Generating Acrostics via Paraphrasing and Heuristic Search

Bruno Soares Fillmann
Fernando Bombardelli da Silva
Jürgen Bauer
William Bombardelli da Silva

Technische Universität Berlin
Datenbanksysteme und Informationsmanagement
DBPRO – Database Projects (WS 2014/2015)

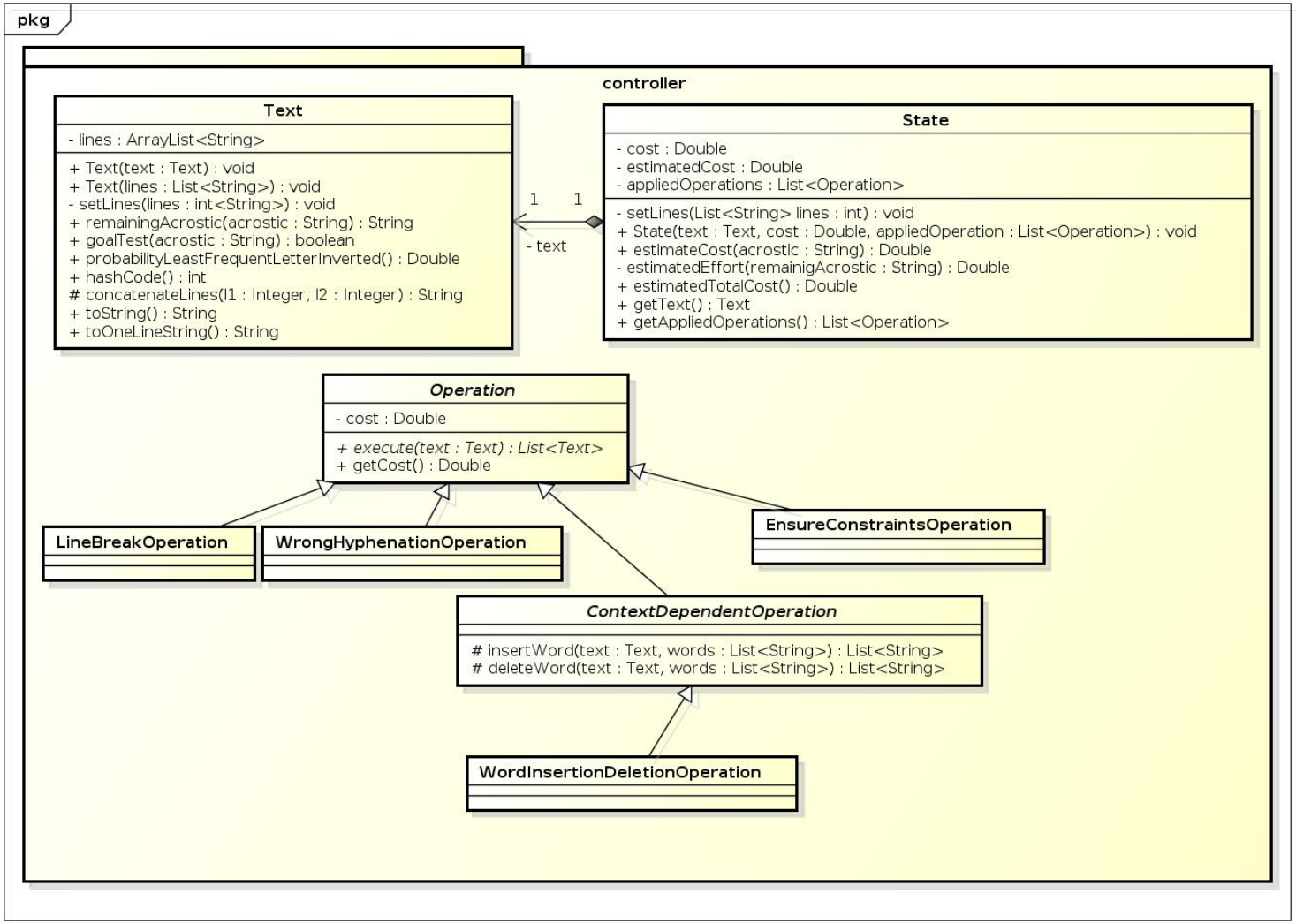
01.12.2014

Organisation

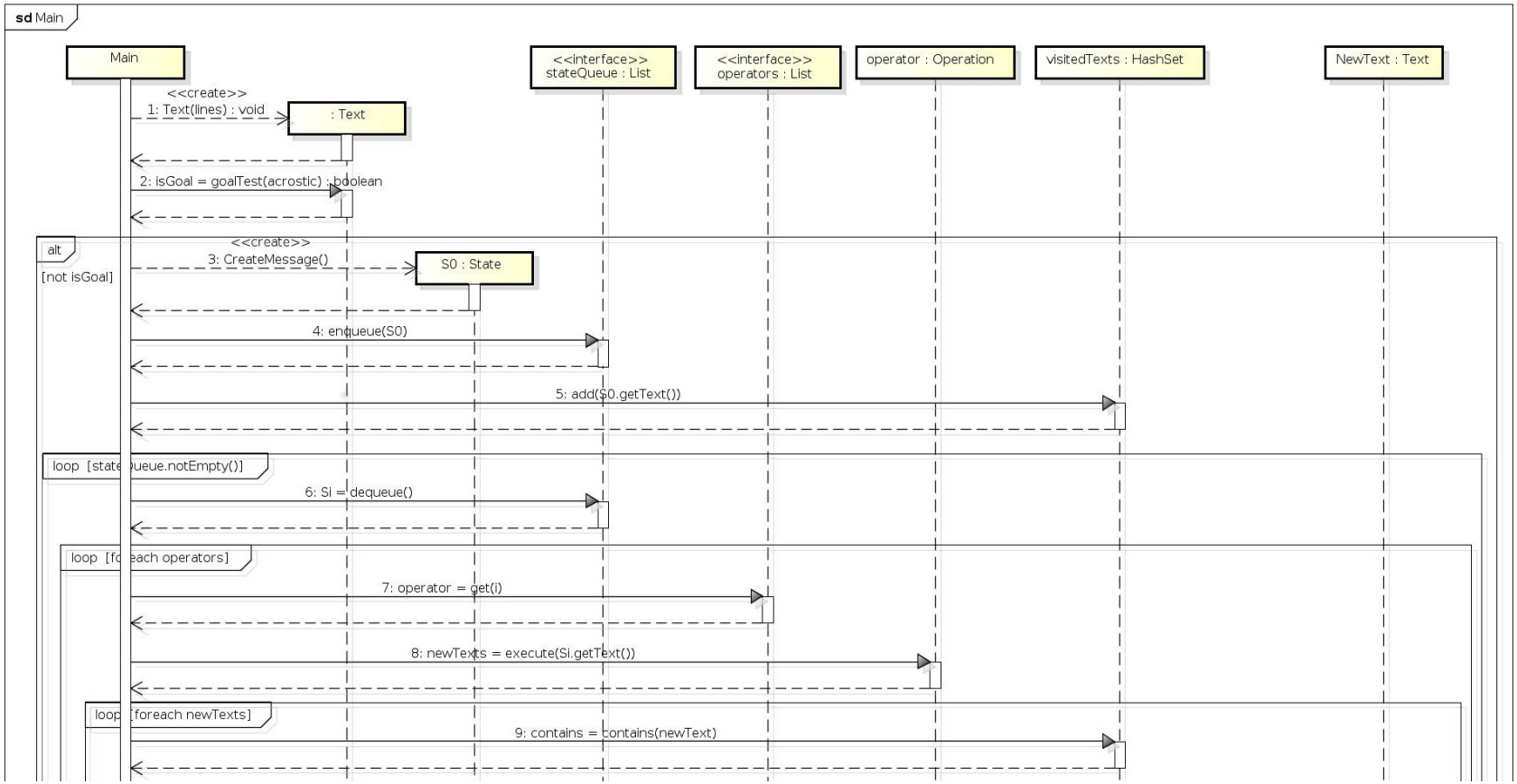
Overview

- **The problem:** Given a text T and an acrostic X, find a paraphased version of T that encodes X in the first letters of each line.
- **Main goal of the project:** Implement the paper for the German language.
- **Main idea of the algorithm:**
  - Modeled as a search problem in a tree.
  - The vertices are states (texts) and the edges are operations over states.
  - Artificial inteligence is applied for the search strategy (A* Algorithm).

