A Survey of Idiomatic Preposition-Noun-Verb Triples on Token Level

Fabienne Fritzinger, Marion Weller and Ulrich Heid

University of Stuttgart
Institute for Natural Language Processing
– Computational Linguistics –
Azenbergstr. 12
D 70174 Stuttgart
[fritzife, wellermn, heid]@ims.uni-stuttgart.de

Motivation

In previous work, we focused on the identification of idiomatic multiword expression (MWE) **types**, like e.g.:

MWE	Idiomatic meaning	Literal meaning
in Leben rufen	to initiate	"to call to life"
unter Teppich kehren	to hide/conceal	"to sweep under carpet"
auf Kopf stellen	to turn sth. inside out	"to place sth. on head"
(sich) aus Staub machen	to leave	"to make oneself out of the dust"

Motivation

In previous work, we focused on the identification of idiomatic multiword expression (MWE) **types**, like e.g.:

MWE	Idiomatic meaning	Literal meaning
in Leben rufen	to initiate	"to call to life"
unter Teppich kehren	to hide/conceal	"to sweep under carpet"
auf Kopf stellen	to turn sth. inside out	"to place sth. on head"
(sich) aus Staub machen	to leave	"to make oneself out of the dust"

... but if we have a closer look at them, we see that

Motivation

In previous work, we focused on the identification of idiomatic multiword expression (MWE) **types**, like e.g.:

MWE	Idiomatic meaning	Literal meaning
in Leben rufen	to initiate	"to call to life"
unter Teppich kehren	to hide/conceal	"to sweep under carpet"
auf Kopf stellen	to turn sth. inside out	"to place sth. on head"
(sich) aus Staub machen	to leave	"to make oneself out of the dust"

... but if we have a closer look at them, we see that
→ some of them can only have an idiomatic meaning

Motivation

In previous work, we focused on the identification of idiomatic multiword expression (MWE) **types**, like e.g.:

MWE	Idiomatic meaning	Literal meaning
in Leben rufen	to initiate	"to call to life"
unter Teppich kehren	to hide/conceal	"to sweep under carpet"
auf Kopf stellen	to turn sth. inside out	"to place sth. on head"
(sich) aus Staub machen	to leave	"to make oneself out of the dust"

... but if we have a closer look at them, we see that
→ some of them can only have an idiomatic meaning
→ while others could also have a literal meaning (theoretically)

Motivation

In previous work, we focused on the identification of idiomatic multiword expression (MWE) **types**, like e.g.:

MWE	Idiomatic meaning	Literal meaning
in Leben rufen	to initiate	"to call to life"
unter Teppich kehren	to hide/conceal	"to sweep under carpet"
auf Kopf stellen	to turn sth. inside out	"to place sth. on head"
(sich) aus Staub machen	to leave	"to make oneself out of the dust"

→ Tools to identify idiomatic MWEs shall be aware of the actual meaning

Motivation

In previous work, we focused on the identification of idiomatic multiword expression (MWE) **types**, like e.g.:

MWE	Idiomatic meaning	Literal meaning
in Leben rufen	to initiate	"to call to life"
unter Teppich kehren	to hide/conceal	"to sweep under carpet"
auf Kopf stellen	to turn sth. inside out	"to place sth. on head"
(sich) aus Staub machen	to leave	"to make oneself out of the dust"

→ Tools to identify idiomatic MWEs shall be aware of the actual meaning
→ We do **not** present a method of how to deal with this!

Motivation

In previous work, we focused on the identification of idiomatic multiword expression (MWE) **types**, like e.g.:

MWE	Idiomatic meaning	Literal meaning
in Leben rufen	to initiate	"to call to life"
unter Teppich kehren	to hide/conceal	"to sweep under carpet"
auf Kopf stellen	to turn sth. inside out	"to place sth. on head"
(sich) aus Staub machen	to leave	"to make oneself out of the dust"

→ Tools to identify idiomatic MWEs shall be aware of the actual meaning
→ We do **not** present a method of how to deal with this!
→ **Instead**, we give an impression of the quantitative dimension of the problem:

Motivation

In previous work, we focused on the identification of idiomatic multiword expression (MWE) **types**, like e.g.:

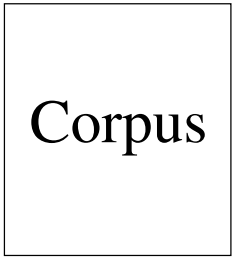
MWE	Idiomatic meaning	Literal meaning
in Leben rufen	to initiate	"to call to life"
unter Teppich kehren	to hide/conceal	"to sweep under carpet"
auf Kopf stellen	to turn sth. inside out	"to place sth. on head"
(sich) aus Staub machen	to leave	"to make oneself out of the dust"

→ Tools to identify idiomatic MWEs shall be aware of the actual meaning
→ We do **not** present a method of how to deal with this!
→ **Instead**, we give an impression of the quantitative dimension of the problem:

we **manually** annotated a huge dataset

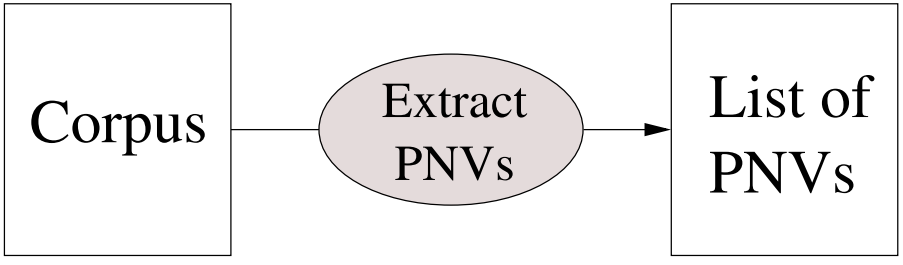
Preprocessing

Extraction of MWEs to be annotated



