ATLAS: Article Tracking, Linking, and Analysis of Swedish Encyclopedias

Albin Andersson, Salam Jonasson, Fredrik Wastring, Pierre Nugues
Lund University
Lund, Sweden
al3146an-s@student.lu.se, ab5015mu-s@student.lu.se,
fr7658wa-s@student.lu.se, pierre.nugues@cs.lth.se

Abstract

The digitization of old encyclopedias represents an important step to improve access to historically structured knowledge. Often, however, this process does not go beyond an optical character recognition, leaving all the underlying structure unexploited. In addition, many encyclopedias had multiple editions reflecting the evolution of knowledge. The lack of structure in the raw text makes it difficult to track changes across these editions. In this work, we built a pipeline to restore the text structure, where we extract the headwords and identify entries; categorize the entities; match entries across editions; and link entries to a Wikidata item. We applied this pipeline to the four major editions of *Nordisk familjebok*, an authoritative Swedish encyclopedia published between 1876 and 1951. We could extract the headwords with an F1 score of 97.8% and we obtained an F1 score of 93.4% on the headword classification. On a small-scale evaluation, we reached a 93% precision on the cross-edition matching, 85% precision and 16.5% recall on the Wikidata linking. This shows that an automated approach to digitized historical knowledge is possible. This should facilitate the preservation of general knowledge and the understanding of knowledge transmission. The datasets and programs are available online.

Keywords: text categorization, named entity recognition, entity resolution, digital humanities

1. Introduction

Old encyclopedias are valuable pieces of historical knowledge, reflecting the life and ideas of their time. However, much of this knowledge remains locked in unstructured text, making it difficult to analyze it systematically and draw usable conclusions.

Nordisk familjebok is the most comprehensive Swedish encyclopedia of its time. It holds an important place in Swedish literature and used to occupy a prominent place in many Swedish home libraries (Frängsmyr, 1991). *Nordisk familjebok* was published between 1876 and 1951 and had four major editions. The first one (E1) comprises 20 volumes (1876-1899), the second one (E2), 38 volumes (1904-1926), the third one (E3), 23 volumes (1923-1937), and the fourth one (E4), 22 volumes (1951). The second edition still has a high cultural status and is often referred to as *Uggleupplagan*, the 'Owl edition' (Nordisk Familjebok, 1876–1951).

These four editions span different periods, reflecting evolving perspectives and knowledge. As a concentrate of their time, they are key sources for the study of intellectual history. Nonetheless, although these editions have been digitized, they still suffer from varying levels of optical character recognition (OCR) quality and inconsistent segmentation markup. This makes them difficult to navigate their content, study the evolution of entries, and relate them to current knowledge.

In this paper, we present ATLAS (Article Tracking, Linking, and Analysis of Swedish encyclopedias), a pipeline for processing historical encyclopedic content. ATLAS aims to manage tasks such as text segmentation into entries and extraction of their headwords; classification of entities into three categories, *Location*, *Person*, and *Other*; matching of a same entry with different versions across editions; and finally linking entries to Wikidata items. Figure 1 shows an overview of the ATLAS pipeline.

The contributions of this work are the following:

1. We scraped the four editions of the encyclopedia. We preprocessed this dataset and cleaned it to get rid of irrelevant content;
2. We created a dataset of segmented entries annotated with their headword. We used it to train models to extract the headwords;
3. We annotated a second dataset of 6000 entries with entity classes and we trained a classifier;
4. We matched entries across editions. We compared each entry of a given edition to all other editions using a sentence embedder;
5. We linked Wikidata items that had a reference to the encyclopedia entries using the same approach as in the previous step.

The datasets and code required to reproduce the experiments are publicly available on Hugging Face at https://huggingface.co/albinandersson/datasets and GitHub at https://github.com/SalamSki/EDAN70.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3566–3576
11–16 May 2026.

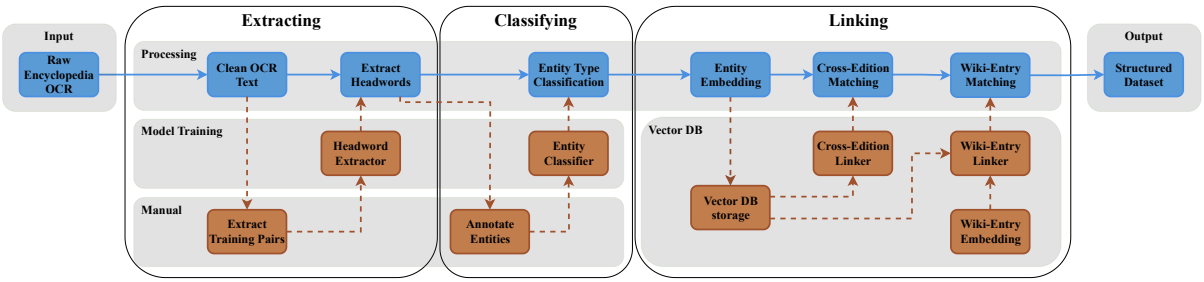

Figure 1: An overview of the ATLAS pipeline.

