Universität Zürich[UZH]

# SNF Horizons Corpus

Programming Project

**Author: Tannon Kew**
Matriculation Nr: 17-717-018

Supervisor: Martin Volk

Contributors: Magdalena Plamada

Submission date: February 2019

Institute for Computational Linguistics

# SNF Horizons Corpus

Tannon Kew

February 20, 2019

## Contents

**1 Introduction** **3**

**2 Horizons** **3**

**3 Horizons Online Corpus** **4**

3.1 Collecting articles . . . . . 4

3.2 Converting HTML to XML . . . . . 5

**4 Horizons PDF Corpus** **5**

4.1 Collecting PDFs . . . . . 5

4.2 PDF text extraction . . . . . 6

4.3 Article alignment . . . . . 7

4.4 XML correction . . . . . 8

4.5 Extracting article content . . . . . 11

4.6 Converting to corpus XML . . . . . 12

**5 Known Issues** **13**

**6 Potential for Future Work** **14**

**7 Corpus statistics** **16**

**Appendix A** **17**

**Appendix B** **18**

**Acknowledgements**

Thank you to Magdalena Plamada for her contributions to this project. Throughout building the corpus, Magdalena was extremely helpful, providing advice and suggestions for improvement. Scripts that were co-authored are credited in the scripts themselves. Previous work done on corpora at our institute, especially Credit Suisse and Text+Berg, also provided a great deal of inspiration for how to approach a number of challenges faced in this project.

# 1 Introduction

Switzerland's diverse linguistic landscape provides a rich source of text documents that are published in multiple languages. These documents, which can cover diverse topics, can be said to be *in parallel* when they contain translations of the same text in two or more languages and thus offer an opportunity to construct new parallel corpora. One such publication that meets this criteria is Horizons, a magazine published by the Swiss National Science Foundation (German: Schweizerischer Nationalfonds (SNF)) and the Swiss Academies of Arts and Sciences which aims to promote scientific research projects in Switzerland. In this project, we have extracted, cleaned and processed the text from all available Horizons issues from the last 13 years in order to build a new parallel corpus, consisting of roughly 1 million tokens in German and French and around 0.4 million tokens in English.

# 2 Horizons

Horizons is a quarterly magazine reporting on scientific research topics in Switzerland. It is printed in German and French and for a short period between 2014 and 2017 it was also printed in English. Majority of the printed versions (starting from 2005) are available to download as PDFs from the SNF archive website[1]. Additionally, since June 2016, the magazine has also been published online[2]. Nowadays, the English version is only available online. Figure 1 provides a timeline overview of the publication formats.

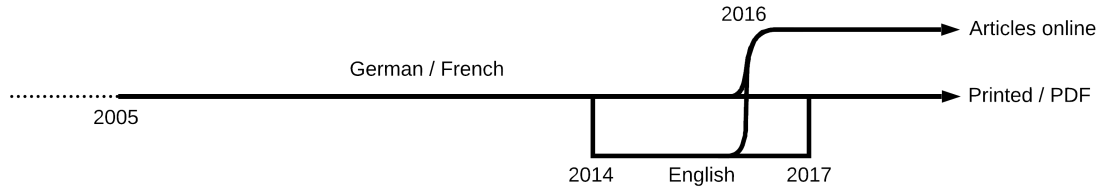



Figure 1: Timeline of Horizons publication formats

[1] `http://www.snf.ch/de/fokusForschung/forschungsmagazin-horizonte/archiv/Seiten/default.aspx` and `https://www.horizonte-magazin.ch/archiv/`

[2] `https://www.horizonte-magazin.ch/`

# 3 Horizons Online Corpus

Articles published online offer the advantage of being simple to collect and easy to process into XML, avoiding many of the problems associated with PDF text extraction or optical character recognition (OCR). This section describes the process of collecting and preparing the articles available on the Horizons website used to build the Horizons Online Corpus.

## 3.1 Collecting articles

In a first step, articles were collected with the web crawling script `SNF_html_crawler.py`. The main challenge with this step is ensuring article alignment between the three languages. An initial attempt to collect the articles relied on the author and publication date as unique ID information for article alignment under the assumption that no author had more than one publication on any given day. However, this assumption proved to be insufficient and resulted in a number of misaligned article files. To resolve this issue, a parallel approach to web crawling was employed which relied on the alternative language links for each German article published on the Horizons website.

**Relevant scripts**

- `SNF_html_crawler.py`

`SNF_html_crawler.py` starts by making a request to the URLs for each German issue in the range specified by the command line arguments, where all relevant article links are collected. Then, further requests are made to each article page, which is parsed for link tags designating the alternative language versions of the article. Once the links are found, the HTML source code for each link is saved accordingly. A counter ensures that the articles in German, French and English receive the same article ID. If an alternative language article is missing, the ID number is simply skipped for that article. In this way, article alignment is guaranteed via the article ID. The HTML files are written to language specific folders created in the output directory given as a command line argument according to the following convention: ID-number_issue-number_language-code_date-of-publication_author's-name_article-ID_article-title, e.g:

```
138_118_de_2018-09-06_Sven-Titz_Lawinengefahr-dank-Radar-
besser-einschätzen
```

Although the author and article titles are not necessary in the article filename, they were kept as they provide an easy way to check the alignment between languages.

In total, 158 articles were collected for German and French, and 157 articles for English.

**How to:** Call:

```
python3 SNF_html_crawler.py -s <starting issue number> -f
<finishing issue number>
```

## 3.2 Converting HTML to XML

The process of web crawling results in a collection of rich HTML files containing all the source code and meta data for each article web page. As a next step, text content needs to be extracted from the HTML files and converted into a well-formed XML format. To do this, HTML elements containing article text content and relevant metadata were first identified and then extracted using the script `convert_html2xml.py`.

**Relevant scripts**

- `convert_html2xml.py`

This script takes a directory of HTML files in one language as input and processes each file in the directory sequentially, writing all articles to a single XML file with article divisions. Therefore, this step must be done for each language. For each file, relevant metadata is extracted from specific HTML tags, and article content text is extracted largely from `p` and `h5` tags. The resulting XML file is then ready for further processing in order to convert it into a specific corpus format (see section 4.6).

**How to:** Call:

```
python3 convert_html2xml.py -i <input directory> -o <output xml
file>
```

# 4 Horizons PDF Corpus

Printed issues of Horizons are available in PDF format on two different websites. Issues published before 2017 are available from the SNF archive website and more recent issues are available on the Horizons website. This section describes the process of collecting the PDF files and extracting and preparing the text content from these documents. A visual overview of the steps involved in the process is provided in figure 3 in appendix A.

## 4.1 Collecting PDFs

The Horizons PDF files that are available on the old SNF archive website were collected using the simple web crawling script `PDF_collector.py`. The few remaining PDFs available on the new Horizons archive website were manually downloaded. In total, There are 53 PDFs in German, 52 in French[3] and 15 in English.

The names of the PDF files were then edited in order to ensure that all names matched the following convention: horizonte_year_issue-number_language, e.g:

```
horizonte_2004_102_de.pdf
```

[3]The French issue No. 71 from December 2016 is missing from the collection. Due to this, the German issue No. 71 was left out for processing in subsequent steps.

