media literacy, algorithmic accountability, and the design of comprehension-oriented news recommendation systems.

2. Related Work

2.1. Recommendation Algorithms

News recommendation systems aim to automatically deliver relevant articles to users by learning both user interests and content representations. Unlike static recommendation domains such as movies or products, news recommendation presents unique challenges due to high content turnover, topic drift, and the need for contextual as well as temporal awareness. As such, methods must model not only user preferences, but also the rapidly evolving semantic structure of news.

Early systems relied on computing similarity between articles using lexical overlap or TF-IDF representations (Pazzani and Billsus, 2007; Lops et al., 2010). These approaches offered limited semantic understanding, often recommending lexically similar but topically divergent articles. Collaborative filtering methods (Sarwar et al., 2001; Koren et al., 2009) later introduced user-based and item-based similarity derived from behavioral data such as clicks or reading histories. However, they struggled with data sparsity and the cold-start problem, both common in dynamic news environments where items quickly become obsolete.

The advent of neural encoders and representation learning marked a major shift in document similarity modeling, from surface-level lexical matching to deep semantic understanding. Embedding-based approaches began representing text in dense vector spaces using models such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), enabling semantic similarity computation between documents beyond exact word overlap. This paradigm was further advanced by contextual encoders such as BERT (Devlin et al., 2019), which capture fine-grained semantic relations through bidirectional attention. Building on these encoder-based representations, subsequent neural architectures have been designed to improve document similarity estimation. For instance, Wu et al. (2019) employed attentive multi-view encoders to align document representations from titles and abstracts, while An et al. (2019) enhanced document matching by modeling hierarchical semantic relations across news content. Overall, neural encoders thus serve as the foundation for computing document similarity, which in turn drives the effectiveness of modern news recommendation systems.

Recent advances have incorporated context-aware modeling to capture interactions among users, news, entities, and topics over time. Graph Neural Networks (GNNs) have been applied to represent user–item–entity relationships (Sheu and Li, 2020; Hu et al., 2020), enabling information propagation across related news items and improving recommendation diversity and temporal coherence. Such approaches better reflect the structural and temporal nature of news consumption, allowing systems to follow evolving storylines rather than isolated articles.

2.2. Event Coreference Resolution

Event Coreference Resolution (ECR) is a discourse-level NLP task that identifies textual mentions referring to the same real or fictional event.

1. Elon Musk [completes]$_{Event-A}$ $44 billion deal to own Twitter after [extended negotiations]$_{Event-B}$

2. Elon Musk's contested and tumultuous Twitter [acquisition]$_{Event-A}$ finalized after [long discussions]$_{Event-B}$

While humans can easily infer that (1) and (2) describe the same pair of real-world events—(*A*)

Cl	Description	IPTC Topic	Articles	CD Events
4	Topic 1: Soyuz Spacecrafts	Science and Technology	15	98
14	Topic 2: Disappearance of MH370	Disaster and Accident	13	124
35	Topic 3: Hedwigepolder (Zeeland, Netherlands)	Environment	9	19
49	Topic 4: 2018 Belgian energy crisis	Society	14	53
99	Topic 5: Facebook Cambridge Analytica	Science and Technology	27	154

Table 1: ENCORE cluster number (Cl), their description, IPTC (International Press Telecommunications Council) topic, number of articles and the total number of CD (Cross-Document) annotated events that span at least 2 documents.

Musk's acquisition and (*B*) the negotiations itself, this level of fine-grained reasoning is difficult for automated systems. Traditional approaches based on document similarity or clustering treat texts as single units, often conflating unrelated content that shares topical or lexical overlap. ECR, by contrast, operates at the mention-level, linking individual event mentions across and within documents. This granularity enables more precise and semantically coherent modeling of event continuity which is crucial for downstream applications such as large-scale news summarization (Liu and Lapata, 2019), information extraction (Humphreys et al., 1997), and event-centric recommendation systems (Vermeulen, 2018).

Despite advances in Large Language Models (LLMs) that demonstrate strong discourse understanding (Ravi et al., 2023; Liu et al., 2023; Zhang et al., 2023), robust large-scale ECR remains difficult, especially in multilingual and cross-document contexts where explicit contextual cues are sparse (Lu and Ng, 2018). Nevertheless, ECR's ability to capture nuanced event relations offers a clear advantage over document-level methods that rely on topic similarity or shared entities rather than event identity. Methodologically, ECR builds on the foundations of entity coreference resolution (Rahman and Ng, 2009), with most systems adopting the mention-pair paradigm, classifying whether two mentions refer to the same event. Early approaches used decision trees (Cybulska and Vossen, 2015), SVMs (Chen et al., 2015), and neural networks (Nguyen et al., 2016), later replaced by transformer-based and span-based architectures such as SpanBERT (Joshi et al., 2020; Cattan et al., 2021; Lu and Ng, 2021). Although neural encoders generally dominate, many classical insights remain relevant: lexical and semantic similarity, discourse proximity, and event distance continue to inform model design (Yao et al., 2023; De Langhe et al., 2025).

3. Knowledge Retention Experiment

ECR remains a difficult task, especially within cross-document content. This is why we decided to first examine the effect of fine-grained content-based news recommendation on Dutch-speaking readers' knowledge retention by comparing a recommendation setting where participants are presented with either a *gold-standard* or *random* set of news articles.

3.1. Data

We relied on the ENCORE corpus (De Langhe et al., 2023) to extract the news articles for this experiment. This corpus consists of a collection of 1,115 Dutch news texts in which coreference between news events has been annotated, both at a within- and cross-document level. Within ENCORE, articles with the same overarching topics are grouped into clusters. For our experiments we manually scrutinized these and chose five clusters (see Table 1 for more details) based on whether sufficient news events within that cluster were actually grouped into cross-document event chains (CD Events) and whether the overarching IPTC News Media Topics[1] were sufficiently diverse. We also made sure that the topics covered both global and local news items. Within the chosen ENCORE clusters the number of articles ranged from 9 to 27. We again manually went through all the articles and selected five articles per cluster, each describing the same news events, but approached from different angles and ranging from more general to very fine-grained descriptions (e.g. the five articles describing the "Hedwigepolder" specifically focus on the last week before the depoldering and mention both nature preservation efforts as well as tips to visit this nature reserve). We also made sure that the articles were not too long given that for the experiments participants would be asked to read all those articles during a restricted period of time. We did not alter the articles in any way.

