



KOKS4CALL
EuroCALL 2001, Nijmegen

Korpusbasierte Kollokationssuche
Corpus based search for collocations
Universität Osnabrück

Contents

Contents

- Introduction

Contents

- Introduction
- Demo

Contents

- Introduction
- Demo
- System Overview

Contents

- Introduction
- Demo
- System Overview
- Phrase Correspondence and Alignment

Contents

- Introduction
- Demo
- System Overview
- Phrase Correspondence and Alignment
- Conclusions



Introduction

About Us

About Us

- participants: 6 students
Arno Erpenbeck, Britta Koch,
Norman Kummer, Philip Reuter,
Patrick Tschorn, Joachim Wagner

About Us

- participants: 6 students
Arno Erpenbeck, Britta Koch,
Norman Kummer, Philip Reuter,
Patrick Tschorn, Joachim Wagner
- 2 advisors:
Dr. habil. Helmar Gust
Dr. Petra Ludewig

About Us

- participants: 6 students
Arno Erpenbeck, Britta Koch,
Norman Kummer, Philip Reuter,
Patrick Tschorn, Joachim Wagner
- 2 advisors:
Dr. habil. Helmar Gust
Dr. Petra Ludewig
- duration: 1 year

About Us

- participants: 6 students
Arno Erpenbeck, Britta Koch,
Norman Kummer, Philip Reuter,
Patrick Tschorn, Joachim Wagner
- 2 advisors:
Dr. habil. Helmar Gust
Dr. Petra Ludewig
- duration: 1 year
- study program:
Computational Linguistics & Artificial Intelligence

Collocations

Collocations

The problem:

Collocations

The problem:

- free combinations vs. collocations vs. idioms

Collocations

The problem:

- free combinations vs. collocations vs. idioms
- meaning not compositionally constructable
heavy smoker - starker Raucher

Collocations

The problem:

- free combinations vs. collocations vs. idioms
- meaning not compositionally constructable
heavy smoker - *starker Raucher*
- replacing with synonyms does not work
die Zähne putzen - **die Zähne bürsten*

Collocations & CALL

Collocations & CALL

Why collocations?

Collocations & CALL

Why collocations?

- difficult for foreign language learners

Collocations & CALL

Why collocations?

- difficult for foreign language learners
- single word lookup in dictionary not helpful

Collocations & CALL

Why collocations?

- difficult for foreign language learners
- single word lookup in dictionary not helpful
- special context/usage:
ins Wasser fallen vs. *in das Wasser fallen*

Collocations & CALL

Why collocations?

- difficult for foreign language learners
- single word lookup in dictionary not helpful
- special context/usage:
ins Wasser fallen vs. *in das Wasser fallen*

Cognitive background:

Collocations & CALL

Why collocations?

- difficult for foreign language learners
- single word lookup in dictionary not helpful
- special context/usage:
ins Wasser fallen vs. *in das Wasser fallen*

Cognitive background:

- collocations among building units of language production

Collocations & CALL

Why collocations?

- difficult for foreign language learners
- single word lookup in dictionary not helpful
- special context/usage:
ins Wasser fallen vs. *in das Wasser fallen*

Cognitive background:

- collocations among building units of language production
- pre-fabricated chunks make processing easier

Idea & Goals

Idea & Goals

How to deal with the problem:

Idea & Goals

How to deal with the problem:

- use parallel corpora (German/English)

Idea & Goals

How to deal with the problem:

- use parallel corpora (German/English)
- search for translation pairs or phrases

Idea & Goals

How to deal with the problem:

- use parallel corpora (German/English)
- search for translation pairs or phrases

Goals of the project:

Idea & Goals

How to deal with the problem:

- use parallel corpora (German/English)
- search for translation pairs or phrases

Goals of the project:

- construct database with collocation pairs

Idea & Goals

How to deal with the problem:

- use parallel corpora (German/English)
- search for translation pairs or phrases

Goals of the project:

- construct database with collocation pairs
- build prototype: aid for text understanding

Idea & Goals

How to deal with the problem:

- use parallel corpora (German/English)
- search for translation pairs or phrases

Goals of the project:

- construct database with collocation pairs
- build prototype: aid for text understanding
- deliver translations for marked phrases

Idea & Goals

How to deal with the problem:

- use parallel corpora (German/English)
- search for translation pairs or phrases

Goals of the project:

- construct database with collocation pairs
- build prototype: aid for text understanding
- deliver translations for marked phrases
- show hints for typical usage

Idea & Goals

How to deal with the problem:

- use parallel corpora (German/English)
- search for translation pairs or phrases

Goals of the project:

- construct database with collocation pairs
- build prototype: aid for text understanding
- deliver translations for marked phrases
- show hints for typical usage

technical info



Demo

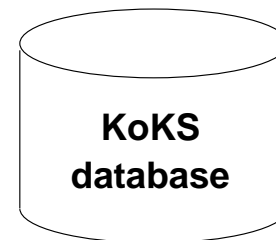
Demo

Screenshots:

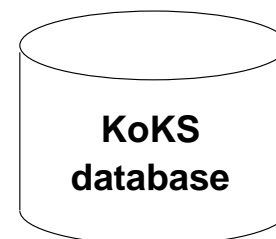
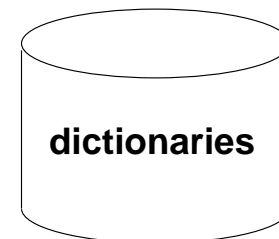
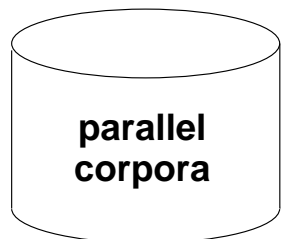


System Overview

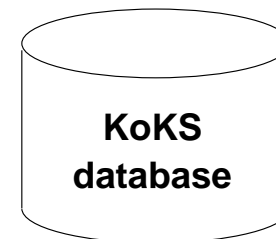
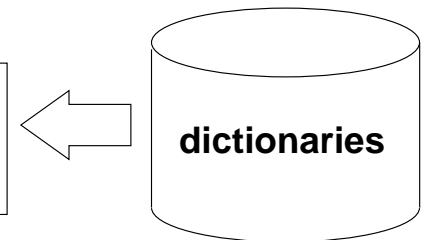
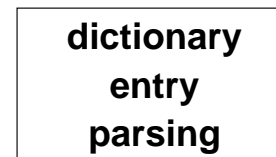
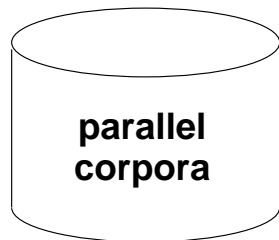
Overview



Overview



Overview



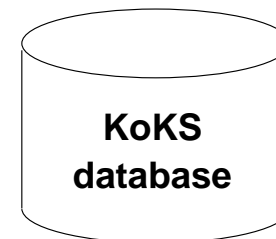
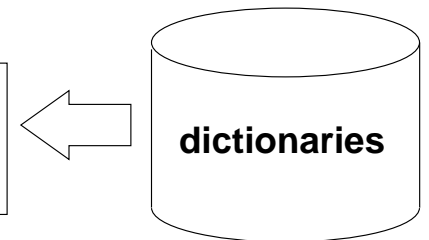
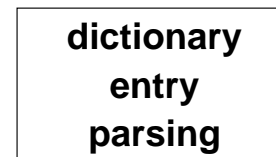
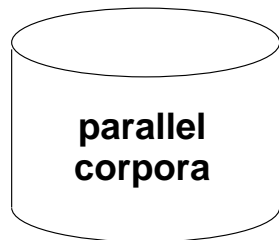
Raw Dictionary Entry

party [n] 1 die Partei; opposing party
die Gegenpartei Politik 2 die Partei,
der/die Beteiligte; to be (a) party in/to
sth. sich an etwas beteiligen Recht 3
die Gruppe, die Gesellschaft, die Partie;
to make one of the party sich der Gruppe
anschließen 4 das Kommando, der Trupp, die
Abteilung Militär 5 die Fete, die Party;
to give/throw a party eine Party geben 6
<ugs> der Typ <ugs>, der Kerl