2. Previous Work

Our system builds on work in three main areas: digitization of historical texts, recognition of named entities (NER) in historical documents, and linking of historical encyclopedias to Wikidata.

2.1. Digitization of Historical Texts

There are now scores of book digitization projects. Project Runeberg has digitized numerous Nordic texts with an OCR and volunteer proofreading (Project Runeberg, 1992–2024). Project Runeberg digitized the four editions of *Nordisk familjebok* and made them accessible online.[1]

While the digitization of *Nordisk familjebok* is highly valuable, the OCR quality varies across volumes, and the segmentation into distinct entries remains incomplete. The proofreading is still in progress with varying degrees of completion depending on the edition. This underscores the need for post-OCR correction methods, such as character sequence-to-sequence models proposed by Ramirez-Orta et al. (2022).

2.2. Entity Recognition

Named entity recognition (NER) has a long history. We can reformulate it as a sequence annotation problem, where we annotate each word to mark it as a part of a named entity or not. As significant milestones, Collobert et al. (2011) applied a unified feed-forward neural network architecture associated with embeddings. Lample et al. (2016) used long short-term memory (LSTM) networks. Finally, Devlin et al. (2019) applied transformer encoders that improved the feed-forward and LSTM scores.

Current NER systems build on embeddings or models often pretrained on contemporary text. Historical documents present unique challenges due to evolving language, shifting spelling conventions, and inconsistent formatting (Ehrmann et al., 2023). Tudor and Pettersson (2024) showed promising results by fine-tuning pretrained models on historical Swedish text to better handle style variations.

[1] https://runeberg.org/nf/

2.3. Linking Encyclopedias to Wikidata

In recent years, entity linking (EL) has advanced significantly, driven by improvements in language models and dense vector embeddings. Wu et al. (2020) introduced a zero-shot approach using dense entity retrieval, while Gillick et al. (2019) demonstrated the effectiveness of dual encoder models for mapping mentions and entities into a shared vector space. Similarly, Logeswaran et al. (2019) developed a zero-shot approach that links textual mentions to a knowledge base without entity-specific training data. These approaches rely on various similarity metrics to compare embeddings. Botha et al. (2020) showed that storing embeddings in vector spaces and performing similarity searches is an efficient strategy for large-scale entity linking. Ayoola et al. (2022a,b) proposed methods to incorporate the type in the disambiguation.

Previous projects have also explored linking historical encyclopedias to the Wikidata entity repository. As examples, Nugues (2022, 2024) linked proper nouns in *Petit Larousse illustré*, a French dictionary, and Diderot's *Encyclopédie*. For *Nordisk familjebok* specifically, Ahlin et al. (2024) explored linking location entities for the second edition and Börjesson et al. (2025) for the first and second editions.

Our work extends these approaches in three key ways: we processed both person and location entities; we covered all four editions of the encyclopedia that were available online before Summer 2025; and we introduced methods for cross-edition entity tracking.

3. Datasets

Nordisk familjebok is organized as a sequence of entries, where the headwords are ordered alphabetically. To recover this structure, we identified the headwords and we segmented the raw text into entries. We then categorized these entries.

To recognize the entries, a possible solution could be to analyze the image layout of the scans as in Wang et al. (2021) or Wang et al. (2022). We decided to only use the text and apply a hybrid method

instead. We first wrote rules that we applied to the corpus to create a "silver standard" dataset of annotated headwords. We then extracted a subset of it that we manually corrected to serve as a test set. We trained a model on the headword training set to recognize the headwords and segment the entries. We evaluated it on the headword test set. We created a second annotated dataset of categorized entries to train the entry classification models.

3.1. Raw Text Collection

Each edition of *Nordisk familjebok* hosted on Project Runeberg corresponds to a set of URLs organized by volumes following the original division of the encyclopedia. Furthermore, each page has a specific URL component that follows the volume URL. We used this structure to scrape the *Nordisk familjebok* editions.

Using the links, we downloaded the HTML content of each page and extracted the OCRed text. The HTML markup structure is regular and makes it trivial to find the beginning and end of the text. The OCRed content contains additional HTML tags. We observed that bold tags, `<b>`, were used to encapsulate headwords at the beginning of an entry. We kept them and removed all the other tags.