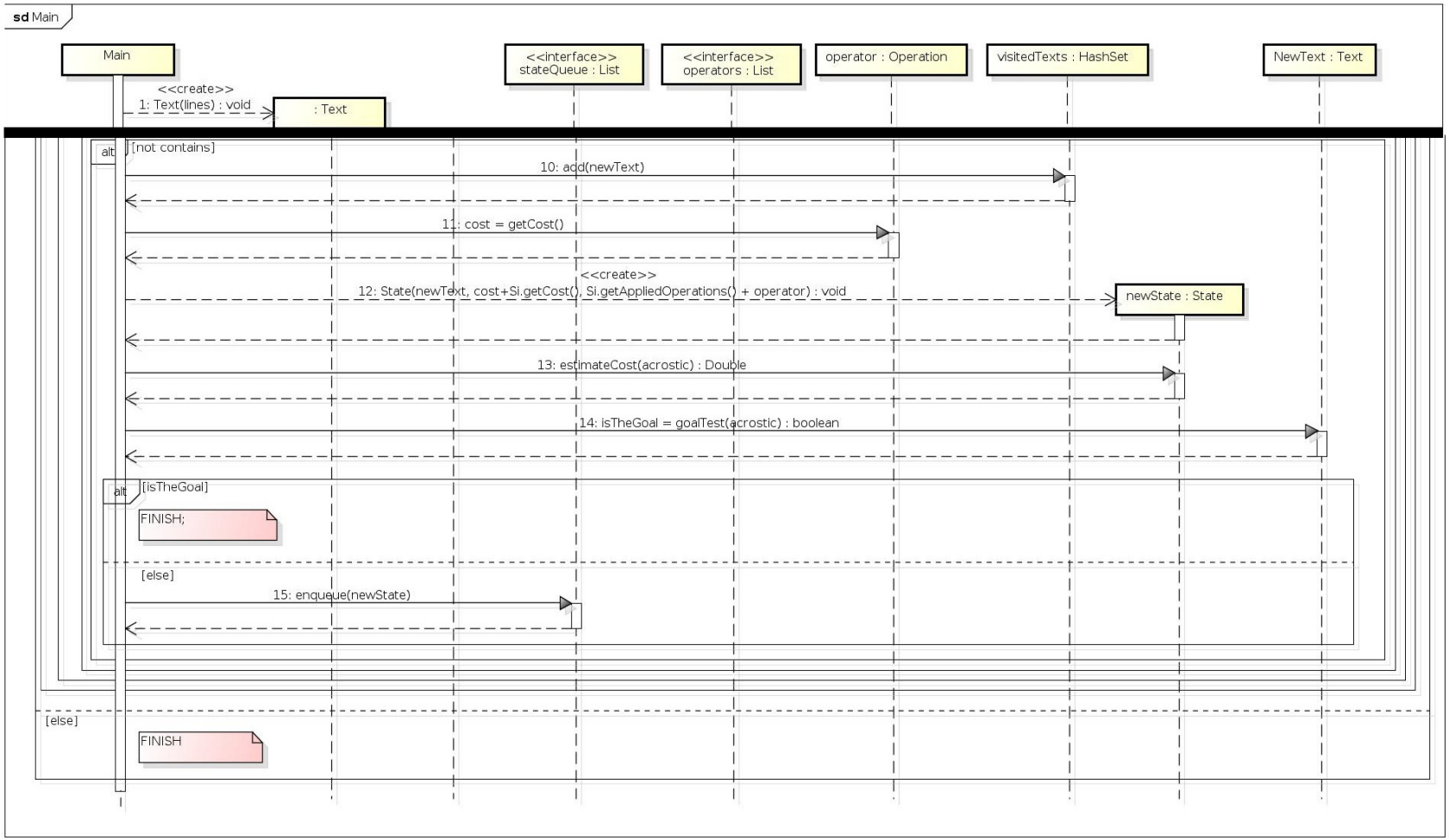
Classes Diagram



Activity Diagram and Search Strategy



Activity Diagram and Search Strategy



Line Break

- Constraints on line length $l_{min} = 50$ and $l_{max} = 70$.
- Two cases:
  - After a word when line length falls in the $[l_{min}, l_{max}]$-window.
  - After a full stop. (i.e., paragraph break)
- Succeeding line break, lines have to be aligned.
- A greedy word wrap algorithm is applied.
- Avoid words of length > 20 in the start text.