In order to test human knowledge retention we subsequently constructed a ten-item multiple-choice (MC) test for each topic to measure readers' recall and understanding of factual content. This was done by iteratively going through the five articles per cluster and asking questions related to the

[1] https://iptc.org/standards/media-topics/

chain of events and entities that were discussed in the articles often targeting outlines, but also details, causes and outcomes were questioned. MC questions ranged in difficulty, with easier True/False questions to harder questions requiring interpretation of the news event. "Don't know" options were also included to minimize guessing. For each question, there is only one correct answer. In total, this resulted in 50 MC questions, 10 per topic, as listed in Appendix A.

3.2. Experimental Setup

Participants were recruited via the online research platform Prolific and randomly and evenly assigned to the two experimental conditions. In the random condition, two fixed articles from the assigned cluster were presented together with three random articles from the other four clusters. In the gold-standard condition, all five articles from the assigned cluster were presented.

Before starting, all participants were informed that they would read five news articles and subsequently answer the ten content-related MC questions, but they were not informed about the purpose of the study nor about which experimental condition they would be assigned to. Informed consent and data privacy compliance were ensured, and participants could withdraw at any time. Prolific supports targeted participant selection and after a pilot test with 20 participants we decided upon the following inclusion criteria: Dutch as the primary language, Belgian or Dutch nationality, and at least a Bachelor's degree to ensure a sufficient reading proficiency. Demographic metadata (such as gender, age and nationality) were collected automatically through Prolific, as well as completion time and quality indicators (e.g., the participant's approval rate of prior submissions).

Given that news consumption is also influenced by a user's personal preferences, we also had all participants fill in a survey on their news consumption habits, the questions of which were inspired by the Flemish imec.digimeter report (De Marez et al., 2024). We also asked the participants whether they had a specific interest in one of the chosen IPTC News Media Topics. See Appendix B for an overview of all surveyed news consumption habits. The estimated completion time of this entire study was established in the same pilot test and set at 20 minutes. All studies submitted by participants were manually verified before payment to exclude incomplete or fraudulent responses. These were easily recognizable by implausible completion times or through Prolific's built-in bot detection.

We hypothesized that participants exposed to the gold-standard news article set would achieve higher scores on the MC test than those in the random condition. Therefore, we mainly investigated the variation in the test results depending on the experimental conditions (random vs. gold-standard). As explained above, the 10-item MC test was designed by manually going through all articles and asking both broad and more detailed questions. Given that the participants in the random setting only had access to two articles of a given topic it is of course to be expected that answering all questions might be more difficult. We therefore made sure that for each topic more than half of the questions would be answerable by only reading through those two articles (and assuming that participants had no prior knowledge of the topic). In total, this was the case for 39 out of the 50 questions.

Linear regression analyses were then conducted with the test scores as the dependent variable and with the experimental condition as independent variable. This was tested using the Ordinary Least Squares (OLS) method (Montgomery et al., 2021) to verify whether a predictor can explain the variance in the dependent variable, namely the MC test score. For each variable, we report the average score, distribution, statistical significance (p-values) and effect size (Cohen's *d*).

3.3. Results

A total of 126 participants took part in the study, the population was well-balanced regarding nationality (Belgium = 61, The Netherlands = 65), gender (male = 62, female = 64) and age (millennial or older = 61, generation z = 65). The average test score per participant on the 10-item MC test was 7 correct answers (mean = 7.3) and we did not notice any notable differences in this average among these three demographics (Table 2). Regarding the news engagement, what draws the attention is that most participants indicated that they barely engage with news (73.8%) and that on average those participants scored higher on the test. We do not really notice an effect in test performance when people are interested in the topic. We give a full overview of all investigated variables in relation to participant scores on the MC test in Appendix C.

If we consider the scores on the 10-item MC test in both experimental settings we observe in Figure 1 that the gold-standard group outperformed the random group (scoring, on average, 1.42 points higher). Looking at the actual answers, we also observed that the random group ticked the option "Don't know" more frequently (17.5% of the answers versus 9.52%). If we take into account the average number of questions that should be answerable by reading only two articles of a certain topic – which amounts to an average score of 7.8 (represented by the dashed line)– we observe that the random group struggles more to answer all those questions, whereas the gold-standard group is more successful, possibly because certain events have

	gender		nationality		>29y		news_engagement			interest_topic		
	♂	♀	BE	NL	no	yes	barely	daily	weekly	neutral	int	no_int
n=	62	64	61	65	65	61	93	25	8	30	44	52
score	7.56	7.05	7.59	7.05	7.51	7.12	7.88	7.20	7.52	7.67	7.46	6.98

Table 2: Selected demographic and news consumption variables in relation to the test scores.

been repeated multiple times.

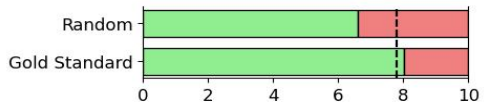

Figure 1: Scores per condition on all 50 questions. The dashed line indicates the threshold of the 39 answerable questions for the random condition.

To further investigate this, we conducted the regression analysis on both the 10-item MC test and the reduced set of answerable questions. Statistical significance and the effect size (Cohen's *d*) are calculated on the premise that the first variable per factor be the baseline state of that factor, meaning that the other variable is treated as a different state. The impact of the transformation from the baseline to these other states is then reported in the statistical significance, Cohen's *d* and the effect size. We use standard significance thresholds, namely: $p < 0.05*$, $p < 0.01**$, and $p < 0.001***$, where smaller values denote increasingly stronger evidence against our observations occurring randomly. We interpret effect size based on Cohen's *d*, with $d < 0.2$ indicating negligible, $0.2 < d < 0.5$ small, $0.5 < d < 0.8$ medium and $0.8 < d$ indicating large effects, to report the magnitude or relevance of observations beyond their statistical significance.

The results in Table 3 show that when all questions are considered, the experimental condition comes out as a strong and highly significant predictor, hinting at better knowledge retention among the group that was each time presented with five semantically coherent news articles within a topic. If we only consider those questions that were answerable by reading the two fixed articles per topic, we notice that the effect is smaller, but nevertheless significant ($p = 0.0156$).

	10-item MC		Answerable	
predictor	random	gold	random	gold
n=	64	62	64	62
score	6.61	8.03	7.36	8.02
stat_sig	***	***	*	*
d	–	0.83	–	0.44
effect	–	large	–	small

Table 3: Results regression analyses.