3.2. Layout Cues

A key challenge of the digital version is the inconsistency in the marking of headwords. Although the first two editions, E1 and E2, have been proofread with a generally coherent use of HTML tags to separate headwords, the third and fourth editions, E3 and E4, have no headword marking. This inconsistency makes it impossible to apply a direct headword extraction.

A quick manual check showed that all occurrences of the bold tag in the first two editions are headword encapsulations. However, this markup is not systematic. In addition, very few headwords in the last two editions are marked this way.

The raw text at this stage is then under or oversegmented depending on whether we use the `<b>` markup or the new lines to delimit the entries:

- The bold tags always correspond to a headword, but many headwords are not marked in the first and second editions, and nearly none in the third and fourth ones.

- Entries always start with a new line, but many long entries consist of two or more paragraphs and may include other types of content such as image and figure captions, footnotes, and page numbers. Therefore, we could not use the new lines as an entry boundary marker.

3.3. Headword Dataset

We created the headword dataset from the headword-annotated entries in E1 and E2. We considered all the paragraphs in the corpus. We identified those starting with the `<b>` tag and we extracted their content using regular expressions.

The structure of the headword dataset consists of two items: The input and the label. As input, we used the raw text of a paragraph from its start and up to 500 characters. As label, we used the extracted headword if we could find one or nothing otherwise. For example, for the *Lund* entry, we have:

Input: Lund, uppstad i Malmöhus län. . . beskaffenhet. I all[mänhet]
"Lund, a city in Malmöhus County. . . nature. In gen[eral]"

Label: Lund

We restricted the paragraphs with no headword to start with a capital letter as with:

Input: Sammanfattningen af dessa nya. . . kan man anse Brandes, hvilken
"The summary of these new. . . one can consider Brandes, who"

Label: None

and we discarded the rest.

We obtained 308,448 paragraphs in total, where about 80% contained a headword. Table 1 shows the detailed counts. Some of the entries were duplicates. We removed them and this resulted into 305,675 entries. We split the dataset into a training set of 300,675 entries and a test set of 5,000.

The simple `<b>` tag rule missed some entries and headwords. We manually curated the test set. Out of its 5000 samples, the `<b>` tag rule had labeled 951 as negatives. Our manual inspection found 320 FNs. As we did not have the means to carry out a manual correction of the training set, we used it unchanged with about 240,000 positive samples and 60,000 negative ones.

We used this dataset to train a sequence annotator and extract the headwords, possibly none, see Sect. 4.1. We also used it to segment the text into entries, where we defined an entry by the presence of a headword in the paragraph.

Editions	Positives	Negatives	**Total**
First	114,770	11,920	126,690
Second	132,264	49,494	181,758
Total	247,034	61,414	308,448

Table 1: Headword dataset, where the positive paragraphs start with a headword.

The resulting dataset, *The Nordisk Familjebok Headword Extraction Dataset*, is publicly available.[2] See Andersson et al. (2026c).

3.4. Headword Category Dataset

After the recognition of entries, we classified them into three categories: person, location, or other. We created a second dataset of 6000 entries annotated with these categories to train a second classifier. Given an entry, we assumed that the category of the entity corresponded to the category of its headword. We carried out the annotation in two steps. We started with a first automatic annotation of the headwords. We then manually checked these annotations.

For the first annotation, we applied an NER model to the entry text in a zero-shot setting. As model, we used KB/bert-base-swedish-cased-ner pretrained on Swedish texts and fined-tuned for sequence annotation tasks on the manually annotated SUC 3.0 corpus (Malmsten et al., 2020). The model uses the tags `PRS` for persons, `LOC` for locations, `ORG` for organizations, `TME` for time, `EVN` for events, or nothing if the token is outside these classes. We applied this model to randomly selected entries. As input, we truncated the entry text to the first 500 characters. We used the headword annotation from the model output to determine the entity category.

When a headword consists of multiple tokens, this method sometimes results in more than one category tag. We applied post-processing rules to merge them into a single label:

1. If they contain at least one *Location* but no *Person*, then the headword is a *Location*.
2. If they contain at least one *Person* but no *Location*, the headword is a *Person*.
3. If they contain at least one *Person* and one *Location*, this indicates an uncertainty. We classify the headword as *Other*.
4. We default to *Other* if neither *Location* nor *Person* is present.

In the second step, following the zero-shot NER predictions, we extracted a balanced subset of 6000 entries that we verified manually. After manually correcting the annotation, the distribution was slightly altered. We used this labeled dataset to fine-tune a transformer encoder pre-trained on Swedish text for the entry classification task.

As with the headword dataset, we have released this classification dataset online, named *The Nordisk Familjebok Category Classification Dataset*.[3] See Andersson et al. (2026a).