Line Break

- **Example:**
- 18 linebreaks,
- No full stop after 28!
- Third often applied operator on a solution path
- Fastest operator

Johann Wolfgang von Goethe wurde am 28. August 1749 im heutigen Goethe-Haus am Frankfurter Großen Hirschgraben geboren. Der Vater Johann Caspar Goethe (1710–1782), der die ursprüngliche Schreibung des Familiennamens von Göthe in Goethe änderte, war promovierter Jurist, übte diesen Beruf jedoch nicht aus, sondern lebte von den Erträgen seines ererbten Vermögens, das später auch dem Sohn ein Leben ohne finanzielle Zwänge ermöglichen sollte. Er war vielseitig interessiert und gebildet, jedoch auch streng und pedantisch, was wiederholt zu Konflikten in der Familie führte.

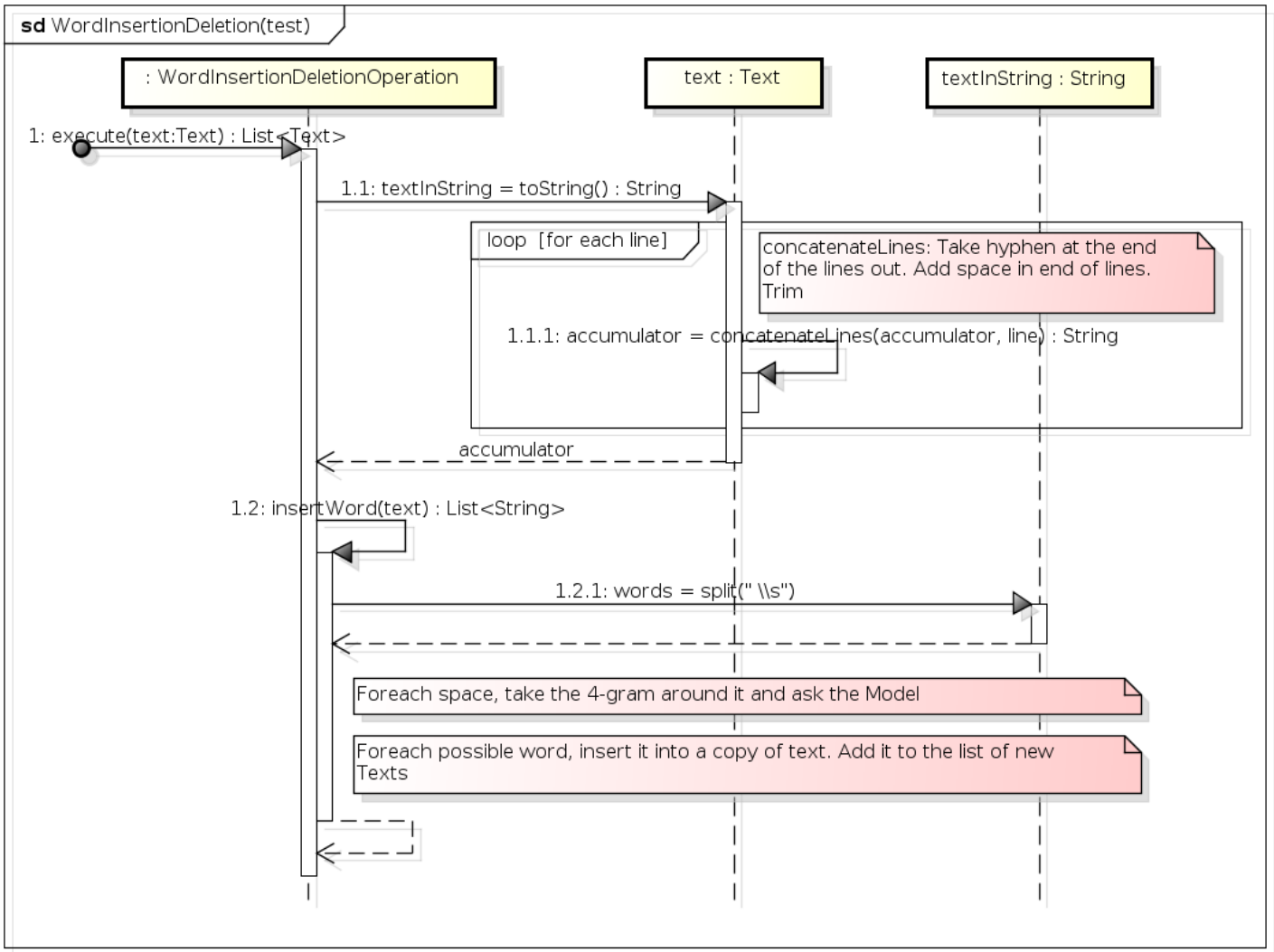
Wrong Hyphenation

- Returns a list of all possible variations of a given text by hyphenating the last word.
- Previous words are hyphenated through the execution of a Line Break operation.
- For every line, find the last word and split it between the current and next line n-1 times, where n is the length of the word.
- Each division adds a new text to the list it returns, so there's a maximun of one new hyphenation per text in the list.
- This way every possible combination is added to the list.
- If a world is already hyphenated in a given line nothing is done at that line.

Ensure Constraints

- Our other operations are not concerned with keeping the text correct
- This operation ensures that all constraints are met
- In order to do this the algorithm goes from top to bottom
- Shifting characters between the line being analyzed and the next line

Word Insertion



Word Insertion – Databases

- Current: Netspeak Requests *(netspeak.org)* [2]
- Possibility: Google NGram Database Manually Handled