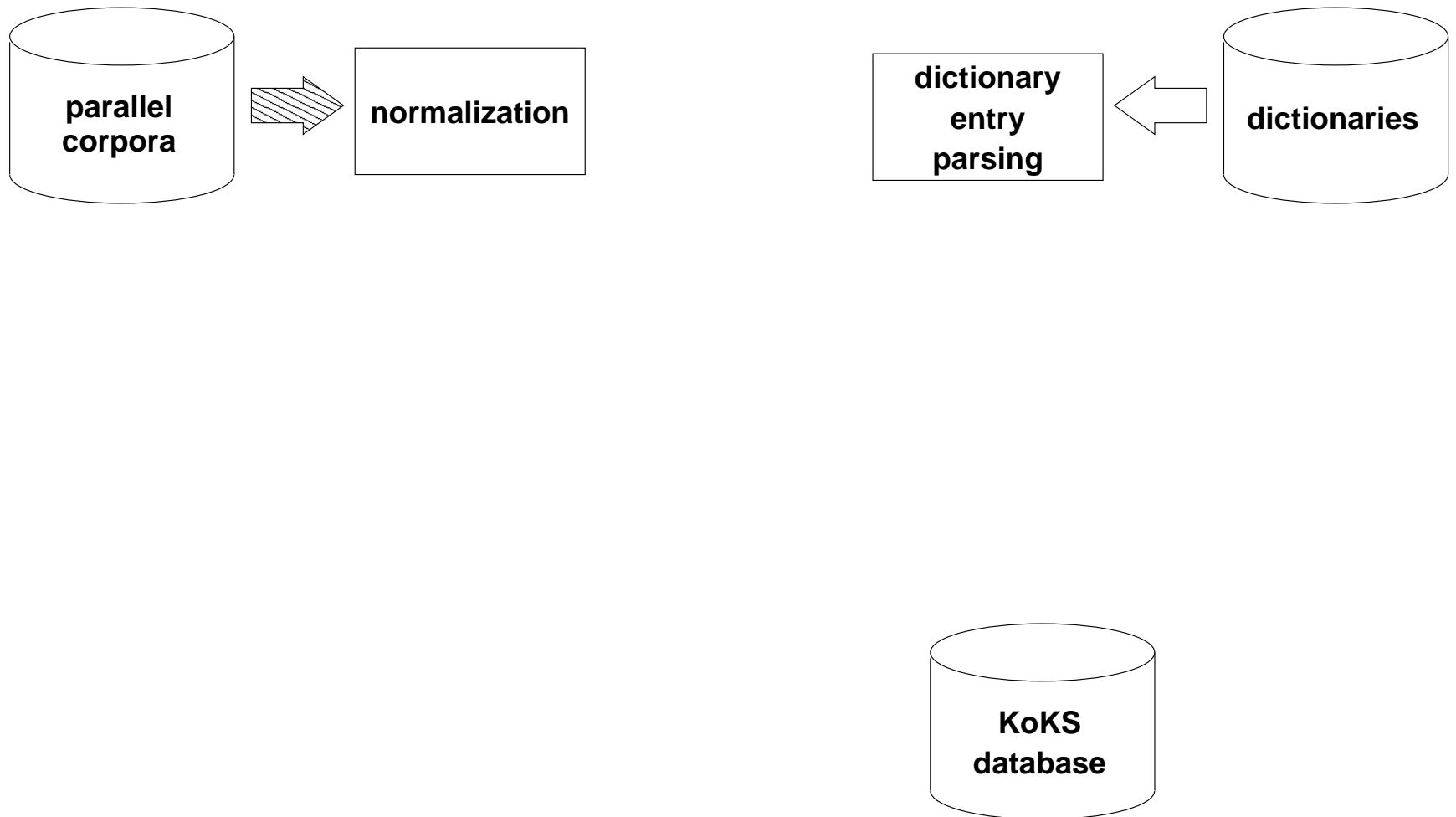
Parsed Dictionary Entry

party	Partei
party	Beteiligte
party	Gruppe
party	Gesellschaft
party	Partie
party	Kommando
party	Trupp
party	Abteilung
party	Fete
party	Party

Overview



Overview



Raw Corpus

<H1>

Mein Wochenende

</H1>

Letztes Wochenende war langweilig. Die Fete zum Ferienbeginn fiel ins Wasser, weil die Disco abgebrannt war. Ausserdem kam auch nichts Anstaendiges im Fernseh.

<H1>

My weekend

</H1>

Last weekend was boring. The school's out party was called off. The club had burned down. Also, there was nothing on the telly.

Normalized Corpus

Mein Wochenende

<ABSATZ>

Letztes Wochenende war langweilig. Die Fete zum Ferienbeginn fiel ins Wasser, weil die Disco abgebrannt war. Ausserdem kam auch nichts Anstaendiges im Fernseh.

