Corpus Design for a Unit Selection Database

Norbert Braunschweiler

Institute for Natural Language Processing (IMS) Stuttgart

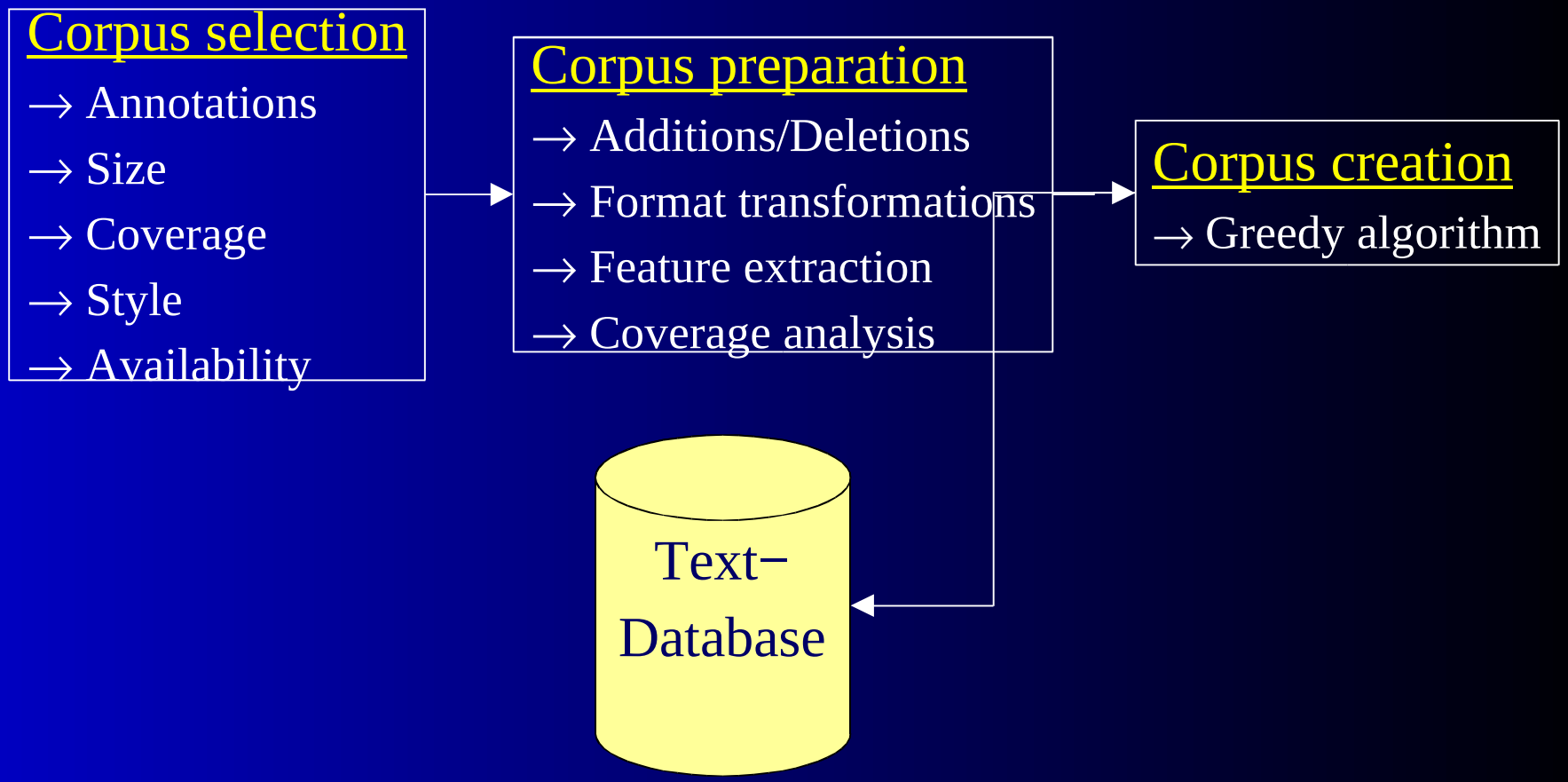
8th–9th October 2002

BITS Workshop, München

Corpus Design – Central Question

Which phonemes, syllables, words, or sentences does the corpus have to include in order to get a sufficient coverage of the sound structure of a given language or of the domain specific utterances?

