Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case

Kremena Ivanova*, Ulrich Heid*, Sabine Schulte im Walde*,
Adam Kilgarriff°, Jan Pomikálek°▷

*Institute for Natural Language Processing, University of Stuttgart, Germany
°Lexical Computing Ltd, Brighton, UK
▷Masaryk University, Brno, Czech Republic

{ivanovka,heid,schulte}@ims.uni-stuttgart.de,
adam@lexmasterclass.com, xpomikal@fi.muni.cz

Marrakech, Morocco, May 28, 2008

The Sketch Engine (Kilgarriff et al. 2004)

A system for corpus exploration

- Input: preprocessed corpora,
  e.g. tokenized, POS-tagged, lemmatized , . . .

The Sketch Engine (Kilgarriff et al. 2004)

A system for corpus exploration

- Input: preprocessed corpora,
  e.g. tokenized, POS-tagged, lemmatized , . . .
- Functions:
  - concordancing
  - collocation extraction with a *sketch grammar*, i.e.
    a set of regular expression search patterns over the corpus

The Sketch Engine (Kilgarriff et al. 2004)

A system for corpus exploration

- Input: preprocessed corpora,
  e.g. tokenized, POS-tagged, lemmatized , . . .
- Functions:
  - concordancing
  - collocation extraction with a *sketch grammar*, i.e.
    a set of regular expression search patterns over the corpus
- Output: *Word sketches*
  Sets of significant word pairs, grouped by grammatical relations, e.g.
  adjective + noun, verb + subject noun, coordinated elements, etc.

The Sketch Engine – word sketches

A sample *word sketch*: collection of cooccurrence data

Node word + 'collocates':

Word sketch for verb *öffnen* 'open':

Lemma of cooccurrence partner – frequency (in BNC) – significance

subj	3017	5.1	**obj-acc**	282	5.9	**adv**	140	5.2
Tür	238	49.37	*Tür*	39	36.24	*täglich*	12	22.68
Pforte	35	35.20	*Auge*	26	26.67	*versehentlich*	3	16.92
Türe	29	33.78	*Pforte*	7	22.71	*leicht*	6	13.89
Tor	62	32.34	*Wohnungstür*	3	21.61	*weit*	13	13.61
Auge	114	32.29	*Türe*	5	19.38	*gleichzeitig*	4	12.37
Fenster	49	28.69	*Datei*	4	12.23	*automatisch*	3	11.42
Schleuse	10	23.27	*Tor*	4	11.7			

Source: *DeWaC*, 10 million words

Sketch Grammars

Regular expression-based: sequence patterns

Sketch Grammars

Regular expression-based: sequence patterns

Example: POS sequences

- Adjective + Noun combination: `2:[tag="ADJA"] 1:[tag=NN"]`

Sketch Grammars

Regular expression-based: sequence patterns

Example: POS sequences

- Adjective + Noun combination: `2:[tag="ADJA"] 1:[tag=NN"]`
  - finds sequences adjective + noun

Sketch Grammars

Regular expression-based: sequence patterns

Example: POS sequences

- Adjective + Noun combination: `2:[tag="ADJA"] 1:[tag=NN"]`
  - finds sequences adjective + noun
  - counts frequency, calculates significance

Sketch Grammars

Regular expression-based: sequence patterns

Example: POS sequences

- Adjective + Noun combination: `2:[tag="ADJA"] 1:[tag=NN"]`
  - finds sequences adjective + noun
  - counts frequency, calculates significance
  - allows for display of pair in
    - list of adjective collocates of a given noun (`1:...`), e.g. *Dorf*

Modifying adjectives		Freq	Sign
klein	'small'	274	37.68
umliegend	'surrounding'	39	37.30
malerisch	'picturesque'	20	28.96
entlegen	'remote'	16	28.58

Sketch Grammars

Regular expression-based: sequence patterns

Example: POS sequences

- Adjective + Noun combination: `2:[tag="ADJA"] 1:[tag=NN"]`
  - finds sequences adjective + noun
  - counts frequency, calculates significance
  - allows for display of pair in
    * list of adjective collocates of a given noun (`1:...`), e.g. *Dorf*

Modifying adjectives		Freq	Sign
klein	'small'	274	37.68
umliegend	'surrounding'	39	37.30
malerisch	'picturesque'	20	28.96
entlegen	'remote'	16	28.58