First Results – Donald Knuth

Knuth ist der Sohn eines Lehrers. Er besuchte die Milwaukee Lutheran High School und begann sein Physikstudium am California Institute of Technology im September 1956. Aus zweierlei Gründen schlug er tatsächlich seinem zweiten Studienjahr jedoch den Weg zur Mathematik ein: Zum einen löste er ein Problem eines seiner Mathematikprofessoren, was ihm eine 1,0 als Note einbrachte, zum anderen fand er wenig Gefallen an den physikalischen Praktika. Er erhielt einen Bachelor- und einen Master-Abschluss 1960 an der Case Western Reserve University. 1963 erhielt er seinen Ph.D. vom California Institute of Technology bei Marshall Hall, wo er dann auch nach der Promotion Assistant Professor und 1966 Associate Professor und schließlich Professor wurde. 1968 wurde er Professor für Informatik an der Stanford University.

Result Text:

Knuth ist der Sohn eines Lehrers. Er besuchte die Milwaukee Luthera- *//Wrong Hyphenation*
nHigh School und begann sein Physikstudium am California Instit- *//Line Break +*
ute of Technology im September 1956. Aus zweierlei Gründen schlug er *//Wrong Hyphenation*
tatsächlich seinem zweiten Studienjahr jedoch den Weg zur Mat- *//Wrong Hyphenation*
hematik ein: Zum einen löste er ein Problem eines seiner Mathematikprofessoren, was ihm eine 1,0 als Note einbrachte, zum anderen fand er wenig Gefallen an den physikalischen Praktika. Er erhielt einen Bachelor- und einen Master-Abschluss 1960 an der Case Western Reserve University. 1963 erhielt er seinen Ph.D. vom California Institute of Technology bei Marshall Hall, wo er dann auch nach der Promotion Assistant Professor und 1966 Associate Professor und schließlich Professor wurde. 1968 wurde er Professor für Informatik an der Stanford University.