When we considered the topics separately, we found some clusters were consistently assessed easier or harder depending on their content. Participants performed best on the *Facebook & Cambridge Analytica* (mean score = 8.22) and *MH370* topics (7.6) and lowest on the more specialized and local Belgian news, i.e. the *Hedwigepolder* (6.71) and the *2018 Belgian energy crisis* (6.76).

In summary, being exposed to semantically coherent article chains seems to have an effect on knowledge retention in this study, though we do want to stress that this was only a small-scale study and that more experiments on larger datasets and with a larger pool of participants would need to be conducted to corroborate these findings.

4. Event Coreference Recommendation

We observed that a carefully curated set of recommendations could significantly enhance news retention and help readers develop a deeper understanding of the content they are engaging with. Such curated recommendations not only reinforce key information but also provide complementary perspectives that enrich the reader's knowledge. However, manually assembling these complementary and diverse recommendation sets, as was done in our study, is not feasible for large-scale or real-time applications. Consequently, in this section, we turn our attention to evaluating how effectively automated recommendation systems can approximate this human-curated gold standard.

Our main hypothesis is that fine-grained event-level approaches, such as event coreference systems, have the potential to generate more effective recommendations than the coarse-grained, similarity-based methods commonly used today. By identifying and linking specific events across articles, these systems can provide readers with targeted, contextually relevant content that goes beyond surface-level topic overlap.

4.1. Experimental Setup

We conceptualize news recommendation as an Information Retrieval (IR) task, where the goal is to provide the reader with a set of relevant and complementary articles to the one they are currently engaging with. Formally, given a collection of articles $\mathcal{C}$ and an anchor article $a \in \mathcal{C}$, the goal is to retrieve a set of related articles $\mathcal{R} = \{r_1, r_2, \ldots, r_n\} \subseteq \mathcal{C} \setminus \{a\}$ that provide complementary information

to a. This facilitates a comprehensive yet relevant understanding of the topic, enabling the identification of fine-grained complementary information that is typically challenging to capture with conventional coarse-grained retrieval methods.

Formally, the retrieval task can be expressed as:

$$\mathcal{R} = \arg \max_{r \in \mathcal{C} \setminus \{a\}} \text{Rel}(a, r)$$

where $\text{Rel}(a, r)$ quantifies an arbitrary degree of (semantic) relevance of article r with respect to a.

In order to retrieve the most relevant and contextualizing documents to the anchor we use the fine-grained end-to-end ECR model introduced by De Langhe et al. (2025). This model builds on a commonly-used joint cross-document coreference system (Cattan et al., 2021), which uses transformer-based encoders to detect and link event mentions across documents via span classification and pairwise scoring. This baseline architecture is extended with three modifications: (1) positional features to reduce lexical over-reliance, (2) within-document-only inference for computational efficiency, and (3) a Variational Graph Auto-Encoder (VGAE) (Kipf and Welling, 2016) that learns graph representations of event chains and connects within-document clusters into cross-document event graphs. This system takes as input multiple documents of raw text and produces a set of intra- and cross-document coreference links between detected textual events. We compute the relevancy score $\text{Rel}(a, r)$ between the anchor *a* and query document *r* by taking the absolute number of cross-document coreference links between those two documents. The final ranking of candidate documents is obtained by ordering all $r \in \mathcal{C} \setminus \{a\}$ in descending order of their relevance scores.

To establish a strong baseline for comparison with our fine-grained recommender, we test two alternative methods of retrieving relevant documents based on a more commonly used technique of coarse-grained document similarity. Specifically, we encode each document $d \in \mathcal{C}$ into an embedding $\mathbf{e}_d \in R^m$ using several encoder baselines: *BERT-base-dutch-cased*, which served as the base model for our ECR recommender, E5-large-trm, which is the current state-of-the-art encoder for the Dutch language (Banar et al., 2025), the widely-used multilingual sentence encoder *paraphrase-miniLM* and *gemini-embedding-001* as well as the current state-of-the-art multilingual embedding model on the MTEB benchmark (Muennighoff et al., 2022). If a document d exceeds the maximum input length L supported by the encoder, it is split into k contiguous chunks $\{d_1, d_2, \ldots, d_k\}$, each of length at most L. Each chunk is embedded separately, and the document embedding is obtained by averaging over all chunk embeddings:

$$\mathbf{e}_d = \frac{1}{k} \sum_{i=1}^{k} \mathbf{e}_{d_i}.$$

Once all document embeddings are computed, we rank all candidate articles based on their semantic similarity to the anchor article a. The relevance score between a and any candidate article r is defined as the cosine similarity between their embeddings. The final ranking is obtained by ordering all $r \in \mathcal{C} \setminus \{a\}$ in descending order of $\text{Rel}(a, r)$.

4.2. Data

As in Section 3.1, the core dataset for these retrieval experiments is the Dutch ENCORE news corpus. Within- and cross-document coreference links in this corpus were manually annotated if two events happen at the same time (temporal alignment), place (spatial alignment) and have the same participants (actorial alignment). The complete dataset consists of 1,115 documents and is split into 91 topical clusters.

We designate all articles belonging to the clusters listed in Table 1 as a held-out test set, and train our end-to-end ECR recommendation model on the remaining topical clusters in the corpus. This setup ensures that there is no topical overlap between the model's training data and the documents used for retrieval, thereby providing a clean evaluation of generalization across unseen topics. Concretely, this means that 87 topical clusters, comprising a total of 1,047 articles are used for training and the remaining 5 topical clusters, comprising 78 articles (cfr. Table 1) are used for testing in the standard experimental setup. For the remainder of this paper, we refer to this newly trained model as ECR-ENCORE.

To approximate a real-world setting where retrieval must be performed over large document collections, we design three experimental scenarios to evaluate the retrieval capabilities of our models. First, in the *cluster-only* setting, the document collection $\mathcal{C}$ is restricted to the articles within the same topical cluster as the anchor. Second, in the *random* setting, we expand $\mathcal{C}$ by randomly selecting 100 additional articles from the remaining corpus. In this case, the end-to-end recommender is re-trained while excluding these specific articles to prevent data leakage. Third, in the *adversarial* setting, we expand $\mathcal{C}$ with articles that are, on average, most semantically similar to each of the test clusters listed in Table 1. Semantic similarity between clusters is computed analogously to the cosine similarity between articles described in Section 4.1. For each test cluster, we add the 100 most similar documents from the collection to create the adversarial dataset.