<ABSATZ>

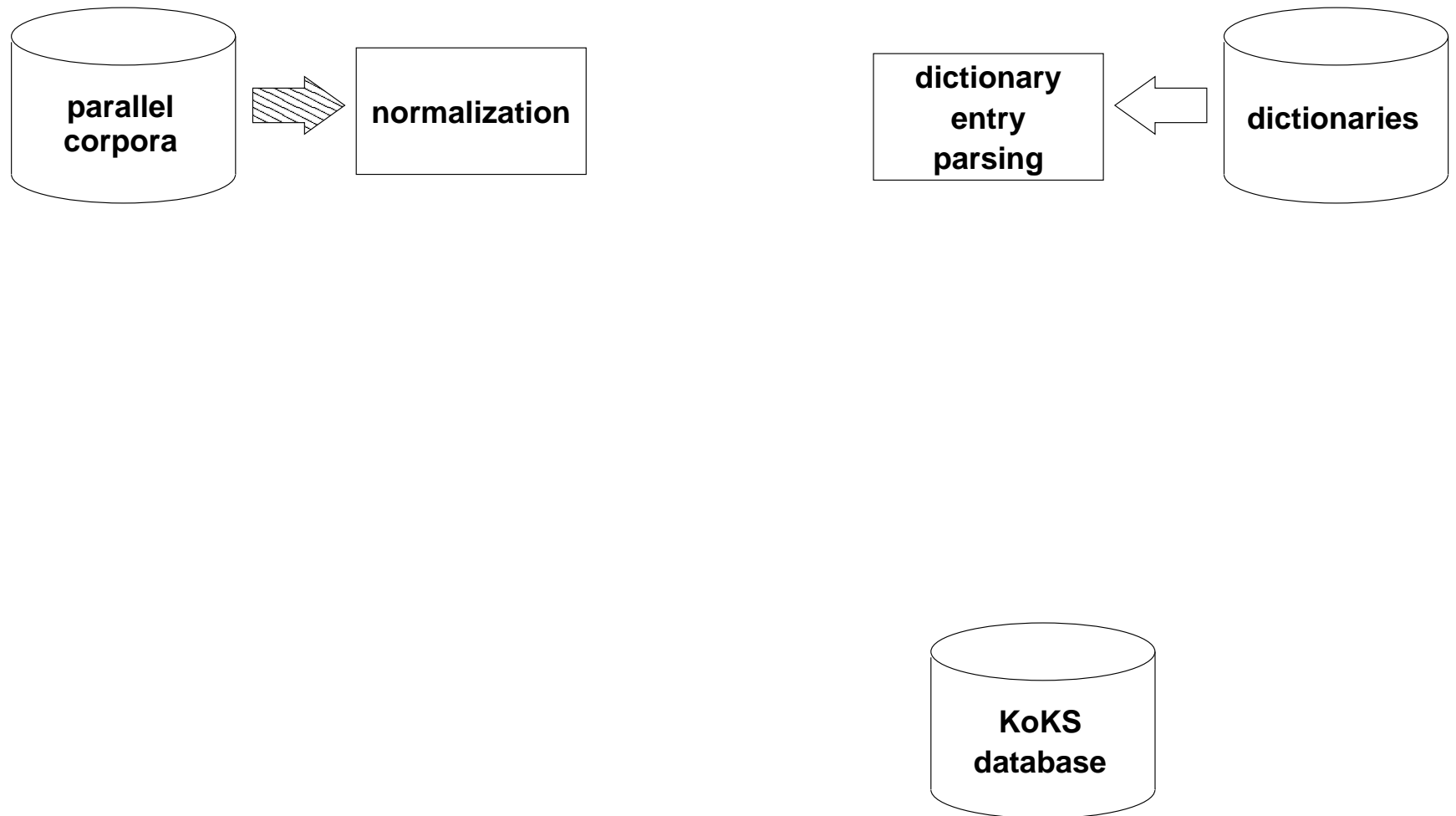
My weekend

<ABSATZ>

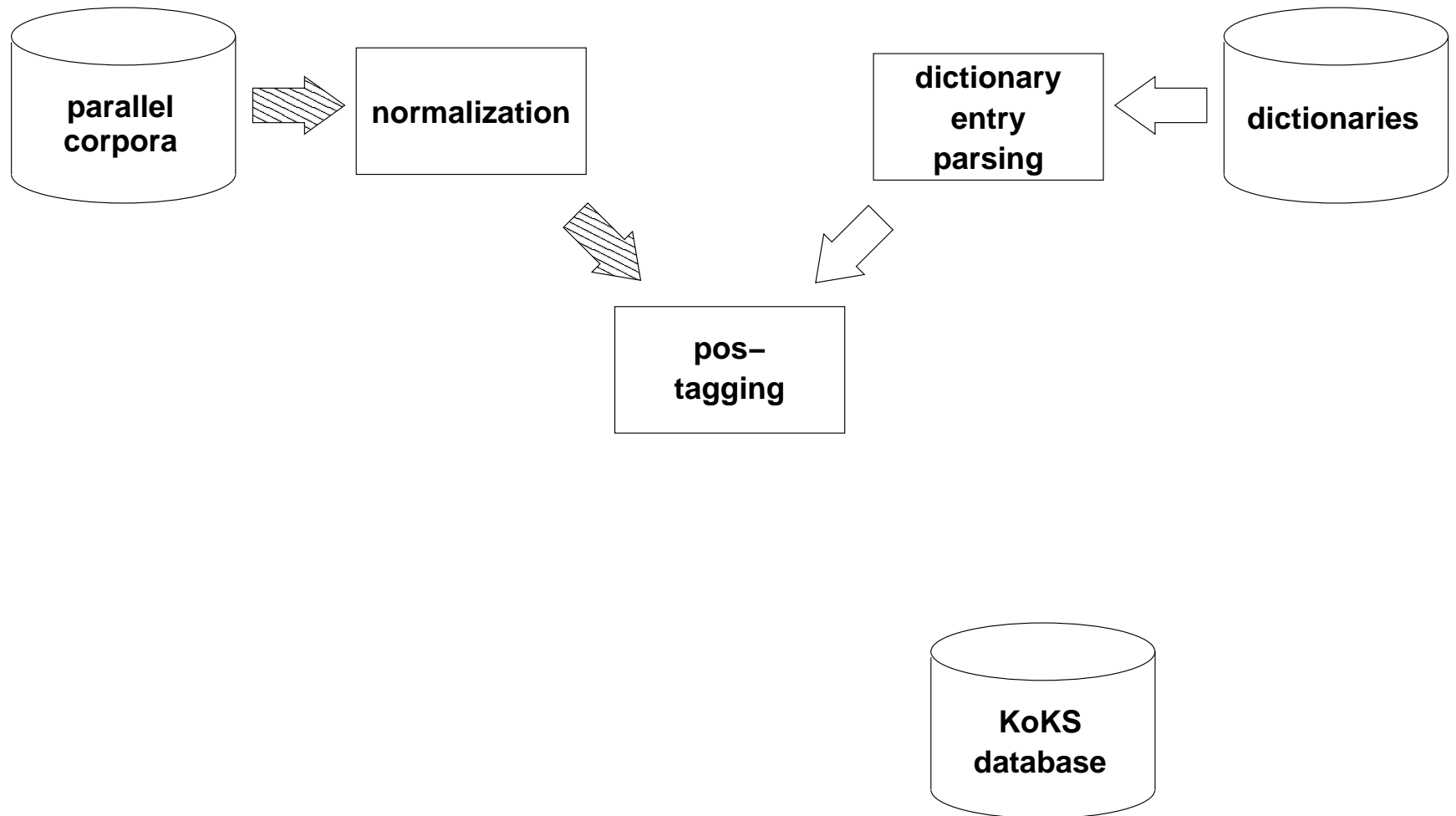
Last weekend was boring. The school's out party was called off. The club had burned down. Also, there was nothing on the telly.