Corpus Design – Task Flow Overview



Corpus Selection – Criteria

- Annotations
  - Part-of-speech tags present?
  - Are the annotations trustable/possibly hand-corrected?
- Size
  - Large, but not necessarily huge
  - Very large corpora are difficult to maintain
- Coverage
  - At least one instance of every phoneme/diphone
  - Smartkom specific words and sentences (can be added later)

Corpus Selection – Criteria

- Style
  - Political, social, humorous, colloquial, etc.?
  - Different prosodic styles (reading news−style, dialog−style, humorous−style, easy−going−style, etc.)?
- Availability
  - Are there any usage restrictions associated with the corpus?

→ Important are a rich vocabulary, trustable annotations, and variation in style

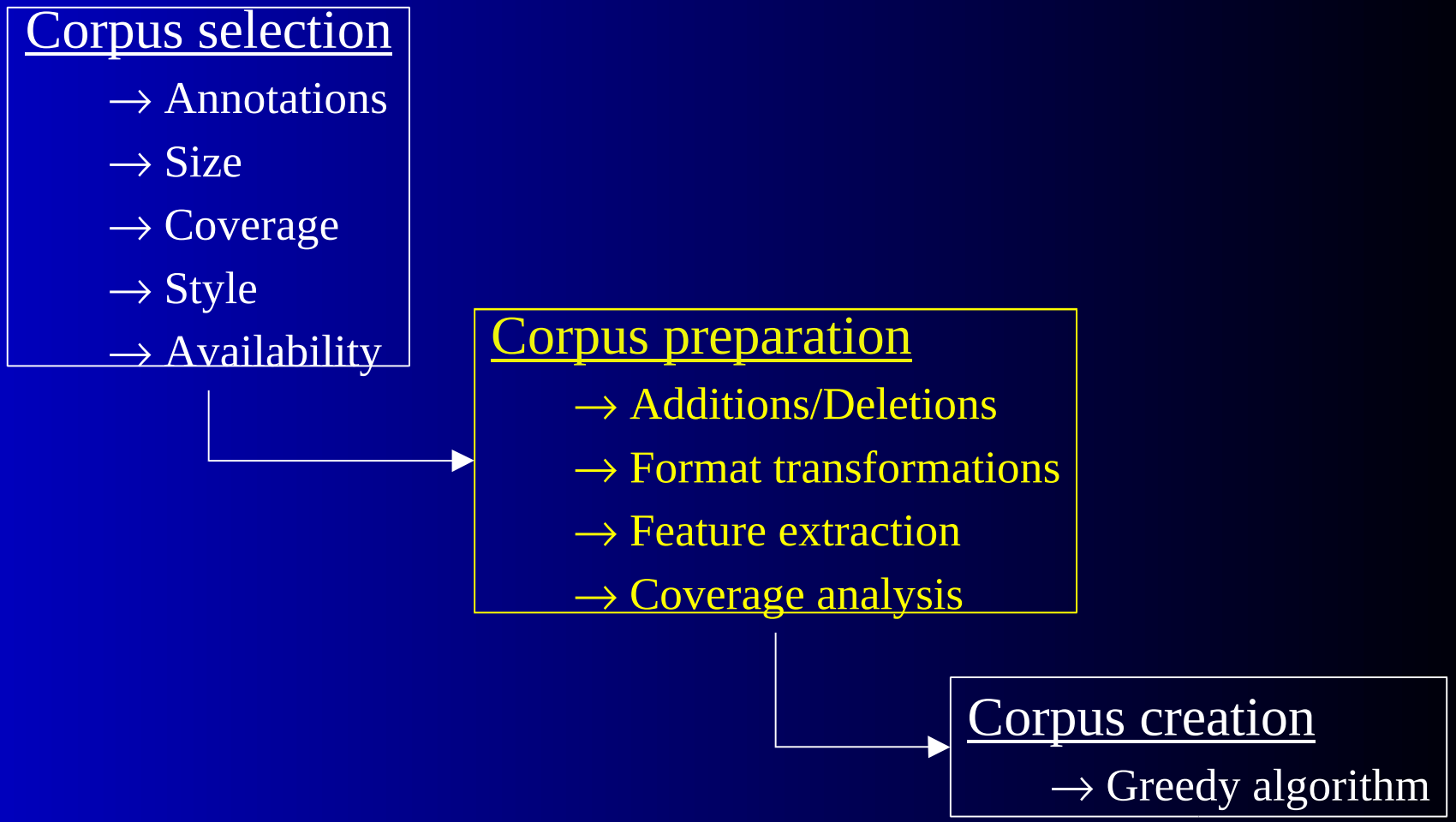
Corpus Selection – Example Corpus

TAZ−corpus

- Includes articles from 6 years (1988−1994)
- 285,000 articles and 76 million words
- Already annotated with part−of−speech−tags
- Available at IMS
- Fairly contemporary style

⇒ As starting point for subsequent processing a sub−corpus was created that had a size of ~40,000 sentences, ~550,000 words, and ~3 million phonemes altogether

Corpus Design – Task Flow



Corpus Preparation

- Additions/Deletions
  - Add Smartkom specific sentences
  - Add acronyms, abbreviations, numbers, etc.
  - Delete problematic entries
- Format transformations
  - Convert corpus to text
  - Split into single articles
- Feature extraction
  - Extract desired features using Festival (e.g. phoneme before/after, syllable accented, position in phrase, etc.)
- Coverage analysis

Corpus Selection – Coverage Analysis

Sub−corpus	Phonemes	Diphones
Nr of different units:	53	2294
Total nr of units:	~3 million	~3 million
Average unit freq.:	~60,000	~1300
Smallest freq.:	1	1
Biggest freq.:	~300,000	~90,000
Smartkom:	48	2795
Missing:	0	~500