Setting	System	MAP@k	Recall@k	Prec@k	Recall@10	Prec@10
Cluster	BERT-base-dutch-cased	0.61	**0.50**	0.50	0.65	0.26
	e5-large-trm	0.57	0.45	0.45	0.70	0.28
	paraphrase-miniLM	0.67	0.45	0.45	**0.75**	0.30
	Gemini-embedding-001	0.63	0.45	0.45	0.60	0.24
	ECR-ENCORE	**0.80**	0.40	**0.65**	0.40	**0.57**
Random	BERT-base-dutch-cased	0.61	**0.50**	0.50	0.65	0.26
	e5-large-trm	0.57	0.45	0.45	0.70	0.28
	paraphrase-miniLM	0.67	0.45	0.45	**0.75**	0.30
	Gemini-embedding-001	0.63	0.45	0.45	0.60	0.24
	ECR-ENCORE	**0.80**	0.40	**0.65**	0.40	**0.57**
Adversarial	BERT-base-dutch-cased	0.57	0.40	0.40	0.60	0.24
	e5-large-trm	0.52	**0.45**	0.45	**0.70**	0.22
	paraphrase-miniLM	0.64	**0.45**	0.45	**0.70**	0.28
	Gemini-embedding-001	0.63	**0.45**	0.45	0.55	0.22
	ECR-ENCORE	**0.80**	0.40	**0.65**	0.40	**0.57**

Table 4: Retrieval performance of three systems across three data settings. Each query has four relevant documents (binary relevance). Bold values indicate the best system per setting.

4.3. Evaluation

For each topical cluster in Table 1 we randomly select one anchor article, based on which the other 4 articles in the cluster should be retrieved. In total, our model evaluation therefore encompasses 5 total queries (one for each of the topical clusters), with each 4 relevant documents in the set $\mathcal{R}$ and a variable size of the document collection $\mathcal{C}$, depending on the topical clusters in question and the setting (cluster, random, adversarial).

To evaluate retrieval performance, we employ standard Information Retrieval metrics: Mean Average Precision at k (MAP@k), Recall@k, and Precision@k. These metrics jointly capture both the accuracy and completeness of the retrieved results. MAP@k is particularly informative, as it accounts not only for whether relevant documents are retrieved but also for their position in the ranked list, thereby rewarding systems that rank relevant items higher. Given that each query in our setting has exactly four relevant documents, we set $k = 4$ to align the evaluation cut-off with the maximum number of relevant items. In addition, we report Precision@10 and Recall@10 to simulate more realistic use cases, reflecting that in practice the total number of relevant documents is typically unknown and users may inspect the top-ranked portion of a larger document collection.

5. Results and Discussion

5.1. Results

The results of our retrieval experiments using the three systems can be found in Table 4. When evaluating retrieval in the cluster-only setting, the results indicate that the ECR-based recommender can retrieve a substantially higher amount of relevant documents compared to its baseline counterparts. Concretely, both Precision@k and Precision@10 are notably higher for the fine-grained system, indicating that in general the model is more accurate in retrieving documents. Note, however, that the higher score for the Precision@10 metric is partly due to the fact that the ECR-based system retrieves fewer-than-k (especially in the case where k = 10) articles for each query. In these cases, the precision is computed by dividing by the actual number of predicted documents rather than 10, inflating the metric somewhat. Additionally, the ECR-based recommender also achieves the highest MAP@k, reflecting its ability to rank relevant documents far more effectively at the top of the list when compared to the coarse-grained baselines. In general, the fine-grained system thus retrieves less documents (hence the lower recall values across the board), but is more accurate in the documents that it does retrieve. It should be noted that the lower recall does not imply worse performance: in practical news recommendation, retrieving a smaller set of highly relevant documents is generally preferable, as users cannot realistically skim through dozens or hundreds of articles to find useful content.

Another insight is that the scores change relatively little when the document collection $\mathcal{C}$ is supplemented with articles that do not belong to the relevant topical cluster. When these additional articles are selected randomly, only the baseline monolingual Dutch encoder suffers a slight performance drop, indicating little impact from the introduced noise. The effect of the noise is more noticeable in the adversarial setting, where we observe a slight drop in performance across the board for the recommenders based on document similarity. How-

ever, this is mostly the case in a setting where ten documents are retrieved, indicating that relevant documents are still ranked favorably. Finally, it is worth highlighting that the fine-grained ECR recommender does not experience any drop in performance in either settings.

5.2. Discussion

5.2.1. Influence of Distractor Articles

To gain further insight into model behavior in the random and adversarial retrieval settings, we visualize the proportion of distractor documents, i.e., retrieved articles that are not in the same topical cluster as the anchor, when each system is asked to return ten documents per query. We focus on $k = 10$ because, as shown in Table 4, the overall performance metrics did not vary much across the adversarial experiments when retrieving only 4 documents, indicating a fairly stable retrieval at this cut-off. Four out of five topical clusters examined for the experiments have at least 10 documents, except for cluster 35, where we set k at 9 (cfr. Table 1).

For each model, we compute the proportion of distractor documents among the ten retrieved items per query and then aggregate these values across all clusters. This yields, for each system, the total number of extra-cluster articles retrieved out of a possible fifty. Although this measure is secondary to performance on the four truly relevant articles linked to the anchor, it nonetheless provides valuable insight into each system's ability to distinguish topically similar but ultimately irrelevant content from genuinely related material.

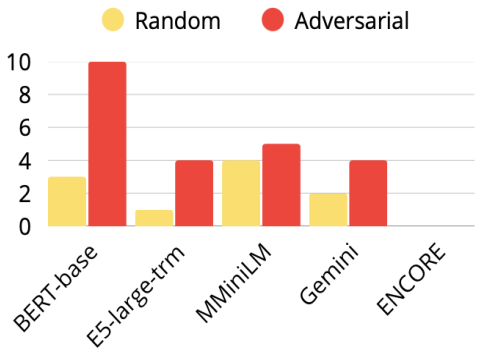

Figure 2: The absolute amount of distractor articles retrieved at k = 10 for each of the retrieval systems.

The results are displayed in Figure 2. As shown in the bar plot, this analysis aligns with our interpretation of the quantitative results presented in Table 4. The baseline Dutch BERT encoder is most affected by the added noise in both the random and adversarial settings, retrieving a higher number of distractor documents. The stronger monolingual and multilingual retrieval systems also produce some topically irrelevant outputs when asked to retrieve ten articles per anchor. However, since the overall performance metrics remain relatively stable, these extra-topical articles appear to be ranked lower than the genuinely relevant ones. Notably, the fine-grained ECR-ENCORE system does not retrieve any distractor article in either setting, underscoring its robustness and precision in fine-grained information retrieval.