<ABSATZ>

Overview



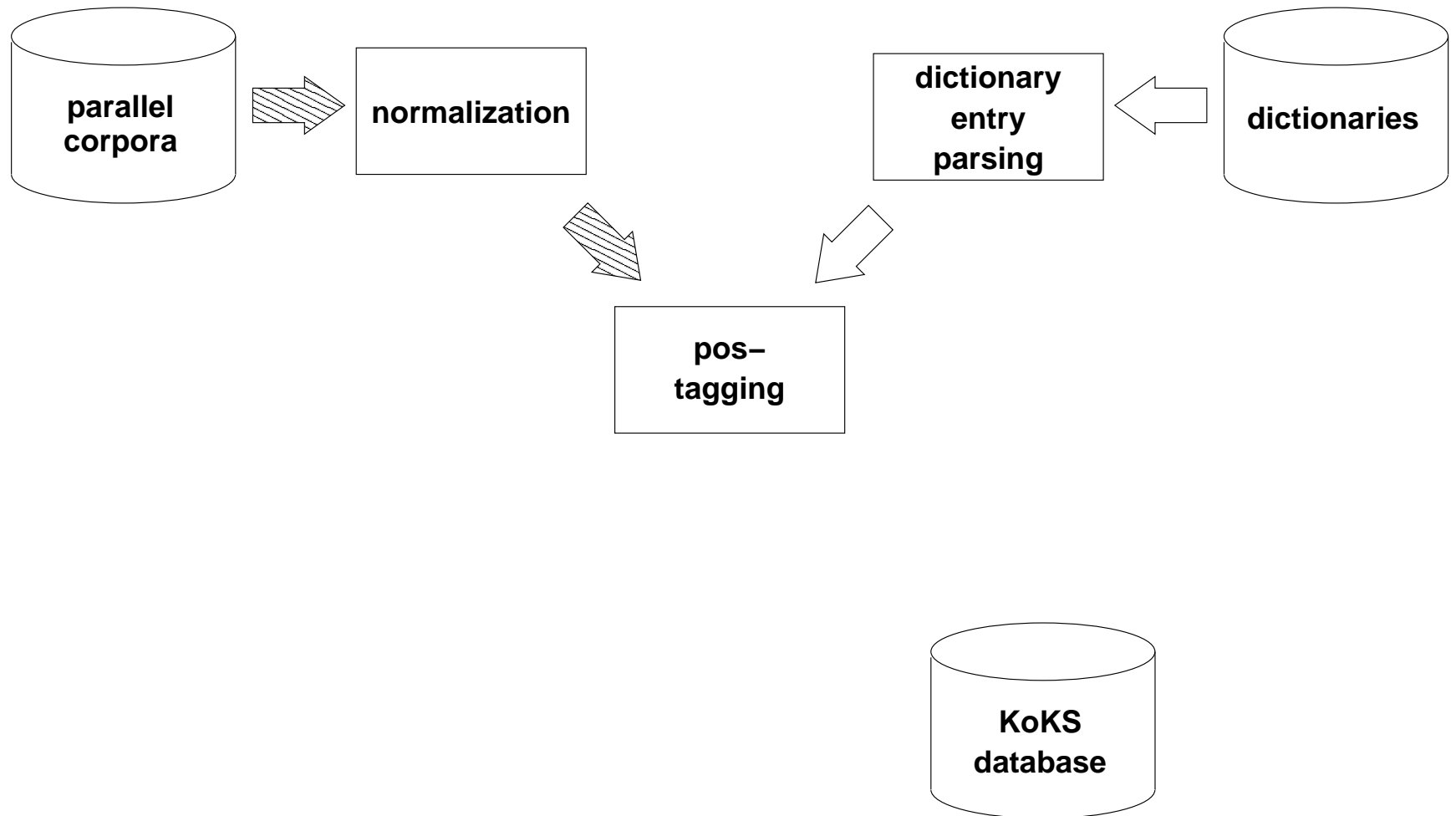
Overview



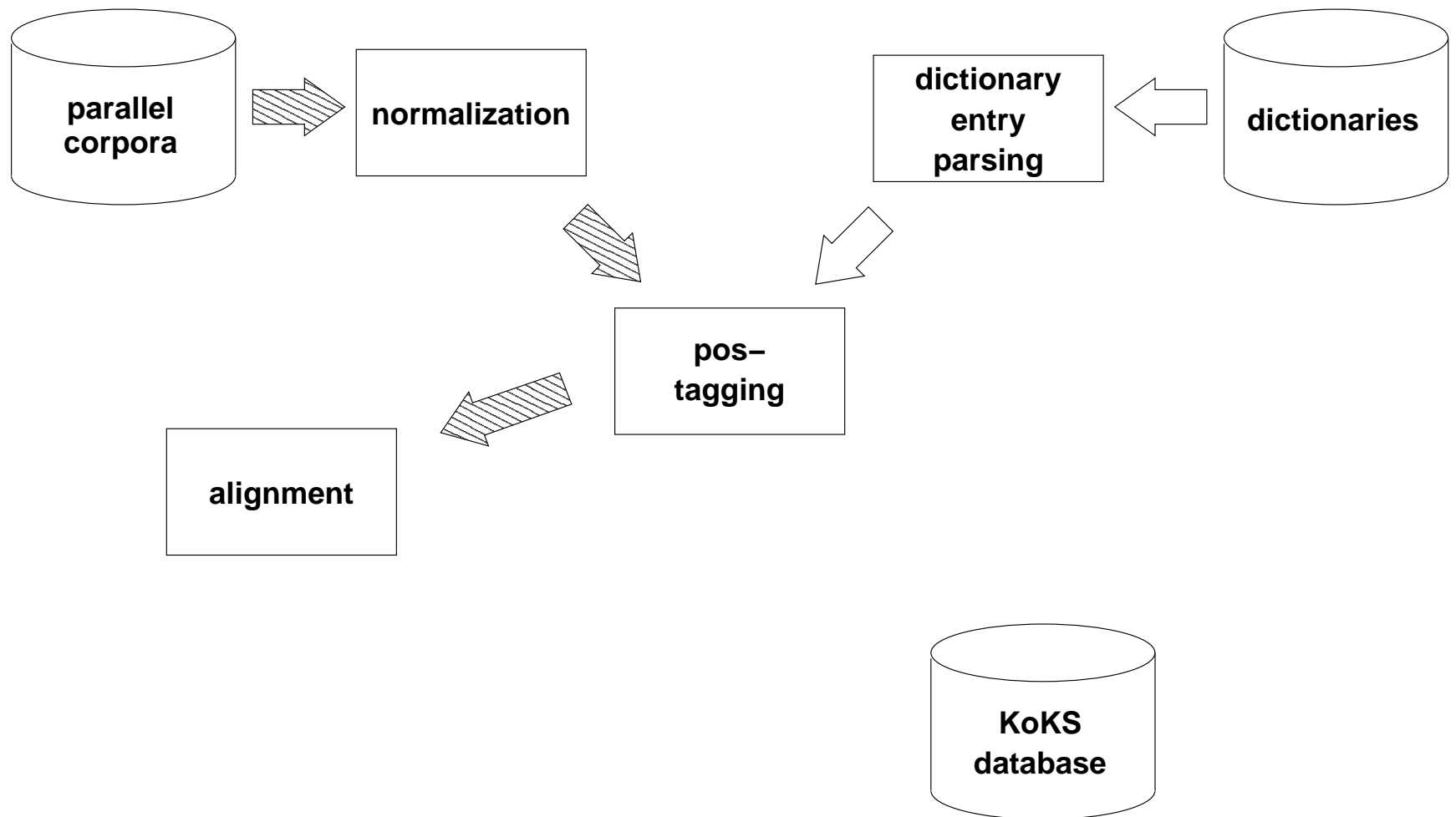
Tagged Corpus

Word	Tag	Lemma	Word	Tag	Lemma
Die	ART	d	The	DT	the
Fete	NN	Fete	school	NN	school
zum	APPRART	zum	's	VBZ	be
Ferienbeginn	NN	Ferienbeginn	out	IN	out
fiel	VVFIN	fallen	party	NN	party
ins	APPRART	ins	was	VBD	be
Wasser	NN	Wasser	called	VCN	call
,	\$,	,	off	RP	off
weil	KOUS	weil	.	SATZ-P	.
die	ART	d	<SATZ>		
Disco	NN	Disco	<segmentgrenze>		
abgebrannt	VVPP	abbrennen	The	DT	the
war	VAFIN	sein	club	NN	club
.	SATZ-P	.	had	VBD	have
<SATZ>			burned	VCN	burn
<segmentgrenze>			down	RP	down
Außerdem	ADV	außerdem	.	SATZ-P	.
kam	VVFIN	kommen	<SATZ>		
auch	ADV	auch	<segmentgrenze>		