**Relevant scripts**

- `PDF_collector.py`

`PDF_collector.py` iterates through a list of URLs, making a request to each and parsing the HTML content for all `div` tags containing PDF documents. Relevant tags can be identified by their class attribute `download-form`. Each PDF file is then saved to the specified output directory given as a command line argument.

**How to:** Call:

```
python3 PDF_crawler.py <output directory>
```

## 4.2 PDF text extraction

PDF text extraction was initially investigated with PyPDF2, Adobe Acrobat and PDFlib TET. PyPDF2 is an easy-to-implement Python module which works quite well on simple text-based PDFs but does not perform well on complex page layouts such as magazines. Adobe Acrobat allows a user to export the text from a PDF in either rich-text or plain-text format to varying results. However, given the number of PDFs, such a manual operation was not suitable. PDFlib TET, on the other hand, offers many customisation options for text extraction, as well as an easy to use command line tool for efficient processing of many files.

TET's customisation options allow the user to specify the level of detail for the extracted text. This ranges from simple utf-8 encoded plain-text to fine-grained glyph information indicating the position, font style and size for every character in XML format. After some experimentation with files based on TET's `page` extraction option, which produces simple XML files with page-level information, it was clear that more fine-grained information was necessary for effective processing. For example, PDFs can contain invisible or hidden text caused by underlying images and maps[4], which is extracted by default with TET[5]. Investigating the issue of invisible text revealed that in many cases problematic text tended to have a much smaller font size than the article content of interest to us. Therefore, TET's `wordplus` extraction option was used to enable unwanted text elements to be filtered out on the basis of font size. The result of this text extraction with PDFLib's TET tool is a collection of TET Markup Language (TETML) files which can be processed with standard XML tools.

[4]Horizons 'Vor Ort' articles generally contain maps and were the typically affected by invisible text.

[5]TET offers an `ignoreinvisibletext` option which ignore text with a specific `textrenderingvalue`, however, experiments with setting this option did not sufficiently remove invisible text in the Horizons PDFs

## 4.3 Article alignment

In order to ensure that articles from DE, FR and EN publications can be aligned, it is necessary to insert article boundaries in the TETML files. To do this, the pages containing the table of contents (ToC) are identified in the TETML file and parsed for article titles and their relevant page numbers. This task is made difficult, however, by the fact that Horizons has had numerous layout changes since 2005. Publications can be split into three distinct 'eras'; OLD issues are the earliest publications, from No. 66–80, MID issues are those between No. 81–95, and NEW issues are from issue No. 96 until now. Furthermore, the complex layout of the ToC in all issues is particularly challenging for TET's extraction tool. There is no guarantee that text boxes are correctly recognised and often times, fully automatic parsing leads to poor results with missing titles, page numbers, or a mixture of both.

To deal with this issue, a semi-automatic approach was adopted which enables the user to validate each article title and its relevant starting page number before article boundaries are inserted into the file. While this approach is obviously more laborious and time-consuming than a fully automatic approach, it is necessary for attaining reliable article alignment and article titles in the language corresponding to the document.

While it may be possible to perform detailed article alignment in just one language, e.g. German, and then assume correspondence in the English and French PDFs, this would result in having German titles for articles in the English and French corpus files. As this could limit the potential for exploiting the language specific parts of this parallel corpus in other tasks, semi-automatic article alignment was performed for each TETML file individually. On average, roughly two to three corrections were performed per document. In general, NEW issues required fewer manual corrections than MID and OLD issues due to a more consistent layout and clearer separation of titles and subtitles.

### Relevant scripts

- `insert_article_boundaries.py`
- `Wordplus_Parser.py`
- `contents_control.py`

The script `insert_article_boundaries.py` accepts either a single TETML file or a directory of TETML files for bulk processing. For each input file, the script `Wordplus_Parser.py` instantiates a `TetmlFile` object to represent the XML tree structure. To help with parsing the document, certain attributes, such as `lang` and `num_of_pages`, are established as instance variables and the TETML namespace is removed from the tree structure in order to simplify processing.

The general approach to processing is as follows: Front matter pages, such as the editorial and contents page(s) are first identified and stored with the help of page headers. The

contents page(s) is then parsed for all potential page numbers according to the document era. Page numbers are assumed to be any digit occurring on the page which is within the range of the document's page count and which has a font size larger than the pre-defined minimum (8.00). Potential titles are then found by looking within the same paragraph node of the page number or its immediate neighbouring node. The result of this step is a list of page number/article title tuples.

If the command line parameter `-wc` is set to `true`, the collected contents list can be validated with the script `contents_control.py`. This script allows the user to confirm the automatically extracted page number/article title pairs. If there is a problem, the user can edit the page number and/or the title directly in the command line. Once each item in the list has been validated, a catch-all feature allows the user to add anything that may have been missed completely[6].

After validation, the extracted ToC list is then returned to `insert_article_boundaries.py` which proceeds to insert article start and end tags at appropriate positions in the XML tree. In addition, the validated contents list is also written to a text file which can be used for future reference or to avoid manually validating every article again in future[7].

**How to:** Call:

```
python3 insert_article_boundaries.py -i <input file/directory>
-o <output filepath> [options]
```

**Options:**

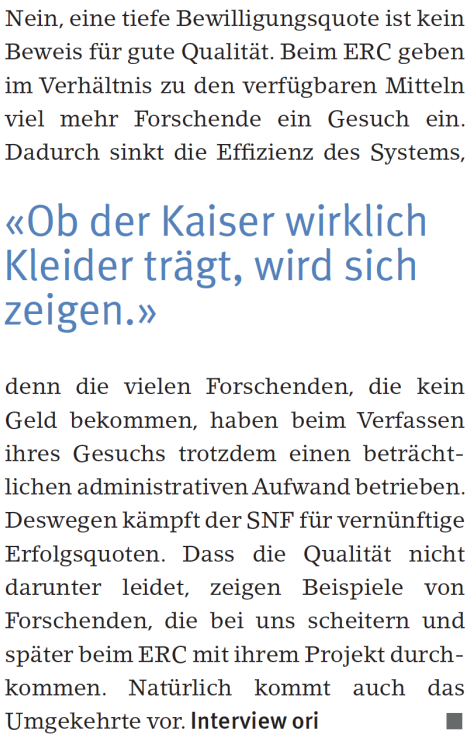
| | |
|---|---|
| `-wc` | if given, the article correction is done fully automatically, i.e. 'without control', however, this is not recommended (default: `false`). |

## 4.4 XML correction

Once article boundaries have been marked, the XML files need to be corrected. Often, paragraphs are split by additional text boxes, images, or even page breaks. Figure 2 shows one such example, where a paragraph break in the PDF layout splits a sentence in two. In addition to these easy-to-spot paragraph breaks, similar paragraph splits can also be caused by the occurrence of invisible text, text associated with map images and certain page features that TET marks as tabular content with a `Table` tag, rather than the regular `Para` tag. These cases need to be dealt with before the article content is extracted in order to ensure that we extract continuous sentences.