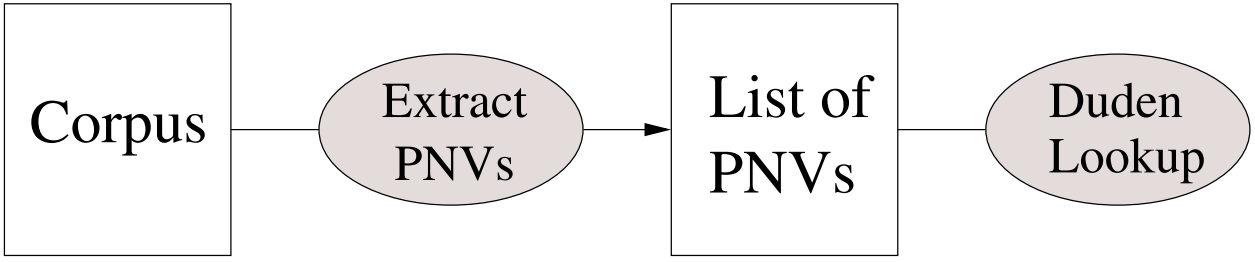
Preprocessing

Extraction of MWEs to be annotated



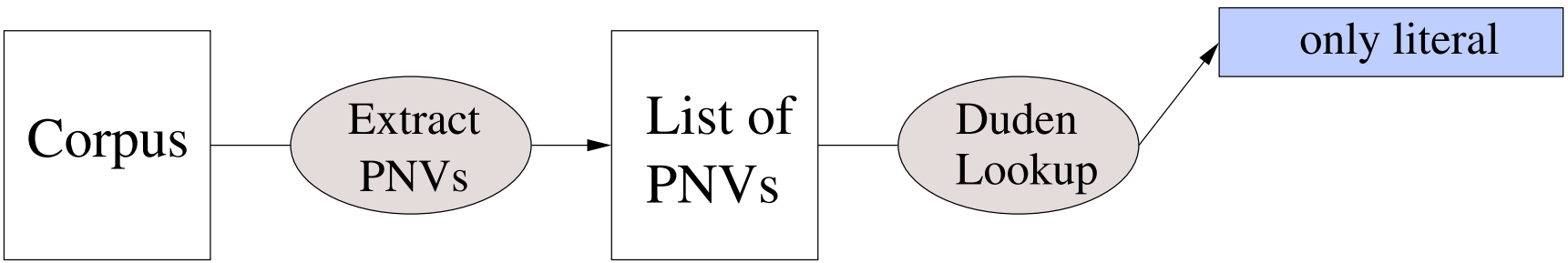
Preprocessing

Extraction of MWEs to be annotated



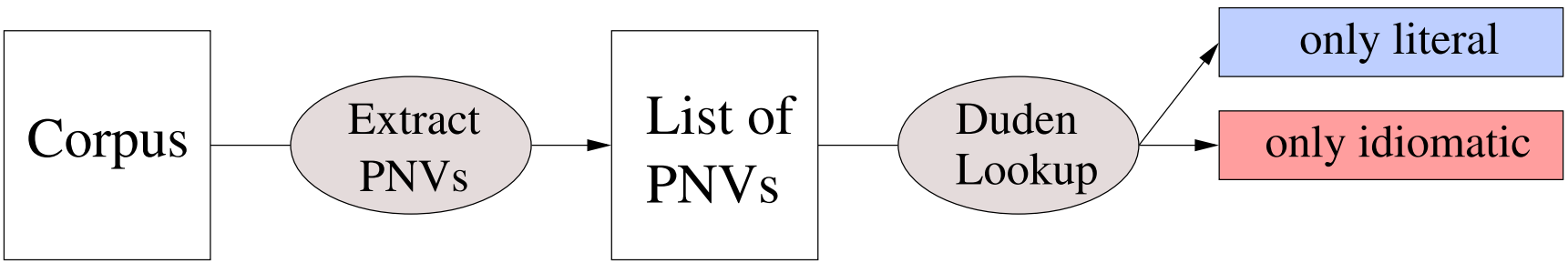
Preprocessing

Extraction of MWEs to be annotated



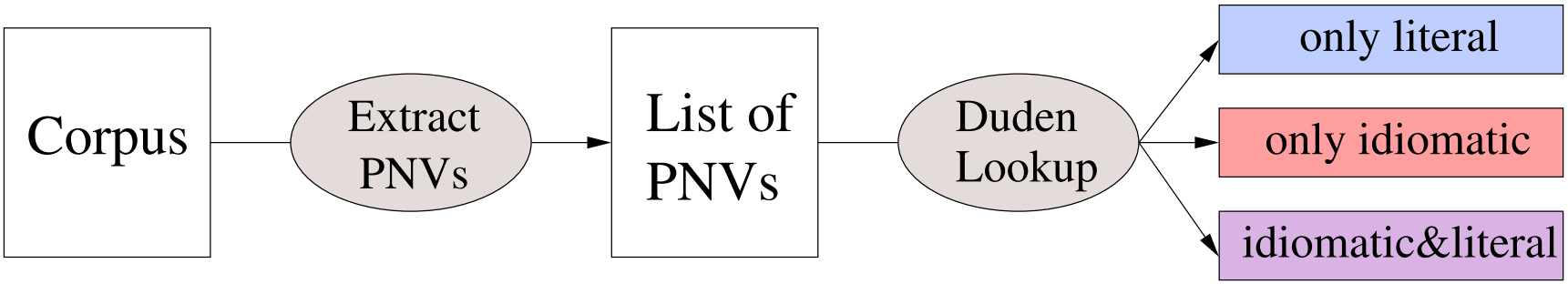
Preprocessing

Extraction of MWEs to be annotated



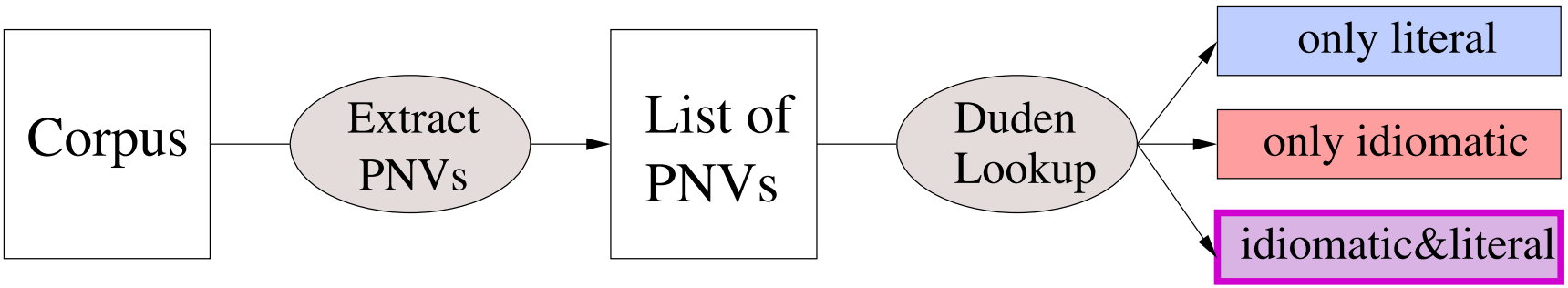
Preprocessing

Extraction of MWEs to be annotated



Preprocessing

Extraction of MWEs to be annotated



Preprocessing

Data and Tools

Data:

corpus	size	years
Frankfurter Allgemeine Zeitung	70Mio	97/98
EUROPARL	35Mio	96-06