Overview



Overview



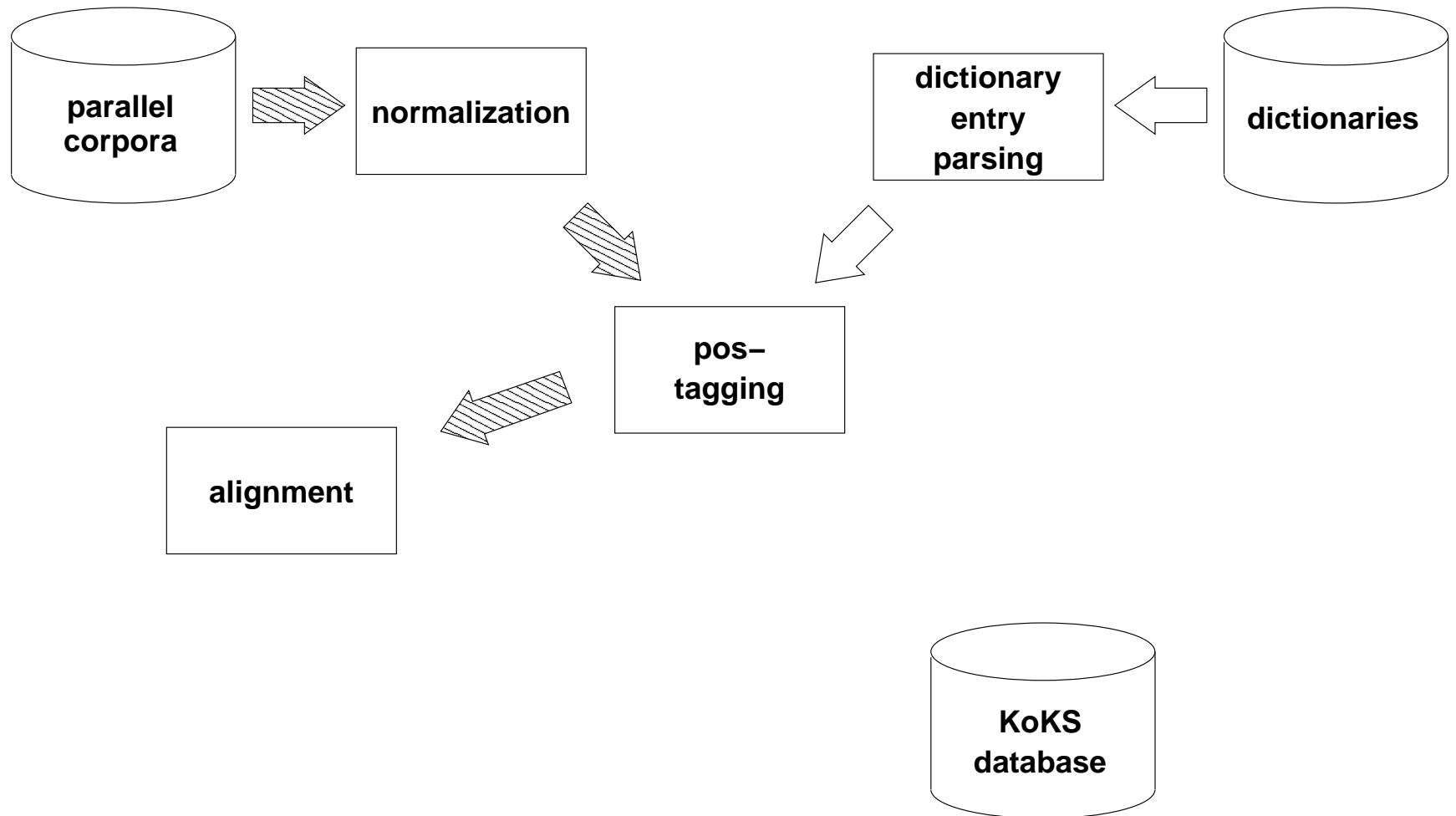
Aligned Corpus

Word	Tag	Lemma	Word	Tag	Lemma
Die	ART	d	The	DT	the
Fete	NN	Fete	school	NN	school
zum	APPRART	zum	's	VBZ	be
Ferienbeginn	NN	Ferienbeginn	out	IN	out
fiel	VVFIN	fallen	party	NN	party
ins	APPRART	ins	was	VBD	be
Wasser	NN	Wasser	called	VCN	call
,	\$,	,	off	RP	off
weil	KOUS	weil	.	SATZ-P	.
die	ART	d	<SATZ>		
Disco	NN	Disco	The	DT	the
abgebrannt	VVPP	abbrennen	club	NN	club
war	VAFIN	sein	had	VBD	have
.	SATZ-P	.	burned	VCN	burn
<SATZ>			down	RP	down
<segmentgrenze>			.	SATZ-P	.
Außerdem	ADV	außerdem	<SATZ>		
kam	VVFIN	kommen	<segmentgrenze>		

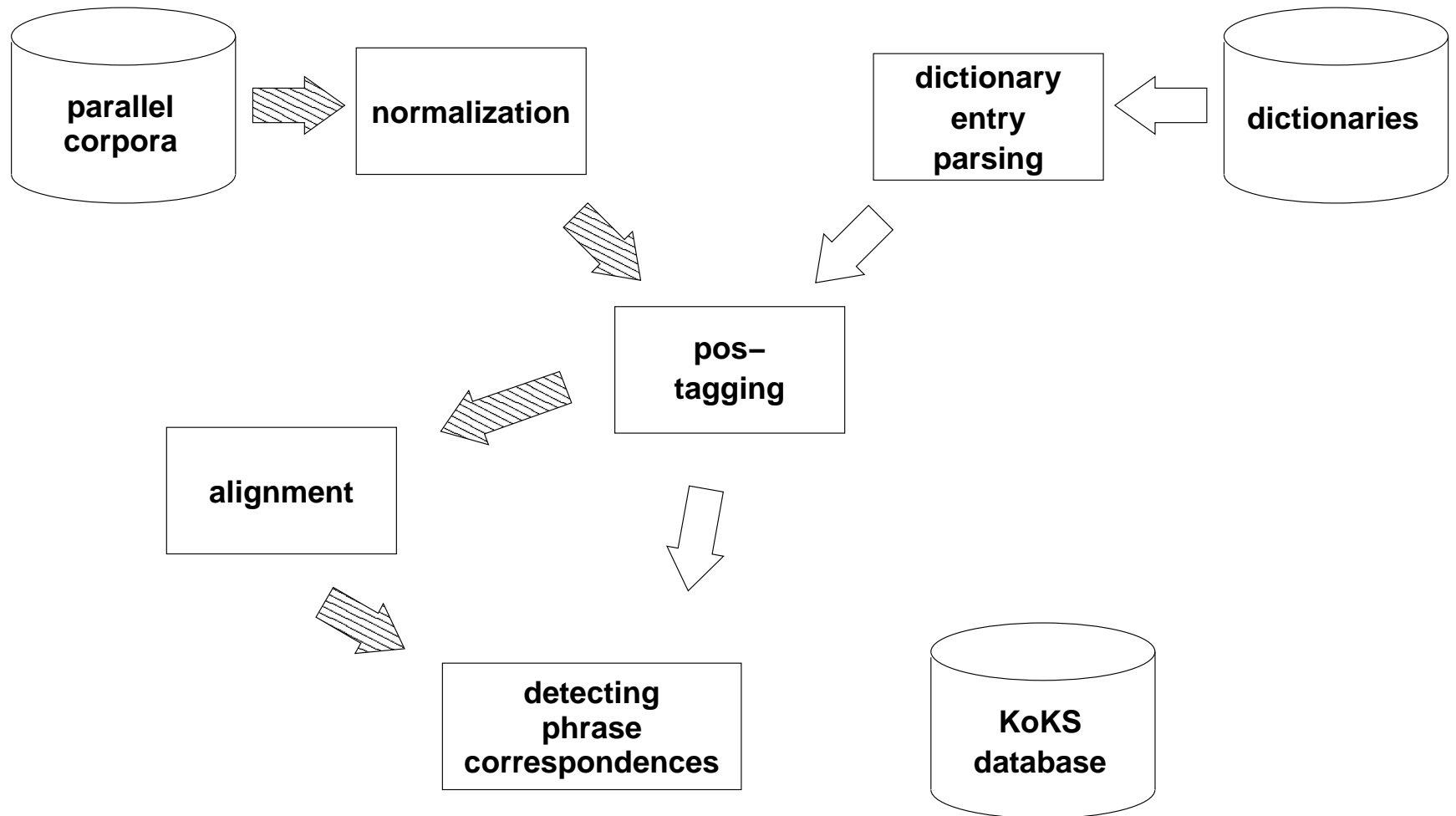
Aligned Corpus

Word	Tag	Lemma	Word	Tag	Lemma
Die	ART	d	The	DT	the
Fete	NN	Fete	school	NN	school
zum	APPRART	zum	's	VBZ	be
Ferienbeginn	NN	Ferienbeginn	out	IN	out
fiel	VVFIN	fallen	party	NN	party
ins	APPRART	ins	was	VBD	be
Wasser	NN	Wasser	called	VCN	call
,	\$,	,	off	RP	off
weil	KOUS	weil	.	SATZ-P	.
die	ART	d	<SATZ>		
Disco	NN	Disco	The	DT	the
abgebrannt	VVPP	abbrennen	club	NN	club
war	VAFIN	sein	had	VBD	have
.	SATZ-P	.	burned	VCN	burn
<SATZ>			down	RP	down
<segmentgrenze>			.	SATZ-P	.
Außerdem	ADV	außerdem	<SATZ>		
kam	VVFIN	kommen	<segmentgrenze>		