[6]This catch-all feature is particularly useful for the French issue No. 77, as the contents listing for the article 'Boîte à outils' on page 26 is missing from the PDF despite still appearing in the magazine and also being listed in the German version. Thanks to this feature, the article boundary can still be marked by manually entering the article's details

[7]The current contents list files are available with the corpus files and follow the naming convention: `horizonte_2005_66_de_validated_contents.txt`



Nein, eine tiefe Bewilligungsquote ist kein Beweis für gute Qualität. Beim ERC geben im Verhältnis zu den verfügbaren Mitteln viel mehr Forschende ein Gesuch ein. Dadurch sinkt die Effizienz des Systems,

«Ob der Kaiser wirklich Kleider trägt, wird sich zeigen.»

denn die vielen Forschenden, die kein Geld bekommen, haben beim Verfassen ihres Gesuchs trotzdem einen beträchtlichen administrativen Aufwand betrieben. Deswegen kämpft der SNF für vernünftige Erfolgsquoten. Dass die Qualität nicht darunter leidet, zeigen Beispiele von Forschenden, die bei uns scheitern und später beim ERC mit ihrem Projekt durchkommen. Natürlich kommt auch das Umgekehrte vor. **Interview ori** ■

```
        <Text>Dadurch</Text>
        <Text>sinkt</Text>
        <Text>die</Text>
        <Text>Effizienz</Text>
        <Text>des</Text>
        <Text>Systems</Text>
        <Text>,</Text>
</Para>
<Para>
        <Text>«</Text>
        <Text>Ob</Text>
        <Text>der</Text>
        <Text>Kaiser</Text>
        <Text>wirklich</Text>
        <Text>Kleider</Text>
        <Text>trägt</Text>
        <Text>,</Text>
        <Text>wird</Text>
        <Text>sich</Text>
        <Text>zeigen</Text>
        <Text>.»</Text>
</Para>
<Para>
        <Text>denn</Text>
        <Text>die</Text>
        <Text>vielen</Text>
        <Text>Forschenden</Text>
        <Text>,</Text>
```

Figure 2: An example of split paragraphs in a PDF and the corresponding (simplified) XML file before paragraph merging is performed. Source: `horizonte_2012_93_de.xml`.

### Relevant scripts

- `correct_XML.py`
- `cleaning_tools.py`
- `merging_tools.py`
- `pre_noun_words_de.py`

The script `correct_XML.py` accepts either a single file or a directory containing XML files with article boundaries as input. For each input file, the XML tree is parsed for `article` tags, which are then processed sequentially with a series of functions that clean the article by removing unwanted or empty text elements and performing paragraph merging operations.

Paragraph merging is divided into two categories, namely, consecutive merging and 'skip' merging. Consecutive merging rejoins broken paragraphs caused by the TET extraction tool incorrectly inserting paragraph breaks where there should not be one. 'Skip' merging, on the other hand, rejoins broken paragraphs caused by additional text boxes and layout features such as images and page breaks.

The general approach to paragraph merging can be described as follows: given a pair of paragraphs $P_i$ and $P_j$, if $P_i$ does not end with a valid sentence final punctuation mark and $P_j$ begins with a lower-case letter, they are considered to be split paragraph candidates.

Naturally, there is a problem with this approach for German due to capitalisation of common nouns. Therefore, when processing German files, a list of some frequent noun-preceding words (`pre_noun_words_de.py`) is used to allow more candidates to be found in cases where $P_j$ begins with a capitalised noun.

Here, text elements such as titles, subtitles, and picture captions pose a challenge as they often do not end with sentence final punctuation. Therefore, rules are applied to consider the character size and font style of paragraph elements before merging. The script `merging_tools.py` contains a number of functions which analyse both article and paragraph-level features required to permit merges. For consecutive merge candidates, if the final character of $P_i$ and the initial character $P_j$ have a matching size and font style, a merge is performed.

Skip merges are slightly less frequent than consecutive merges and mainly concern the article content itself and not other text elements such as subtitles and picture captions. Therefore, skip merges also consider the character size and font style of the article content. Here, the assumption is made that majority of the characters in an article belong to the article content. Under this assumption, the most frequent character size and font style occurring in an article is calculated and stored. If a skip candidate pair is found, the final character of $P_i$ and the initial character of $P_j$ must match the most frequent character size and font style in the article for a merge to be permitted. Figure 5 in appendix B shows an example of a page containing subtitles and author credits not ending with sentence final punctuation. Applying the most frequent character size and font style restriction ensures that these are ignored as skip merging candidates, while the article paragraphs are still eligible for skip merging operations with any number of intermediary paragraph nodes.

Both merging functions are recursive, so that when merges are performed, the article is searched again for potential new merges. Once no more merges can be performed, the process finishes.

In addition to paragraph merges, a number of article-level 'cleaning' operations need to be performed. These are taken care of with the functions provided in `cleaning_tools.py`. Tags containing out-of-article text, such as page headers, text inadvertently extracted from images and invisible text are removed on the basis of font size and the occurrence of diagonal coordinates at character level. 'Floating' words, which don't have a `Para` parent tag, are appended to the previous `Para` tag. Also, dropped capitals, which usually appear in paragraph initial position, are sometimes isolated in their own `Para` node by TET. These are identified using the TETML attribute `dropcap` and are appended to the first word in the following paragraph.

Incorrect paragraph placement occurs in some articles, leading to mixed sentences and incoherent text. This can be due to random paragraph breaks caused by sections of invisible text[8] and/or TET's extraction tool, which sometimes confuses mid article paragraphs with article initial paragraphs[9]. Therefore, articles containing potential issues are flagged with the function `mark_potential_errors()`. Since it is difficult to automatically identify occurrences of incorrect paragraph placement caused by TET, an assumption is made on the

[8]See figure 4 in appendix B.
[9]See figure 6 in appendix B.

basis of dropped capitals. If there is no dropped capital in the first 5 paragraphs, but there is one elsewhere in the article, the `potential_errors` flag is set. Occurrences of tabular formats and the resulting floating words are easier to detect as the flag can be set whenever a correction operation is performed. The cause of the potential error is also specified as the value of the `potential_errors` attribute in order to assist with error analysis.

**How to:** Call:

```
python3 correct_XML.py -i <input file/directory> -o <output
filepath> [options]
```

**Options:**

| | |
|---|---|
| `-v` | allows the user to the console to inspect corrections by printing verbose print statements to the console (default: `false`). |

## 4.5 Extracting article content

Once the XML file has been marked for article boundaries and corrected it is time to extract the article content text. By default, the TET extraction tool performs hard tokenization, splitting word boundaries at almost all punctuation marks[10]. For example, English "it's" is split by TET into three distinct tokens "it", "'" and "s". In order to be able to perform reliable tokenization with a rule based tokenizer such as `Cutter`, punctuation first needs to be restored.