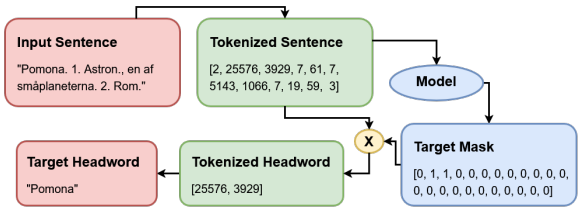

Figure 2: The input tokenization and output mask.

4. Method

The ATLAS system consists of a pipeline of components. Figure 1 shows its architecture that we describe now.

4.1. Headword Extractor and Entry Segmenter

We modeled the headword extraction task as a sequence annotation task, where each token in an input sentence is classified as either part of the headword (1) or not (0). This enables the model to identify headwords in entries while producing empty masks for the rest of the paragraphs. To segment the entries, we used these results with the rule that an entry always starts with a headword.

We compared two different architectures for this task: A bidirectional LSTM-based model trained from scratch and a fine-tuned transformer encoder trained on Swedish text, KB-BERT (Malmsten et al., 2020), with selective layer unfreezing.

Data preprocessing. For both models, we tokenized the input sentences of the training set and their corresponding headwords using KB-BERT's tokenizer with a maximum sequence length of 100 tokens. We generated the binary target mask by comparing the tokenized sentence with the headword sequences. We annotated the first occurrence of the headword tokens with ones and the rest with zeros, see Figure 2.

Additionally, we applied a post-processing to recover the words from the subwords marked with `##` prefixes. In case the annotation of the subwords of a same token was contradictory, positive and negative, we propagated positive classifications backward to all its subwords.

[2] https://huggingface.co/datasets/albinandersson/nf-headword-extraction

[3] https://huggingface.co/datasets/albinandersson/nf-category-classification

Model architectures. The LSTM-based architecture consists of an embedding layer (128-dimensional) followed by a bidirectional LSTM (128-dimensional hidden states). The model processes the embedded sequence and outputs token-level binary classifications.

In the transformer-based approach, we fine-tuned KB-BERT with a task-specific classification head. We employed an unfreezing strategy, where we experimented with different configurations to determine the best number of trainable layers. We applied these models to the raw text of the four editions. Both methods enabled us to determine the headwords and segment the text into entries.

4.2. Headword and Entry Classification

We then classified the resulting entries into three categories: *Location*, *Person*, or *Other*. We used the pretrained KBLab/bert-base-swedish-cased transformer encoder that we fine-tuned on the headword category dataset described in Sect. 3.4. We added a classification layer with three outputs corresponding to our target types.

To determine the optimal architecture, we used a systematic layer freezing strategy, where all transformer layers were frozen except for the final N layers. We experimented with $N = 0, 2, 4 \ldots, 12$, and selected the best performing model based on the test set performance.

We split the dataset into 70/15/15 train/validation/test sets. We used the AdamW optimizer with a learning rate of 2×10^{-5}, and batch size of 16. We implemented early stopping with a patience of 3 on validation accuracy to prevent overfitting.

4.3. Cross-edition Matching

Tracking entities across different editions is not straightforward. While entries referring to the same entity share similar content across editions, their definitions evolve to reflect temporal changes (e.g., adding death dates for persons or updated population statistics for locations). Additionally, headwords alone are insufficient for matching since they often only contain surnames for persons and some terms can refer to either locations or persons (e.g., *Lund* is both a city and surname).

We embedded the entries in dense vector representations with a sentence transformer model pretrained on Swedish text: S-BERT Swedish Cased (Rekathati, 2021). We assigned each embedding a unique identifier consisting of the edition number and a sequential index. For instance, the identifier `E2_622` corresponds to the statistician Gottfried Achenwall who has entry number 622 in the second edition.

We stored these embeddings in the Qdrant vector database. For each entry, we ranked the most similar embeddings in other editions using a cosine similarity. We filtered the results by edition prefix to ensure one candidate match per edition. We applied a conservative threshold of 0.75 and we removed the matches that were not symmetrical. We also ensured that, in a match, both editions had the same headword.

4.4. Wikidata Linking

After matching entries across editions, we linked them to corresponding Wikidata items. To make the task computationally feasible, we started from the observation that many articles in the Swedish Wikipedia reused entries of the *Nordisk familjebok*. We thus restricted our scope to Wikipedia articles that reference *Nordisk familjebok* as their source. We retrieved these articles and their Wikidata items using SPARQL queries that extracted entries containing a "described in" property, `P1343`, and the *Nordisk familjebok* QID as object:

```
SELECT ?item
WHERE {
  ?item wdt:P1343 wd:Q678259 .
}
```

We found about 11,550 such items. A quick inspection showed us that many of the corresponding Wikipedia articles reuse almost identically the text in *Nordisk familjebok*.