→ Assumption: literal instances in newspaper,
but more rarely in EUROPARL

Preprocessing

Data and Tools

Data:

corpus	size	years
Frankfurter Allgemeine Zeitung	70Mio	97/98
EUROPARL	35Mio	96-06

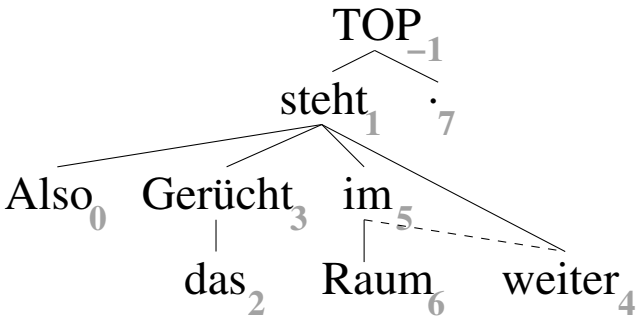
→ Assumption: literal instances in newspaper,
but more rarely in EUROPARL

Tools:

- FSPAR, German dependency parser
- PERL scripts

Extraction of Preposition-Noun-Verb Triples

Fspar, (Schiehlen, 2003)



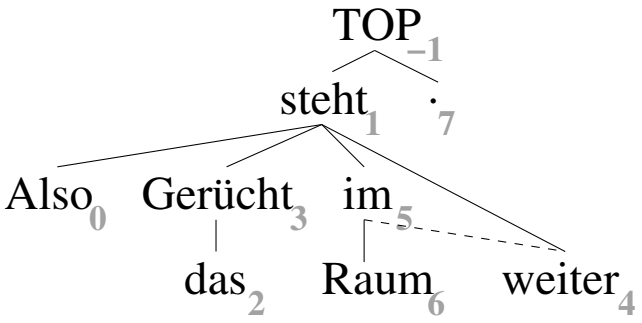
PNV LEMMA	DET	FUS	NUM
in Raum stehen	def	+	Sg

0	Also	ADV	also	|	1	ADJ
1	steht	VVFIN	stehen	3:Sg:Pres:Ind*	-1	TOP
2	das	ART	d	|	3	SPEC
3	Gerücht	NN	Gerücht	Nom:N:Sg	1	NP:1
4	weiter	ADV	weiter	|	1||5	ADJ
5	im	APPRART	in	Dat:M:Sg	1	ADJ
6	Raum	NN	Raum	Dat:M:Sg	5	PCMP
7	.	$.	.	|	-1	TOP

Thus, the rumour is still to be dealt with.

Extraction of Preposition-Noun-Verb Triples

Fspar, (Schiehlen, 2003)



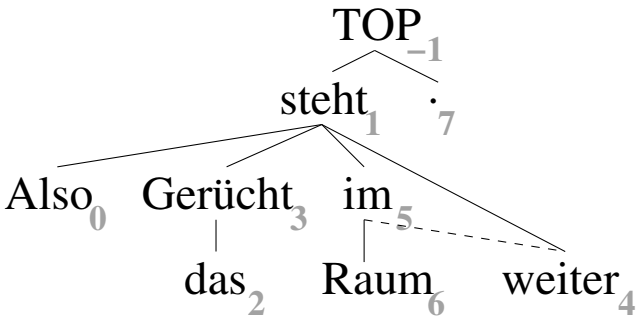
PNV LEMMA	DET	FUS	NUM
in Raum stehen	def	+	Sg

0	Also	ADV	also	|	1	ADJ
1	steht	VVFIN	stehen	3:Sg:Pres:Ind*	-1	TOP
2	das	ART	d	|	3	SPEC
3	Gerücht	NN	Gerücht	Nom:N:Sg	1	NP:1
4	weiter	ADV	weiter	|	1||5	ADJ
5	im	APPRART	in	Dat:M:Sg	1	ADJ
6	Raum	NN	Raum	Dat:M:Sg	5	PCMP
7	.	$.	.	|	-1	TOP

Thus, the rumour is still to be dealt with.