### Relevant scripts

- `text_extractor.py`

The script `text_extractor.py` accepts either a single file or a directory containing corrected XML files as input. For each input file, the XML tree is parsed and for each `article` tag, paragraph nodes are processed as a list of tokens. De-tokenization is performed on this list according to a series of rules. French is handled by a language-specific function to accommodate for differences in punctuation usage[11]. In addition, URLs and email addresses which are also taken apart by TET's hard tokenization are corrected and page headers and footers containing page numbers that may have been missed in `correct_xml.py` are also removed in this step. The output of this step is a simplified XML file with articles consisting of continuous text contained within paragraph `div` tags.

[10] TET's `wordplus` extraction method offers an optional content analysis setting `punctuationbreaks=false` which does not perform this hard tokenization, however, this was only discovered after the de-tokenization script was written. It is highly recommended that this option be used for text extraction in future projects.

[11] In German and English, an apostrophe usually represents a missing character in the second word of a contraction and therefore generally needs to appended to the following token. In French, on the other hand, an apostrophe often replaces a character(s) in the first word of a contraction and, therefore, should be appended to its preceding token.

**How to:** Call:

```
python3 text_extractor.py -i <input file/directory> -o <output
filepath>
```

## 4.6 Converting to corpus XML

Once the XML files are reduced to continuous text elements contained within `article` tags it is possible to convert this format to a verticalised corpus XML format, as is commonly used to store other corpora at our institute, such as Text+Berg and the Credit Suisse corpora.

### Relevant scripts

- `corpusXML.py`

Once again, input to the script is either a single XML file (in the case of the online articles) or a directory of valid XML files (such as the PDF corpus). Each article in the input file is processed as follows: Sentences are split using Spacy's language specific sentence tokenizer which employs a neural model[12]. Before tokenizing each sentence, language detection is performed with a mixed dictionary/rule-based approach and a statistical approach using the Python module `langdetect`. Experiments showed that `langdetect` outperformed the alternative module `langid`, however, it still made mistakes concerning shorter sentences and those consisting largely of URLs, email addresses and postal addresses. Therefore, the dictionary/rule-based approach was added in order to improve language detection results for such sentences. Once the sentence language is detected, the rule-based, language-specific tokenizer `Cutter`, written by Johannes Gräen, is employed to split the sentence into a list of tokens. Unlike the pipeline used for the Credit Suisse corpus, here, `Cutter` is called using its Python wrapper. Once sentences have been tokenized, part-of-speech tagging and lemmatization is performed with TreeTagger, which is also called using the Python wrapper `treetaggerwrapper`, written by Laurent Pointal. In order for TreeTagger to process files with the correct parameter files, it is recommended to specify the full path to the directory containing the relevant language parameter files using the optional command line argument when running the script. In a final step, special characters and punctuation marks are normalised and TreeTagger's default <unknown> lemma tag is replaced with the value `'unk'`, in order to be consistent with other corpora at our institute.

**How to:** Call:

```
python3 corpusXML.py -i <input file/directory> -o <output
filepath> [options]
```

[12] It is also possible to use NLTK's rule-based sentence splitter by specifying it as a command line argument when calling the script, however, it is expected that a neural model performs better as it avoids some of the pitfalls of a rule-based approach when there is lacking sentence final punctuation in some texts.

**Options:**

| | |
|---|---|
| `-T` | allows the user to specify the path to a directory containing TreeTagger's language-specific parameter files: `german.par`, `french.par` and `english.par` (default: `/Applications/Tree-Tagger/lib/`). |
| `-S` | allows the user to specify the method of sentence segmentation. A rule based approach can be selected by giving the argument `nltk` or a neural model can be used by specifying the argument as `spacy` (default: `spacy`). |

# 5 Known Issues

Constructing a corpus from PDF documents poses numerous challenges regarding the extraction of clean and consistent linguistic data. This section describes the known issues in the current corpus release.

- **Split sentences** – Despite efforts to repair broken sentences through the removal of invisible text and by merging split paragraphs with the XML correction script, as described in section 4.4, broken sentences continue to be a major problem in the Horizons PDF corpus.

  Inspecting the corpus files reveals that many of the split sentences are the result of poor layout detection by TET's extraction tool which incorrectly orders text boxes and article paragraphs in the extracted TETML files. The TET command line tool offers a list of optional arguments which can improve layout detection for documents such as magazines, however, experiments showed that this optional setting did not fully eliminate the problem. Due to the difficulty of automatically detecting broken and incoherent sentences, accurately identifying problematic articles is rather challenging and therefore, assessing improvement with the optional layout detection setting is also difficult.

- **Detecting potential errors** – The current rules used to set the `potential_errors` flag do not always successfully detect problematic articles. For example, article `a15` in `horizonte_2005_67_de.xml` reveals a problematic case where the `potential_errors` flag is set to `false` despite the occurrence of split sentences in the extracted article content due to mixed paragraph placement by TET's extraction tool. In addition, 'invisible text' or text belonging to maps and other image elements remains an issue for some articles. For example, article `a8` in `horizonte_2010_85_de.xml` contains text elements from a map image with character sizes greater than the predetermined threshold of 8.00. Therefore, the assumption for removing invisible text does not hold in this case. As a result, no deletions are performed based on this assumption and the `potential_errors` flag is not set.

- **Paragraph merging** – At present, paragraph merging does not correctly handle words split over two lines with a hyphen. For example, `horizonte_2012_92_de.xml`, article `a12` contains the sentence, "Die text-genetisch-kritische Schule dagegen will den Werdegang des Textes möglichst genau abbilden;", which is split over two pages at the first hyphen in the compound adjective "text-genetisch-kritische". Ideally, instances such as these should handled in such a way that the hyphenated parts of a word split over a layout paragraph break are glued back together, in order to form a single token.

- **Misleading article titles** – Some article titles do not accurately represent the contents of the article as depicted in the magazine's ToC.

  In cases where multiple articles appear on a single page, the corpus XML article title is often reduced to include only the first title of the group of articles. For example, in article `a5` of `horizonte_2007_75_de.xml` the title is "Die Genetik des emotionalen Gedächtnisses", however, this article element actually encompasses the short articles "Als Erik der Rote grün sah" and "Der federnde Gang der ersten Europäer" as well as "Die Genetik des emotionalen Gedächtnisses".

  Article titles also often appear together with a subtitle or miniature abstract which are not separated by punctuation marks in the PDF's ToC. As a result, these elements are often extracted as a continuous sentence string for the article title. For example `horizonte_2016_109_de.xml`, article `a22` has the title "Nützliches Rauschen der Neuronen Neuronen verarbeiten elektrische Signale unterschiedlich", which consists of the article title, "Nützliches Rauschen der Neuronen" together with the subtitle "Neuronen verarbeiten elektrische Signale unterschiedlich".

- **Sentence tokenization** – Inspecting the corpus XML files reveals that some web addresses are not correctly recognised and as a result, end up being interrupted by sentence boundaries. For example, the URL path "blogs.philosophie.ch/mensch", which appears in article `a10` of `horizonte_2016_109_de.xml` is split into separate sentences at each period. Additionally, Spacy's sentence tokenization tends to separate opening quotation marks which should be identified and reattached to the sentence they belong to.