Similarly to the cross-edition matching, we stored the first 500 characters of these Swedish Wikipedia articles and the *Nordisk familjebok* entries in a vector database. For each *Nordisk familjebok* entry, we computed cosine similarity scores against the Wikipedia candidate embeddings. When the similarity exceeded the threshold of 0.75, and the Wikidata label contained the entry headword, we considered it a match and stored the corresponding Wikidata QID.

We stored the results in a table, where, for each entry, we indicated the headword, its type, the corresponding matching entries in the other editions, if we could find any, as well as the QID. See Table 2 for an example.

5. Results

We broke down the results of each step in our pipeline, namely scraping, headword extraction, NER classification, cross-edition matching, and Wikidata linking.

5.1. Headword Extraction

We trained and evaluated different model architectures on the headword dataset of Sect. 3.3. We

Entry ID	headword	Type	Edition	E1_match	E2_match	E3_match	E4_match	QID
E1_385	Achenwall	2	E1	–	E2_622	E3_416	E4_473	Q215933
E1_386	Acheron	1	E1	–	E2_623	E3_417	E4_476	–
E1_387	Acherontia	0	E1	–	–	–	–	–
E1_388	Acherusia	1	E1	–	E2_625	–	–	–

Table 2: Dataset structure with four entries from the first edition: E1_385 to E1_388. The columns contain the entry headword, type (0: Other, 1: Location, and 2: Person), the matches we found in the three other editions, if any, and the QID identifier in Wikidata.

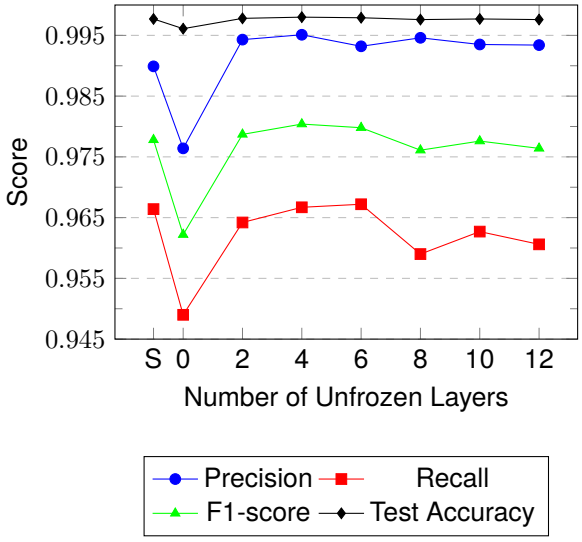

Figure 3: Performance metrics across different model configurations for headword extraction. On the x-axis, "S" represents the LSTM model while the figures indicate the number of unfrozen layers in the fine-tuned KB-BERT model.

Confusion matrix		Performance metrics	
486,211	236	Accuracy	0.9977
		Precision	0.9899
905	12,648	Recall	0.9664
		F1-score	0.9778

Table 3: LSTM score on the manually validated test set with 5000 entries (and 500,000 tokens).

compared the LSTM model with various configurations of the fine-tuned KB-BERT model. Figure 3 shows the results with different numbers of unfrozen layers. While the fine-tuned model with 4-6 unfrozen layers achieved marginally higher scores, the more lightweight LSTM model had a comparable performance. This minimal performance trade-off combined with the LSTM's lower computational requirements led us to select it for this task. Table 3 shows the results, where we reach an F1 score of 0.9778. Please note that the evaluation is token-based and therefore has a support number of 500,000 values, 100 tokens for each test sample.

Editions	Scraped	Extracted		Discarded	
E1	133,857	117,473	0.88	16,384	0.12
E2	247,563	185,063	0.75	62,500	0.25
E3	43,003	26,464	0.62	16,539	0.38
E4	131,530	89,221	0.68	42,309	0.32
Total	555,953	418,221	**0.75**	137,732	**0.25**

Table 4: Scraped paragraphs, extracted entries, and discarded paragraphs for each edition.

5.2. Entry segmentation

Scraping the four editions into text files and splitting them by new lines results in more than 550,000 potential entries, as shown in Table 4. The column Scraped shows that the first and fourth editions are fairly similar in size, with the largest edition being the second and most popular. The incomplete digitization of the third edition before Summer 2025 is also reflected by the smaller number of entries.

Applying the LSTM headword extractor on the scraped files results in 418,221 headword classified entries. This means that approximately 75% of all paragraphs correspond to the first paragraph of an entry. Table 4 also shows the disparity in extraction percentage between the first and second half of the encyclopedia, resulting in more paragraphs per entry for the third and fourth editions. The reason might be that the headword extractor is trained on the first and second editions.