Extraction of Preposition-Noun-Verb Triples

Fspar, (Schiehlen, 2003)



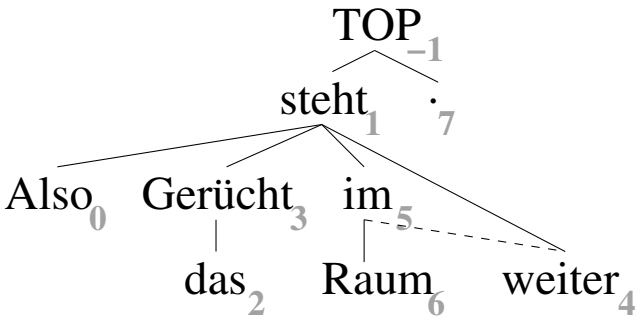
PNV LEMMA	DET	FUS	NUM
in Raum stehen	def	+	Sg

0	Also	ADV	also	|	1	ADJ
1	steht	VVFIN	stehen	3:Sg:Pres:Ind*	-1	TOP
2	das	ART	d	|	3	SPEC
3	Gerücht	NN	Gerücht	Nom:N:Sg	1	NP:1
4	weiter	ADV	weiter	|	1||5	ADJ
5	im	APPRART	in	Dat:M:Sg	1	ADJ
6	Raum	NN	Raum	Dat:M:Sg	5	PCMP
7	.	$.	.	|	-1	TOP

Thus, the rumour is still to be dealt with.

Extraction of Preposition-Noun-Verb Triples

Fspar, (Schiehlen, 2003)