# 6 Potential for Future Work

Given the known issues mentioned in the previous section, it is clear that work still remains on the Horizons PDF corpus. While the web corpus is easily expandable as new articles are published online, the problems related to text extraction in the PDF corpus require more attention in order to ensure accurate and reliable linguistic data. Therefore, this section outlines some of the possible tasks that could be carried out to develop and improve the PDF corpus in future.

Overcoming the persistent errors caused by incorrect paragraph placement and interference from non-content related text is crucial for attaining clean and reliable text data. One

possible solution could be to use OCR to process problematic articles. This could potentially detect page layout features more accurately and eliminate interference from text boxes containing non-article content. However, automatically identifying these problematic articles is still an unresolved issue. Therefore, further methods should be developed to successfully detect problematic articles. If these methods are developed, comparing the results of processing a corpus with TET's optional layout argument that is supposed to improve layout detection for magazines would also be useful to understand where other improvements could be made and to assist with future projects.

Post-processing steps for German, such as lemmatization correction for elliptical composite nouns (e.g. Vor- und Nachteile) and separable prefix verbs, which are performed in the Credit Suisse processing pipeline, are still required.

Finally, sentence and word alignment should be performed in order to be able to exploit the potential of this parallel corpus and to be able to incorporate the linguistic data in applications such as Multilingwis.

# 7 Corpus statistics

As a result of this work, we have created a new parallel corpus from the SNF Horizons magazine. The tables below provide some basic statistics, indicating the current condition of both the Horizons PDF and Online corpus.

Table 1: Horizons PDF Corpus

| | German | French | English |
|---|---|---|---|
| Magazines | 52 | 52 | 15 |
| Articles | 1,239 | 1,239 | 395 |
| Potential error articles detected | 189 | 161 | 29 |
| Tokens | 1.02 million | 1.19 million | 0.39 million |
| Token types | 85,221 | 51,563 | 24,794 |
| Lemma types | 35,578 | 17,557 | 14,210 |
| Unknown lemmas | 52,705 | 47,876 | 8,958 |

Table 2: Horizons Online Corpus

| | German | French | English |
|---|---|---|---|
| Issues | 10 | 10 | 10 |
| Articles | 158 | 158 | 157 |
| Tokens | 114,084 | 126,333 | 131,146 |
| Token types | 19,318 | 15,011 | 13,324 |
| Lemma types | 10,609 | 7,315 | 8,404 |
| Unknown lemmas | 6,040 | 5,193 | 2,970 |

**A**

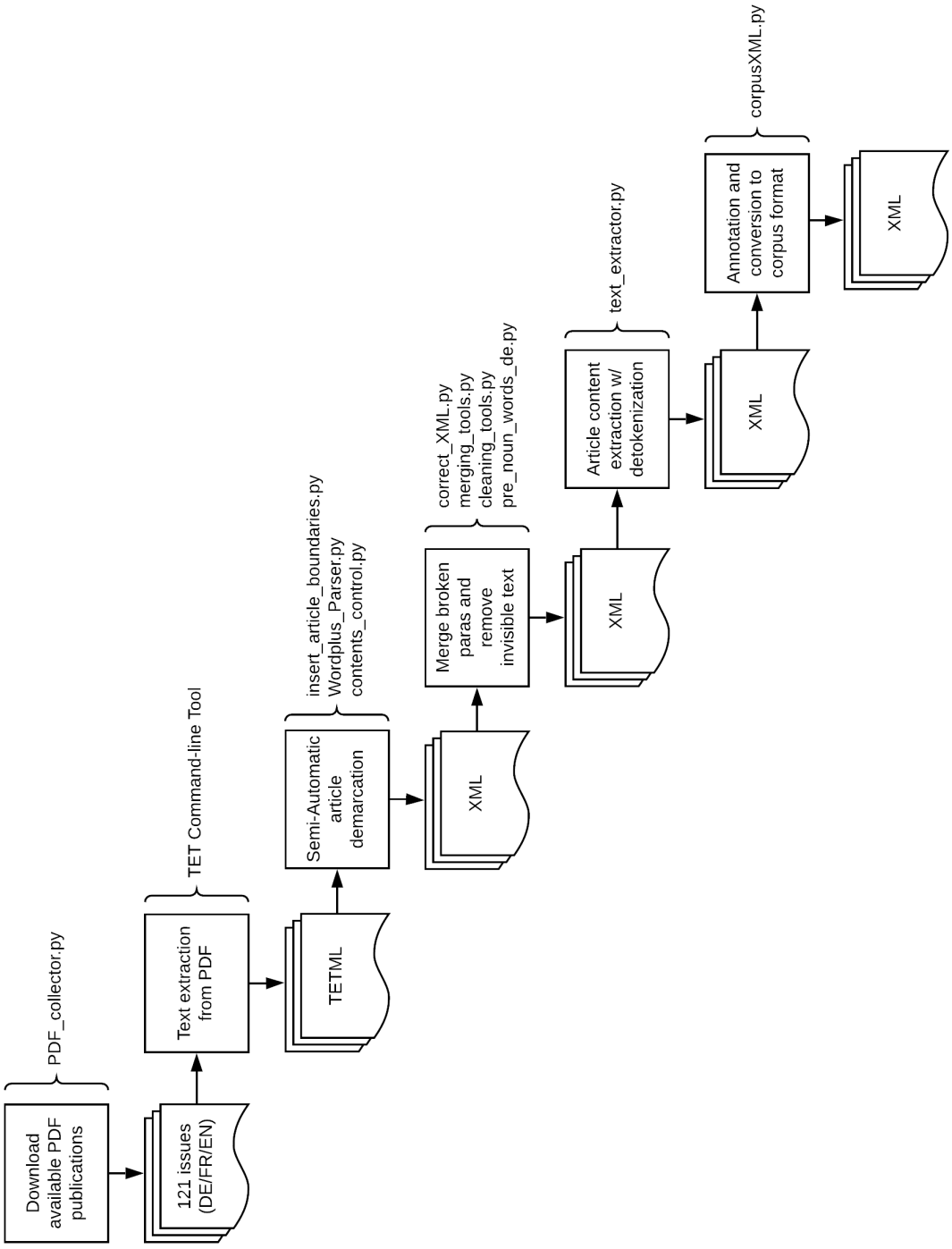
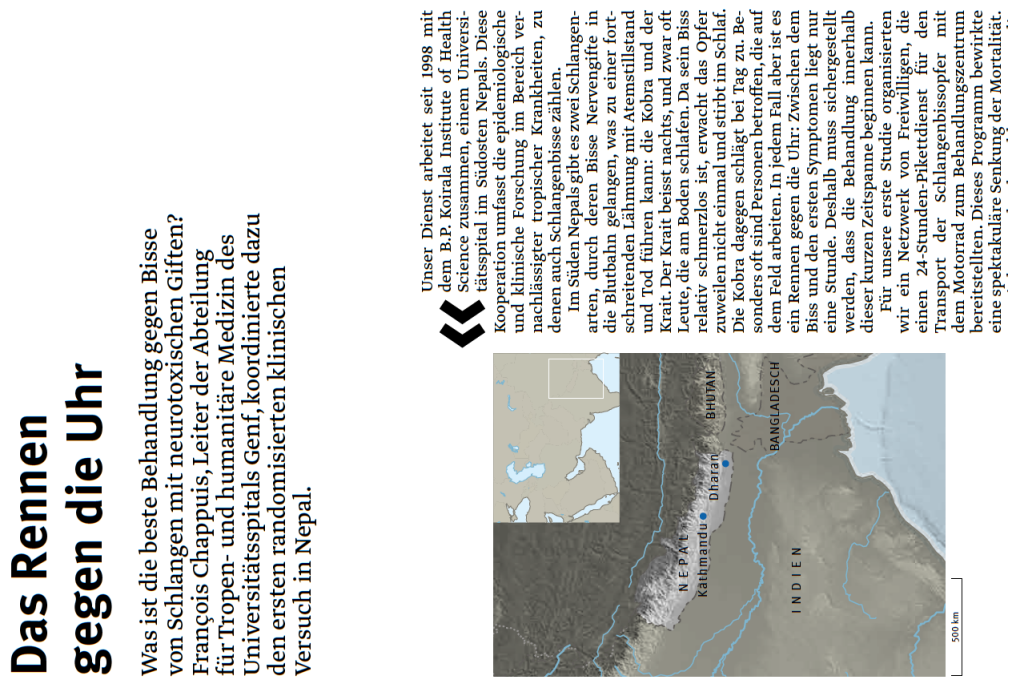



