Efficient Wikipedia Entity Linking and Co-Reference Resolution in C++

STUDENT

Benjamin Dietrich
4907020

BETREUERIN

Natalie Prange

Mittwoch, den 23.11.2022

Fakultät für Informatik der Universität Freiburg
Lehrstuhls für Algorithmen und Datenstrukturen

I Definition: 'Named Entity' und 'Co-Reference'

Eigennamen	Nomen	Pronomen	Entitätstyp
Marie Curie	scientist	she	the human
Snickers	chocolate bar	it	the dessert
Titanic	ship	it	the ocean liner
Named Entity	~~Named Entity~~	Co-Reference	

Voyager 1

Voyager 1 is a space probe launched by NASA on September 5, 1977, as part of the Voyager program to study the outer Solar System and interstellar space beyond the Sun's heliosphere. Launched 16 days after its twin Voyager 2, Voyager 1 has been operating for 45 years, 2 months and 9 days as of November 14, 2022 UTC. It communicates through NASA's Deep Space Network to receive routine commands and to transmit data to Earth. Real-time distance and velocity data is provided by NASA and JPL. At a distance of 158.79 AU (14.760 billion mi) from Earth as of November 7, 2022, it is the most distant man-made object from Earth. The probe made flybys of Jupiter, Saturn, and Saturn's largest moon, Titan. NASA had a choice of either doing a Pluto or Titan flyby.

Legende: Synonym Hyperlink Ignorierte Hyperlink Entity Linking Co-Refererence Resolution

Voyager 1

Voyager 1 is a space probe launched by NASA on September 5, 1977, as part of the Voyager program to study the outer Solar System and interstellar space beyond the Sun's heliosphere. Launched 16 days after its twin Voyager 2, Voyager 1 has been operating for 45 years, 2 months and 9 days as of November 14, 2022 UTC. It communicates through NASA's Deep Space Network to receive routine commands and to transmit data to Earth. Real-time distance and velocity data is provided by NASA and JPL. At a distance of 158.79 AU (14.760 billion mi) from Earth as of November 7, 2022, it is the most distant man-made object from Earth. The probe made flybys of Jupiter, Saturn, and Saturn's largest moon, Titan. NASA had a choice of eith r doing a Pluto or Titan flyby.

Legende: Synonym Hyperlink Ignorierte Hyperlink Entity Linking Co-Refererence Resolution

Voyager 1

Voyager 1 is a space probe launched by NASA on September 5, 1977, as part of the Voyager program to study the outer Solar System and interstellar space beyond the Sun's heliosphere. Launched 16 days after its twin Voyager 2, Voyager 1 has been operating for 45 years, 2 months and 9 days as of November 14, 2022 UTC. It communicates through NASA's Deep Space Network to receive routine commands and to transmit data to Earth. Real-time distance and velocity data is provided by NASA and JPL. At a distance of 158.79 AU (14.760 billion mi) from Earth as of November 7, 2022, it is the most distant man-made object from Earth. The probe made flybys of Jupiter, Saturn, and Saturn's largest moon, Titan. NASA had a choice of either doing a Pluto or Titan flyby.

Legende: Synonym Hyperlink Ignorierte Hyperlink Entity Linking Co-Refererence Resolution

Ausgangslage:

Es existiert der **Wiki Entity Linker**, ein Wikipedia Entity Linking und Co-Reference Resolution System mit dem **Fokus auf Qualität**. Er wurde in **Python** entwickelt und hat eine **Laufzeit von 30 Tagen** für einen kompletten **Wikipedia-Dump**.

Ausgangslage:

Es existiert der **Wiki Entity Linker**, ein Wikipedia Entity Linking und Co-Reference Resolution System mit dem **Fokus auf Qualität**. Er wurde in **Python** entwickelt und hat eine **Laufzeit von 30 Tagen** für einen kompletten **Wikipedia-Dump**.

Problemstellung:

- **Neuentwicklung** des Wiki Entity Linkers in **C++ mit dem Fokus auf Effizienz**

Ausgangslage:

Es existiert der **Wiki Entity Linker**, ein Wikipedia Entity Linking und Co-Reference Resolution System mit dem **Fokus auf Qualität**. Er wurde in **Python** entwickelt und hat eine **Laufzeit von 30 Tagen** für einen kompletten **Wikipedia-Dump**.

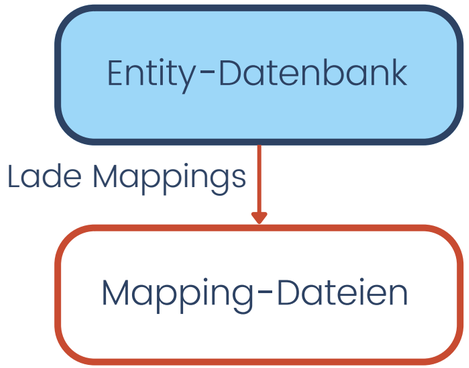
Problemstellung:

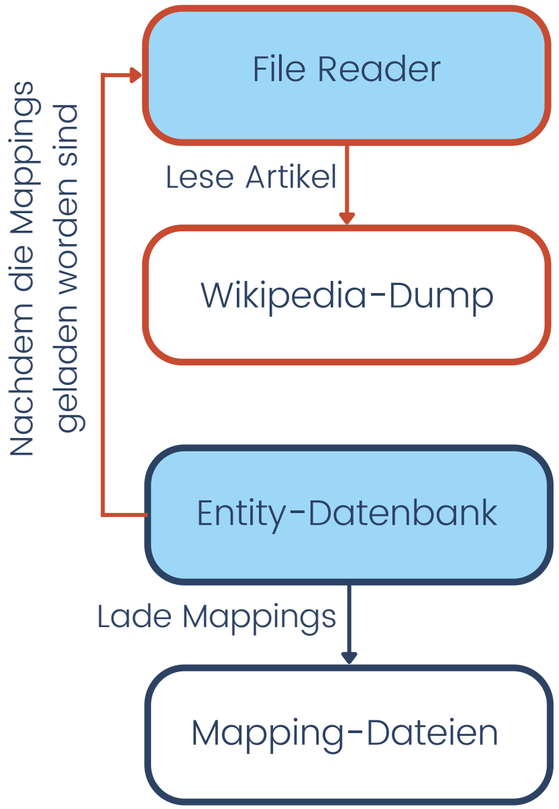
- **Neuentwicklung** des Wiki Entity Linkers in **C++ mit dem Fokus auf Effizienz**

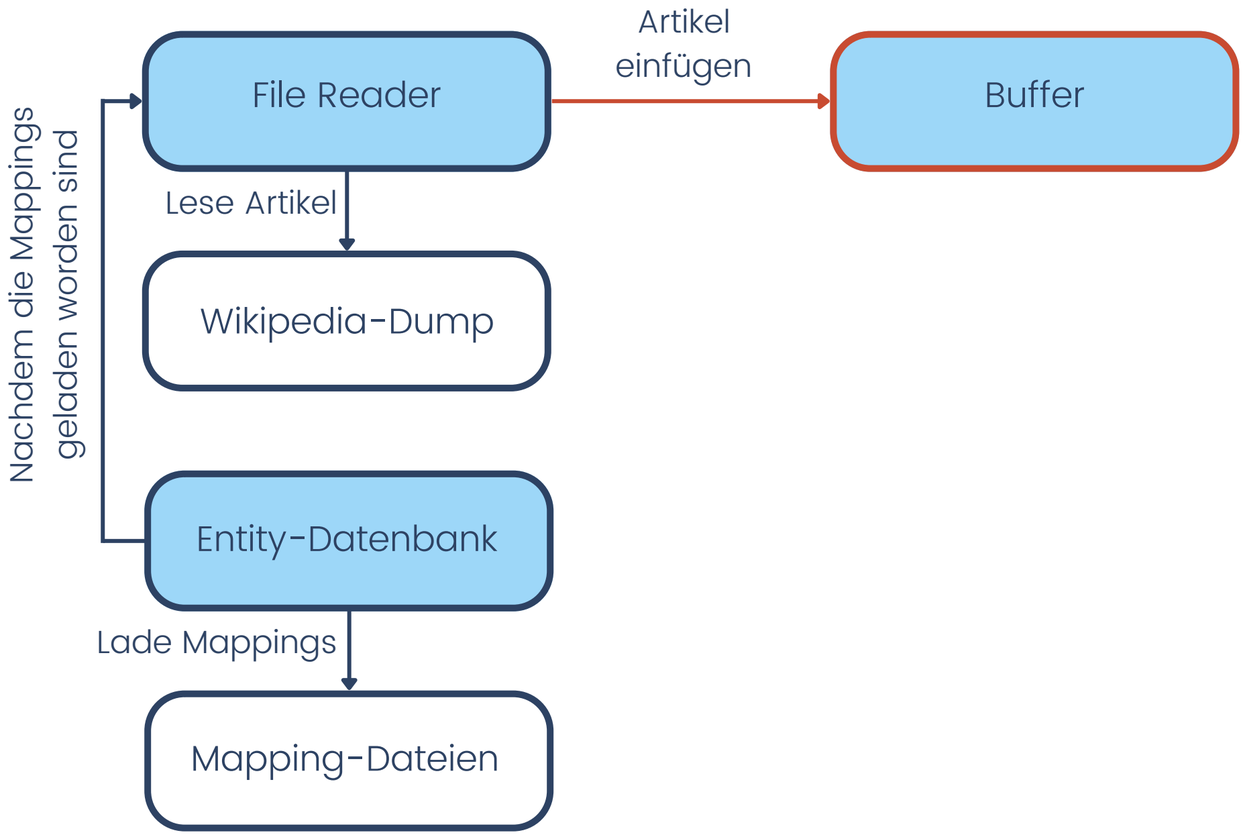
Zielsetzung:

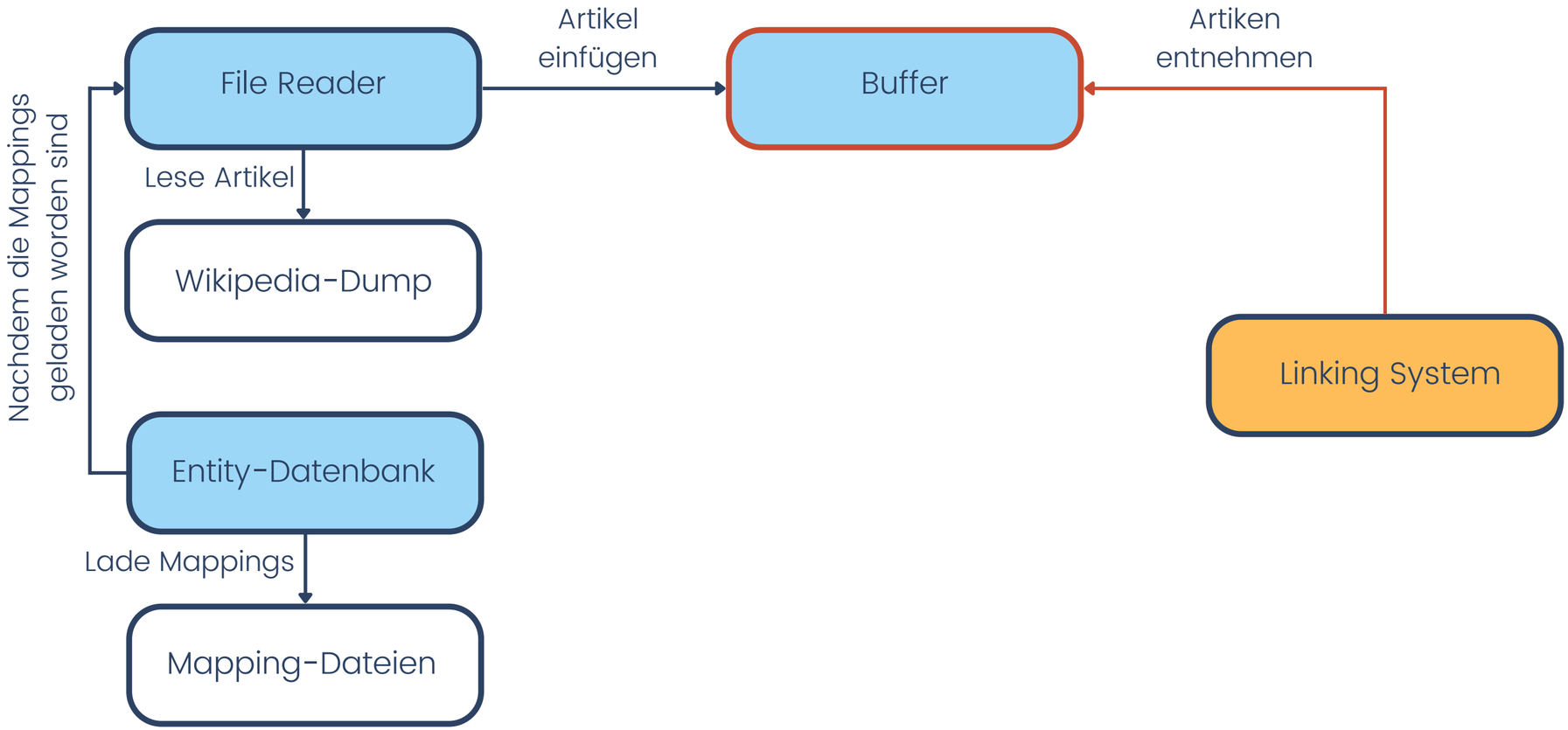
- Eine **Laufzeitverkürzung** von 30 Tagen auf einige Stunden
- Die **Ergebnisqualität** des Wiki Entity Linkers sollte **erhalten** bleiben