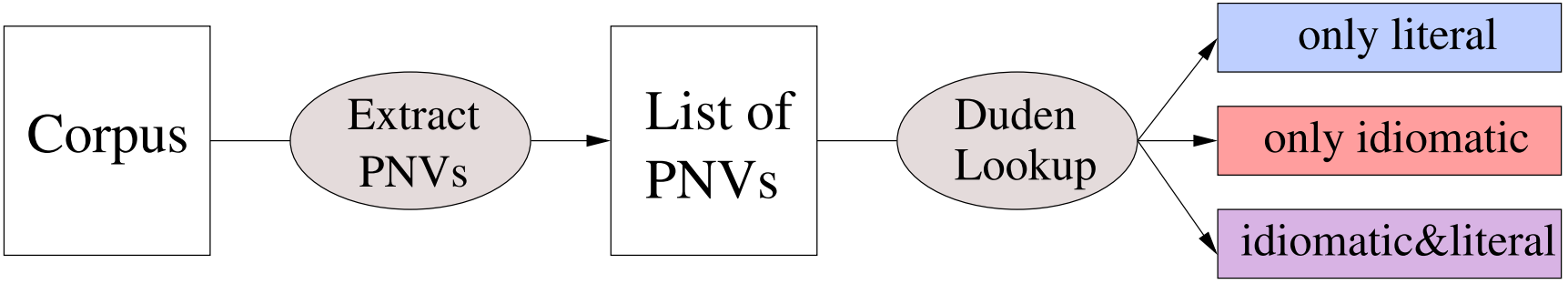


PNV LEMMA	DET	FUS	NUM
in Raum stehen	def	+	Sg

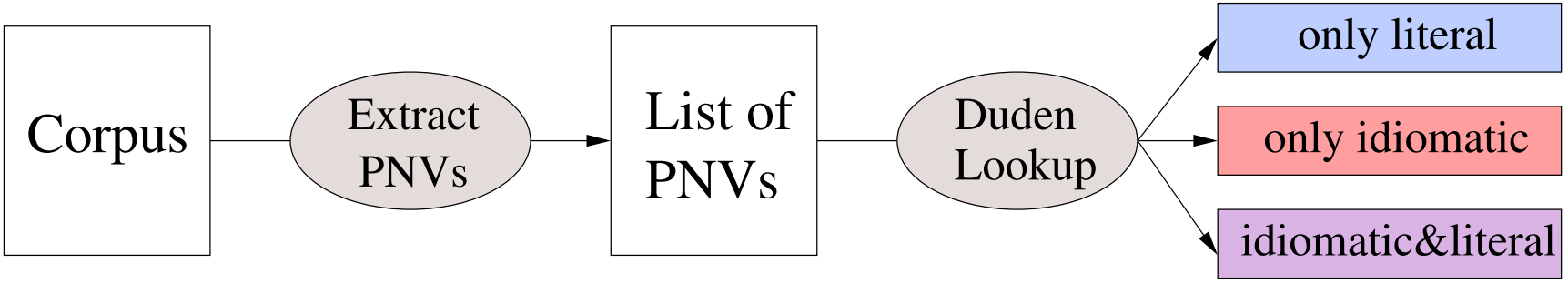
0	Also	ADV	also	|	1	ADJ
1	steht	VVFIN	stehen	3:Sg:Pres:Ind*	-1	TOP
2	das	ART	d	|	3	SPEC
3	Gerücht	NN	Gerücht	Nom:N:Sg	1	NP:1
4	weiter	ADV	weiter	|	1||5	ADJ
5	im	APPRART	in	Dat:M:Sg	1	ADJ
6	Raum	NN	Raum	Dat:M:Sg	5	PCMP
7	.	$.	.	|	-1	TOP

Thus, the rumour is still to be dealt with.

Duden Lookup



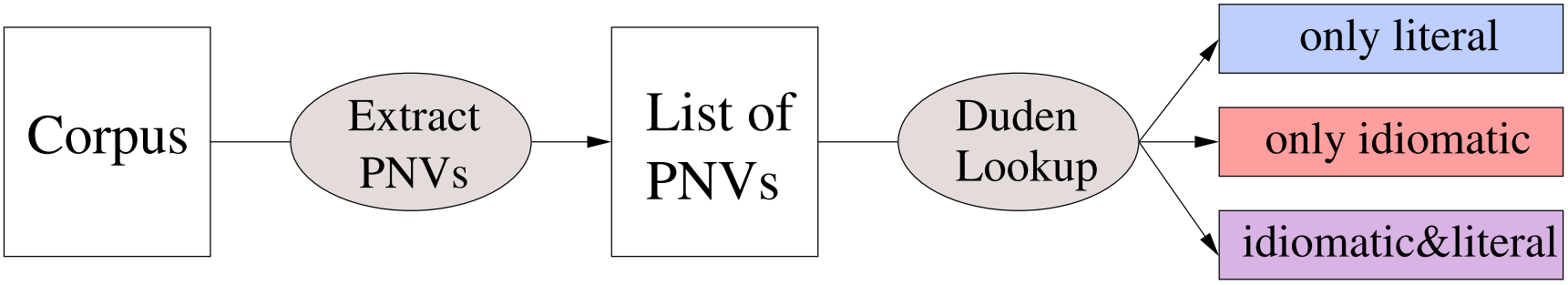
Duden Lookup



Idiomatic PNVs amongst 1,000 most frequent:

Corpus	in *Duden*
FAZ	155
EUROPARL	108
Total	196

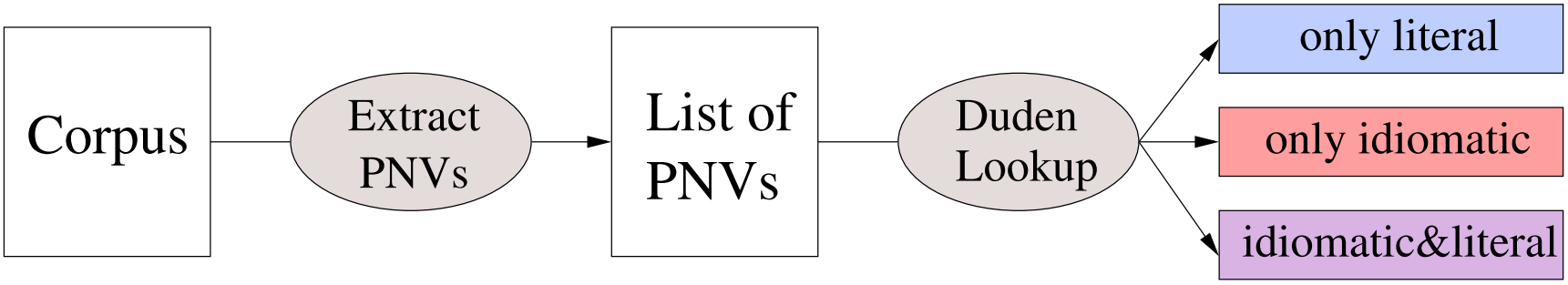
Duden Lookup



Idiomatic PNVs amongst 1,000 most frequent:

Corpus	in *Duden*	only idiom.	idiom.&lit.
FAZ	155	86	69
EUROPARL	108	73	35
Total	196	119	77

Duden Lookup

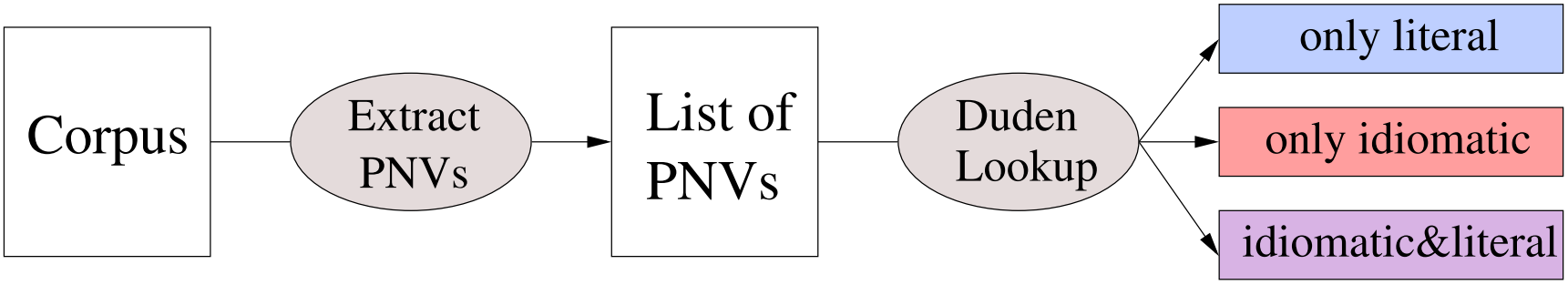


Idiomatic PNVs amongst 1,000 most frequent:

Corpus	in *Duden*	only idiom.	idiom.&lit.
FAZ	155	86	69
EUROPARL	108	73	35
Total	196	119	77

ins Leben rufen ("to call to life", to create sth.)

Duden Lookup



Idiomatic PNVs amongst 1,000 most frequent:

Corpus	in *Duden*	only idiom.	idiom.&lit.
FAZ	155	86	69
EUROPARL	108	73	35
Total	196	119	**77**

ins Leben rufen ("to call to life", to create sth.)

auf Schlauch stehen ("to stand on a hose", to have a mental block)

Manual Annotation

- 100 sentences for each PNV-Triple, randomly extracted
- annotated by two German native speaker (independently)
  → agreement in 6,550 of 6,690 cases of FAZ (97.9%)
  → remaining 140 discussed with a third native speaker

corpus	all
FAZ	6,690
EUROPARL	3,050
total	**9,740**

Manual Annotation

- 100 sentences for each PNV-Triple, randomly extracted
- annotated by two German native speaker (independently)
  → agreement in 6,550 of 6,690 cases of FAZ (97.9%)
  → remaining 140 discussed with a third native speaker
- annotated categories: idiomatic (I), literal (L), ambiguous (A), parsing/extraction error (X).

corpus	all	I	L	A	X
FAZ	6,690	6,176	345	75	94
EUROPARL	3,050	2,937	31	14	68
total	**9,740**	9,113	376	89	162

Selected Examples

PNV	FAZ			EUROPARL		
	L	I	A	L	I	A
auf Seite stehen to stand on (so.'s) side	16	75	3	12	86	1
zu Einsatz kommen to come into operation	1	97	1	1	98	0
zu Fall bringen to bring down sth./sb.	17	82	1	0	47	0

Selected Examples

PNV	FAZ			EUROPARL		
	L	I	A	L	I	A
auf Seite stehen to stand on (so.'s) side	16	75	3	12	86	1
zu Einsatz kommen to come into operation	1	97	1	1	98	0
zu Fall bringen to bring down sth./sb.	17	82	1	0	47	0