    * list of noun nodes of a given adjective (`2:...`), e.g. *klein*

Modified nouns		Freq	Sign
Ausschnitt	'extract'	188	37.49
Junge	'boy'	325	33.91
Dorf	'village'	274	32.80
Meerjungfrau	'mermaid'	46	31.19

Sketch Grammars

Regular expression-based: sequence patterns

Example: POS sequences

- Adjective + Noun combination: `2:[tag="ADJA"] 1:[tag=NN"]`
  - finds sequences adjective + noun
  - counts frequency, calculates significance
  - allows for display of pair in
    - list of adjective collocates of a given noun (`1:...`), e.g. *Dorf*

Modifying adjectives		Freq	Sign
klein	'small'	274	37.68
umliegend	'surrounding'	39	37.30
malerisch	'picturesque'	20	28.96
entlegen	'remote'	16	28.58

    - list of noun nodes of a given adjective (`2:...`), e.g. *klein*

Modified nouns		Freq	Sign
Ausschnitt	'extract'	188	37.49
Junge	'boy'	325	33.91
Dorf	'village'	274	32.80
Meerjungfrau	'mermaid'	46	31.19

- Simple model of a noun phrase as a POS sequence:
  `DET? ADV* ADJA* NOUN`

Sketch Grammars

Identifying grammatical relations, e.g. verb + object noun

Sketch Grammars

Identifying grammatical relations, e.g. verb + object noun

- EN (configurational): by position wrt the verb:
  Subject $<$ Verb $<$ Object (Kilgarriff et al. 2004)

Sketch Grammars

Identifying grammatical relations, e.g. verb + object noun

- EN (configurational): by position wrt the verb:
  Subject $<$ Verb $<$ Object (Kilgarriff et al. 2004)
- CHI: by position and particles (Kilgarriff 2005)

Sketch Grammars

Identifying grammatical relations, e.g. verb + object noun

- EN (configurational): by position wrt the verb:
  Subject $<$ Verb $<$ Object (Kilgarriff et al. 2004)
- CHI: by position and particles (Kilgarriff 2005)
- CZ, SLO (inflecting): by inflectional affixes:
  SLO *lépa híša* ("beautiful house"): NOM-SG
  lépi híši: DAT-SG | LOC-SG (+ Prep.) (Kilgarriff et al. 2004, Krek/Kilgarriff 2006)

Sketch Grammars

Identifying grammatical relations in German texts

Sketch Grammars

Identifying grammatical relations in German texts

- not via word order:
  den Mitarbeiter$_{\text{Acc}}$ *lobt der Chef*$_{\text{Nom}}$
  ("the boss speaks highly of the collaborator")
  Constituent order is relatively free in German

Sketch Grammars

Identifying grammatical relations in German texts

- not via word order:
  den Mitarbeiter$_{\text{Acc}}$ *lobt der Chef*$_{\text{Nom}}$
  ("the boss speaks highly of the collaborator")
  Constituent order is relatively free in German
- not often via inflection:
  Hans$_{\text{Nom/Acc}}$ *lobt Maria*$_{\text{Nom/Acc}}$
  weil der Chef$_{\text{Acc}}$ *der Firma*$_{\text{Gen/Dat}}$ *in Berlin*$_{\text{PP}}$ *empfahl, . . . zu . . .*
  Only ca. 21 % of all NPs are unambiguous wrt case (Evert 2004)

Sketch Grammars

Identifying grammatical relations in German texts

- not via word order:
  den Mitarbeiter$_{\text{Acc}}$ *lobt der Chef*$_{\text{Nom}}$
  ("the boss speaks highly of the collaborator")
  Constituent order is relatively free in German
- not often via inflection:
  Hans$_{\text{Nom/Acc}}$ *lobt Maria*$_{\text{Nom/Acc}}$
  weil der Chef$_{\text{Acc}}$ *der Firma*$_{\text{Gen/Dat}}$ *in Berlin*$_{\text{PP}}$ *empfahl, . . . zu . . .*
  Only ca. 21 % of all NPs are unambiguous wrt case (Evert 2004)

$\Rightarrow$ harder than in other languages

A Sketch Grammar for German

Knowledge for the identification of grammatical relations

1. {gender, number, case} of nouns $\leftrightarrow$ inflectional affixes

A Sketch Grammar for German

Knowledge for the identification of grammatical relations

1. {gender, number, case} of nouns $\leftrightarrow$ inflectional affixes
2. Preferential constituent ordering:
   verb-final constituent order model is more regular than others