Corpus Selection – Coverage Analysis

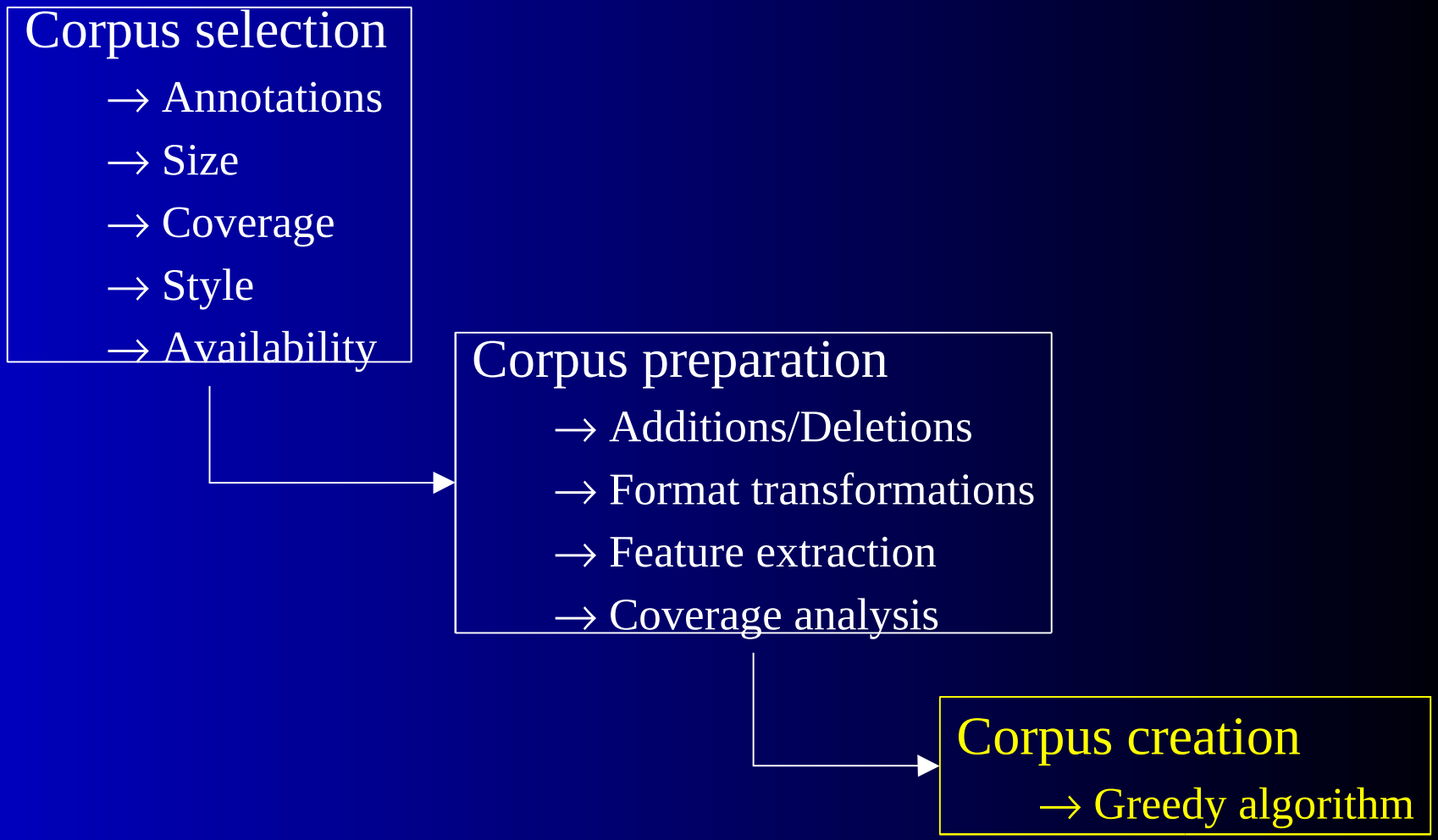
Sub−corpus

Phonemes	#	Diphones	#	
eI	1	2:−?−U	1	
u	1	2:−?−o:		1
o~	9	2:−S	1	
...				
@	172016	E−6		41748
t	228575	n−t		46298
n	316364	@−n		91380

Corpus Preparation

- Decide what features to include
- Transformation of phonemes into diphones in order to prepare the input format for the greedy algorithm
- Input to greedy algorithm consists of 3 columns:
  - Sentence
  - Phoneme transcription
  - Diphone transcription

Corpus Design – Task Flow



Greedy algorithm

- Greedy algorithm is used for the creation of a sub−corpus that fulfills a number of conditions that are of one's choice

- Algorithm works step−by−step: a first sentence is selected according to a criterion; the sentence is added to the cover, and the covered units are removed from the set of units to cover. The process starts again: the second sentence, in this example, contains a maximum of non−already covered units. The process stops when all units are covered.

Corpus Creation – Applying the Greedy Algorithm

How many sentences/words does a sub−corpus have to include in order to have at least one occurrence of

- each phoneme and
- each diphone?

⇒ Sub−corpus with ~700 sentences and ~13,500 words

Input corpus: ~40,000 sentences and ~550,000 words.

Corpus Creation – Some Combinations

Corpus	Sentences	Words	‘Phonemes’	Diphones
Input	41,000	550,000	53	2294
Each phoneme + each diphone once	691	13,415	53	2294
Phonemes + Diphones + stressed/unstressed	691	13,415	101	2294
Phonemes + Diphones + stressed/unstressed + Position in sentence	699	13,359	228	2294
?	?	?		

Corpus Creation – Conclusion

- Selection of input corpus based on aspects of annotations, coverage, size, style, and availability
- Check phoneme and diphone coverage of input corpus
- Add missing or desired phonemes/diphones
- Add domain specific utterances to the corpus, e.g. both Smartkom demo dialogs
- Decide what features to include
- Manual correction of input corpus
- Test corpus with a working unit selection algorithm