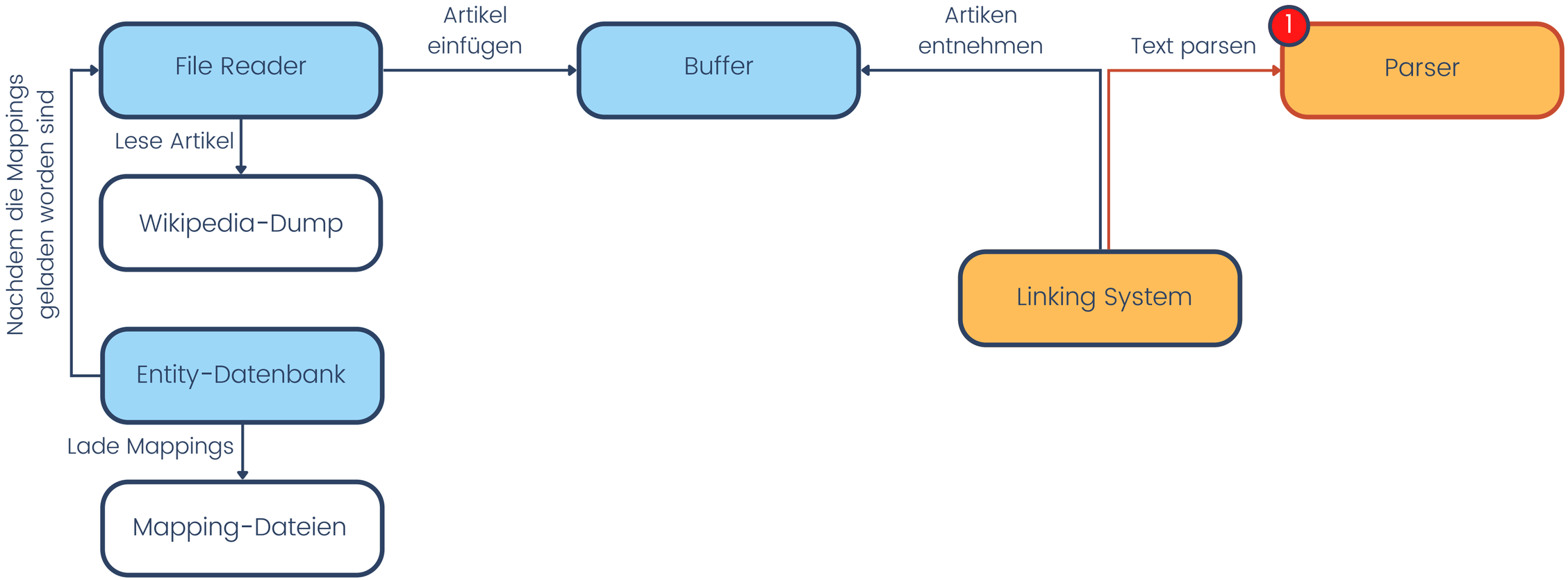
Fragen oder Unklarheiten?

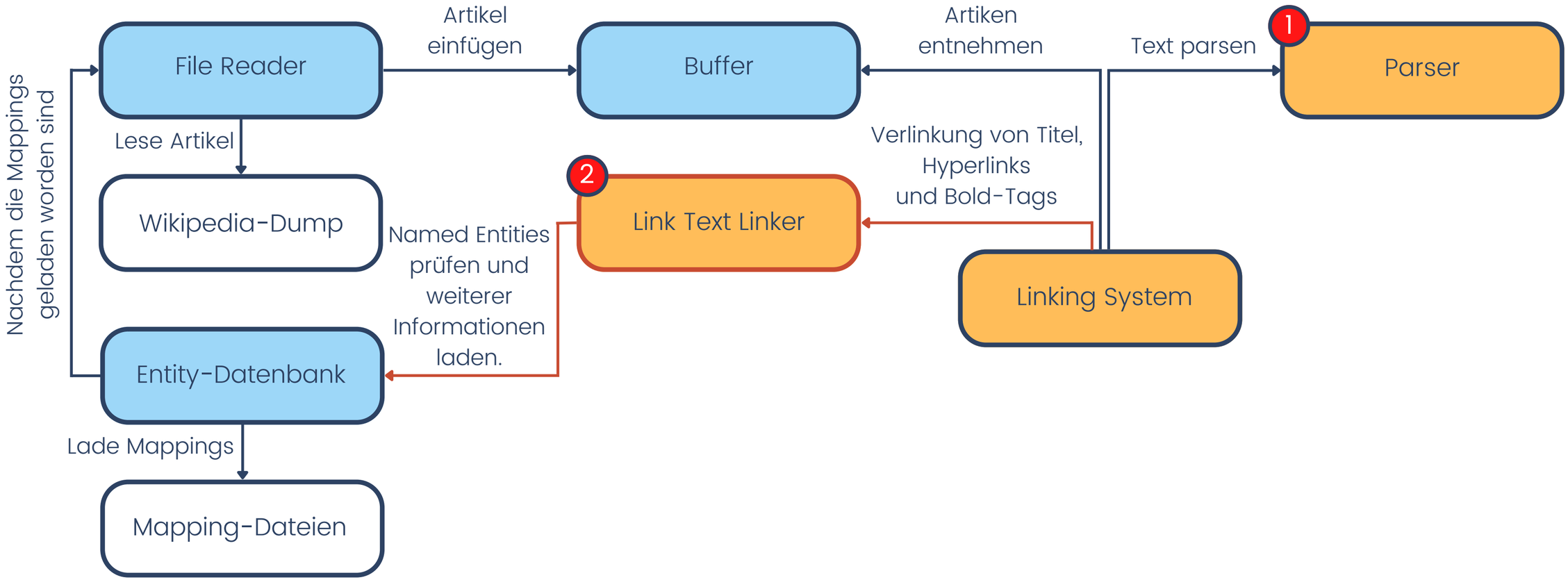

Entity-Datenbank
Lade Mappings
Mapping-Dateien

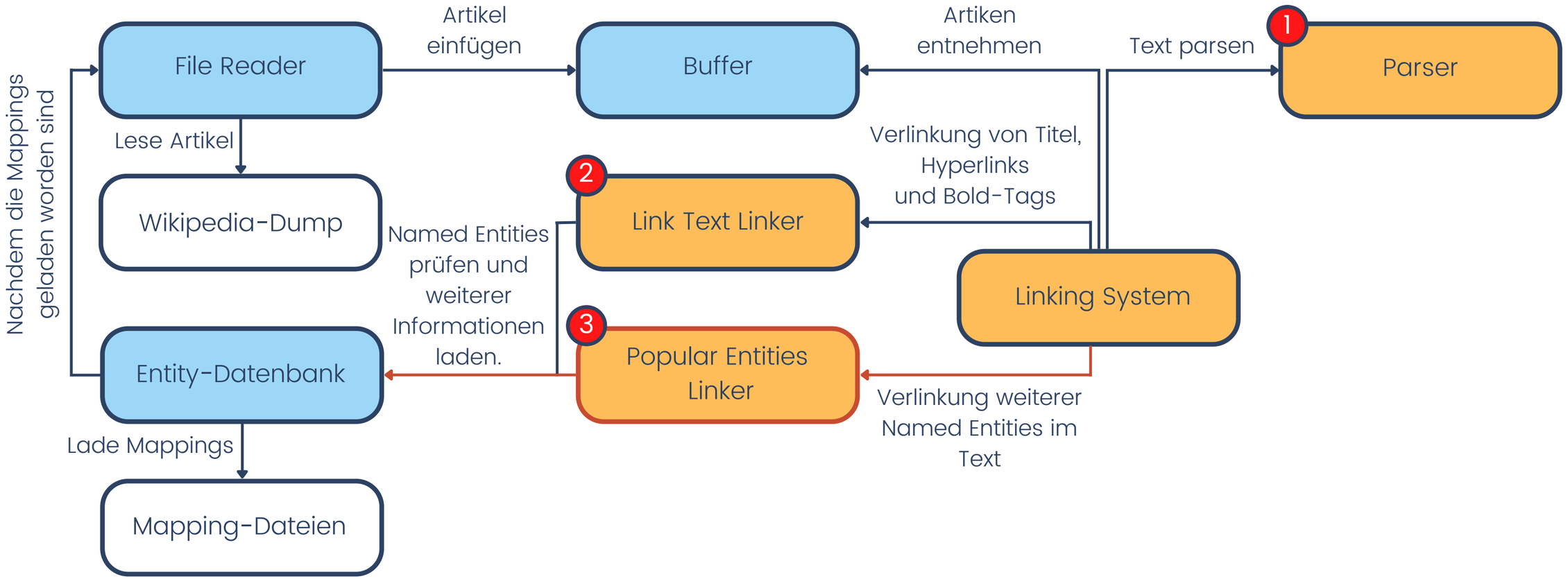

File Reader
Lese Artikel
Wikipedia-Dump
Nachdem die Mappings geladen worden sind
Entity-Datenbank
Lade Mappings
Mapping-Dateien

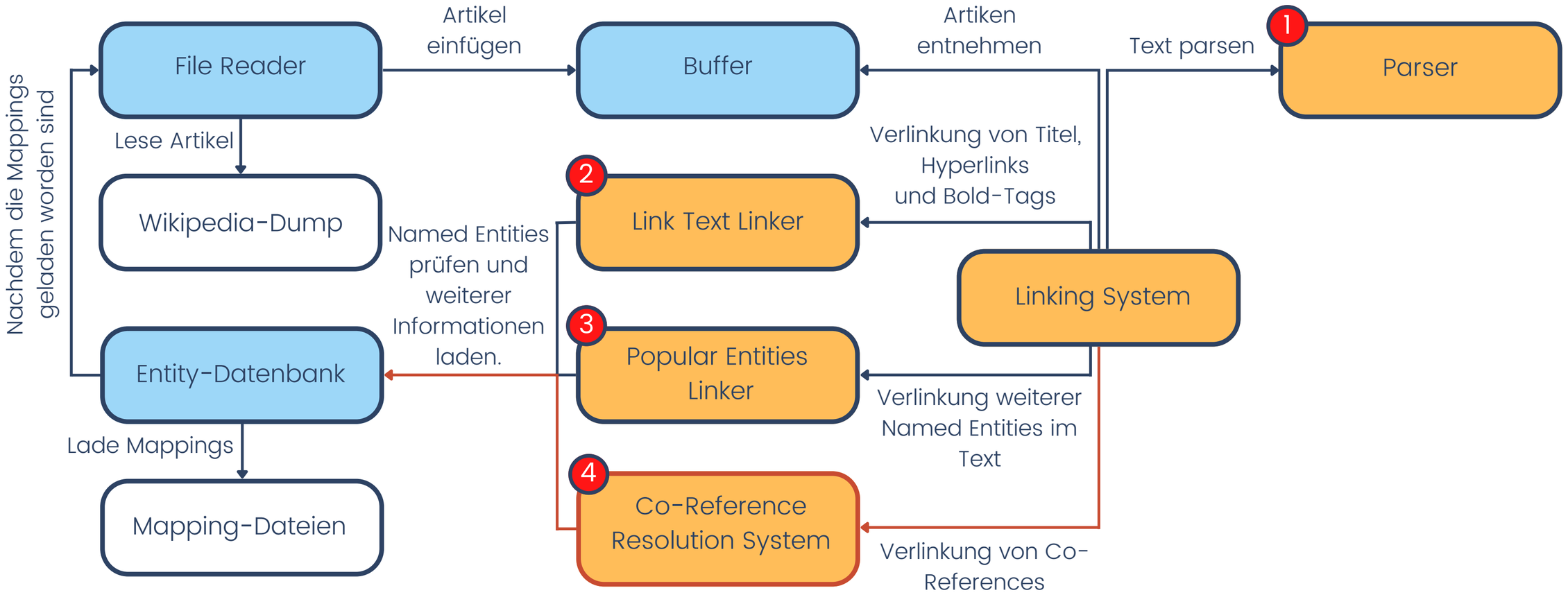

File Reader
Artikel einfügen
Buffer
Lese Artikel
Wikipedia-Dump
Nachdem die Mappings geladen worden sind
Entity-Datenbank
Lade Mappings
Mapping-Dateien

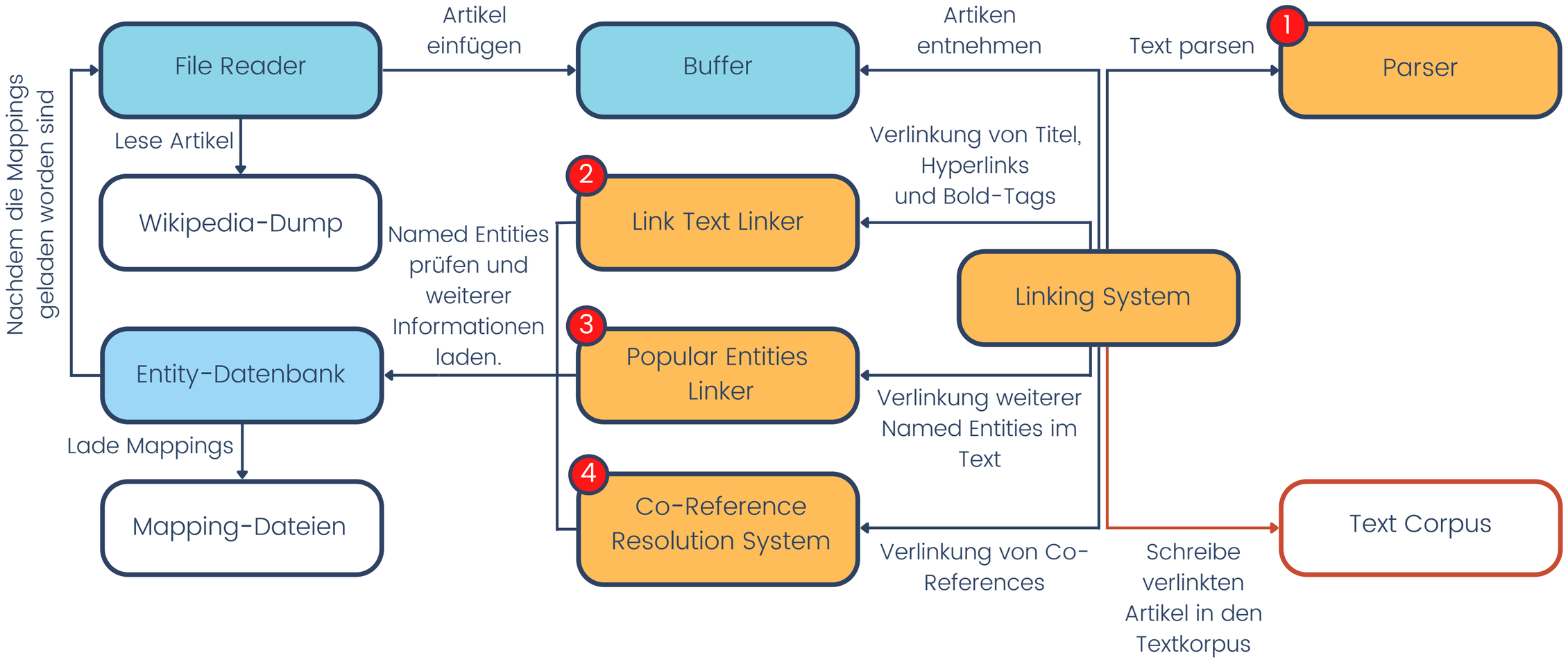

File Reader
Artikel einfügen
Buffer
Artiken entnehmen
Lese Artikel
Wikipedia-Dump
Linking System
Nachdem die Mappings geladen worden sind
Entity-Datenbank
Lade Mappings
Mapping-Dateien