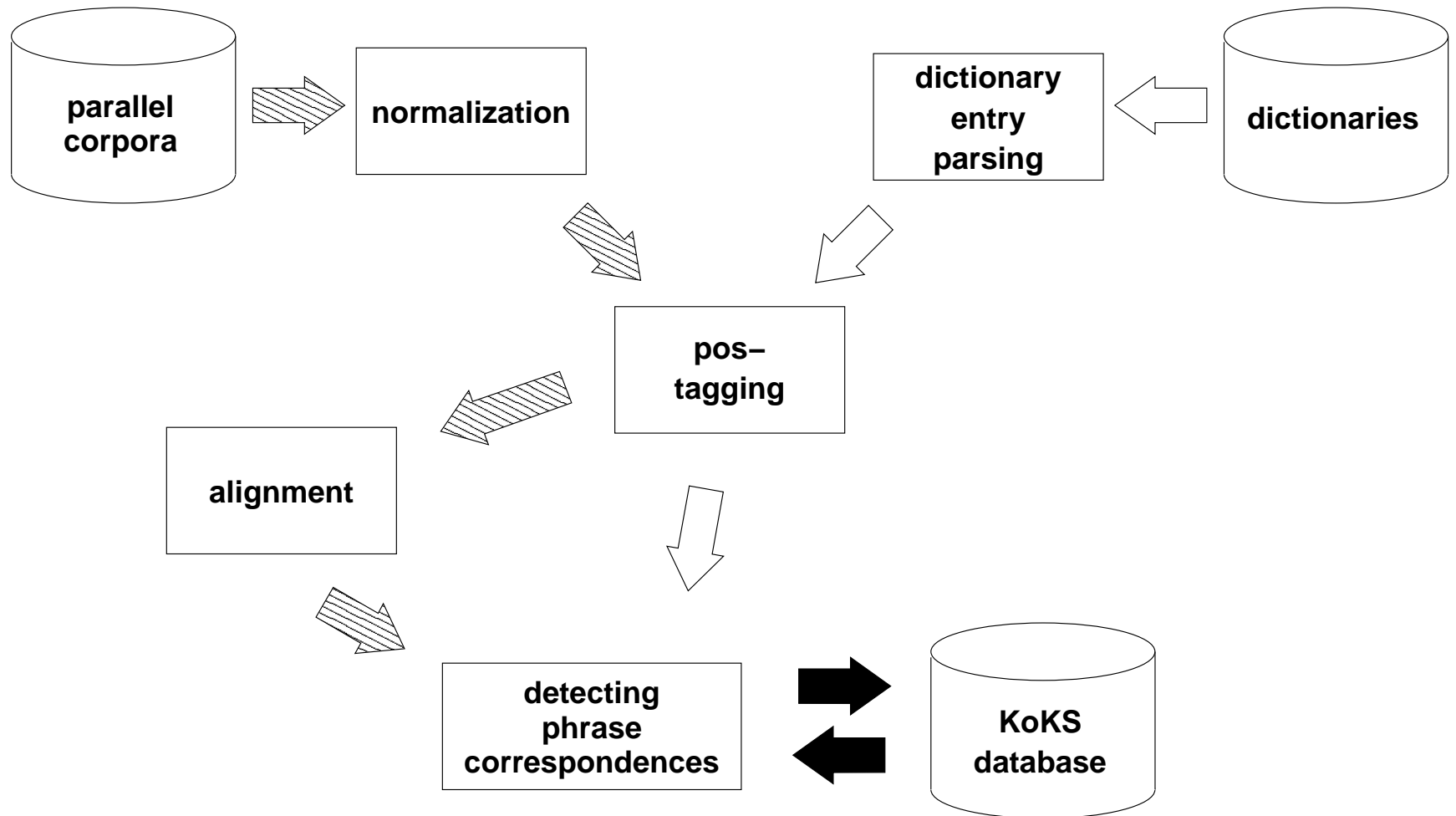
Overview



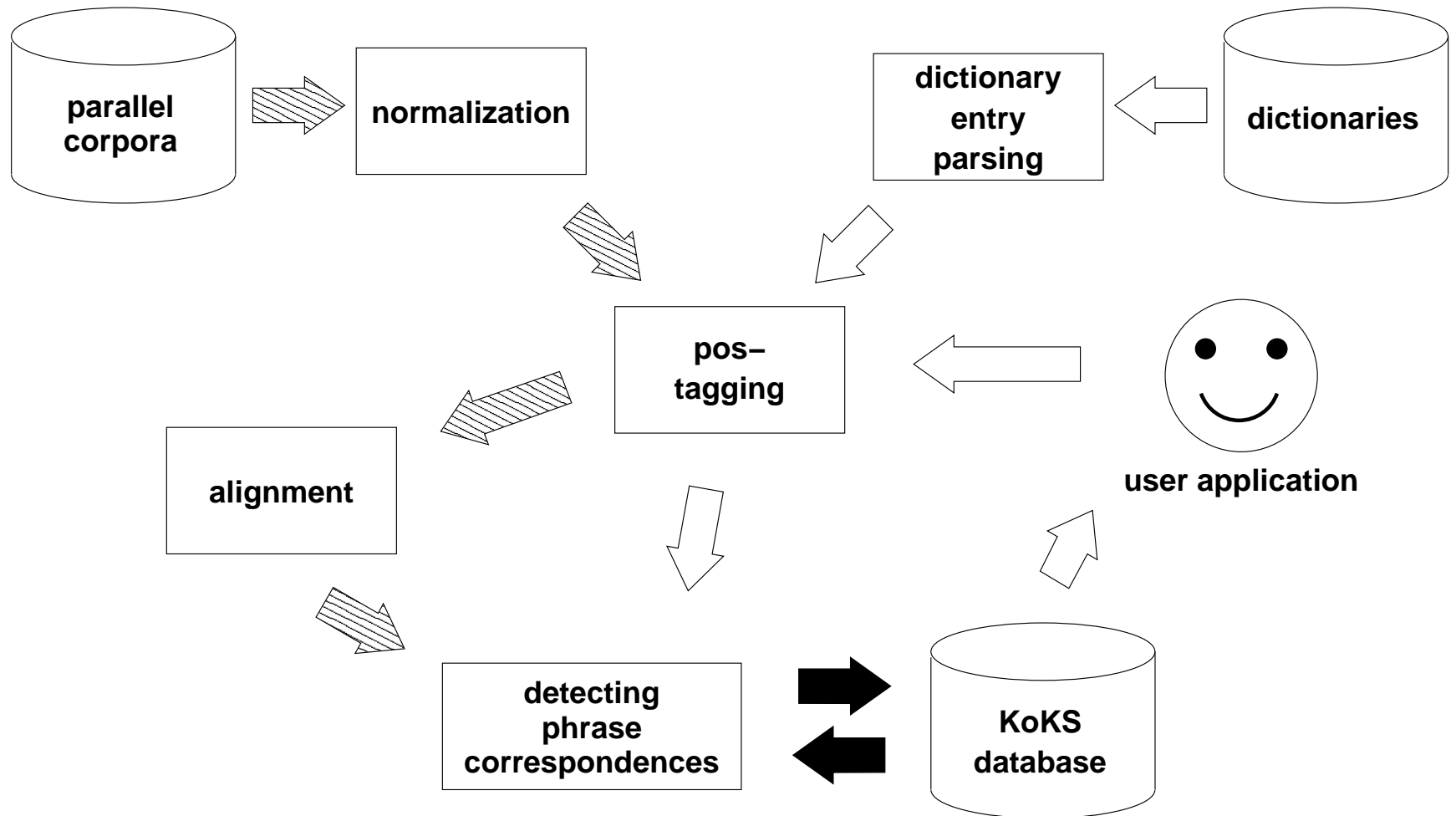
Overview



Overview



Overview





Phrase Correspondence and Alignment

Phrase Correspondence

DT NN VBZ IN NN VBD VBN RP
The school 's out party was called off.