5.2.2. Influence of Cluster Size

To investigate the impact of cluster size on model performance, we analyze the relationship between the number of documents in each test cluster and the corresponding number of retrieved relevant documents by different systems. Table 5 below details the number of retrieved relevant documents for each system in the cluster-only setting.

Cluster	Total	BERT	Gemini	ECR-ENCORE
4	15	8	7	4
14	13	9	8	3
35	9	6	6	5
49	14	10	9	4
99	27	5	5	2

Table 5: Correctly retrieved documents per cluster for each system, with total cluster size.

Spearman's rank correlation coefficient is used to quantify whether larger clusters tend to yield higher or lower precision. This analysis allows us to identify which systems are more sensitive to cluster length and how this affects retrieval quality in the form of Precision@k. We find that for BERT-base-dutch-cased, the correlation is moderately negative ($\rho = -0.700$) ($p = 0.188$), while Gemini-embedding-001 shows a stronger negative correlation ($\rho = -0.821$), suggesting a tendency for precision to decrease slightly in larger clusters. In contrast, ECR-ENCORE exhibits a very weak negative correlation ($\rho = -0.154$), indicating that its precision is largely unaffected by cluster size. This aligns with the design of ECR-ENCORE, which selectively predicts only a small number of highly relevant documents per cluster, making it robust to variations in cluster length. Overall, these results suggest that ECR-ENCORE maintains consistent top-ranked precision regardless of cluster size, whereas the baseline encoders show at least some sensitivity to larger clusters.

6. Conclusion

In this paper, we provide a substantiated argument for incorporating fine-grained NLP methods in news recommendation systems. Through a user study, we first demonstrated the impact of gold-standard human curation on the quality and informativeness of recommended articles through knowledge retention experiments. We then explored whether such curated recommendations could be approximated computationally using Event Coreference Resolution (ECR), a task that identifies and links mentions of the same events across documents to generate a nuanced, interconnected representation of news content. Our results show that ECR not only provides tangible benefits in this practical setting, but also outperforms current state-of-the-art embedding-based retrieval methods and this both in precision and contextual relevance.

In our knowledge retention experiments, curated sets of articles were carefully assembled to provide coherent and in-depth coverage of a topic, contrasted with a random selection of articles. Participants exposed to curated sets demonstrated higher comprehension and retention, highlighting the importance of structured content delivery. While human curation is clearly beneficial, it is impractical at scale. Our experiments show that ECR-based recommendation can approximate curated sets more accurately than LLM embeddings: it predicts fewer, but more relevant articles across diverse topics, reducing noise and improving the user experience. Qualitative analysis further reveals that event-based recommendations are robust against semantically similar but irrelevant articles and maintain higher accuracy even in large search spaces. These findings suggest that integrating event-level semantic modeling into recommendation pipelines can substantially improve precision, interpretability, and user engagement.

In future work we see several directions to enhance these benefits: incorporating temporal, geographic, or cross-lingual signals to better capture complementary and follow-up articles, combining ECR with user modeling to personalize recommendations without sacrificing semantic precision and evaluating more long-term effects on knowledge retention, user engagement, and diversity across news sources.

Acknowledgements

This work was supported by the Research Foundation–Flanders (FWO) under project grant number FWO.OPR.2020.0014.01.

Limitations

While the findings of this study suggest the potential of fine-grained, event-based content modeling to foster comprehension-oriented news consumption, we acknowledge the following limitations. First, the knowledge retention experiment was conducted on a relatively small and homogeneous pool (n=126) of individuals with a higher educational background. Although this ensured linguistic and reading proficiency, this limits the generalizability of the results to broader audiences. Future research should focus on including larger and more demographically diverse populations to assess whether similar effects hold across languages, educational levels and cultural contexts. Second, the experimental design necessarily simplified real-world news consumption. Participants were asked to read a fixed number of articles under time constraints and to complete a test immediately after. This setup does not fully capture natural reading behavior which depends on interest, emotion and prior knowledge. Third, the computational comparison between the ECR system and embedding-based retrieval methods was constrained by the available resources and corpus in Dutch. The generalizability of ECR-based retrieval to larger, more heterogeneous news corpora or other languages remains to be validated.

Bibliographical References

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 336–345.

Nikolay Banar, Ehsan Lotfi, Jens Van Nooten, Cristina Arhiliuc, Marija Kliocaite, and Walter Daelemans. 2025. Mteb-nl and e5-nl: Embedding benchmark and models for dutch. *arXiv preprint arXiv:2509.12340*.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. *arXiv preprint arXiv:2106.01210*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 167–176.

Camiel Colruyt, Orphée De Clercq, Thierry Desot, and Véronique Hoste. 2023. Eventdna: a dataset for dutch news event extraction as a basis for news diversification. *Language Resources and Evaluation*, 57(1):189–221.

Agata Cybulska and Piek Vossen. 2015. Translating Granularity of Event Slots into Features for Event Coreference Resolution. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Loic De Langhe, Orphée De Clercq, and Veronique Hoste. 2023. Constructing a cross-document event coreference corpus for dutch. *Language Resources and Evaluation*, 57(2):819–848.

Loic De Langhe, Orphée De Clercq, and Veronique Hoste. 2025. Position-aware end-to-end cross-document event coreference resolution for dutch. *Natural Language Processing Journal*, page 100184.

Lieven De Marez, R Sevenhant, F Denecker, A Georges, G Wuyts, and D Schuurman. 2024. Imec. digimeter. 2023. *Digitale trends in Vlaanderen. Imec*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Linmei Hu, Siyong Xu, Chen Li, Cheng Yang, Chuan Shi, Nan Duan, Xing Xie, and Ming Zhou. 2020. Graph neural news recommendation with unsupervised preference disentanglement. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4255–4264.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Proceedings of the ACL/EACL Workshop on Operational Factors in Practical, Robus Anaphora Resolution for Unrestricted Texts*, pages 75–81.

Glen Joris, Camiel Colruyt, Judith Vermeulen, Stefaan Vercoutere, Frederik De Grove, Kristin Van Damme, Orphée De Clercq, Cynthia Van Hee, Lieven De Marez, Veronique Hoste, et al. 2019. News diversity and recommendation systems: Setting the interdisciplinary scene. In *IFIP International Summer School on Privacy and Identity Management*, pages 90–105. Springer.