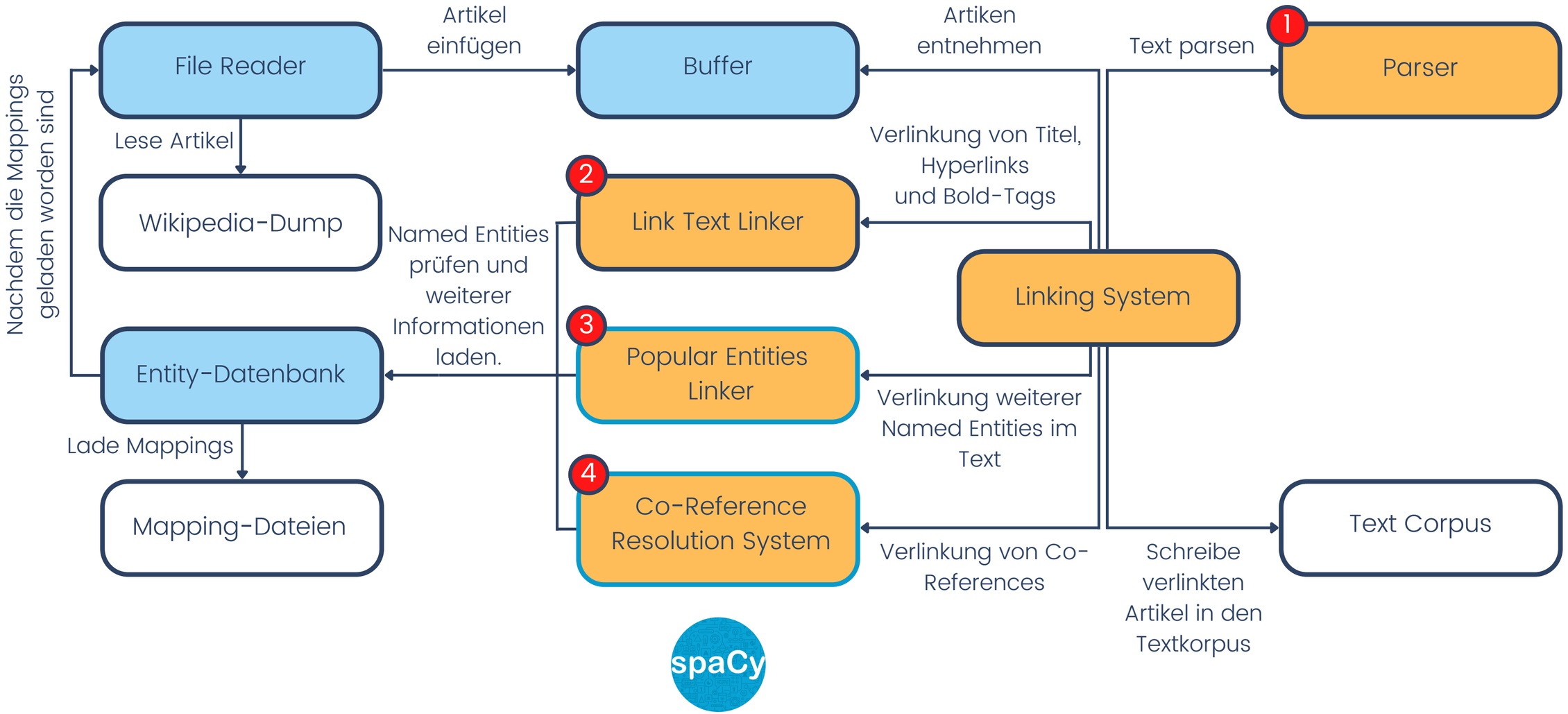

File Reader
Artikel einfügen
Buffer
Artiken entnehmen
Text parsen
1
Parser
Lese Artikel
Wikipedia-Dump
Nachdem die Mappings geladen worden sind
Linking System
Entity-Datenbank
Lade Mappings
Mapping-Dateien


File Reader
Artikel einfügen
Buffer
Artiken entnehmen
Text parsen
1
Parser
Lese Artikel
Wikipedia-Dump
2
Link Text Linker
Verlinkung von Titel, Hyperlinks und Bold-Tags
Named Entities prüfen und weiterer Informationen laden.
Linking System
Nachdem die Mappings geladen worden sind
Entity-Datenbank
Lade Mappings
Mapping-Dateien

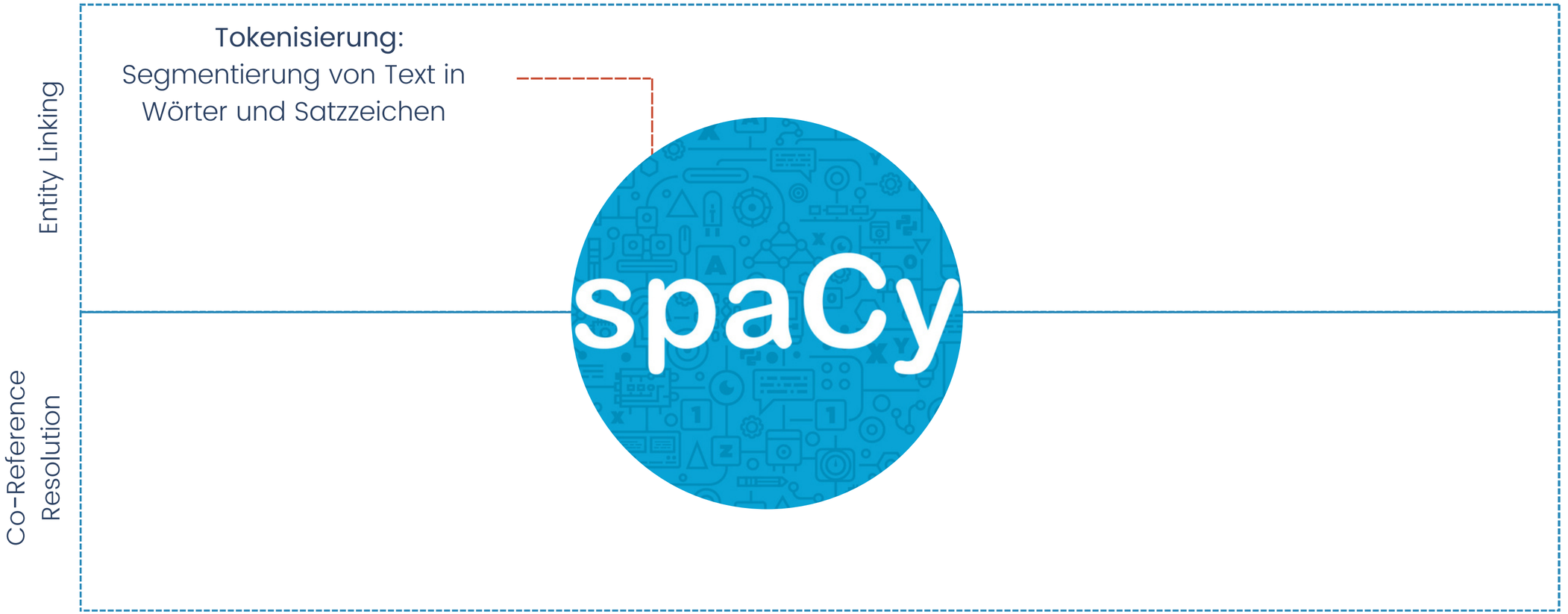

File Reader
Artikel einfügen
Buffer
Artiken entnehmen
Text parsen
1
Parser
Lese Artikel
Wikipedia-Dump
2
Link Text Linker
Verlinkung von Titel, Hyperlinks und Bold-Tags
Named Entities prüfen und weiterer Informationen laden.
Linking System
3
Popular Entities Linker
Nachdem die Mappings geladen worden sind
Entity-Datenbank
Verlinkung weiterer Named Entities im Text
Lade Mappings
Mapping-Dateien

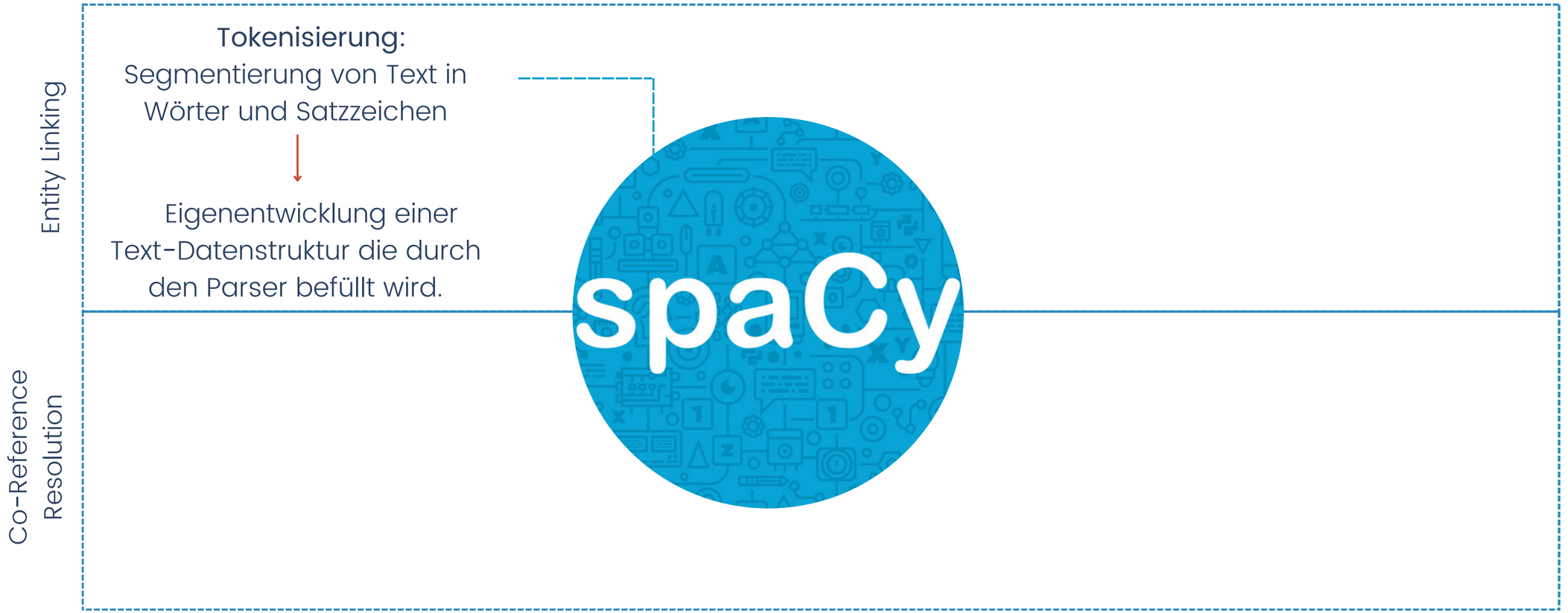

File Reader
Artikel einfügen
Buffer
Artiken entnehmen
Text parsen
1
Parser
Lese Artikel
Wikipedia-Dump
Verlinkung von Titel, Hyperlinks und Bold-Tags
2
Link Text Linker
Named Entities prüfen und weiterer Informationen laden.
Linking System
Nachdem die Mappings geladen worden sind
Entity-Datenbank
3
Popular Entities Linker
Verlinkung weiterer Named Entities im Text
Lade Mappings
Mapping-Dateien
4
Co-Reference Resolution System
Verlinkung von Co-References