ART NN APPRART NN VVFIN APPRART NN
Die Fete zum Ferienbeginn fiel ins Wasser.

Phrase Correspondence

- mark words with irrelevant tags (yellow)

DT NN VBZ IN NN VBD VBN RP
The school 's out party was called off.

ART NN APPRART NN VVFIN APPRART NN
Die Fete zum Ferienbeginn fiel ins Wasser.

Phrase Correspondence

- mark words with irrelevant tags (yellow)
- mark words that have translations in other sentence (blue)

DT NN VBZ IN NN VBD VBN RP
The school 's out party was called off.

ART NN APPRART NN VVFIN APPRART NN
Die Fete zum Ferienbeginn fiel ins Wasser.

Phrase Correspondence

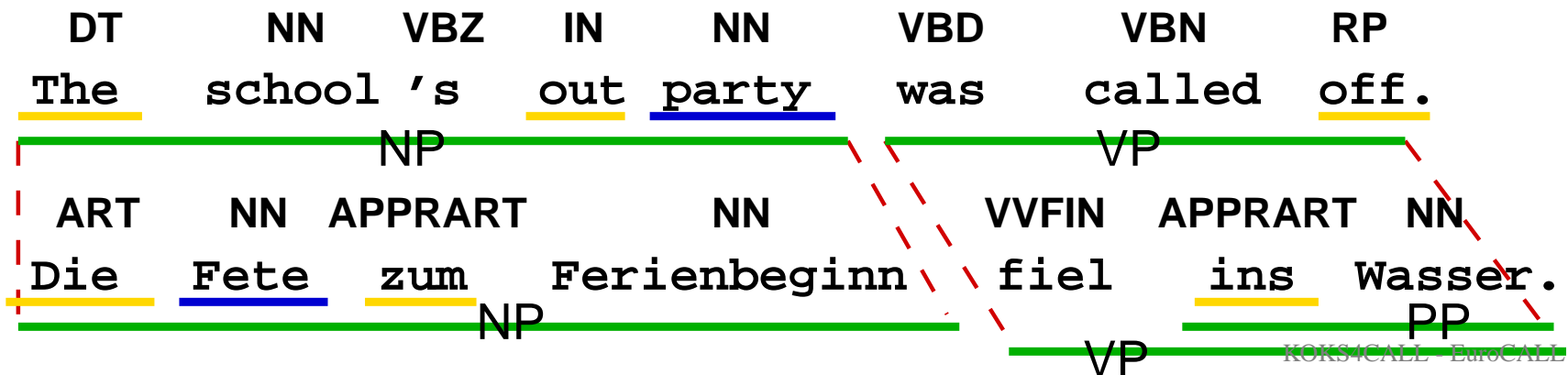
- mark words with irrelevant tags (yellow)
- mark words that have translations in other sentence (blue)
- construct tag sequence crystals by category (green)

DT	NN	VBZ	IN	NN	VBD	VBN	RP
<u>The</u>	school	's	<u>out</u>	<u>party</u>	was	called	<u>off.</u>
NP				VP			

ART	NN	APPRART	NN	VVFIN	APPRART	NN
<u>Die</u>	<u>Fete</u>	<u>zum</u>	Ferienbeginn	fiel	<u>ins</u>	Wasser.
NP				PP		
				VP		

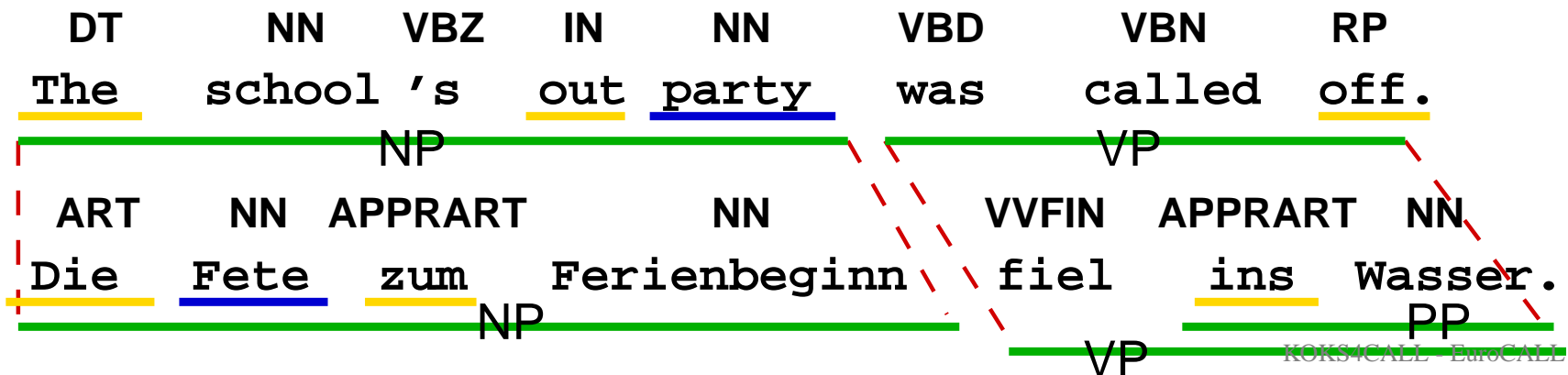
Phrase Correspondence

- mark words with irrelevant tags (yellow)
- mark words that have translations in other sentence (blue)
- construct tag sequence crystals by category (green)
- pair all tag sequences with equal categories (red)



Phrase Correspondence

- mark words with irrelevant tags (yellow)
- mark words that have translations in other sentence (blue)
- construct tag sequence crystals by category (green)
- pair all tag sequences with equal categories (red)
- pair leftover words



Alignment

Alignment

- sentence alignment important

Alignment

- sentence alignment important
- use of lexical knowledge

Alignment

- sentence alignment important
- use of lexical knowledge
- distance calculation based on dictionary

Alignment

- sentence alignment important
- use of lexical knowledge
- distance calculation based on dictionary
- quality of correspondence depends on whether sentences really are aligned

Alignment

- sentence alignment important
- use of lexical knowledge
- distance calculation based on dictionary
- quality of correspondence depends on whether sentences really are aligned

Examples in context:

Letztes Wochenende war langweilig. Die Fete zum Ferienbeginn fiel ins Wasser, weil die Disco abgebrannt war.

Last weekend was boring. The school's out party was called off. The club had burned down.

Phrase Recognition

Der Ausflug fiel ins Wasser.

Phrase Recognition

- tag input sentence

ART NN VVFIN APPRART NN
Der Ausflug fiel ins Wasser.

Phrase Recognition

- tag input sentence
- mark words with irrelevant tags (yellow)

ART NN VVFIN APPRART NN
Der Ausflug fiel ins Wasser.