First Results – Goethe

```
Johann Wolfgang von Goethe wurde am 28. August 1749 im heutigen
Goethe-Haus am Frankfurter Großen Hirschgraben geboren. Der Vater
Johann Caspar Goethe (1710{1782), der die ursprüngliche Schreibung des
Familiennamens von Göthe in Goethe änderte, war promovierter Jurist,
übte diesen Beruf jedoch nicht aus, sondern lebte von den Erträgen
seines ererbten Vermögens, das später auch dem Sohn ein Leben ohne
finanzielle Zwänge ermöglichen sollte. Er war vielseitig interessiert
und gebildet, jedoch auch streng und pedantisch, was wiederholt zu
Konflikten in der Familie führte.
```

Result Text:

```
Johann Wolfgang von Goethe wurde am 28. August 1749 im he-      //Wrong Hyphenation
utigen Goethe-Haus am Frankfurter Großen Hirschgraben gebore-   //LineBreak + Wrong Hyphenation
n. Der Vater Johann Caspar Goethe (1710{1782), der die ursprün- //Wrong Hyphenation
gliche Schreibung des Familiennamens von Göthe in Goethe änderte, war
promovierter Jurist, übte diesen Beruf jedoch nicht aus, sondern lebte
von den Erträgen seines ererbten Vermögens, das später auch dem Sohn
ein Leben ohne finanzielle Zwänge ermöglichen sollte. Er war
vielseitig interessiert und gebildet, jedoch auch streng und
pedantisch, was wiederholt zu Konflikten in der Familie führte.
```

First Results – Berlin

```
Berlin ist die Hauptstadt und zugleich ein Land der Bundesrepublik
Deutschland.  Der Stadtstaat Berlin bildet eine Einheitsgemeinde und
ist mit über 3,4 Millionen Einwohnern die bevölkerungsreichste und mit
knapp 892 Quadratkilometern die flächengrößte Kommune Deutschlands
sowie nach Einwohnern die zweitgrößte der Europäischen Union.  Zudem
ist Berlin mit rund 3800 Einwohnern je Quadratkilometer die am
drittdichtesten bevölkerte Gemeinde Deutschlands.  Berlin ist eine
Enklave im Land Brandenburg und bildet das Zentrum der Metropolregion
Berlin/Brandenburg (6 Millionen Einwohner) sowie der Agglomeration
Berlin (4,4 Millionen Einwohner).  Der Stadtstaat unterteilt sich in
zwölf Bezirke.
```

Result Text:

```
Berlin ist die Hauptstadt und zugleich ein Land der Bundesrepubl-      //Wrong Hyphenation
ik Deutschland.  Der Stadtstaat Berlin bildet eine                     //Line Break + Line Break
Einheitsgemeinde und ist mit über 3,4 Millionen Einwohne-              //Wrong Hyphenation
rn die bevölkerungsreichste und mit knapp 892 Quadratkilometern die
flächengrößte Kommune Deutschlands sowie nach Einwohnern die
zweitgrößte der Europäischen Union.  Zudem ist Berlin mit rund 3800
Einwohnern je Quadratkilometer die am drittdichtesten bevölkerte
Gemeinde Deutschlands.  Berlin ist eine Enklave im Land Brandenburg und
bildet das Zentrum der Metropolregion Berlin/Brandenburg (6 Millionen
Einwohner) sowie der Agglomeration Berlin (4,4 Millionen Einwohner).
Der Stadtstaat unterteilt sich in zwölf Bezirke.
```

Next Steps

- Optimization of the Algortithm
- Word insertion and deletion
- Implementing spelling operation
- Synonyms
- Hyphenation

References

Benno Stein, Matthias Hagen, and Christof Bräutigam. *Generating Acrostics via Paraphrasing and Heuristic Search*.
In Junichi Tsujii and Jan Hajic, editors, 25th International Conference on Computational Linguistics (COLING 14), pages 2018-2029, August 2014. Association for Computational Linguistics.

Martin Potthast, Martin Trenkmann, and Benno Stein. *Netspeak: Assisting Writers in Choosing Words*.
In Cathal Gurrin et al, editors, Advances in Information Retrieval. 32nd European Conference on Information Retrieval (ECIR 10) volume 5993 of Lecture Notes in Computer Science, pages 672, Berlin Heidelberg New York, March 2010. Springer.

Questions?