UTACLIR @ CLEF 2001

New features for handling compound words and untranslatable proper names

The UTA-CLEF Research Group

University of Tampere

Turid Hedlund, Heikki Keskustalo, Ari Pirkola, Eija Airio and Kalervo Järvelin

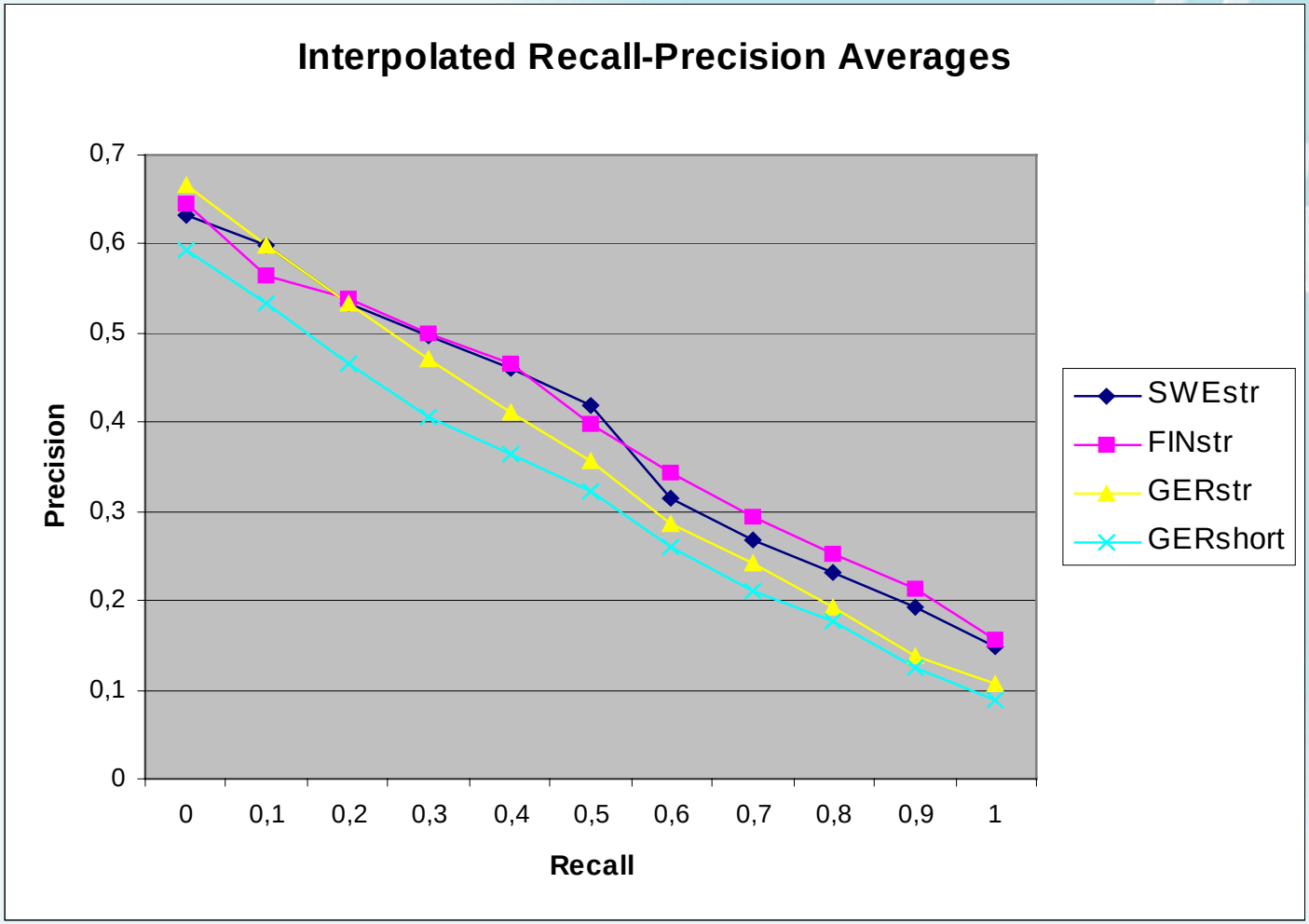
The runs

- Four automated bilingual runs (three language pairs)
  - Finnish - English
  - Swedish - English
  - German - English
- Test with two types of dictionaries (German - English)
  - comprehensive dictionary
  - limited dictionary, where direct translation of compounds is eliminated