A Sketch Grammar for German

Knowledge for the identification of grammatical relations

1. {gender, number, case} of nouns $\leftrightarrow$ inflectional affixes
2. Preferential constituent ordering:
   verb-final constituent order model is more regular than others
3. Constraints on subcategorization patterns, e.g.
   'No two identical grammatical functions in one sentence'
   (cf. 'coherence' in LFG)

A Sketch Grammar for German

Proportion between preprocessing (offline) and query (online)

1. Gender, number, case:
   not annotated: STTS: `"NN"` (UPenn: `"NNS"` – `"NNP"`)
   $\longrightarrow$ Need to identify these within the sketch grammar
2. Preferential constituent ordering under V-final:
   $\longrightarrow$ Search in a subset of the corpus sentences
3. Constraints on subcategorization patterns:
   $\longrightarrow$ Implementation as patterns in the sketch grammar

A Sketch Grammar for German

Proportion between preprocessing (offline) and query (online)

1. Gender, number, case:
   not annotated: STTS: "NN" (UPenn: "NNS" – "NNP")
   → Need to identify these within the sketch grammar
2. Preferential constituent ordering under V-final:
   → Search in a subset of the corpus sentences
3. Constraints on subcategorization patterns:
   → Implementation as patterns in the sketch grammar

⇒ To assess usefulness of these types of information:
Different versions of the sketch grammar
which include the different types of information

A Sketch Grammar for German

Versions of the grammar with different types of information (1/2)

Conditions for the evaluation

Morphological restrictions: alternatives

A Sketch Grammar for German

Versions of the grammar with different types of information (1/2)

Conditions for the evaluation

Morphological restrictions: alternatives

- *inflection*:
  case guessing from the form of affixes (affix sequences)
  dem_{Dat} $kleinen_{\text{Dat}}$ $Haus_{\text{Nom/Dat/Acc}}$

A Sketch Grammar for German

Versions of the grammar with different types of information (1/2)

Conditions for the evaluation

Morphological restrictions: alternatives

- *inflection*:
  case guessing from the form of affixes (affix sequences)
  dem_{Dat} $kleinen_{\text{Dat}}$ $Haus_{\text{Nom/Dat/Acc}}$
- *affix-gender*:
  case and gender guessing
  from derivational affixes and inflectional affixes
  $den_{\text{ACC-SG-MASC/DAT-PL-FEM}}$ $Schwierigkeiten_{\text{ANY-PL-FEM}}$
  $\Rightarrow$ subset of nouns with known agreement properties

A Sketch Grammar for German

Versions of the grammar with different types of information (2/2)

Conditions for the evaluation

Structural restrictions: alternatives

A Sketch Grammar for German

Versions of the grammar with different types of information (2/2)

Conditions for the evaluation

Structural restrictions: alternatives

- *no-structure(-constraints)*:
  extraction without any structural constraints

A Sketch Grammar for German

Versions of the grammar with different types of information (2/2)

Conditions for the evaluation

Structural restrictions: alternatives

- *no-structure(-constraints)*:
  extraction without any structural constraints
- *verb-final*:
  extraction only from verb-final sentences (= subclauses),
  according to constraints on subcategorization patterns

A Sketch Grammar for German

Versions of the grammar with different types of information (2/2)

Conditions for the evaluation

Structural restrictions: alternatives

- *no-structure(-constraints)*:
  extraction without any structural constraints
- *verb-final*:
  extraction only from verb-final sentences (= subclauses),
  according to constraints on subcategorization patterns
- *all-clauses*:
  extraction from an explicit model of all verb position models
  (V1, V2, Vlast), according to subcategorization patterns

Evaluation: comparing versions of the Sketch Grammar

Combining the restrictions

no affix-gender		no structure
	×	verb-final (R)
with affix-gender (R)		all-clauses (R)

inflection = minimum knowledge

(1) inflection + no-structure
(2) inflection + affix-gender + no-structure
(3) inflection + verb-final
(4) inflection + affix-gender + verb-final
(5) inflection + all-clauses
(6) inflection + affix-gender + all-clauses

- fewest restrictions (R)
- structural restrictions (R)
- most restr. (R)

Evaluation: comparing versions of the Sketch Grammar

Gold standard corpus

- 1000 randomly selected sentences from DeWaC

Evaluation: comparing versions of the Sketch Grammar

Gold standard corpus

- 1000 randomly selected sentences from DeWaC
- Manual annotation for NP (one annotator):
  - start and end point
  - case
- Example:
  [Ich]$_{\text{NPnom}}$ *musste [meine Arbeit]*$_{\text{NPakk}}$ *schon sehr gut machen, um anerkannt zu werden .*
  'I had to do my work really well to be approved.'