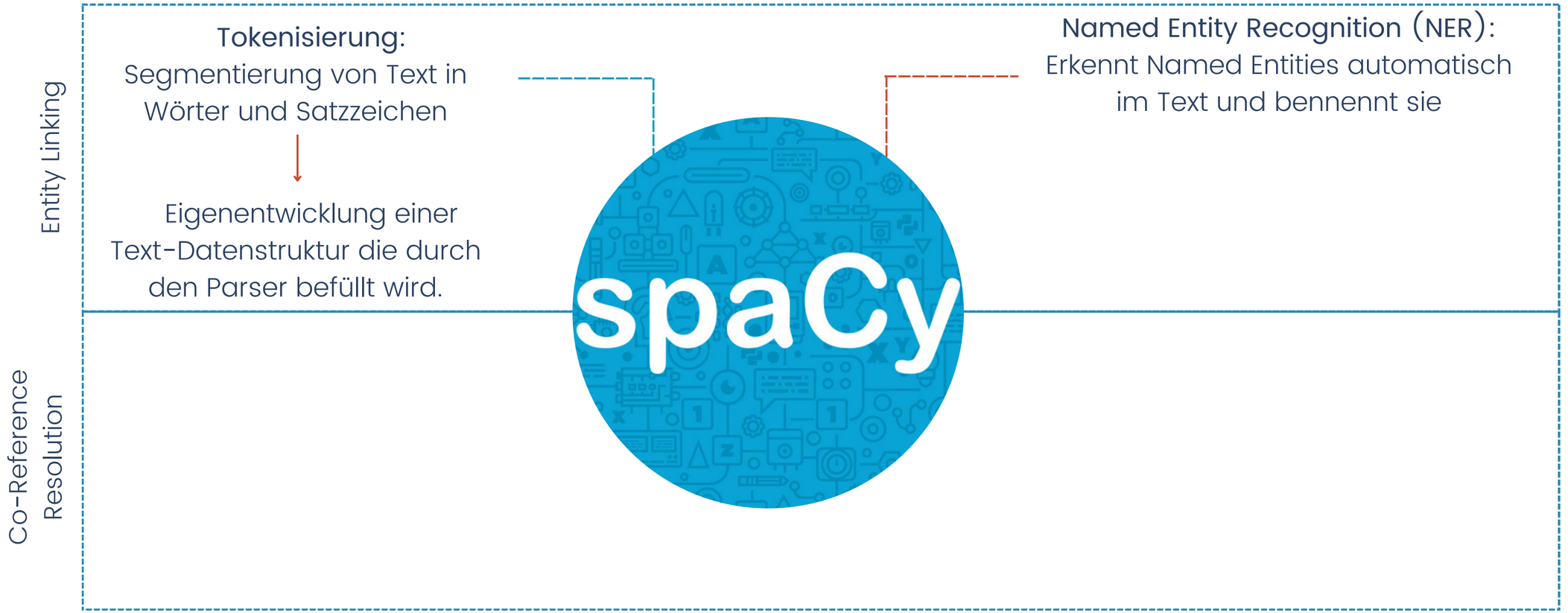

File Reader
Artikel einfügen
Buffer
Artiken entnehmen
Text parsen
1
Parser
Lese Artikel
Wikipedia-Dump
Verlinkung von Titel, Hyperlinks und Bold-Tags
2
Link Text Linker
Named Entities prüfen und weiterer Informationen laden.
Linking System
Nachdem die Mappings geladen worden sind
Entity-Datenbank
3
Popular Entities Linker
Verlinkung weiterer Named Entities im Text
Lade Mappings
Mapping-Dateien
4
Co-Reference Resolution System
Verlinkung von Co-References
Text Corpus
Schreibe verlinkten Artikel in den Textkorpus

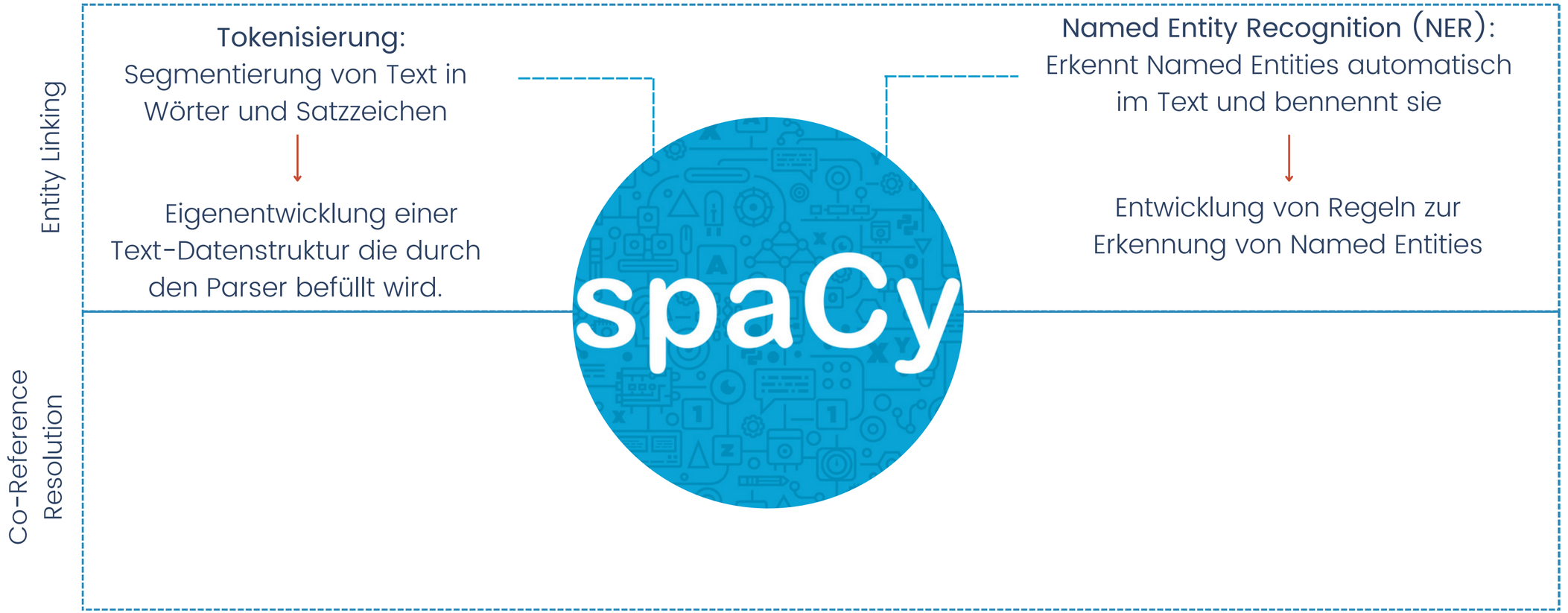

File Reader
Artikel einfügen
Buffer
Artiken entnehmen
Text parsen
1
Parser
Lese Artikel
Wikipedia-Dump
Verlinkung von Titel, Hyperlinks und Bold-Tags
2
Link Text Linker
Named Entities prüfen und weiterer Informationen laden.
Linking System
Nachdem die Mappings geladen worden sind
Entity-Datenbank
3
Popular Entities Linker
Verlinkung weiterer Named Entities im Text
Lade Mappings
Mapping-Dateien
4
Co-Reference Resolution System
Verlinkung von Co-References
Text Corpus
Schreibe verlinkten Artikel in den Textkorpus
spaCy

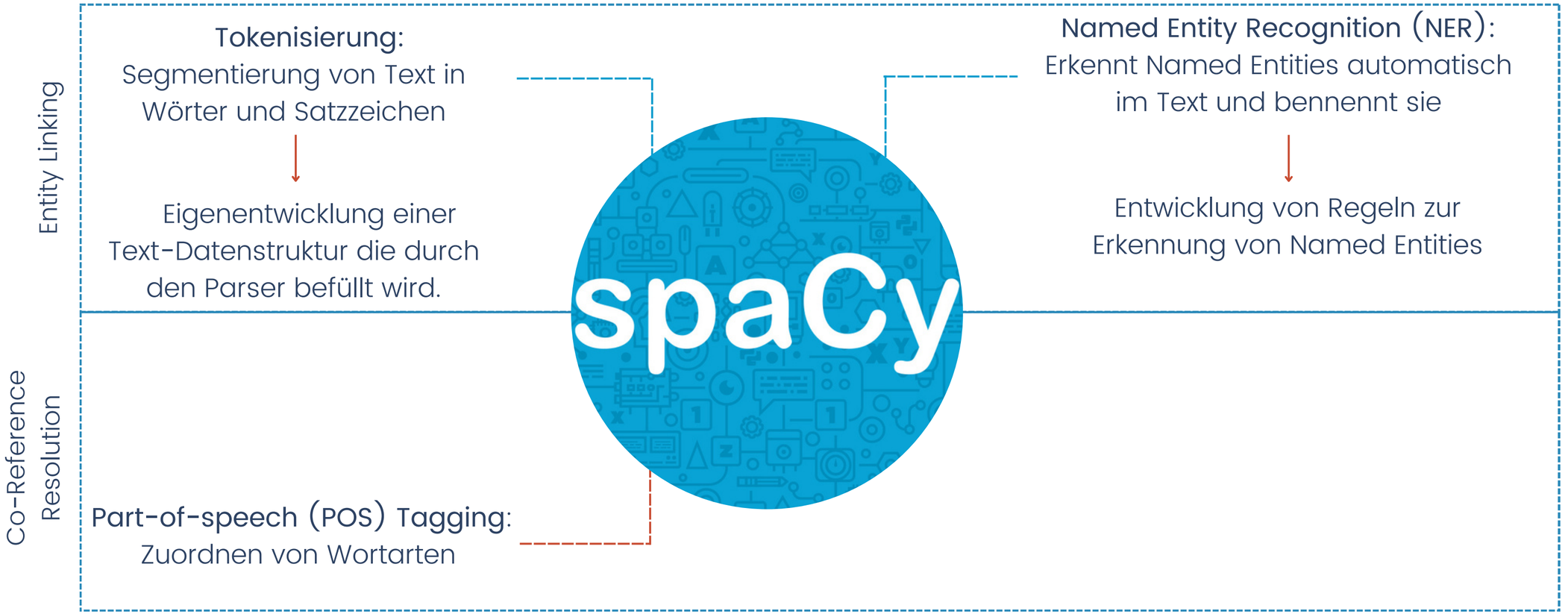
Ersetzung von spaCy

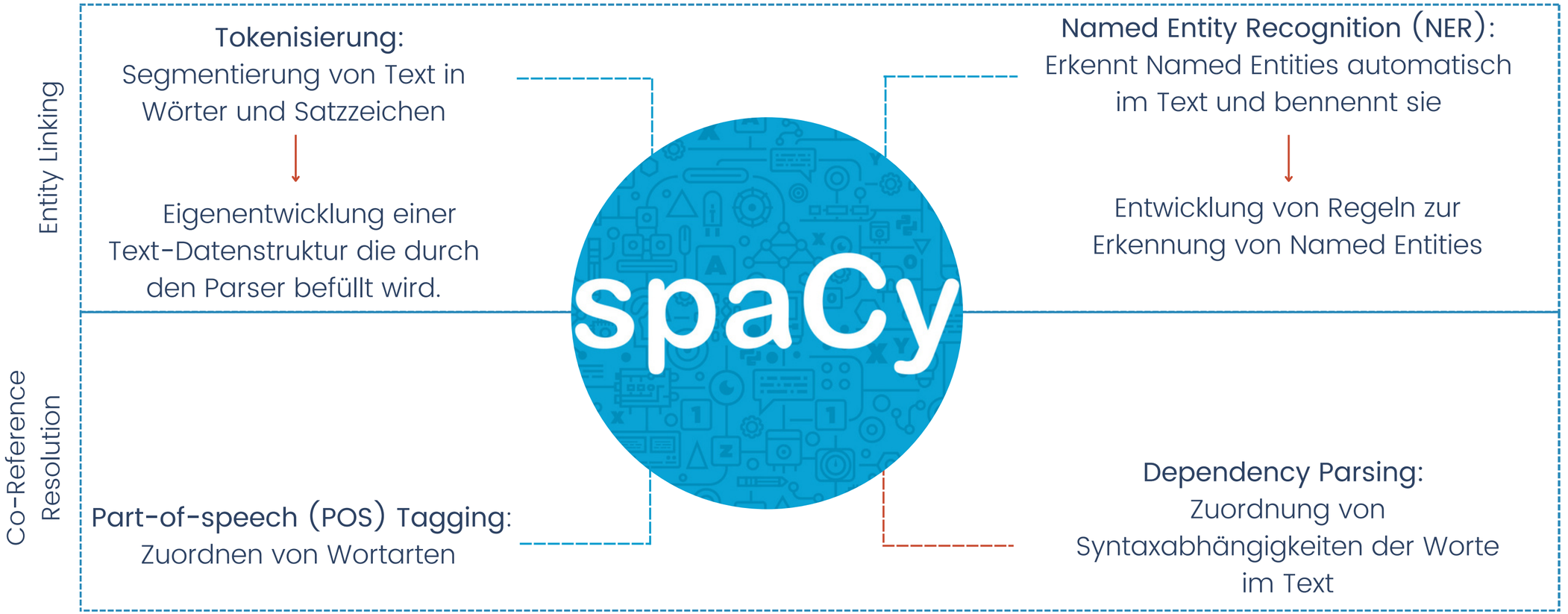

Tokenisierung:
Segmentierung von Text in
Wörter und Satzzeichen
Entity Linking
Co-Reference
Resolution
spaCy