The results

Testrun	Average precision	
	CLEF 2001	**Clef 2000**
TAYfinstr	0,3894	0,2275
TAYswestr	0,3769	0,2540
TAYgerstr	0,3474	0,2665
TAYgershort	0,3054	-------

Interpolated Recall-Precision Averages



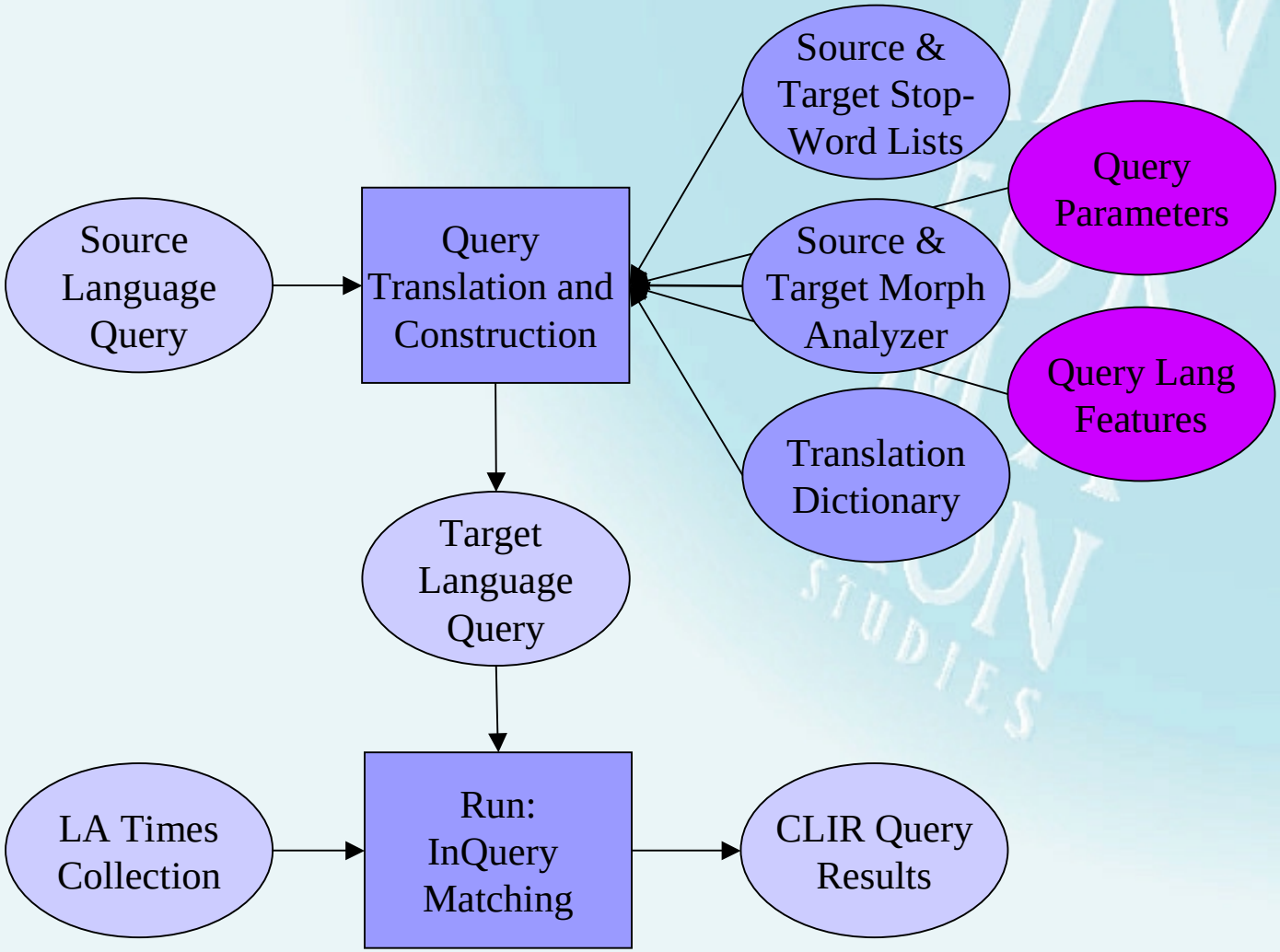
The General Approach

- Bilingual dictionaries
- Word normalisation in indexing and topic words
- Stopword lists
- Splitting of compounds
- Recognition of the right component
- Handling of non-translated words
- Phrase composition in the target language
- Structuring of queries

New Features

- Unified process (Swedish and German) focusing on compound words
  - For Finnish the old process is used except for a more flexible proximity operator and a broader window size. The n-gram matching technique is also in use.
- A new process for matching proper names and other non-translatable words