Literal instance (FAZ):
Emmerling hatte im Strafraum Benedyk zu Fall gebracht .
"Emmerling had brought down Benedyk in the penalty area "

Selected Examples

PNV	FAZ			EUROPARL		
	L	I	A	L	I	A
auf Seite stehen to stand on (so.'s) side	16	75	3	12	86	1
zu Einsatz kommen to come into operation	1	97	1	1	98	0
zu Fall bringen to bring down sth./sb.	17	82	1	0	47	0

Literal instance (FAZ):
Emmerling hatte im Strafraum Benedyk zu Fall gebracht .
"Emmerling had brought down Benedyk in the penalty area "

Idiomatic instance (EP):
Der Versuch der Tabaklobby , dieses Gesetz zu Fall zu bringen , ist absurd .
"The attempt of the tabacco lobby to kill the law is absurd.

Distribution of Morpho-Syntactic Features

Example taken from FAZ

in Raum stehen

	#PNVs	DET			FUS		NUM	
		no	def	quant	+	-	sg	pl
literal	33	2	24	1	19	14	28	5
idiomatic	53	0	53	0	53	0	53	0
ambiguous	2	0	2	0	2	0	2	0

Distribution of Morpho-Syntactic Features

Example taken from FAZ

in Raum stehen

	#PNVs	DET			FUS		NUM	
		no	def	quant	+	-	sg	pl
literal	33	2	24	1	19	14	28	5
idiomatic	53	0	53	0	53	0	53	0
ambiguous	2	0	2	0	2	0	2	0

Literal use

In *den besseren* **Räumen standen** *Öfen, an denen er sich wärmte.*
"In the superior rooms stood stoves, at which he himself warmed."
In the superior rooms were stoves at which he warmed himself.

Distribution of Morpho-Syntactic Features

Example taken from FAZ

in Raum stehen

	#PNVs	DET			FUS		NUM	
		no	def	quant	+	-	sg	pl
literal	33	2	24	1	19	14	28	5
idiomatic	53	0	53	0	53	0	53	0
ambiguous	2	0	2	0	2	0	2	0

Idiomatic use

Widersprüche **stehen** *ungelöst* **im Raum**.
"Contradictions stand unresolved in the room"
There are unresolved contradictions to be dealt with.

Distribution of Morpho-Syntactic Features

Example taken from FAZ

in Raum stehen

	#PNVs	DET			FUS		NUM	
		no	def	quant	+	-	sg	pl
literal	33	2	24	1	19	14	28	5
idiomatic	53	0	53	0	53	0	53	0
ambiguous	2	0	2	0	2	0	2	0

Ambiguous use

Sie **steht** *als Manifest einer neuen Wohnkultur* **im Raum** *und macht heute jeden Besucher zum potentiellen Bewohner.*
"It stands as manifest of a new living home decor in the room and makes today every visitor to a potential resident."

Conclusion

- We presented a dataset of 9,700 manually analysed instances of idiomatic preposition-noun-verb triples, provided along with their morpho-syntactic properties.

Conclusion

- We presented a dataset of 9,700 manually analysed instances of idiomatic preposition-noun-verb triples, provided along with their morpho-syntactic properties.
- We intend to make this dataset available to the research community.

Conclusion

- We presented a dataset of 9,700 manually analysed instances of idiomatic preposition-noun-verb triples, provided along with their morpho-syntactic properties.
- We intend to make this dataset available to the research community.
- It can be used to train supervised classification approaches (cf. Diab and Bhutada, 2009)

Conclusion

- We presented a dataset of 9,700 manually analysed instances of idiomatic preposition-noun-verb triples, provided along with their morpho-syntactic properties.
- We intend to make this dataset available to the research community.
- It can be used to train supervised classification approaches (cf. Diab and Bhutada, 2009)
- Similar resources available for English (cf. Cook, 2008; Sporleder, 2010)

Conclusion

- We presented a dataset of 9,700 manually analysed instances of idiomatic preposition-noun-verb triples, provided along with their morpho-syntactic properties.
- We intend to make this dataset available to the research community.
- It can be used to train supervised classification approaches (cf. Diab and Bhutada, 2009)
- Similar resources available for English (cf. Cook, 2008; Sporleder, 2010)

$\rightarrow$ but no similar resource for German!