Glen Joris, Stefaan Vercoutere, Orphée De Clercq, Kristin Van Damme, Peter Mechant, and Lieven De Marez. 2024. Nudging towards exposure diversity: Examining the effects of news recommender design on audiences' news exposure behaviours and perceptions. *Digital Journalism*, 12(8):1118–1139.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, pages 1–43.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *arXiv preprint arXiv:1905.13164*.

Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2010. Content-based recommender systems: State of the art and trends. *Recommender systems handbook*, pages 73–105.

Jing Lu and Vincent Ng. 2018. Event Coreference Resolution: A Survey of Two Decades of Research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5479–5486, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Jing Lu and Vincent Ng. 2021. Conundrums in event coreference resolution: Making sense of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1380.

Katharina Ludwig, Alexander Grote, Andreea Iana, Mehwish Alam, Heiko Paulheim, Harald Sack, Christof Weinhardt, and Philipp Müller. 2023. Divided by the algorithm? the (limited) effects of content-and sentiment-based news recommendation on affective, ideological, and perceived polarization. *Social Science Computer Review*, 41(6):2188–2210.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their com-

positionality. *Advances in neural information processing systems*, 26.

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. 2021. *Introduction to linear regression analysis*. John Wiley & Sons.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Nic Newman. 2025. Digital news report 2025: Executive summary. Technical report, University of Oxford.

Thien Huu Nguyen, Adam Meyers, and Ralph Grishman. 2016. New york university 2016 system for kbp event nugget: A deep learning approach. In *TAC*.

Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web - WWW '14*, pages 677–686, Seoul, Korea. ACM Press.

Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*, pages 325–341. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 968–977.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwarz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. *arXiv preprint arXiv:2302.09715*.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295.

Heng-Shiou Sheu and Sheng Li. 2020. Context-aware graph embedding for session-based news recommendation. In *Proceedings of the 14th ACM conference on recommender systems*, pages 657–662.

Karthik Shivaram, Ping Liu, Matthew Shapiro, Mustafa Bilgic, and Aron Culotta. 2022. Reducing cross-topic political homogenization in content-based news recommendation. In *Proceedings of the 16th ACM conference on Recommender Systems*, pages 220–228.

Tursynkhan Tursunov, Dinara Kaibassova, and Nurzhamal Kashkimbayeva. 2025. Comparative analysis of recommendation algorithms: Collaborative, content-based and hybrid approaches. In *2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST)*, pages 1–5. IEEE.

Judith Vermeulen. 2018. newsdna : promoting news diversity : an interdisciplinary investigation into algorithmic design, personalization and the public interest (2018-2022).

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6389–6394.

Yao Yao, Zuchao Li, and Hai Zhao. 2023. Learning event-aware measures for event coreference resolution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13542–13556.

Hang Zhang, Wenjun Ke, Jianwei Zhang, Zhizhao Luo, Hewen Ma, Zhen Luan, and Peng Wang. 2023. Prompt-based event relation identification with constrained prefix attention mechanism. *Knowledge-Based Systems*, page 111072.

Appendix

A. Multiple-Choice Tests

Here follow the MC tests per topic translated into English used in the Knowledge Retention Experiment. For each question, 'no idea' was a possible answer which we here omit. Correct answers are in italics.

Topic 1: Soyuz Spacecrafts

1. The emergency landing of the Soyuz rocket occurred because a micrometeorite struck the capsule.
   A) True
   B) *False*
2. What is the reason that crewed launches are temporarily suspended?
   A) The rocket was unstable
   B) *Safety procedures were not followed*
   C) There was insufficient fuel
   D) The capsule was incorrectly attached
3. Which two agencies will independently investigate the incident?
   A) NASA and ESA
   B) *Roscosmos and NASA*
   C) ESA and SpaceX
   D) SpaceX and Blue Origin
4. How many G-forces did the astronauts experience during the emergency landing?
   A) 3G
   B) 9G
   C) *6G*
   D) 1G
5. How does a ballistic landing of space capsules proceed?
   A) The capsule actively uses its engines to slow down
   B) The capsule slowly glides horizontally toward Earth
   C) *The capsule follows a passive trajectory*
   D) The capsule always lands at sea to reduce impact
6. Which factors precisely determine a ballistic landing?
   A) The astronauts' muscle strength and heat shield technology
   B) *Air density, wind, gravity, and humidity*
   C) The speed of the ISS and solar activity
   D) Only gravity
7. How long does it take to reach the ISS?
   A) Approximately 24 hours
   B) *Approximately six hours*
   C) Two days
   D) 180 days
8. Why was it important that the Soyuz rocket remained operational despite the incidents?
   A) Because Russia no longer had another transport vehicle for its military satellites
   B) *Because there were no other crewed vehicles available for transport to the ISS*
   C) Because commercial spaceflight companies were still in their testing phase
   D) Because the Soyuz deal with NASA was financially crucial for Roscosmos
9. Where was the mysterious hole in the Soyuz found?
   A) In the navigation system
   B) In a solar panel
   C) *Behind a toilet in the living module*
   D) In the heat shields
10. What would have happened if the leak in the ISS had not been sealed?
    A) The ISS would have fallen out of its orbit around Earth
    B) *The air would have run out, causing the crew to suffocate*
    C) The Soyuz capsule would have exploded

Topic 2: Disappearance of MH370

1. Which factor did Professor Kristensen use in his mathematical model to determine the possible crash area of MH370?
   A) The number of passengers and their mobile phone signals
   B) The satellite signals of the onboard radio
   C) *The Doppler shift of the satellite signal*
   D) The air pressure measurements at cruising altitude
2. Why would MH370 have deliberately flown through the Intertropical Convergence Zone?
   A) To save fuel
   B) Because it was the shortest route to Beijing
   C) *Because that area makes detection by radar and satellites more difficult*
   D) To bring the aircraft closer to the U.S. naval base
3. What strengthens the theory that MH370 made a controlled landing at sea?
   A) The satellite pings showed a sudden end of the signal

B) The wreckage was largely untraceable

C) *The recovered debris was relatively intact*

D) The pilot had previously simulated how to land on water

4. What is a striking similarity between MH370 and the search vessel Seabed Constructor?

   A) *Both switched off their tracking systems in sensitive zones*

   B) Both were simultaneously located by satellites

   C) Both had defective black boxes on board

   D) Both sailed through the Indian Ocean without permission

5. What is an important consequence of switching off MH370's tracking system?

   A) The aircraft automatically returned to Kuala Lumpur

   B) *It became invisible to both military and civilian radars*

   C) The black box transmitted incorrect signals

6. What makes the ship Seabed Constructor particularly suitable for searching for aircraft wreckage?

   A) It has a nuclear reactor on board

   B) *It has underwater drones capable of searching the deep sea*

   C) It was designed by former NASA engineers

   D) It contains a mini-submarine for crewed missions

7. Which of the following elements was NOT explicitly mentioned as a possible explanation in the official reports on MH370?