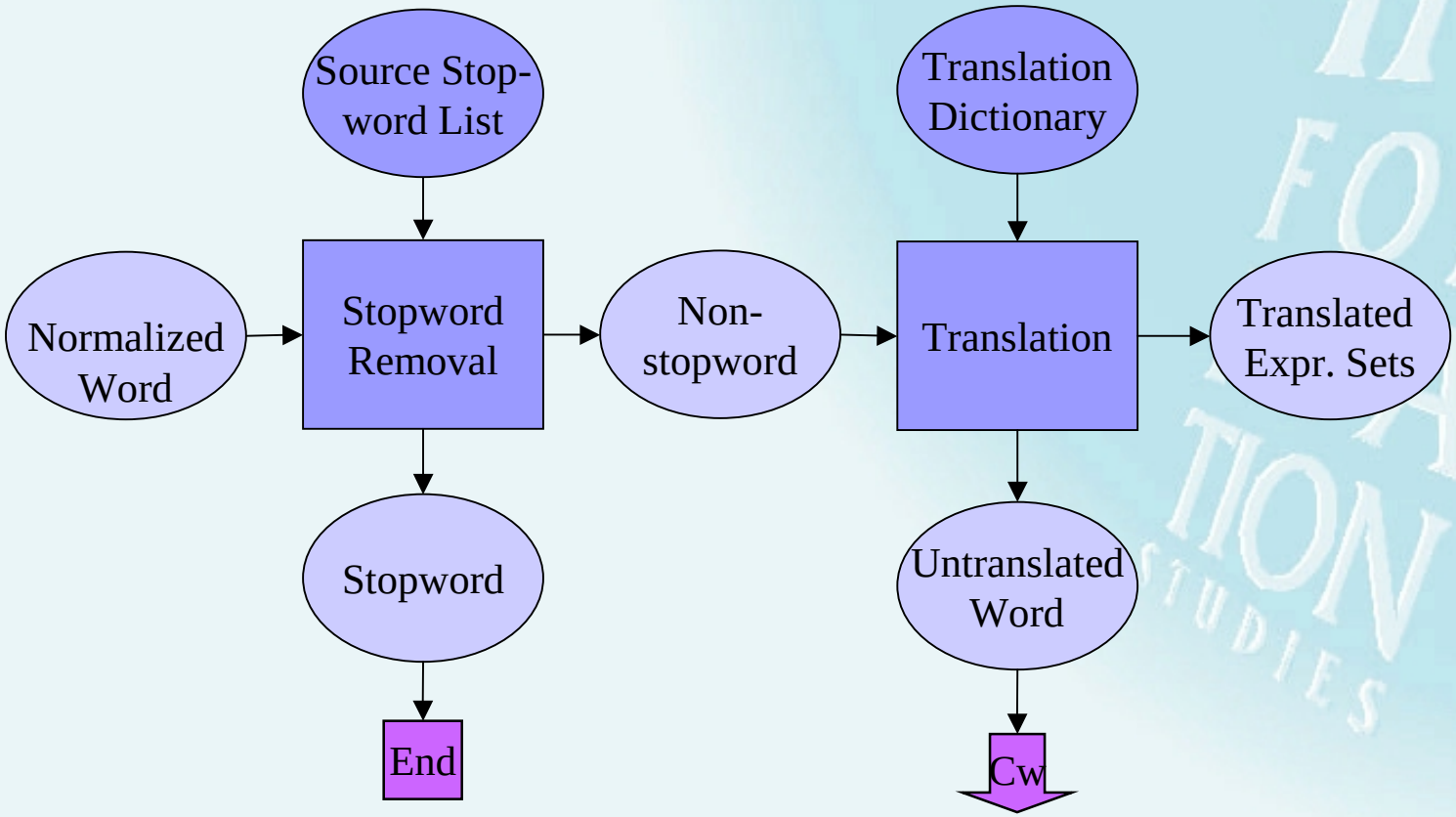
Query Construction Process