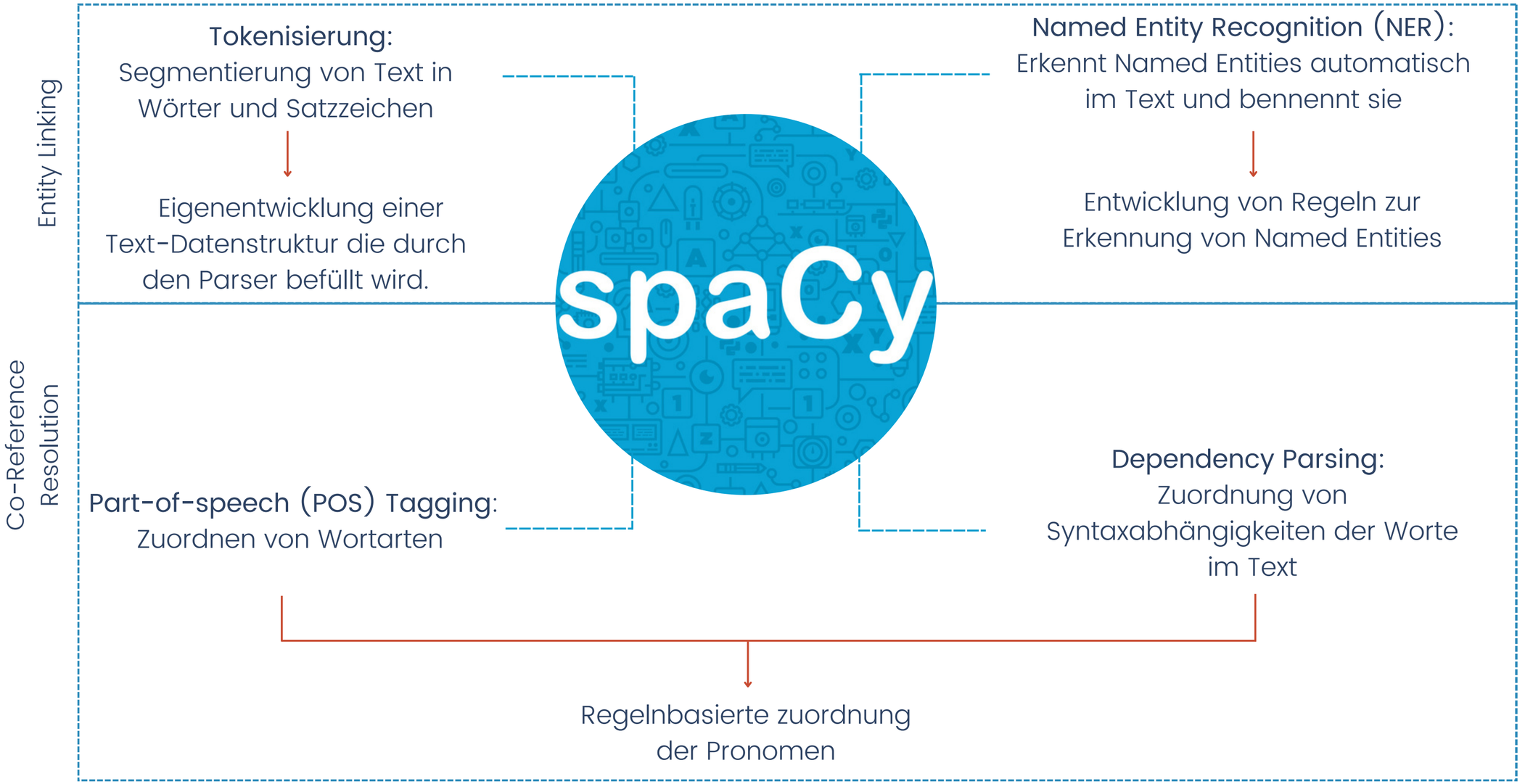

Tokenisierung:
Segmentierung von Text in Wörter und Satzzeichen
Eigenentwicklung einer Text-Datenstruktur die durch den Parser befüllt wird.
spaCy
Entity Linking
Co-Reference Resolution


Tokenisierung:
Segmentierung von Text in Wörter und Satzzeichen
Eigenentwicklung einer Text-Datenstruktur die durch den Parser befüllt wird.
Named Entity Recognition (NER):
Erkennt Named Entities automatisch im Text und bennennt sie
spaCy
Entity Linking
Co-Reference Resolution

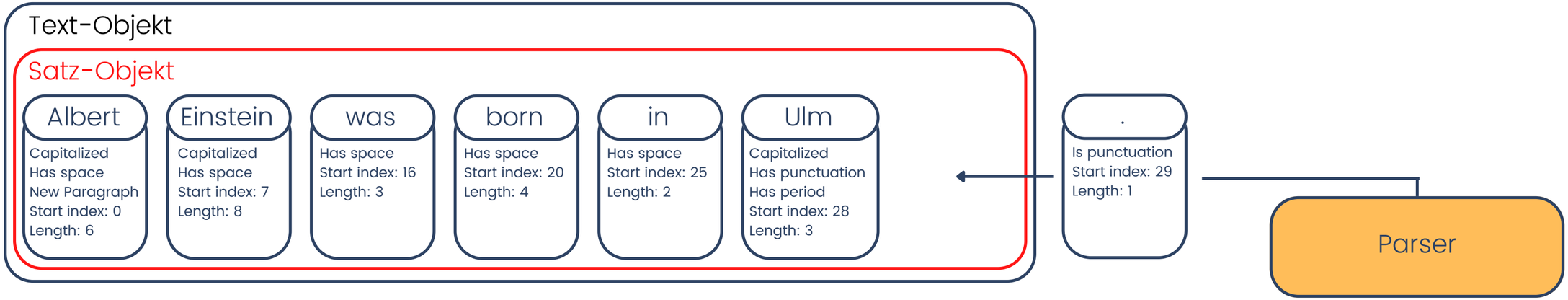

Tokenisierung:
Segmentierung von Text in Wörter und Satzzeichen
Eigenentwicklung einer Text-Datenstruktur die durch den Parser befüllt wird.
Named Entity Recognition (NER):
Erkennt Named Entities automatisch im Text und bennennt sie
Entwicklung von Regeln zur Erkennung von Named Entities
spaCy
Entity Linking
Co-Reference Resolution


Tokenisierung:
Segmentierung von Text in Wörter und Satzzeichen
Eigenentwicklung einer Text-Datenstruktur die durch den Parser befüllt wird.
Named Entity Recognition (NER):
Erkennt Named Entities automatisch im Text und bennennt sie
Entwicklung von Regeln zur Erkennung von Named Entities
spaCy
Entity Linking
Co-Reference Resolution
Part-of-speech (POS) Tagging:
Zuordnen von Wortarten

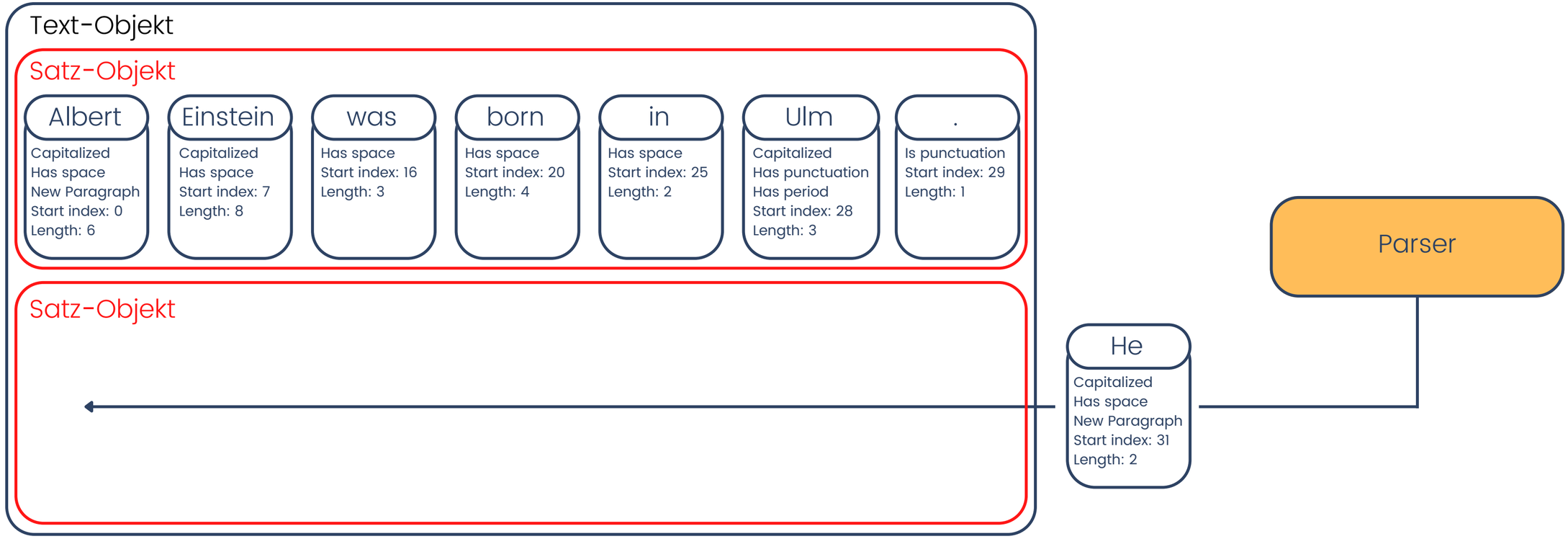

Entity Linking
Co-Reference Resolution
Tokenisierung:
Segmentierung von Text in Wörter und Satzzeichen
Eigenentwicklung einer Text-Datenstruktur die durch den Parser befüllt wird.
Named Entity Recognition (NER):
Erkennt Named Entities automatisch im Text und bennennt sie
Entwicklung von Regeln zur Erkennung von Named Entities
spaCy
Part-of-speech (POS) Tagging:
Zuordnen von Wortarten
Dependency Parsing:
Zuordnung von Syntaxabhängigkeiten der Worte im Text


Entity Linking
Co-Reference Resolution
Tokenisierung:
Segmentierung von Text in Wörter und Satzzeichen
Eigenentwicklung einer Text-Datenstruktur die durch den Parser befüllt wird.
Named Entity Recognition (NER):
Erkennt Named Entities automatisch im Text und bennennt sie
Entwicklung von Regeln zur Erkennung von Named Entities
spaCy
Part-of-speech (POS) Tagging:
Zuordnen von Wortarten
Dependency Parsing:
Zuordnung von Syntaxabhängigkeiten der Worte im Text
Regelnbasierte zuordnung der Pronomen

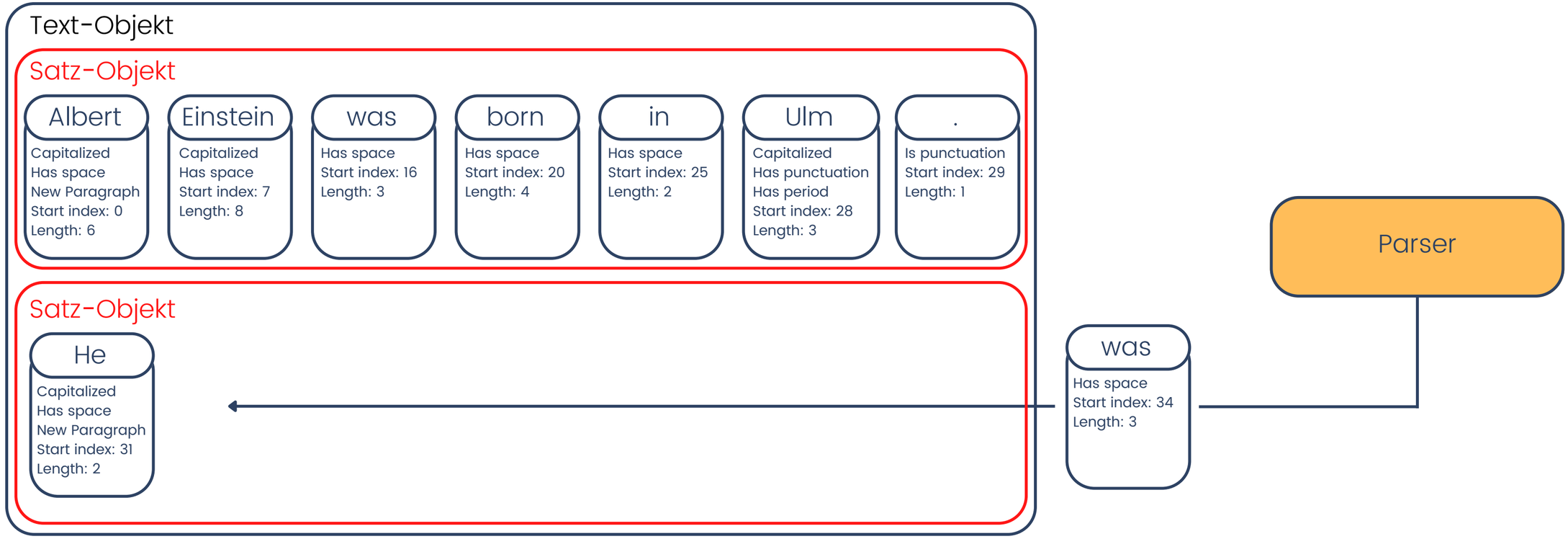
Text-Datenstruktur zur Ersetzung des Tokenizers



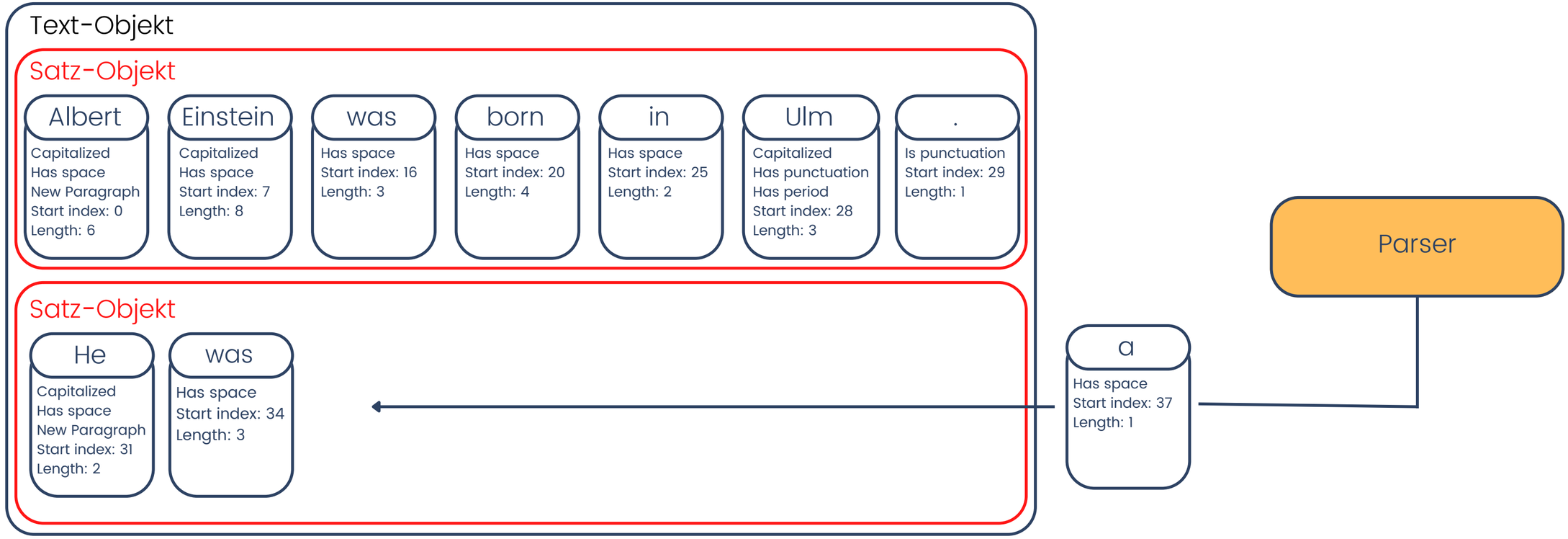
Text-Datenstruktur zur Ersetzung des Tokenizers