5.3. Entity Classification

To fine-tune the KB-BERT models for entity classification, we split the headword category dataset of Sect. 3.4 into a 70:15:15% split, resulting in a 4200:900:900 sample count for the training, validation, and test sets respectively. Table 5 shows the results with varying numbers of unfrozen layers.

We notice an almost linear steady increase in all metrics as more layers are unfrozen, reaching the peak with an F1 score of 0.9339 at the maximum of 12 unfrozen layers. Table 6 shows its confusion matrix. We applied this model to classify the entities of all extracted entries.

Table 7 shows the category breakdown we obtained. We notice a somewhat general pattern where for each edition, around 50-60% of the articles are classified as *Other*, 18-23% as *Locations*,

UFL	Accuracy	Precision	Recall	F1
0	0.8078	0.8111	0.8074	0.8072
2	0.9256	0.9248	0.9293	0.9261
4	0.9167	0.9152	0.9201	0.9171
6	0.9267	0.9268	0.9282	0.9271
8	0.9322	0.9328	0.9333	0.9328
10	0.9267	0.9282	0.9278	0.9277
12	**0.9333**	**0.9344**	**0.9359**	**0.9339**

Table 5: Classification performance of the KB-Bert fine-tuned models. UFL means unfrozen layers.

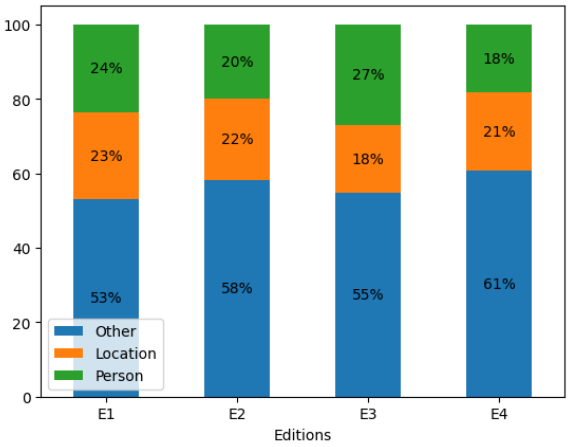

Figure 4: The resulting entity recognition on the extracted entries for each edition.

and 20-30% as *Persons*. Figure 4 shows the relative proportions for each category per edition.

5.4. Cross-edition Matching and Wikidata Linking

We applied the entry matching method to the four editions. Figure 5 shows the number of additions and removals for person and location entries, respectively, in the E1, E2, and E4 editions. We did not include E3 as its digitization was incomplete before Summer 2025 and hence does not have the same coverage.

The linking process successfully connected 10,964 entries to Wikidata items. The distribution of links varied across the editions: We could create 3766 links to E1 entries, 5088 links to E2, 704 links to E3, and 1406 links to E4. This distribution roughly correlates with the relative sizes of each edition, with E2 being the largest edition overall.

		Predicted		
	12 layers	Other	Location	Person
True	Other	**0.8727**	0.0848	0.0424
	Location	0.0149	**0.9731**	0.0119
	Person	0.0213	0.0170	**0.9617**

Table 6: Confusion matrix of the 12-layer model.

Edition	Locations	Persons
E1	27,554	27,760
E2	40,423	36,802
E3	4850	7127
E4	18,794	16,216

Table 7: Category breakdown by edition.

	Quads	Distinct	Match	True QID
All	1498	514	486	80
QID	267	101	94	80

Table 8: Number of quadruples, quadruples with a QID (quintuples), distinct quadruples, correctly matched quadruples, and correct QIDs assigned to a quadruple match.

As there is no comparable annotated dataset, we could not extensively evaluate the matching and linking results. We created a limited test set consisting of the person entries with a match in the four editions as in the first row in Table 2. We can easily verify that two definitions in two different editions correspond to a same person using the name as well as the dates of birth and death. This simplifies considerably the annotation.

Using this selection criterion, we found 1498 entries describing a person with a matching entry in the three other editions forming thus quadruples. Of them, 267 had a Wikidata link as the first row in Table 2. This row corresponds to the quintuple:

```
(E1_385, E2_622, E3_416, E4_473, Q215933)
```

Most quadruples have duplicates, for instance when an entry in E1 has matches in E2, E3, and E4, and the entry in E2, has the same matching entries. Table 8 shows the number of quadruples, distinct quadruples, and quadruples with a QID (quintuples).