Conclusions

Conclusions

- construction of collocation dictionary based on parallel corpora

Conclusions

- construction of collocation dictionary based on parallel corpora
- prototype learner application using the dictionary and examples

Conclusions

- construction of collocation dictionary based on parallel corpora
- prototype learner application using the dictionary and examples
- simple chunk parsing as a first effort

Conclusions

- construction of collocation dictionary based on parallel corpora
- prototype learner application using the dictionary and examples
- simple chunk parsing as a first effort
- further ideas: better parsing, better statistics

Conclusions

- construction of collocation dictionary based on parallel corpora
- prototype learner application using the dictionary and examples
- simple chunk parsing as a first effort
- further ideas: better parsing, better statistics
- good, extensive parallel corpora needed



Discussion

Contact Information

KoKS - Korpusbasierte Kollokationssuche
Corpus based search for collocations

<http://www.cl-ki.uni-osnabrueck.de/~koks>

Institut für Kognitionswissenschaft
Katharinenstr. 24
49078 Osnabrück

Technical Information

Technical Information

- Linux, MySQL DB

Technical Information

- Linux, MySQL DB
- scripting languages: Python, Perl, Ruby

Technical Information

- Linux, MySQL DB
- scripting languages: Python, Perl, Ruby
- Java applet

Technical Information

- Linux, MySQL DB
- scripting languages: Python, Perl, Ruby
- Java applet
- POS tagger: IMS Stuttgart

Technical Information

- Linux, MySQL DB
- scripting languages: Python, Perl, Ruby
- Java applet
- POS tagger: IMS Stuttgart
- dictionaries from WWW, results from institute

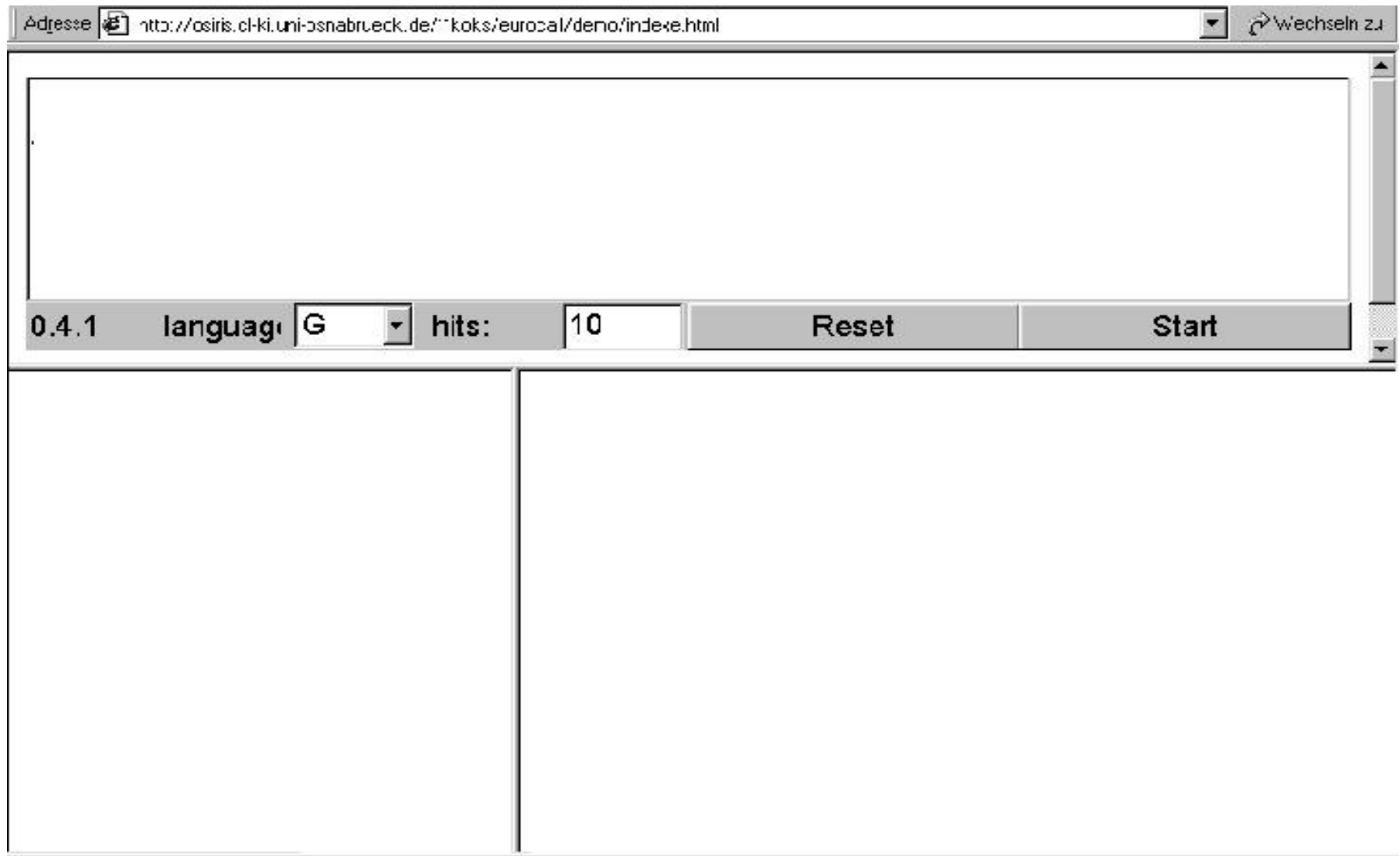
Technical Information

- Linux, MySQL DB
- scripting languages: Python, Perl, Ruby
- Java applet
- POS tagger: IMS Stuttgart
- dictionaries from WWW, results from institute
- corpora: mostly from WWW (EU, Nato, DE-News)

Technical Information

- Linux, MySQL DB
- scripting languages: Python, Perl, Ruby
- Java applet
- POS tagger: IMS Stuttgart
- dictionaries from WWW, results from institute
- corpora: mostly from WWW (EU, Nato, DE-News)
- back

Screenshot 1





Screenshot 2

The screenshot shows a web browser window with the following elements:

- Address Bar:** Displays the URL `http://osiris.ch-ki.uni-bonnabueck.de/~koks/eurocal/deno/indexe.html` and a "Wechseln zu" (Change to) button.
- Search Results:** A text box containing the German sentence: "Leider war Herr Krause krank geworden. Der Ausflug fiel ins Wasser. Aller waren darueber sehr traurig." The word "fiel" is highlighted in black.
- Search Controls:** A horizontal bar containing:
 - The text "0.4.1" on the left.
 - A "language" dropdown menu currently set to "G".
 - A "hits:" label followed by a text input field containing the number "10".
 - A "Reset" button.
 - A "Start" button.
- Output Area:** Two large empty rectangular boxes at the bottom of the page. The left one contains the text "Found phrases will appear here." and the right one contains "Examples with translations will appear here."

Screenshot 3

Adresse  http://osiris.cl-ki.uni-osnabrueck.de/~koks/eurocal/deno/indexe.html  [Wechseln zu](#)

Leider war Herr Krause krank geworden. **Der Ausflug fiel ins Wasser.**
Alle waren darüber sehr traurig.

0.4.1 language: hits:

phrases found:

according to examples

- fiel ins Wasser




according to the dictionary

- fallen
- fallen
- fallen
- fallen

count of hits:

Examples with translations will appear here.

Screenshot 4

Adresse  http://osiris.ch-ki.uni-bonnabueck.de/~kok/s/eurocal/demo/indexe.html   Wechseln zu

Leider war Herr Krause krank geworden. **Der Ausflug fiel ins Wasser.**
Alle waren darüber sehr traurig.

0.4.1 language: **G** hits: **10** **Reset** **Start**

phrases found:

according to examples

- fiel ins Wasser

according to the dictionary

- fallen
- Fallen
- fallen

count of hits: **Ink ar it**

translations found:

german: Die erste Berlinparade der Skater in diesem Jahr fiel ins Wasser . Die Veranstalter sagten wegen des schlechten Wetters die Skater-Demonstration ab .
english: Due to bad weather the first parade of skaters of this year has been cancelled .

german: Die Silvesterparty zum Jahrtausendwechsel fiel ins Wasser , da viele Beschäftigte wegen der befürchteten Computerprobleme Bereitschaft oder Präsenzzeiten hatten .
english: The New Year #s Eve party was called off . Many employees were on standby or had to be at their workplace because computer problems had been expected at the dawn of the new millenium .

Screenshots

back