Figure 3: An overview of the process a extracting text content from the SNF Horizonte magazines.

**B**



# Das Rennen gegen die Uhr

Was ist die beste Behandlung gegen Bisse von Schlangen mit neurotoxischen Giften? François Chappuis, Leiter der Abteilung für Tropen- und humanitäre Medizin des Universitätsspitals Genf, koordinierte dazu den ersten randomisierten klinischen Versuch in Nepal.

und Tod führen kann: die Kobra und der relativ schmerzlos ist, erwacht das Opfer ein Rennen gegen die Uhr: Zwischen dem Im Süden Nepals gibt es zwei Schlangenarten, durch deren Bisse Nervengifte in die Blutbahn gelangen, was zu einer fortschreitenden Lähmung mit Atemstillstand Krait. Der Krait beisst nachts, und zwar oft Leute, die am Boden schlafen. Da sein Biss zuweilen nicht einmal und stirbt im Schlaf.

«Unser Dienst arbeitet seit 1998 mit dem B. P. Koirala Institute of Health Science zusammen, einem Universitätsspital im Südosten Nepals. Diese Kooperation umfasst die epidemiologische und klinische Forschung im Bereich vernachlässigter tropischer Krankheiten, zu denen auch Schlangenbisse zählen.

Die Kobra dagegen schlägt bei Tag zu. Besonders oft sind Personen betroffen, die auf dem Feld arbeiten. In jedem Fall aber ist es ein Rennen gegen die Uhr: Zwischen dem Biss und den ersten Symptomen liegt nur eine Stunde. Deshalb muss sichergestellt werden, dass die Behandlung innerhalb dieser kurzen Zeitspanne beginnen kann.

Für unsere erste Studie organisierten wir ein Netzwerk von Freiwilligen, die einen 24-Stunden-Pikettdienst für den Transport der Schlangenbissopfer mit dem Motorrad zum Behandlungszentrum bereitstellten. Dieses Programm bewirkte eine spektakuläre Senkung der Mortalität.

Im Süden Nepals gibt es zwei Schlangenarten, durch deren Bisse Nervengifte in die Blutbahn gelangen, was zu einer fortschreitenden Lähmung mit Atemstillstand und Tod führen kann: die Kobra und der Krait. Der Krait beisst nachts, und zwar oft Leute, die am Boden schlafen. Da sein Biss relativ schmerzlos ist, erwacht das Opfer zuweilen nicht einmal und stirbt im Schlaf.

Die Kobra dagegen schlägt bei Tag zu. Besonders oft sind Personen betroffen, die auf dem Feld arbeiten. In jedem Fall aber ist es ein Rennen gegen die Uhr: Zwischen dem Biss und den ersten Symptomen liegt nur eine Stunde. Deshalb muss sichergestellt werden, dass die Behandlung innerhalb dieser kurzen Zeitspanne beginnen kann.

Für unsere erste Studie organisierten wir ein Netzwerk von Freiwilligen, die einen 24-Stunden-Pikettdienst für den Transport der Schlangenbissopfer mit dem Motorrad zum Behandlungszentrum bereitstellten. Dieses Programm bewirkte eine spektakuläre Senkung der Mortalität.




```
<article title="Das Rennen gegen die Uhr" potential_errors="floating_words" id="
    a9">
  <div>Das Rennen gegen die Uhr</div>
  <div>Was ist die beste Behandlung gegen Bisse von Schlangen mit neurotoxischen
      Giften? François Chappuis, Leiter der Abteilung für Tropen- und humanitäre
      Medizin des Universitätsspitals Genf, koordinierte dazu den ersten
      randomisierten klinischen Versuch in Nepal.</div>
  <div>und Tod führen kann: die Kobra und der relativ schmerzlos ist, erwacht das
      Opfer ein Rennen gegen die Uhr: Zwischen dem Im Süden Nepals gibt es zwei
      Schlangenarten, durch deren Bisse Nervengifte in die Blutbahn gelangen, was
      zu einer fortschreitenden Lähmung mit Atemstillstand Krait. Der Krait beisst
       nachts, und zwar oft Leute, die am Boden schlafen. Da sein Biss zuweilen
      nicht einmal und stirbt im Schlaf.</div>
  <div>«Unser Dienst arbeitet seit 1998 mit dem B. P. Koirala Institute of Health
      Science zusammen, einem Universitätsspital im Südosten Nepals. Diese
      Kooperation umfasst die epidemiologische und klinische Forschung im Bereich
      vernachlässigter tropischer Krankheiten, zu denen auch Schlangenbisse zählen
      .</div>
  <div>Die Kobra dagegen schlägt bei Tag zu. Besonders oft sind Personen betroffen
      , die auf dem Feld arbeiten. In jedem Fall aber ist es Biss und den ersten
      Symptomen liegt nur eine Stunde. Deshalb muss sichergestellt werden, dass
      die Behandlung innerhalb dieser kurzen Zeitspanne beginnen kann.</div>
  <div>Für unsere erste Studie organisierten wir ein Netzwerk von Freiwilligen,
      die einen 24-Stunden-Pikettdienst für den Trans port der Schlangenbissopfer
      mit dem Motorrad zum Behandlungszentrum bereit stellten. Dieses Programm
      bewirkte eine spektakuläre Senkung der Mortalität.</div>
```

Figure 4: An example of a page containing a map with invisible text and `Table` tags that interrupt the article text and it's corresponding extracted XML file after correction. Here, the `potential_errors` flag has successfully been set, showing the cause for errors as 'floating words'. Source: 'Vor Ort' article from `horizonte_2014_103_de.pdf` (p.24).