Evaluation: comparing versions of the Sketch Grammar

Gold standard corpus

- 1000 randomly selected sentences from DeWaC
- Manual annotation for NP (one annotator):
  - start and end point
  - case
- Example:
  [Ich]$_{\text{NPnom}}$ *musste [meine Arbeit]*$_{\text{NPakk}}$ *schon sehr gut machen, um anerkannt zu werden .*
  'I had to do my work really well to be approved.'
- Figures: NPs in the 1000 sentences

Nominative	1.709
Genitive	437
Dative	149
Accusative	618

Evaluation: comparing versions of the Sketch Grammar

Results: recall and precision

Evaluated per case and per condition:

Exception: Genitive not implemented under conditions 3 + 4:

No verb with genitive object in the corpus, we only consider genitives in NPs

Case	*N*	Conditions											
		incl. *inflection*						incl. *inflection + affix-gender*					
		1		3		5		2		4		6	
		R	P	R	P	R	P	R	P	R	P	R	P
Nominative	1,709	85	28	7	76	26	65	43	53	9	81	28	60
Accusative	618	64	24	6	37	18	41	51	30	6	35	14	45
Dative	149	62	9	21	34	41	35	55	13	25	59	40	74
Genitive	437	78	34			65	79	57	44			60	82

Evaluation: comparing versions of the Sketch Grammar

Recall vs. precision

Case	*N*	Conditions											
		incl. *inflection*						incl. *inflection* + *affix-gender*					
		1		3		5		2		4		6	
		R	P	R	P	R	P	R	P	R	P	R	P
Nominative	1,709	85	28	7	76	26	65	43	53	9	81	28	60
Accusative	618	64	24	6	37	18	41	51	30	6	35	14	45
Dative	149	62	9	21	34	41	35	55	13	25	59	40	74
Genitive	437	78	34			65	79	57	44			60	82

Evaluation: comparing versions of the Sketch Grammar

Recall vs. precision

Case	*N*	Conditions											
		incl. *inflection*						incl. *inflection + affix-gender*					
		1		3		5		2		4		6	
		R	P	R	P	R	P	R	P	R	P	R	P
Nominative	1,709	85	28	7	76	26	65	43	53	9	81	28	60
Accusative	618	64	24	6	37	18	41	51	30	6	35	14	45
Dative	149	62	9	21	34	41	35	55	13	25	59	40	74
Genitive	437	78	34			65	79	57	44			60	82

- Condition 1 vs. condition 2: $\oplus$ precision $\ominus$ recall
  Adding derivation-based gender-guessing

Evaluation: comparing versions of the Sketch Grammar

Recall vs. precision

Case	*N*	Conditions											
		incl. *inflection*						incl. *inflection + affix-gender*					
		1		3		5		2		4		6	
		R	P	R	P	R	P	R	P	R	P	R	P
Nominative	1,709	85	28	7	76	26	65	43	53	9	81	28	60
Accusative	618	64	24	6	37	18	41	51	30	6	35	14	45
Dative	149	62	9	21	34	41	35	55	13	25	59	40	74
Genitive	437	78	34			65	79	57	44			60	82

- Condition 1 vs. condition 2: $\oplus$ precision $\ominus$ recall
  Adding derivation-based gender-guessing
- Condition 1 vs. 3, 2 vs. 4: $\oplus$ precision $\ominus$ recall
  Verb-final clauses: ca. 20 % of all corpus sentences
  Stronger changes than in condition 1 vs. 2

Evaluation: comparing versions of the Sketch Grammar

Recall vs. precision

Case	*N*	Conditions											
		incl. *inflection*						incl. *inflection + affix-gender*					
		1		3		5		2		4		6	
		R	P	R	P	R	P	R	P	R	P	R	P
Nominative	1,709	85	28	7	76	26	65	43	53	9	81	28	60
Accusative	618	64	24	6	37	18	41	51	30	6	35	14	45
Dative	149	62	9	21	34	41	35	55	13	25	59	40	74
Genitive	437	78	34			65	79	57	44			60	82

- Condition 1 vs. condition 2: $\oplus$ precision $\ominus$ recall
  Adding derivation-based gender-guessing
- Condition 1 vs. 3, 2 vs. 4: $\oplus$ precision $\ominus$ recall
  Verb-final clauses: ca. 20 % of all corpus sentences
  Stronger changes than in condition 1 vs. 2
- Cond. 4 vs. 6: better precision (!) and increased recall
  –recall: *all-clauses* is less restrictive than *verb-final*
  –precision: usefulness of explicit modelling?