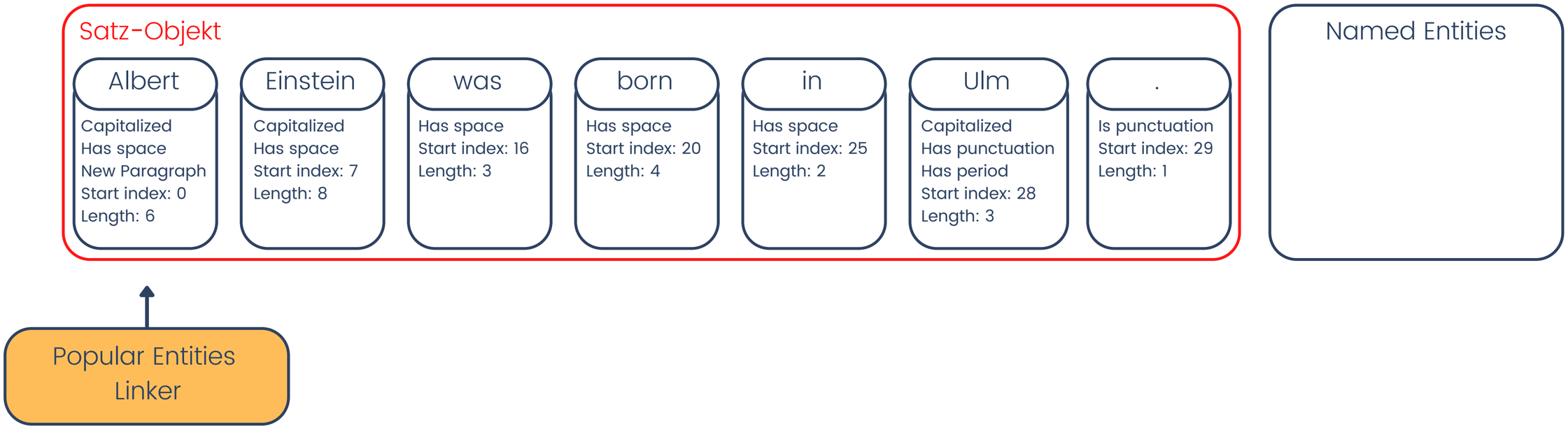
Text-Datenstruktur zur Ersetzung des Tokenizers



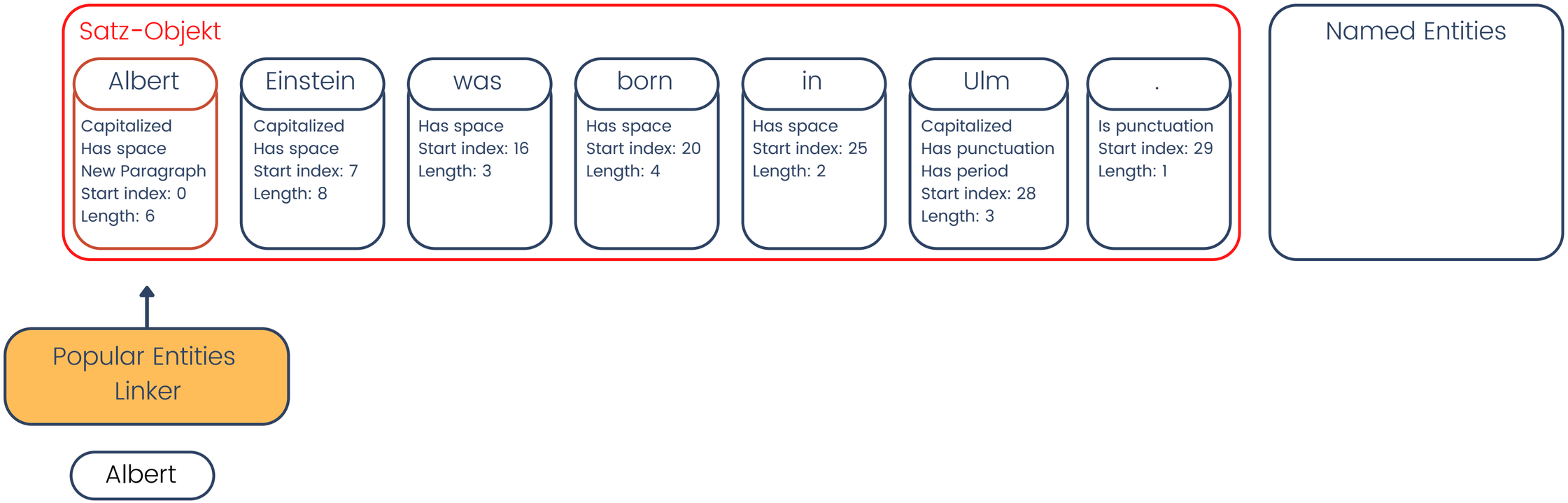
Text-Datenstruktur zur Ersetzung des Tokenizers

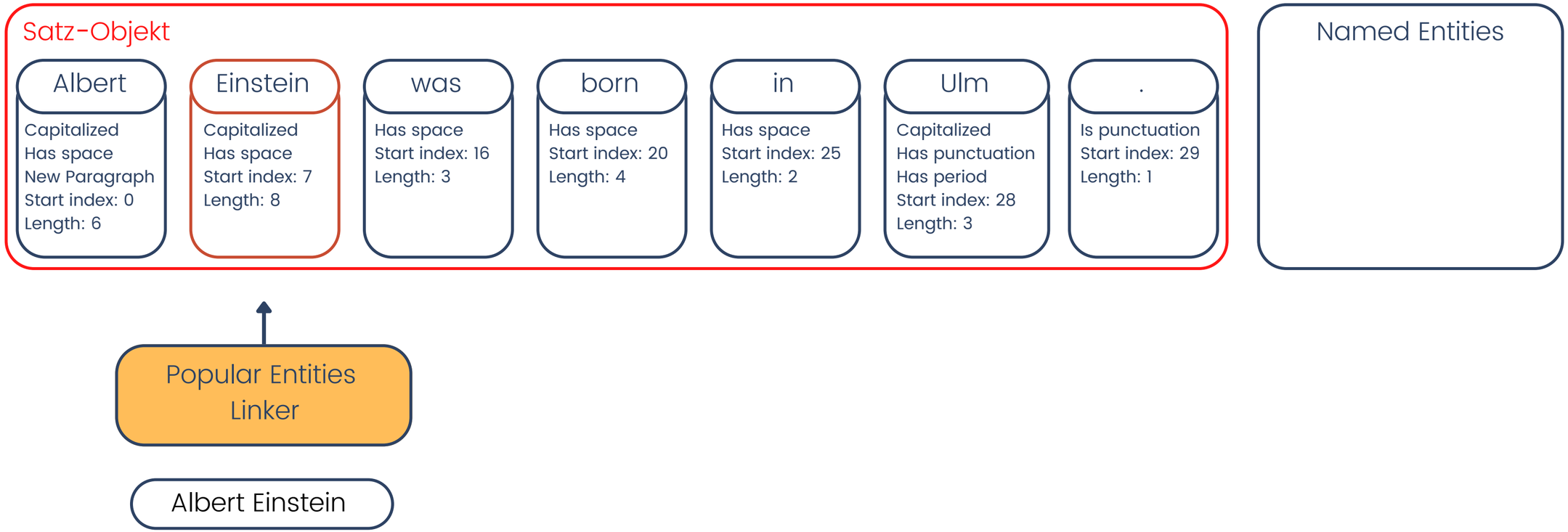


Regelbasierte Erkennung von Named Entities

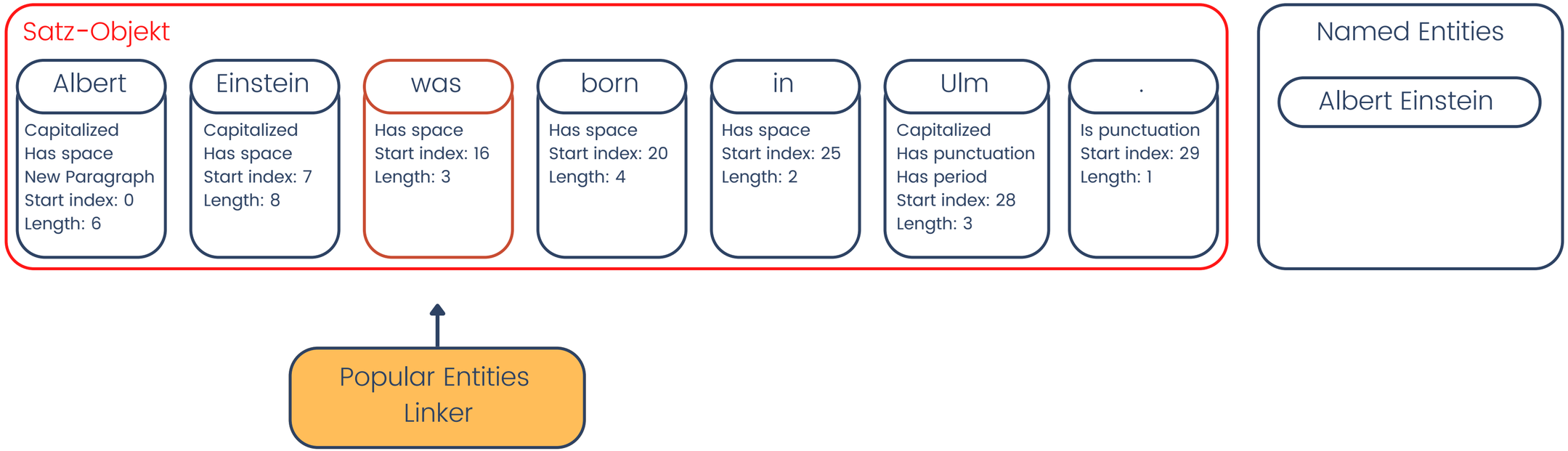


Regelbasierte Erkennung von Named Entities

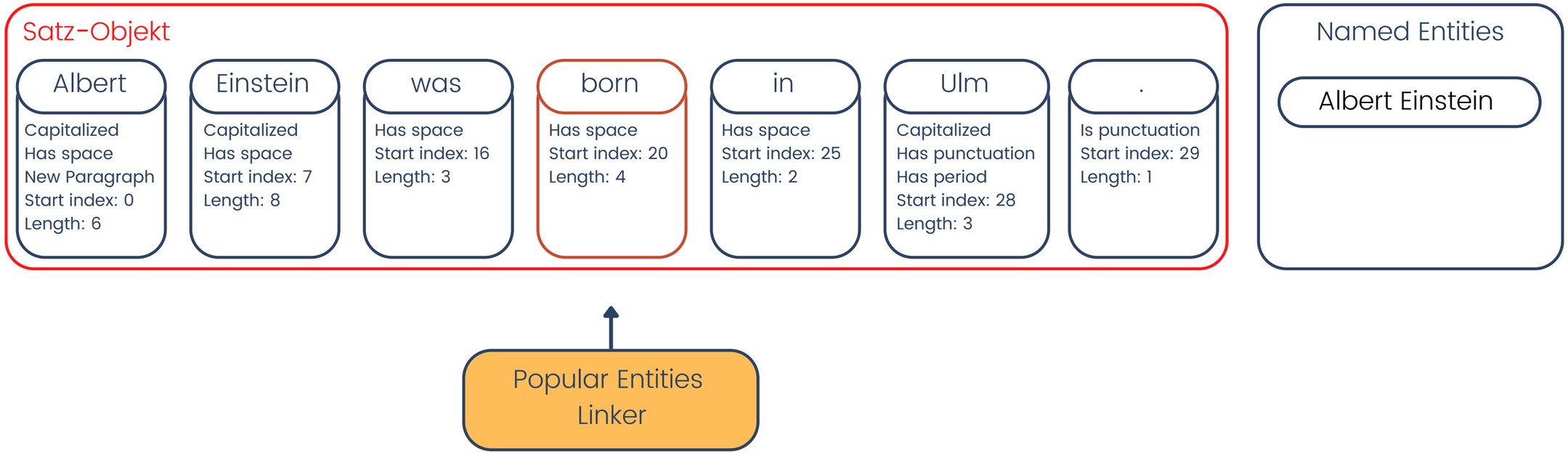



Satz-Objekt
Albert
Capitalized
Has space
New Paragraph
Start index: 0
Length: 6
Einstein
Capitalized
Has space
Start index: 7
Length: 8
was
Has space
Start index: 16
Length: 3
born
Has space
Start index: 20
Length: 4
in
Has space
Start index: 25
Length: 2
Ulm
Capitalized
Has punctuation
Has period
Start index: 28
Length: 3
.
Is punctuation
Start index: 29
Length: 1
Named Entities
Popular Entities Linker
Albert Einstein