Using these quadruples, we measured the matching precision and, for the quintuples, the linking precision and recall. We decided that a quadruple was correct if it consisted of the same person. Of the 514 distinct quadruples, 486 had four times the same person, yielding a precision of 94.6%. In this dataset, 101 quadruples also have a QID (quintuples). For this part of the dataset, 94 quadruples describe the same person. This results in a precision of 93.1% which is about the same as for the quadruples with no QID.

We defined the recall of the linking task as the percentage of links the system could find from the matches. Taking all quintuples, we had 80 correct QIDs. To compute F1, we considered the whole test set and the set of correct matches. We obtained a precision of about 85% and a recall of 16.5%, see Table 9. It has long been noted that there is a trade-off between precision and recall (van Rijsbergen, 1979). We had a conservative matching procedure

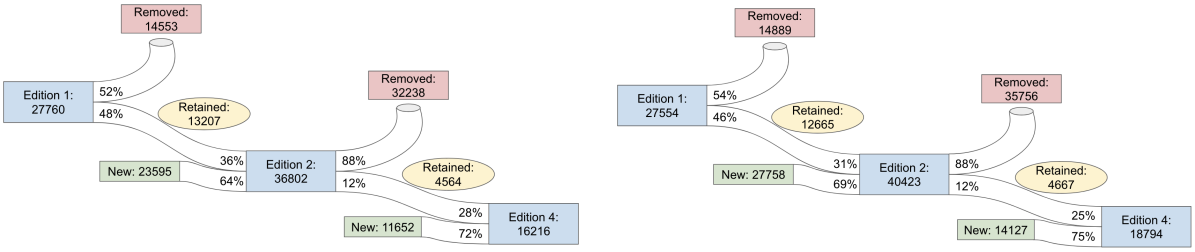

Figure 5: Additions and removals for person (left) and location (right) entries across editions.

	Precision	Recall	F1
Distinct quads	79.21	15.56	26.02
Correct matches	85.11	16.46	27.59

Table 9: Link evaluations considering all distinct quadruples and only the correct matches.

with a high threshold that favors precision.

The final structured dataset containing headwords, entity types, cross-edition matches, and Wikidata identifiers is publicly available.[4] See Andersson et al. (2026b).

6. Discussion

Table 4 shows there is a significant difference in the extraction results between E1-E2 and E3-E4: around 20% roughly. We explain this with the structure of the training set, which mainly contains entries from the first two editions. The high extraction percentage of the first edition (88%) could be due to an overfit. However, the results might stem from factors other than the simple fact that most <b> tag occurrences are retrieved from E1-E2:

- The evolution of the encyclopedia across editions, where the patterns to define headwords have changed. Thus the same entry in two different editions may have some information discarded that was previously deemed important. The same logic applies to differences in spelling and general writing methods. Therefore, the resulting predictions may potentially differ for slightly varying entries from two different editions, mostly E1 and E4.
- Even if an entry retains most information and structure across editions, it is still prone to a varying model prediction in the last two editions. This is mainly due to their low proofreading rate.

The matching results show how people and locations covered in the encyclopedia changed between editions. Each new edition removed and added a considerable number of entries. This suggests editors actively chose which entries to include based on what was relevant at the time, rather than just adding to previous content.

Some entries do appear across all editions from 1 to 4, likely representing people and places that remained historically important throughout this period. This kind of analysis could help us understand how encyclopedias reflect what society considered important at different times.

7. Limitations and Future Work

We used a semi-automatic labeling to build the training set of headwords and segmented entries. The initial rule posits that headwords are marked with <b> tags in E1 and E2. Unfortunately, it creates a few false negatives. This can be even more confusing when two identical entries are marked differently. Cascading this problem over the rest of our 300 thousand entries in the dataset creates a question of the model's reliability if trained on a large number of false negatives.

We would like to explore this segmentation area in future work. Should we retain the automatic construction of the dataset, we would like to lower the number of false negatives or evaluate to which degree the model is affected by potential confusion.

The headword extractor is a sequence annotator that considers only one word. Moreover, we limited the entry categories to three types. We could include more types such as *ORG* and *TME*. We would also like to recover the complete annotation of the entries. This would enable us to enrich the definitions with more information such as mentions of dates, persons, and locations. We could thus extend the linking step and create relations either internally between mentions and entries of *Nordisk familjebok* or externally to Wikidata.

We linked nearly 11,000 entries of *Nordisk familjebok* to Wikidata. Our evaluation is nonetheless limited to a much smaller dataset. An improved evaluation would use more manually annotated data. In addition, we restricted Wikidata linking to items explicitly referencing *Nordisk familjebok*. We could apply this procedure to a larger subset of the Swedish Wikipedia, or explore any relevant

[4] https://huggingface.co/datasets/albinandersson/nf-headword-linked

information contained in the linked Wikidata objects for extraction and further analysis.