Evaluation: comparing versions of the Sketch Grammar

Which German sketch grammar to choose?

So far: developer evaluation:

Case	*N*	Conditions											
		incl. *inflection*						incl. *inflection* + *affix-gender*					
		1		3		5		2		4		6	
		R	P	R	P	R	P	R	P	R	P	R	P
Nominative	1,709	85	28	7	76	26	65	43	53	9	81	28	60
Accusative	618	64	24	6	37	18	41	51	30	6	35	14	45
Dative	149	62	9	21	34	41	35	55	13	25	59	40	74
Genitive	437	78	34			65	79	57	44			60	82

- Best recall: condition 1: least constrained
- Best precision: condition 6: morph. + structural constraints

Evaluation: comparing versions of the Sketch Grammar

Which German sketch grammar to choose?

So far: developer evaluation:

Case	*N*	Conditions											
		incl. *inflection*						incl. *inflection + affix-gender*					
		1		3		5		2		4		6	
		R	P	R	P	R	P	R	P	R	P	R	P
Nominative	1,709	85	28	7	76	26	65	43	53	9	81	28	60
Accusative	618	64	24	6	37	18	41	51	30	6	35	14	45
Dative	149	62	9	21	34	41	35	55	13	25	59	40	74
Genitive	437	78	34			65	79	57	44			60	82

- Best recall: condition 1: least constrained
- Best precision: condition 6: morph. + structural constraints

User evaluation: "Clients" would have to decide (ongoing work)

- Lexicographers: need high-precision data ($\rightarrow$ condition 6)
- NLP researchers: may prefer large amounts of candidates ($\rightarrow$ cond. 1)

But: decision to be taken on Word Sketches, not on precision/recall

Evaluation for lexicography

Sample word sketch

Word sketch for noun *Pflanze* 'plant'

attr-adj	1566	2.0	**subj-of**	905	2.5
gentechnisch	94	47.14	*wachsen*	26	24.45
verändert	100	42.3	*gedeihen*	6	18.46
genmanipuliert	30	39.44	*anbauen*	5	18.30
fleischfressend	16	35.93	*werden*	73	15.91
transgenen	16	34.59	*können*	44	15.15
exotisch	24	30.00	*sollen*	30	15.03
transgener	8	28.45	*gießen*	4	14.52

Beyond the current state

We have presented

- a methodology for testing and evaluating (sketch) grammars
  for data extraction from corpora: applicable also to other languages
- a draft sketch grammar for German
  with different types and portions of linguistic knowledge

Beyond the current state

We have presented

- a methodology for testing and evaluating (sketch) grammars for data extraction from corpora: applicable also to other languages
- a draft sketch grammar for German with different types and portions of linguistic knowledge

Next

Beyond the current state

We have presented

- a methodology for testing and evaluating (sketch) grammars
  for data extraction from corpora: applicable also to other languages
- a draft sketch grammar for German
  with different types and portions of linguistic knowledge

Next

- further restrict the grammar, to improve precision,
  with a view to lexicographic use

Beyond the current state

We have presented

- a methodology for testing and evaluating (sketch) grammars
  for data extraction from corpora: applicable also to other languages
- a draft sketch grammar for German
  with different types and portions of linguistic knowledge

Next

- further restrict the grammar, to improve precision,
  with a view to lexicographic use
- integrate lexical resources (e.g. on noun gender),
  to improve precision and to compensate for flat tagset

Beyond the current state

We have presented

- a methodology for testing and evaluating (sketch) grammars
  for data extraction from corpora: applicable also to other languages
- a draft sketch grammar for German
  with different types and portions of linguistic knowledge

Next

- further restrict the grammar, to improve precision,
  with a view to lexicographic use
- integrate lexical resources (e.g. on noun gender),
  to improve precision and to compensate for flat tagset
- possibly use more deeply preprocessed data

Beyond the current state

We have presented

- a methodology for testing and evaluating (sketch) grammars for data extraction from corpora: applicable also to other languages
- a draft sketch grammar for German with different types and portions of linguistic knowledge

Next

- further restrict the grammar, to improve precision, with a view to lexicographic use
- integrate lexical resources (e.g. on noun gender), to improve precision and to compensate for flat tagset
- possibly use more deeply preprocessed data
- evaluate quality of word sketches from a lexicographic viewpoint