   A) Technical malfunction of the aircraft

   B) Controlled action by someone on board

   C) Sabotage by an external party

   D) *Natural disaster caused by lightning strike*

8. What did the Malaysian final report conclude about the pilot's role in the disappearance of MH370?

   A) The pilot was completely exonerated

   B) The pilot was definitively identified as responsible

   C) *The pilot's involvement could not be ruled out*

   D) There was insufficient information about the pilot to say anything

9. Why was the search for MH370 officially terminated by Malaysia in 2018?

   A) The United States asked them to stop

   B) They suspected sabotage of the search team

   C) *There was insufficient concrete evidence to continue searching*

   D) New leads had emerged that disrupted the investigation

10. Why could the wreckage of MH370 not be located despite years of searching?

    A) The black box was never activated

    B) The aircraft likely exploded in mid-air

    C) *The satellite pings were too vague to determine the exact location*

    D) The ocean floor was completely flat, without obstacles

Topic 3: Hedwigepolder

1. According to research, what is the long-term expectation for the Hedwigepolder after depoldering?

   A) It will become a saltwater marsh with reduced biodiversity

   B) It will turn into a muddy sludge basin

   C) *It will become a diverse salt marsh and mudflat landscape with many water birds and benthic organisms*

   D) It will become less polluting agricultural land

2. What was the official reason for the depoldering according to the Flemish representative Axel Buyse?

   A) Compensation for dredging the Western Scheldt

   B) The plan was part of the Flemish Sigma Plan

   C) *Necessary nature compensation under European guidelines*

   D) Safety for surrounding villages in case of flooding

3. What was the long-standing disagreement between the Netherlands and Belgium about?

   A) Whether the Hedwigepolder should become part of the Oosterschelde National Park

B) *Whether fertile agricultural land could be sacrificed for nature development*

4. Which animals are explicitly mentioned as future inhabitants of the nature area in the Hedwigepolder?

   A) Otters, deer and beavers
   B) Storks, hares and starlings
   C) *Common terns, redshanks and curlews*
   D) Seagulls, sheep and foxes

5. What was the ruling of the Supreme Court in the lawsuit of owner Géry De Cloedt?

   A) The expropriation was annulled
   B) The case was referred back to a lower court
   C) *The expropriation was definitively approved*
   D) The case is still pending before the Council of State

6. With which symbolic action did the committee "Save Our Polders" conclude their protest in the Hedwigepolder?

   A) Planting new trees
   B) *Burning protest banners*
   C) A march to the town hall of Hulst
   D) Submitting a petition to the Flemish government

7. Which treaties formed the basis for the depoldering of the Hedwigepolder?

   A) *The Scheldt Treaty of 2005 and the European Birds and Habitats Directives*
   B) The Paris Climate Agreement and the Antwerp Port Treaty
   C) No treaties; it was purely an ecological decision

8. What was the amount of the buyout sum that the Dutch state offered to De Cloedt?

   A) 800 million euros
   B) *Approximately 15 million euros*
   C) 50 million euros
   D) No amount was mentioned

9. What is the ultimate goal of the Hedwigepolder within the larger project?

   A) To expand the port of Antwerp
   B) *To contribute to the largest brackish water marsh area in Europe*
   C) To convert it into a recreational area
   D) To form a buffer zone between Belgium and the Netherlands

10. How did project leaders attempt to compensate for the loss of fauna, such as little owls and bats?

    A) By transferring animals to a rescue center
    B) *By providing alternative habitats and nesting boxes outside the area*
    C) By temporarily suspending the project
    D) By housing all animals in nature reserves

Topic 4: 2018 Belgian energy crisis

1. What is a disadvantage of Electrabel's turbojets for generating electricity?

   A) They produce too little electricity to make a difference
   B) They operate only on solar energy
   C) *They are expensive, inefficient and polluting*
   D) They can only be used once per year

2. What was remarkable about the plans of the Luxembourg company EGL (BTK)?

   A) They would build nuclear power plants instead of gas plants
   B) They would refuse subsidies from the Belgian government
   C) *They would build climate-neutral gas power plants*
   D) They would exclusively use coal

3. What statement did Minister Marghem make about selling the Belgian nuclear power plants?

   A) It was a visionary decision
   B) It led to too many investments
   C) *It may not have been the best decision*
   D) It made Belgium energy independent

4. What legal concern does Marghem express about the behavior of Engie Electrabel?

   A) They may have deliberately spread false information
   B) *They may be abusing their dominant market position*
   C) They may have obtained subsidies unlawfully
   D) They may have concluded secret foreign contracts

5. Who is responsible for Belgium's electricity supply according to energy expert André Jurres?

   A) Only Electrabel is responsible
   B) It is the government's task
   C) *Elia must guarantee the electricity supply*
   D) All responsibility lies with foreign investors

6. According to EGL, what is the advantage of their gas power plants?

   A) They are mobile and can be placed anywhere
   B) *They emit no CO_2 or NOx*
   C) They run entirely on renewable energy
   D) They are partly financed by Electrabel

7. What criticism did the Belgian government receive regarding the sale of the nuclear power plants?

   A) The plants were sold without public notification
   B) *Foreign owners invested insufficiently in maintenance*
   C) The sale led to several small-scale nuclear incidents
   D) It was a temporary lease, not a sale

8. What triggered the increased vigilance around energy supply in Belgium during the autumn months?

   A) A major European blackout had been predicted
   B) *Six of the seven nuclear power plants in Belgium were shut down*
   C) France refused to supply electricity to Belgium
   D) Wind turbines were operating below capacity

9. What consequences do temporary solutions such as mobile generators and emergency power plants entail?

   A) CO_2 emissions will decrease in the short term but increase in the long term
   B) Dependence on foreign energy imports will decrease
   C) *There is a risk that structural, sustainable investments will be postponed*
   D) The energy supply will be definitively secured for the next 20 years