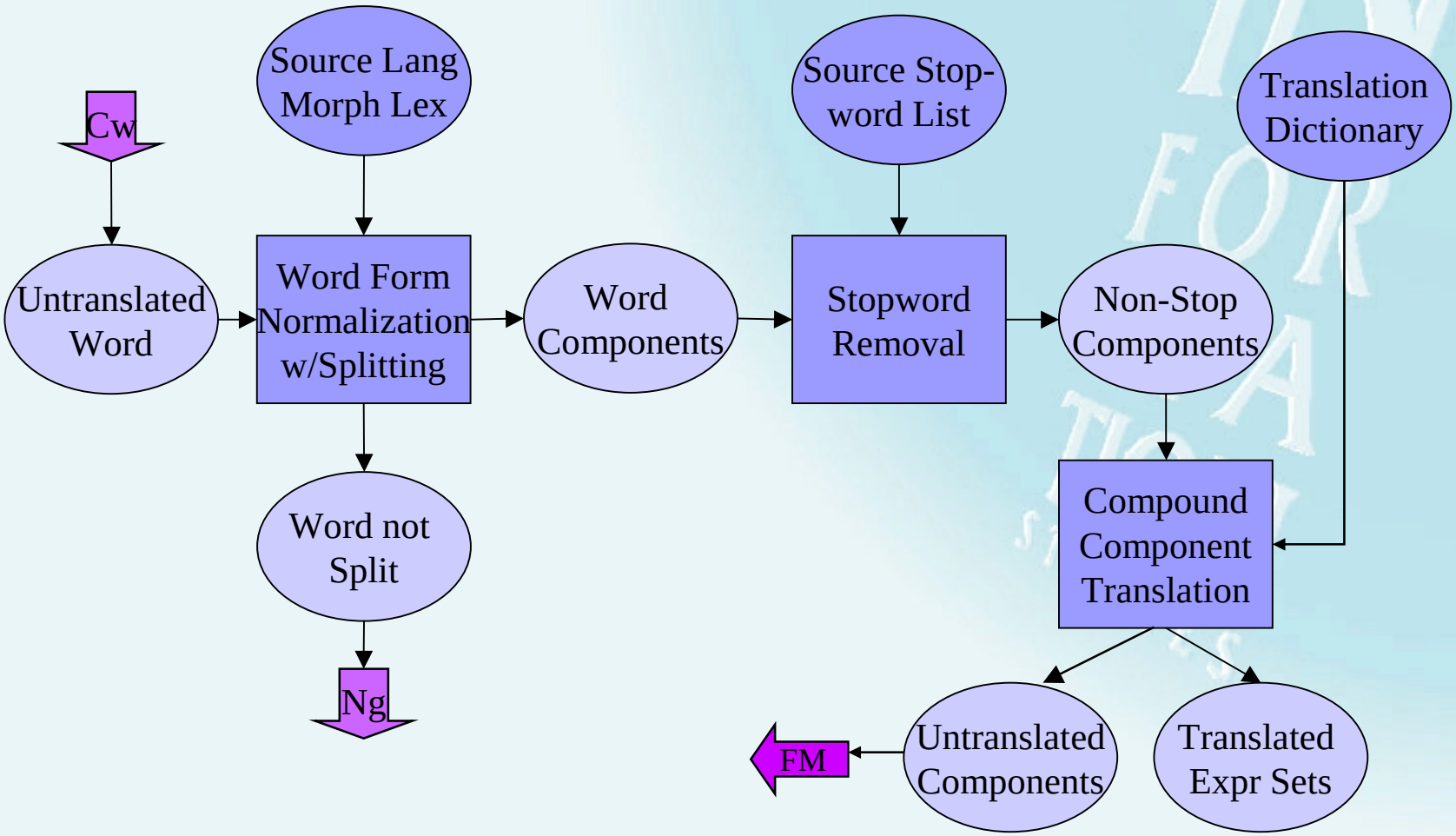
The Unified Process for Handling Compounds