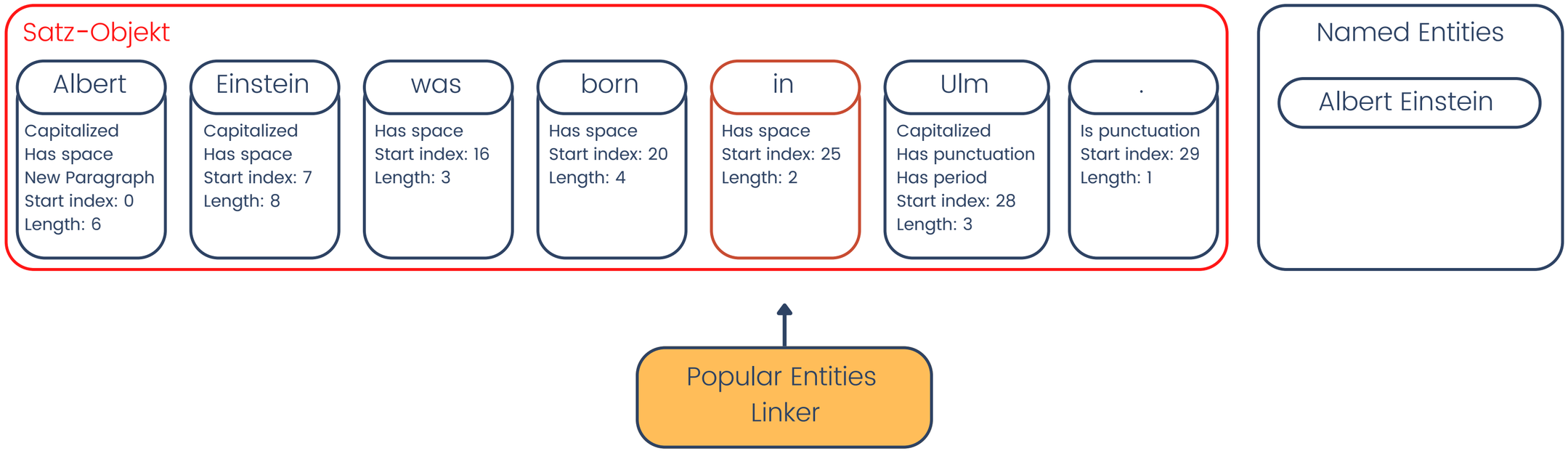
Regelbasierte Erkennung von Named Entities



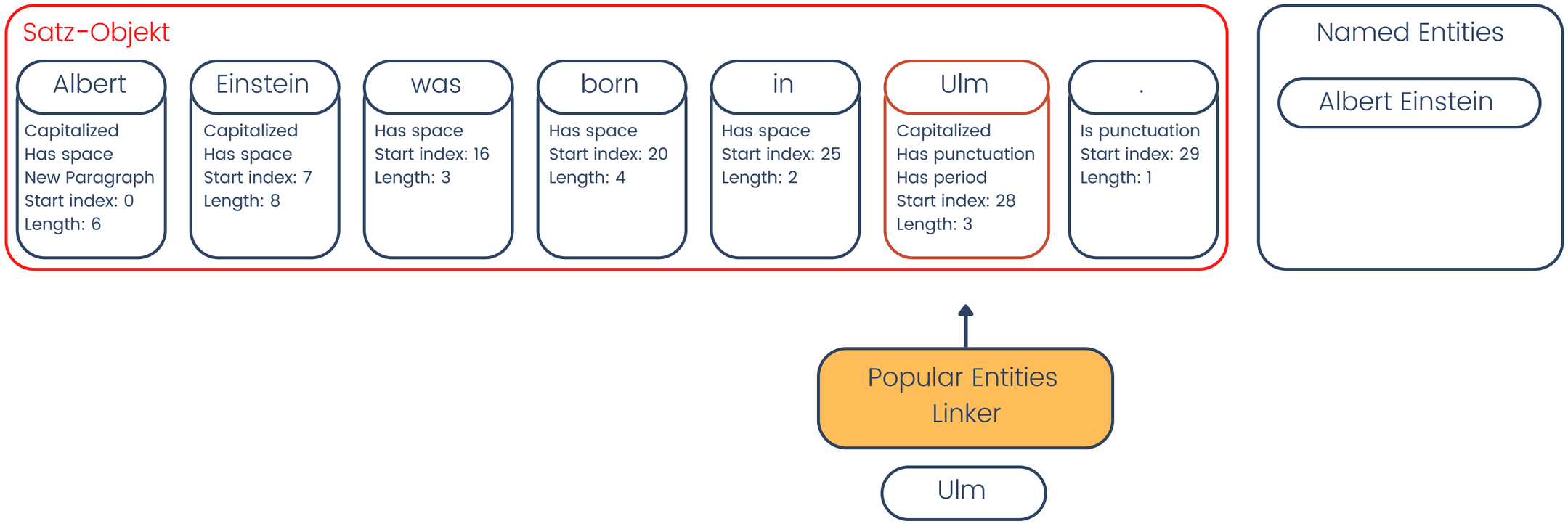
Regelbasierte Erkennung von Named Entities



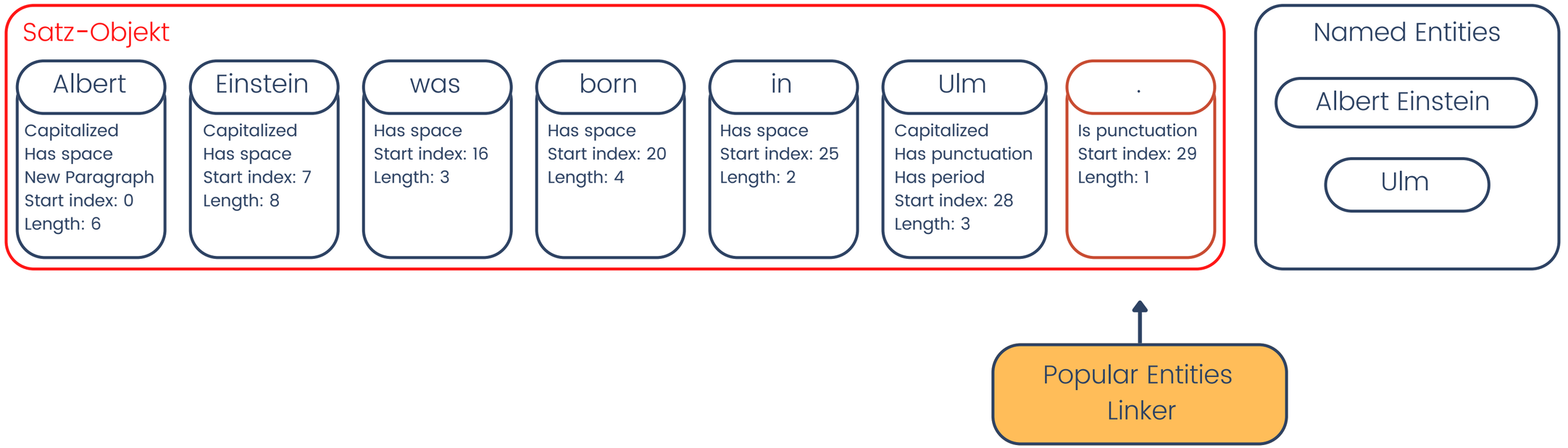
Regelbasierte Erkennung von Named Entities



Regelbasierte Erkennung von Named Entities



Regelbasierte Erkennung von Named Entities

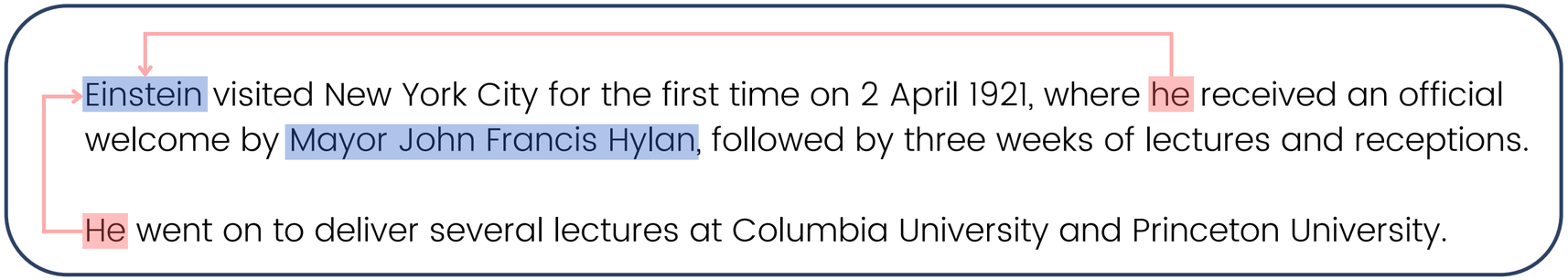


Regel:

Die Named Entities im Text werden einem der drei folgenden Geschlechter zugeordnet:

männlich, weiblich oder neutral

Immer die erste Named Entity eines Geschlechts in einem Satz, wird als das Subjekt des Satzes angesehen. Auf das Subjekt werden die darauf folgenden Pronomen verlinkt.



Vorteile der Ersetzung von spaCy:

1. Möglichkeit Multithreading zu verwenden
2. Laufzeit pro Artikel von spaCy kann eingespart werden

Gibt es Fragen oder Unklarheiten?

Verwendete Dateien:

Die Laufzeitanalyse wurde mit 1000 zufällig ausgewählten Wikipedia-Artikeln durchgeführt.

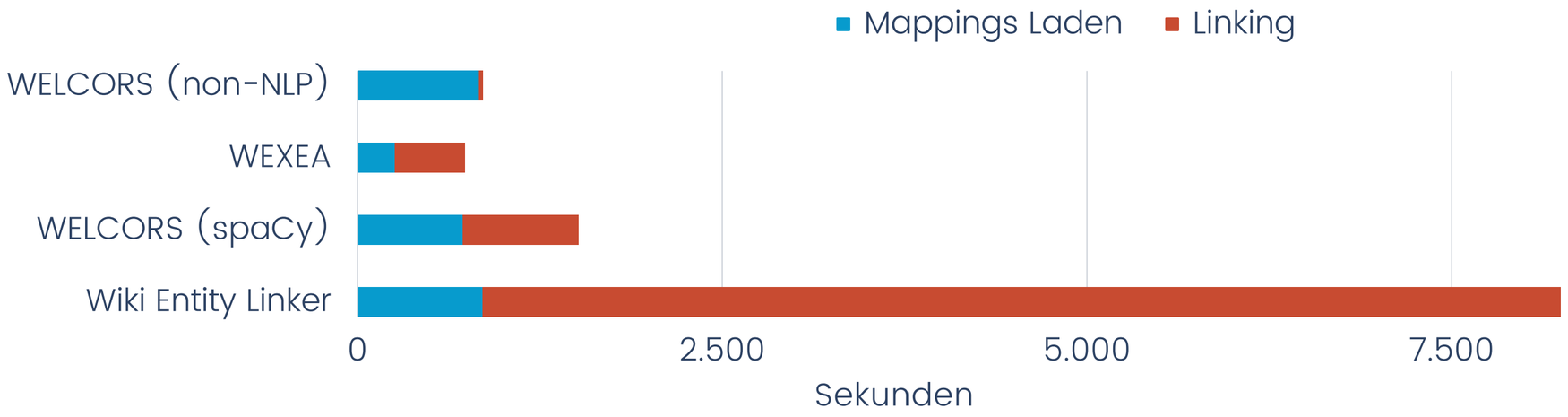
Beispiel für einen Artikel im JSON-Format:

{"id": "21357",
"url": "https://en.wikipedia.org/wiki?curid=21357",
"title": "Telecommunications in New Zealand",
"text": "Telecommunications in New Zealand\n\nTelecommunications in New Zealand are fairly typical for an industrialised country.\n\nFixed-line broadband and telephone services are largely provided through copper-based networks, although fibre-based services are increasingly common. Spark New Zealand, Vodafone New Zealand, and 2degrees provide most services.\n\nMobile telephone services are provided by Spark, Vodafone and 2degrees, although a number of smaller mobile virtual network operators also exist.\n\nSection::::History.\nThe first telegraph opened in New Zealand between the port of Lyttelton and Christchurch on 16 June 1862. "}

Evaluation der Laufzeit: 1000 Artikel

Systeme	Mappings	Linking-Laufzeit	Gesamt
WELCORS (non-NLP)	00:13:48	00:00:27	00:14:15
WEXEA	00:04:12	00:07:58	00:12:10
WELCORS (spaCy)	00:11:58	00:13:12	00:25:10
Wiki Entity Linker	00:14:13	02:03:11	02:17:24

Format der Laufzeitmessung: hh:mm:ss



Benchmarks:

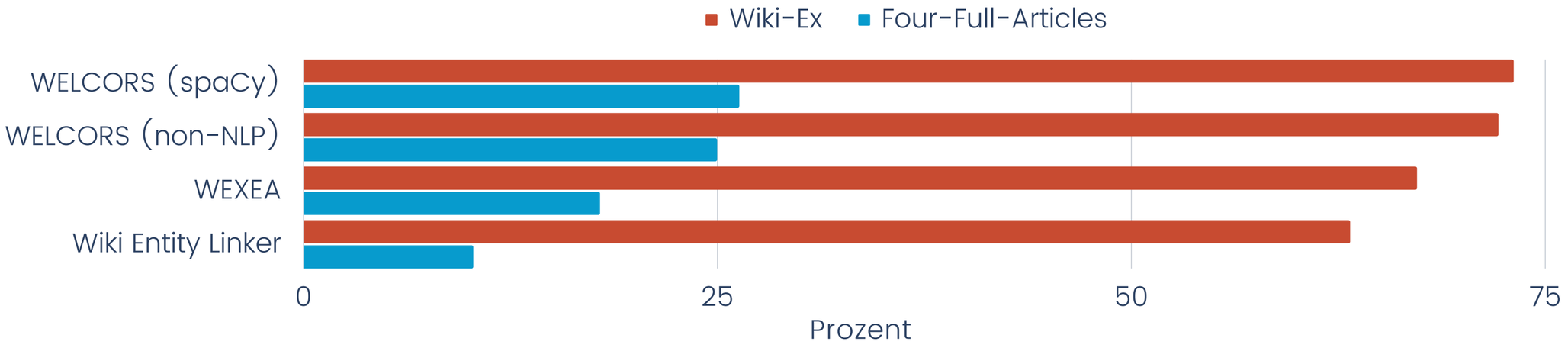
Zwei Benchmarks wurden für die Evaluation der Qualität verwendet:

1. Wiki-Ex Benchmark von Natalie Prange und Matthias Hertel
2. Four-Full-Articles Benchmark von Johanna Götz

Beispiel für einen Benchmarkartikel im JSON-Format:


```
{"id": 38413800,
 "title": "Toxomerus politus",
 "text": "Toxomerus politus\n\nToxomerus politus is a species of hoverfly (Diptera: Syrphidae).
 "hyperlinks": [[[53, 61], "hoverfly"]],
 "title_synonyms": [[19, 36]],
 "url": "https://en.wikipedia.org/wiki?curid=38413800",
 "evaluation_span": [19, 312],
 "labels": [{"id": 0, "span": [19, 36], "entity_id": "Q7830454", "name": "Toxomerus politus", "parent": null,
             "children": [], "optional": false, "type": "Q16521"},
             {"id": 1, "span": [53, 61], "entity_id": "Q217905", "name": "Syrphidae", "parent": null, "children":
             [], "optional": false, "type": "Q16521"}]}
```

Systeme	Wiki-Ex			Four-Full-Articles		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
WELCORS (spaCy)	80.85 %	66.64 %	73.06 %	52.35 %	17.57 %	26.31 %
WELCORS (non-NLP)	77.67%	67,35 %	72.14 %	47.38 %	16.95 %	24.97 %
Wiki Entity Linker	72.48%	62.67 %	67.22 %	38.90 %	11.63 %	17.90 %
WEXEA	76.12%	54.00 %	63.18 %	33.47 %	6.05 %	10.25 %

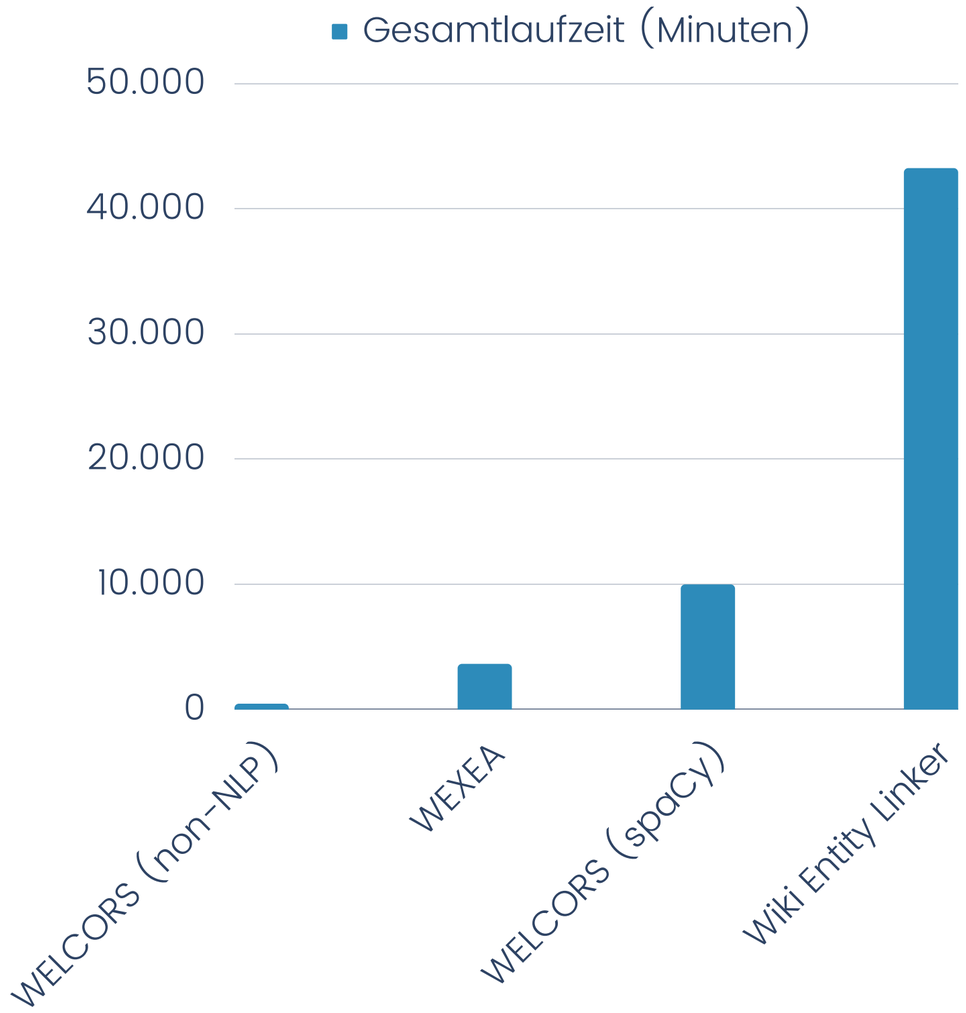


Vielen Dank für Ihre Aufmerksamkeit

Evaluation der Laufzeit: Wikipedia-Dump

Eigenname	Cores	Gesamt
WELCORS (non-NLP)	1	00:06:57:39
WEXEA	6	02:12:00:00
WELCORS (spaCy)	1	06:21:46:44
Wiki Entity Linker	1	30:00:00:00

Format der Laufzeitmessung:
dd:hh:mm:ss



Zusatzfolie

Evaluation der Laufzeit des WELCORS (non-NLP): Wikipedia Dump

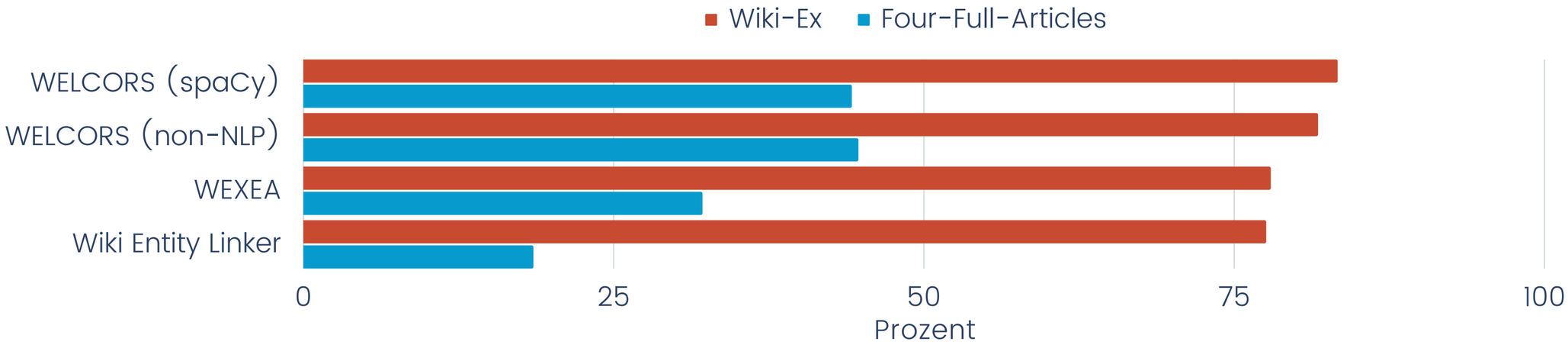
Cores	Laden der Mappings	Linking-Laufzeit	Merging	Gesamt
1	00:10:56	06:18:03	00:06:12	06:57:39
2	00:10:56	03:17:00	00:06:29	03:33:26
4	00:10:14	01:32:39	00:06:17	02:01:10
8	00:10:08	00:47:45	00:06:18	01:16:21
16	00:09:55	00:28:51	00:06:16	00:56:53
24	00:10:21	00:25:01	00:06:08	00:53:14

Format der Laufzeitmessung: hh:mm:ss

Zusatzfolie

Evaluation: Entity Linking

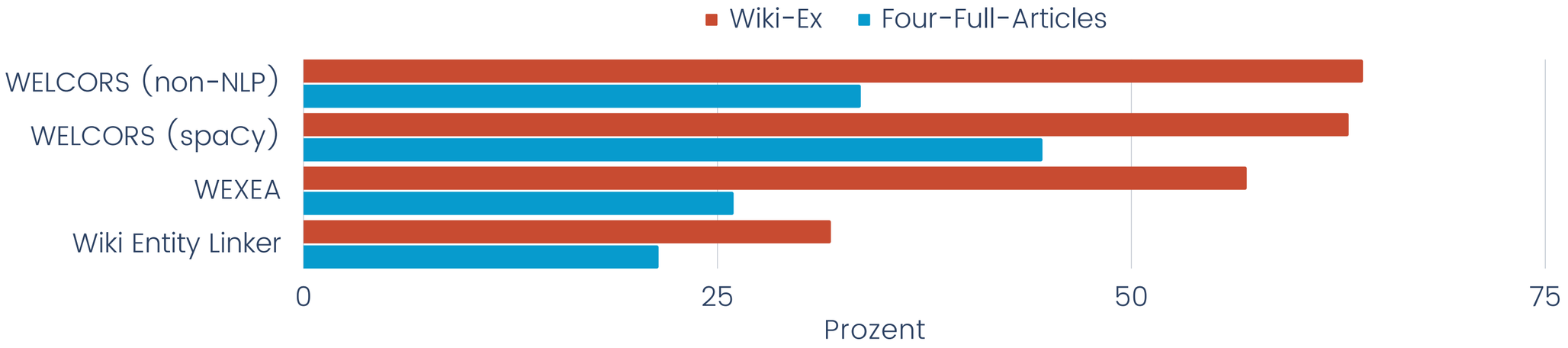
Systeme	Wiki-Ex			Four-Full-Articles		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
WELCORS (spaCy)	85.36 %	81.36 %	83.31 %	53.28 %	37.73 %	44.18 %
WELCORS (non-NLP)	80.11 %	81.36 %	81.73 %	56.94 %	36.80 %	44.71 %
Wiki Entity Linker	78.80 %	77.05 %	77.92 %	40.86 %	26.49 %	32.15 %
WEXEA	81.08 %	74.32 %	77.55 %	32.31 %	12.99 %	18.53 %



Zusatzfolie

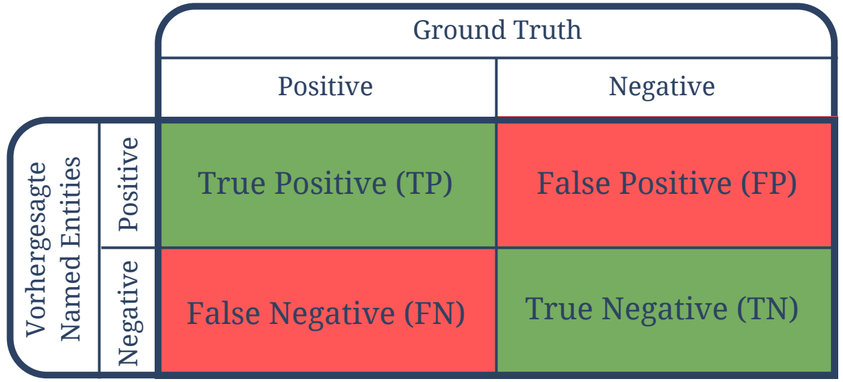
Evaluation: Co-Reference Resolution

Systeme	Wiki-Ex			Four-Full-Articles		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
WELCORS (non-NLP)	69.44 %	**59.29 %**	**63.97 %**	31.07 %	36.71 %	33.65 %
WELCORS (spaCy)	**72.08 %**	56.13 %	63.11 %	**51.85 %**	**39.16 %**	**44.62 %**
Wiki Entity Linker	64.50 %	50.99 %	56.95 %	34.09 %	20.98 %	25.97 %
WEXEA	66.25 %	20.95 %	31.83 %	50.00 %	13.64 %	21.43 %



Zusatzfolie

Konfusionsmatrix:



Precision:

Pronzentualer Anteil der Korrekt vorhergesagten Named Entities von
allen vorhergesagten Named Entities.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall:

Pronzentualer Anteil der Korrekt vorhergesagten Named Entities von
allen im Text enthaltenen Named Entities.

$$\text{Precision} = \frac{TP}{TP + FN}$$

F1-Score:

Setzt sich aus Precision und Recall zusammen und kann als das harmonische Mittel dieser Werte begriffen werden.

$$\text{Precision} = \frac{TP}{TP + FP + FN}$$

Zusatzfolie

Auflistung aller Stellen im System an denen besonders auf Effizienz geachtet wurde:

- Laden der Entity Datenbank (Multithreading)
- Befüllen des Buffers (Producer/Consumer Pattern)
- Ersatz von spaCy
- Verlinkung der Named Entities (Prefix Search Tree)
- Verarbeitung von UTF-8 Kodierung (Bit-Arithmetik)
- Ausgabe der Linkingergebnisse (Multi-Outputstreaming mit anschließendem Merging)
- Multithreading des gesamten Linkingprozesses

Künftige Arbeiten

- Weiterentwicklung der Regeln zur Erkennung von Named Entities beim Entity Linking und der Zuweisung von Pronomen in der Co-Reference Resolution.
- Verwendung einer schnellen auf C++ basierten NLP-Bibliothek (Derzeit nicht verfügbar)
- Reduktion der konstanten Zeiten des Systems durch Vorverarbeitung der Mappings oder Abbau der Mappings nachdem sie nicht mehr gebraucht werden.
- Analyse der abnehmenden Reduktion der Laufzeit bei wachsender Anzahl von CPU-Cores.

Zusatzfolie