10. What conclusion can be drawn about the role of the government in the Belgian energy situation?

    A) The government plays only a limited role because the market regulates everything
    B) *A lack of timely and clear decisions contributed to the capacity problem*
    C) The government proactively invested in green alternatives such as methanol
    D) Belgium largely follows France's example, reducing its risk

Topic 5: Facebook Cambridge Analytica

1. What role did Cambridge Analytica play in the Facebook user data scandal?

   A) It was the first company to collect data through advertisements
   B) *It created psychological profiles of users to influence their voting behavior*
   C) It attempted to hack Facebook to gain access to data
   D) It conducted legitimate market research through authorized Facebook channels

2. Which misconception about technology is emphasized in the article about Plato and Facebook?

   A) Technology makes people too independent from others
   B) *Technology automatically leads to wisdom and moral progress*
   C) Technology endangers political power
   D) Technology has a stronger influence on younger generations

3. Why is Facebook criticized for its response to the data scandal?

   A) Because Facebook acted immediately and too strictly against Cambridge Analytica
   B) Because Facebook made the information about the abuse public before it could be investigated
   C) *Because Facebook reacted slowly and lacked transparency and control over app developers*
   D) Because Facebook refused to cooperate with European privacy regulators

4. How did Cambridge Analytica collect data from millions of Facebook users?

   A) By directly hacking Facebook accounts

B) *Through a personality quiz that also involved participants' friends*

C) By paying Facebook for full access to its database

D) Through cooperation with governments that passed on data

5. How did Facebook attempt to restore its image after the revelations about data misuse?

A) By immediately taking legal action against all parties involved

B) By massively purchasing advertisements to clear its name

C) *By publicly apologizing and promising to better protect users*

D) By temporarily taking the Facebook website offline

6. What broader political impact may the Facebook data scandal have had?

A) *The data misuse may also have played a role in the Brexit referendum*

B) Facebook helped governments run campaigns against disinformation

C) There were no political consequences; it was purely a marketing issue

7. How does Plato's story about criticism of the technological invention of writing relate to modern technology such as Facebook?

A) Technology brings us closer to the truth

B) Written information leads to true knowledge

C) *Technology can create a false sense of knowledge and connectedness*

D) Technology is the only way to preserve information sustainably

8. What stance did Facebook take toward Cambridge Analytica after being informed of the data theft in 2015?

A) Facebook immediately warned users and reported it to authorities

B) Facebook launched a public investigation into Cambridge Analytica's activities

C) *Facebook demanded that the data be deleted but did not follow up and did not inform anyone*

D) Facebook denied any contact with Cambridge Analytica

9. Why is Facebook's slogan "bringing the world closer together" questioned because of the data scandal?

A) Because Facebook was banned worldwide in multiple countries

B) *Because the platform causes division and isolation rather than connectedness*

C) Because users massively switched to other networks

D) Because the slogan was never officially used by Facebook

10. What did Christopher Wylie do after leaving his position at Cambridge Analytica?

A) He helped cover up the data misuse

B) He founded a competing data company

C) *He exposed the data theft as a whistleblower through the media*

D) He became an advisor in Hillary Clinton's campaign team

B. News Consumption Habits

1. How many times do you follow the news?
– Daily/weekly/barely/never

2. Order the modality you usually use to follow the news?
– Reading, watching, listening

3. Which news channels do you usually use?
– Television, radio, news paper, news websites, news apps, social media

4. Which news carrier do you usually use?
– News paper, smartphone, computer, tablet

5. Do you have a paying news subscription?
– Yes/no

6. Order the following news topics according to your interests.
– Science, disaster, environment, society, technology

7. Are you older than 29 years?
– Yes/no

C. Knowledge Retention Experiment Results

	condition[a]		topic[a]					news_engagement[b]			subscribed[b]	
predictor	random	topic	cl_35	cl_99	cl_49	cl_14	c_4	barely	daily	weekly	no	yes
n=	64	62	21	27	29	25	24	93	25	8	95	31
score	6,61	8,03	6,71	8,22	6,76	7,60	7,17	7,88	7,20	7,52	7,26	7,45
stat_sig	***	***	***	**	ns	ns	ns	***	ns	ns	***	ns
d	–	0,83	–	0,92	0,02	0,60	0,24	–	-0,36	-0,21	–	0,10
effect	–	large	–	large	neglig	medium	small	–	small	small	–	neglig

	news_device[b]				read_news[b]			watch_news[b]			listen_news[b]		
predictor	phone	pc	paper	tablet	yes	±	no	yes	±	no	yes	±	no
n=	101	16	5	4	91	27	8	68	29	29	89	31	6
score	7,38	7,00	7,20	7,00	7,34	7,15	7,50	7,10	7,19	7,79	7,83	7,71	7,13
stat_sig	***	ns	ns	ns	***	ns	ns	***	ns	ns	***	ns	ns
d	–	-0,21	-0,09	-0,20	–	-0,10	0,09	–	0,05	0,40	–	-0,08	-0,35
effect	–	small	neglig	neglig	–	neglig	neglig	–	neglig	small	–	neglig	small

	gender[c]		nationality[c]		>29y[c]		degree[c]			duration_minutes[c]				
predictor	♂	♀	BE	NL	no	yes	BA	MA	PhD	15-20	5-10	10-15	20-25	25+
n=	61	64	61	65	65	61	58	57	11	45	8	49	9	15
score	7,56	7,05	7,59	7,05	7,51	7,12	7,02	7,56	7,55	7,53	7,75	7,08	7,56	7,00
stat_sig	***	ns	***	ns	***	ns	***	ns	ns	***	ns	ns	ns	ns
d	–	-0,28	–	-0,30	–	-0,21	–	0,29	0,26	–	0,12	-0,25	0,01	-0,28
effect	–	small	–	small	–	small	–	small	small	–	neglig	small	neglig	small

	interest_topic[c]			accepted_studies[c]				
predictor	neutral	interest	no_interest	0-100	101-200	201-300	301-500	500+
n=	30	44	52	28	25	17	26	30
score	7,67	7,46	6,98	6,87	7,56	7,47	7,81	6,67
stat_sig	***	ns	ns	***	ns	ns	ns	ns
d	–	-0,12	-0,38	–	0,37	0,29	0,53	-0,11
effect	–	neglig	small	–	small	small	medium	neglig

Table 6: Experimental[a], news consumption related[b] and personal[c] variables as predictors.