- Direct dictionary look-up and translation of compounds if possible
- Compound splitting, forming of consecutive component pairs, if the compound is untranslatable
- Dictionary look-up and translation of compound components and component pairs
- Stop word removal
- Fogemorpheme algorithm for Swedish and German
- N-gram matching technique for unrecognised and untranslatable words (proper name component)

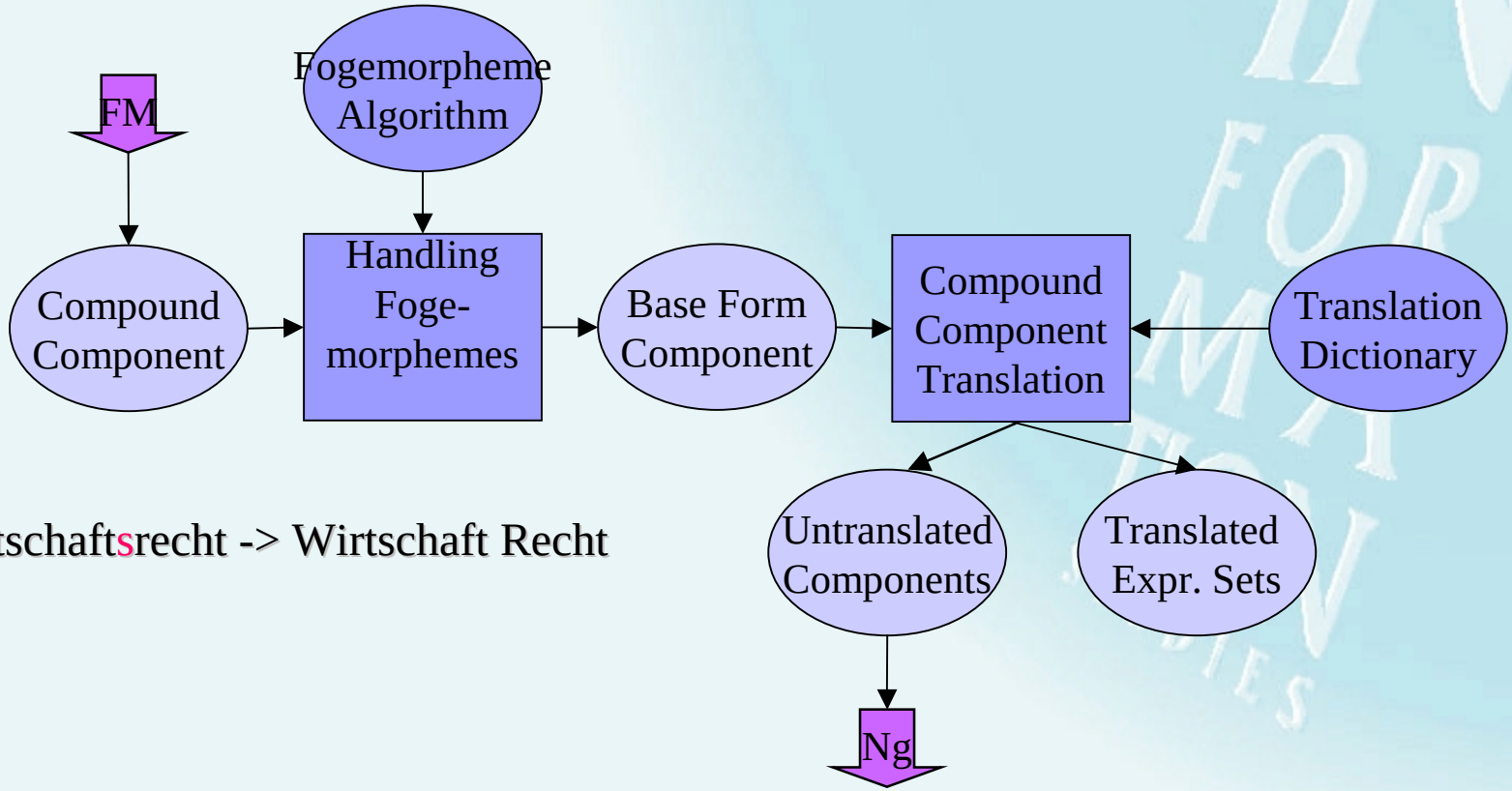
Direct Dictionary Look-up



Splitting of Compounds

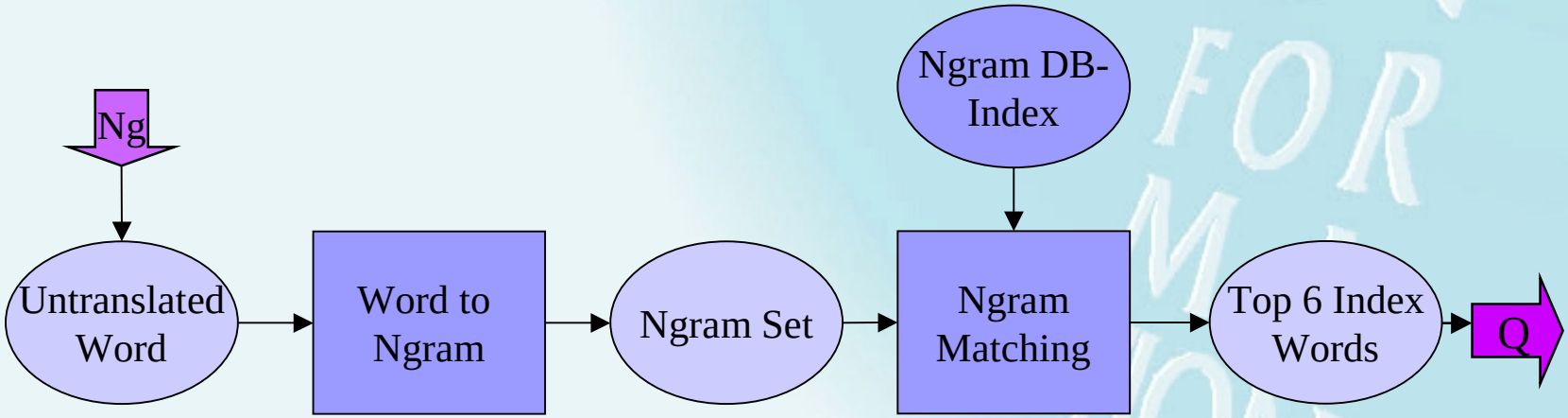


Handling of Fogemorphemes



Wirtschaftsrecht -> Wirtschaft Recht

N-gram Matching for Unrecognised / Untranslated Words



Query Construction

- Query parameters
  - structuring of queries
  - phrase construction for compounds
  - proximity operators and window size

Conclusions

- **Improvement from last year**
  - the new methods for compounds and proper names seem account for the improvements
  - the use of n-gram technique for unidentified words was successful especially for Finnish (proper names in inflected form)
  - but it also adds noise to queries when used for all unidentified words and compound components (Swedish and German). Proper names could be identified and the use may be restricted to these.

Conclusions 2

- **Interesting results for the tests with two dictionaries for German**
  - the advantage of direct translation is inevitable
  - Our method for handling compounds works as a good and necessary complement, no dictionary holds entries for all compounds
  - The UTACLIR process works well also with a limited dictionary