## Wissenschaft als Beruf

An der Karriere zu arbeiten und Zeit für Partnerschaft und Familie zu haben: Das wünschen sich auch junge Wissenschaftlerinnen und Wissenschaftler, wie sich in einer Studie der Soziologin Ulle Jäger von der Universität Basel zeigt, in der sie 40 Interviews aus der Schweiz und Deutschland auswertete. Doch in der Realität, in der für eine wissenschaftliche Karriere Mobilität und uneingeschränkte zeitliche Verfügbarkeit verlangt wird, ist dieses Ziel schwer zu erreichen. «Für mich sind zwei Szenarien denkbar», so Ulle Jäger. Zugespitzt formuliert lautet die erste Zukunftsvision: Der Wissenschaftsapparat läuft weiter wie bisher. Diejenigen Männer und die geringere Anzahl an Frauen, denen privat «der Rücken frei gehalten» wird, können Professuren leichter besetzen als ihre Kolleginnen und Kollegen in egalitären Partnerschaften. Im zweiten Szenario stellen zeitliche Verzögerungen in der Karriereplanung, beispielsweise durch eine Familiengründung oder die Pflege älterer Familienmitglieder, keinen Nachteil dar.

Damit das zweite Szenario annähernd Realität würde, müssten heutige Anforderungen stärker reflektiert und hinterfragt werden: Ist ein Auslandsaufenthalt tatsächlich für alle Positionen in der Wissenschaft unverzichtbar? Wie wichtig ist die Anzahl der Publikationen für eine Lehrtätigkeit? Ulle Jäger: «Statt nur um Exzellenzkriterien sollte es darum gehen, die beruflichen Ansprüche so zu gestalten, dass eine Person ‹gut genug› sein kann und nicht über ihre Grenzen gehen muss, wenn sie Berufliches und Privates miteinander in Einklang bringen möchte.» *Nora Heinicke*

Photographee.eu/Shutterstock

**Einklang zwischen Beruflichem und Privatem?**

Universitätsbibliothek Basel

National-Zeitung

Organ für Handel und Industrie – Amtsgeblatt der Stadt Basel

Das politische Gespräch zwischen den englischen und italienischen Staatsmännern abgebrochen

Stalin und Dynamik in der Weltpolitik

**Die «National-Zeitung» aus Basel: «Das letzte freie Wort in deutscher Sprache».**

## Als die Basler Zeitung liberal war

«Kein Blatt wird in den Prager Cafés jetzt eifriger verlangt als ihre Nationalzeitung, auch auf der Strasse wird ihr Blatt viel gekauft. Als das letzte freie Wort in deutscher Sprache hat es eine Sonderstellung.» Diese Zeilen schrieb Max Brod im Winter 1939 an den Feuilletonredaktor Otto Kleiber in Basel. Während mehr als drei Jahrzehnten leitete Kleiber von 1919 bis 1953 das Feuilleton der Basler «National-Zeitung» und bot der deutschen Exilliteratur im Nationalsozialismus einen sicheren Hafen. Bekannte Leute wie Bertolt Brecht und Erika Mann, aber auch unbekannte Publizisten veröffentlichten in der Rubrik «Unter dem Strich» ihre Texte, die sie in Hitler-Deutschland nicht publizieren konnten.

Die Literaturwissenschaftlerin Bettina Braun von der Universität Zürich hat vor drei Jahren begonnen, die noch weitgehend unbekannte Bedeutung der «National-Zeitung» für die Exilliteratur zwischen 1933 und 1940 aufzuarbeiten. Die Durchsicht von rund 5000 Ausgaben – damals erschienen die Tageszeitungen noch in einer Früh- und einer Spätausgabe – ergab rund 3500 Veröffentlichungen von Exilanten. Braun hat die Texte in einer Datenbank erfasst, die der Forschung zugänglich gemacht werden soll. Die Textsammlung bildet die Grundlage ihrer Dissertation zur Gattungsgeschichte des Feuilletons in der Schweiz während dieser Zeit. Als Adresse für Exilliteratur sticht die «National-Zeitung», die 1977 mit den «Basler Nachrichten» zur «Basler Zeitung» fusioniert wurde, heraus: Die Zürcher «NZZ» wollte damals die kritischen Texte der Exilanten nicht abdrucken. Liberal war damals die Basler Zeitung. *Stefan Stöcklin*

B. Braun (2012): Das literarische Feuilleton des Exils in der Schweiz – Die Basler «National-Zeitung». Zeitschrift für Germanistik, Heft 3/2012: 667–669.

## Afrika altert schnell

Die demografische Entwicklung beschäftigt nicht nur Industriestaaten, sondern auch Entwicklungs- und Schwellenländer. Ein Forschungsteam vom Ethnologischen Seminar der Universität Basel hat das Älterwerden in Afrika am Beispiel von Tansania untersucht. Wichtigste Erkenntnis: «Altwerden in Afrika ist mit vielen Unsicherheiten verbunden», sagt Studienleiterin Brigit Obrist. Formelle Unterstützungssysteme wie eine staatliche Altersvorsorge oder öffentliche Pflegeheime gibt es kaum, dabei leiden auch in Afrika alternde Menschen zunehmend an chronischen Krankheiten und werden vermehrt pflegeanfällig. Die wichtigsten Stützen sind Familie, Verwandtschaft und die Gemeinschaft, doch diese traditionellen Netzwerke sind «brüchig und durchlässig» geworden, so Projektleiter Piet van Eeuwijk. Allmählich entstehen als zusätzliche Absicherung auch neue Beziehungsmuster, etwa durch Sozialkontakte mit Mobiltelefon, durch Geldüberweisungen von im Ausland lebenden Kindern oder durch Mitgliedschaften in Altersvereinigungen.

Trotz vieler Unsicherheiten streben Afrikanerinnen und Afrikaner ein Altern in Würde an. Die meisten von ihnen, insbesondere die Männer, arbeiten, solang es geht. Wer keiner Erwerbsarbeit mehr nachgeht, hat oft noch Aufgaben und Funktionen in der Familie und im sozialen Umfeld. Brigit Obrist: «Auch wenn junge Generationen heute nach anderem Wissen streben, gelten die Alten noch immer als einflussreiche Instanzen im privaten und öffentlichen Leben.» So seien alte Menschen eine wichtige soziale und politische Stütze für die Gesellschaften Afrikas. *Irène Dietschi*

Sandra Staudacher-Preite

**Eine Studienteilnehmerin aus Sansibar zusammen mit der Frau ihres Enkels und der Urenkelin.**



Figure 5: An example of a challenging page for skip merging paragraphs. Subtitles and author credits not ending with sentence final punctuation are ignored as skip merging candidates under the assumption that their character size and font style differ from that of the main article content. Source: 'Kultur und Gesellschaft' article from `horizonte_2014_103_de.pdf` (p.35).

porträt

# Stefania Xella Hansen: Passionierte Teilchenbeschleunigerin

VON ROLAND FISCHER
BILDER STUDIO25

**Italienerin, Teilchenphysikerin, zweifache Mutter, Jungforscherin: Stefania Xella Hansen vereint vermeintliche Gegensätze ganz selbstverständlich. Für die Diplomarbeit hatte es sie einst ans Cern verschlagen. Nun ist sie über einige Umwege wieder in der Schweiz gelandet.**

Ihr Name lässt einen stutzen. Xella ist, vermutet sie, griechischen Ursprungs, und Hansen ist der Name ihres dänischen Mannes. Aus Bologna kommt sie, aber wenn man ihr so zuhört, wie sie ruhig und jeden Satz abwägend erzählt, da würde man ihr die Italienerin nicht geben. Es ist nun auch fast zehn Jahre her, seit sie ihre Heimat verlassen hat und sich aufgemacht hat auf eine Tour d'horizon durch die wichtigen Labors der Teilchenphysik. Seit zwei Jahren forscht sie an der Universität Zürich.