8. Conclusion

In this work, we described a comprehensive pipeline for processing historical encyclopedias. It consists of four major steps, notably an automated headword extraction, where we achieved an F1 score of 97.8% and an entity type classification with an F1 score of 93.4%. In a small-scale evaluation of the cross-edition matching, we obtained an accuracy better than 93%. We linked approximately 11,000 Wikidata items across all editions with a precision of 85% on our test set at the expense of recall that was of 16.5%.

We hope that our work will offer clearer, quantifiable insights into the perspectives and viewpoints of people living during the time the encyclopedias were published. We processed the four editions and we created tools to match entries. This should improve their comparison and the understanding of how the editors and readers perceived the evolution of the world and society. Finally, it should contribute to the digital preservation of historical knowledge, making this resource more accessible for future research and analysis. Beyond the methodological contributions, this work releases three publicly available datasets that may support future research in historical NLP, entity linking, and digital humanities.

9. Ethics Statement

The collection of *Nordisk familjebok* editions is in the public domain. Our work contributes to the development of tools for language resources and their annotation. We hope it can improve the understanding of human knowledge transmission through the extraction of versions of biographies and locations. Nonetheless,

1. The corpus we used contains dated and possibly false information. This can notably be the case for scientific theories or technological developments.

2. The Swedish historical context and ideas of years 1870-1950 may convey biases and old-fashioned viewpoints, possibly offensive. Users must be informed of this context.

10. Acknowledgements

This work was partially supported by *Vetenskaprådet*, the Swedish Research Council, registration number 2021-04533.

11. References

Axel Ahlin, Alfred Myrne Blåder, and Pierre Nugues. 2024. Mapping the past: Geographically linking an early 20th century Swedish encyclopedia with Wikidata. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11040–11048, Torino, Italia. ELRA and ICCL.

Tom Ayoola, Joseph Fisher, and Andrea Pierleoni. 2022a. Improving entity disambiguation by reasoning over a knowledge base. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2899–2912, Seattle, United States. Association for Computational Linguistics.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022b. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Simon Börjesson, Erik Ersmark, and Pierre Nugues. 2025. Matching and linking entries in historical Swedish encyclopedias. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 1–10, Albuquerque, New Mexico. Association for Computational Linguistics.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76):2493–2537.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56.

Tore Frängsmyr. 1991. encyklopedi. In *Nationalencyklopedin*, volume 5, pages 477–479. Bokförlaget Bra Böcker.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the National Library of Sweden – making a Swedish BERT. *CoRR*, abs/2007.01658.

Nordisk Familjebok. 1876–1951. *Nordisk Familjebok*. Project Runeberg. https://runeberg.org/nf.

Pierre Nugues. 2022. Connecting a French dictionary from the beginning of the 20th century to Wikidata. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2548–2555, Marseille, France. European Language Resources Association.

Pierre Nugues. 2024. Linking named entities in diderot's encyclopÃ©die to wikidata. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10610–10615, Torino, Italy. ELRA and ICCL.

Juan Antonio Ramirez-Orta, Eduardo Xamena, Ana Maguitman, Evangelos Milios, and Axel J. Soto. 2022. Post-OCR document correction with large ensembles of character sequence-to-sequence models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11192–11199.

Faton Rekathati. 2021. The KBLab blog: Introducing a Swedish sentence transformer.

Crina Tudor and Eva Pettersson. 2024. People and places of the past – named entity recognition in Swedish labour movement documents from historical sources. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 185–195, St. Julians, Malta. Association for Computational Linguistics.

Cornelis J van Rijsbergen. 1979. *Information Retrieval*. Butterworth.

Renshen Wang, Yasuhisa Fujii, and Ashok C. Popat. 2022. Post-OCR Paragraph Recognition by Graph Convolutional Networks . In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2533–2542, Los Alamitos, CA, USA. IEEE Computer Society.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

12. Language Resource References

Andersson, Albin and Jonasson, Salam and Wastring, Fredrik and Nugues, Pierre. 2026a. *Nordisk Familjebok Category Classification Dataset*. Hugging Face.

Andersson, Albin and Jonasson, Salam and Wastring, Fredrik and Nugues, Pierre. 2026b. *Nordisk Familjebok Headword Classified Matched Linked Dataset*. Hugging Face.

Andersson, Albin and Jonasson, Salam and Wastring, Fredrik and Nugues, Pierre. 2026c. *Nordisk Familjebok Headword Extraction Dataset*. Hugging Face.

Project Runeberg. 1992–2024. *Project Runeberg – Nordic Literature*. Linköping University. PID https://runeberg.org/. Accessed: 2024-12-02.