Teilchenphysik ist das Gegenstück zur Astronomie. Wo die Astrophysiker mit ihren Teleskopen nach der Struktur des Kosmos im Grossen suchen, da tun es die Teilchenphysiker mit ebenso gewaltigen Apparaten im Kleinen. Ein schöner Zufall, dass ihr Mann gerade diese physikalische Gegenwelt erforscht, sie findet es «sehr angenehm, dass ich mich einfach an ihn wenden kann, wenn ich eine Frage zur Astrophysik habe». Zusammen gearbeitet haben sie aber nie, dafür sind die Gebiete doch zu verschieden.

Die elementaren Teilchen im Atom sind nicht einfach Proton, Elektron und Neutron, diese bestehen wiederum aus so genannten Quarks, und dazu bevölkern noch eine Unzahl weiterer seltsamer Teilchen einen veritablen Teilchenzoo. Den Verhaltensweisen dieser Teilchen auf die Spur zu kommen und so die grundsätzlichsten Gesetze des Universums aufzuklären, das ist die Aufgabe der Teilchenphysik. Wie erklärt sie ihrem kleinen Sohn, womit sie sich beschäftigt im Büro? Gar nicht, wie sollte sie es ihm auch erklären, meint sie, er sei ja erst vierjährig, und ohnehin, ihn interessiere höchstens, warum sie denn nicht den ganzen Tag zu Hause bleibe, um mit ihm zu spielen. Sie hat noch eine Tochter, etwas über ein Jahr alt, und damit wäre ja eigentlich schon ein volles Mutter- und Hausfrauenpensum beisammen. «Die Familie ist mir das Wichtigste», sagt sie, aber natürlich möchte sie auf die interessante Arbeit an der Universität nicht verzichten, ebenso wenig wie ihr Mann. «Für den Moment kommen wir mit der Situation noch klar, aber es ist schwierig, mit dieser familiären Situation eine Festanstellung zu finden.» Sie wird unterstützt vom Marie-Heim-Vögtlin Programm des Schweizerischen Nationalfonds, das Frauen mit meist familienbedingt verzögerten Karrieren ermöglicht, ihre wissenschaftliche Laufbahn weiterzuverfolgen. Sie vernachlässigt die Forschung durchaus nicht: Die Kinder sind tagsüber in der Krippe, und wenn sie dann im Bett sind, arbeiten sie und ihr Mann mitunter eben noch eine Stunde oder zwei zu Hause weiter. Insofern böte die wissenschaftliche Arbeit ja die nötige Flexibilität für unkonventionelle Lösungen des Familie- Karriere-Problems. Doch Festanstellungen sind in der akademischen Jobbörse rar. Und wo sich einmal eine Jobperspektive auftut, da muss es für ihren Mann natürlich auch eine interessante Stelle geben.

**Paar-Synchrontanz**

Diesen Synchrontanz vollführt das Paar nun schon fast zehn Jahre. Sie hatten sich am Cern kennen gelernt, als sie dort ihre Diplomarbeit schrieb. Nach einem Aufenthalt in Hamburg am deutschen Teilchenbeschleuniger DESY setzte sie ihr PhD-Studium in Kopenhagen fort, wo auch ihr zukünftiger Mann beschäftigt war. Zusammen zog es sie dann für gut drei Jahre nach Oxford, wo sie am berühmten Rutherford Lab arbeitete. Sie hätte eigentlich eine Fünfjahresposition gehabt, doch ihr Mann fand in Oxford keine weitere Anstellung, und so landeten sie eben gemeinsam in Zürich. Sie mag die Stadt, doch feste Wurzeln hat sie noch keine geschlagen. Ortsveränderungen gehören für sie zum Forscherberuf dazu: «Indem man in verschiedenen Labors arbeitet, kann man sich einen Überblick über sein Gebiet verschaffen.» Doch mit Kindern ist man nicht mehr so flexibel. Sie erwägt höchstens nochmals einen Ortswechsel, dann will sie die Kinder nicht mehr aus ihrem sozialen Umfeld herausreissen.

Stefania Xella Hansen ist im Gebiet der Teilchenphysik Experimentalphysikerin. Sie entwickelt die riesigen Apparate, die registrieren, was passiert, wenn Elementarteilchen mit Höchstgeschwin-

> «Schwer einzuordnende Resultate sind das Salz in der Suppe der Experimentalphysik.»



```
<article title="Stefania Xella Hansen: Passionierte Teilchenbeschleunigerin"
    potential_errors="odd_dropcaps" id="a9">
  <div>Stefania Xella Hansen: Passionierte Teilchenbeschleunigerin</div>
  <div>Italienerin, Teilchenphysikerin, zweifache Mutter, Jungforscherin: Stefania
      Xella Hansen vereint vermeintliche Gegensätze ganz selbstverständlich. Für
      die Diplomarbeit hatte es sie einst ans Cern verschlagen. Nun ist sie über
      einige Umwege wieder in der Schweiz gelandet.</div>
  <div>lichsten Gesetze des Universums aufzuklären, das ist die Aufgabe der
      Teilchenphysik. Wie erklärt sie ihrem kleinen Sohn, womit sie sich beschä
      ftigt im Büro? Gar nicht, wie sollte sie es ihm auch erklären, meint sie, er
       sei ja erst vierjährig, und ohnehin, ihn interessiere höchstens, warum sie
      denn nicht den ganzen Tag zu Hause bleibe, um mit ihm zu spielen. [...].
      Insofern böte die wissenschaftliche Arbeit ja die nötige Flexibilität für
      unkonventionelle Lösungen des Familie- Karriere-Problems. Doch Festanstellun
      -</div>
  <div>Ihr Name lässt einen stutzen. Xella ist, vermutet sie, griechischen
      Ursprungs, und Hansen ist der Name ihres dänischen Mannes. Aus Bologna kommt
       sie, aber wenn man ihr so zuhört, wie sie ruhig und jeden Satz abwägend erz
      ählt, da würde man ihr die Italienerin nicht geben. [...].</div>
```

Figure 6: An example of a problematic article layout where TET's extraction tool falsely identifies the middle column as the first paragraph in the article. In the content extracted XML, it is clear that this has a negative effect on the integrity of the article content, as the paragraphs are out of order and paragraph merging fails to join the split word 'Festanstellungen'. Here, the `potential_errors` flag has successfully been set using the out-of-place dropcap rule. Source: 'Porträt' article from `horizonte_2005_66_de.pdf` (p.17).