May the Best Team Win: Language Resources in Cross-Language Information Retrieval

Anne Diekema

Syracuse University

CLEF 2000

Workshop on Cross-Language Information Retrieval and Evaluation

ECDL 2000, 21 September

Lisbon, Portugal

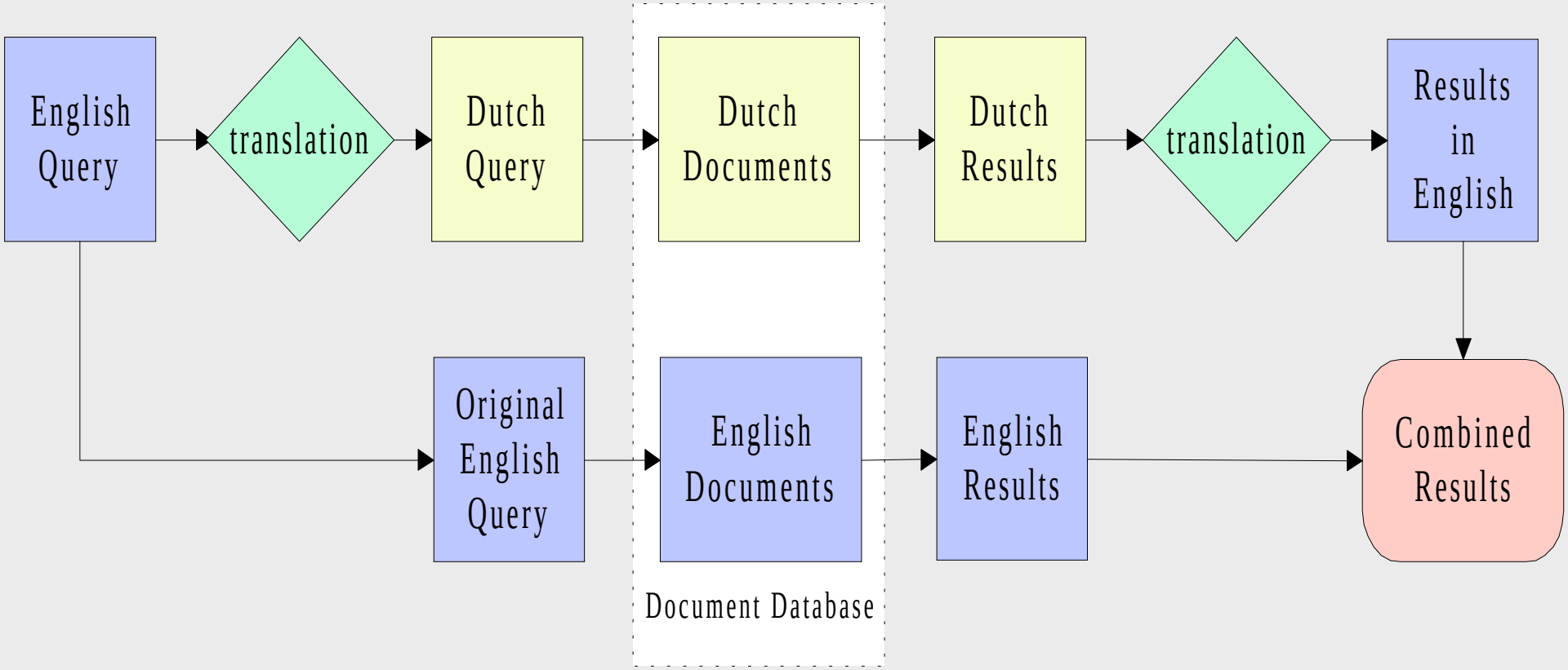
Cross-Language Information Retrieval

- CLIR is a special case of Information Retrieval
- retrieval is not restricted to the query language
- queries in one language retrieve documents in multiple languages
- simplifies searching by multilingual users
- allows monolingual searchers to judge relevance based on machine translated results and/or to allocate expensive translation resources to the most promising foreign language documents

Complexities of CLIR

- dealing with intricacies of multiple natural languages
- matching between nearly distinct vocabularies which requires some form of translation
- merging of different result sets
- documents from result set might need to be translated
- translation inherently difficult

CLIR system using query translation



What resources are needed for a CLIR system?

- multiscript text processing
- language identification tool
- tokeniser
- compound splitter
- pos-tagger
- stemmer
- stoplist
- translation resource for matching
- translation resource for result set presentation

FC Barcelona 1999-2000

http://www.fcbarcelona.es

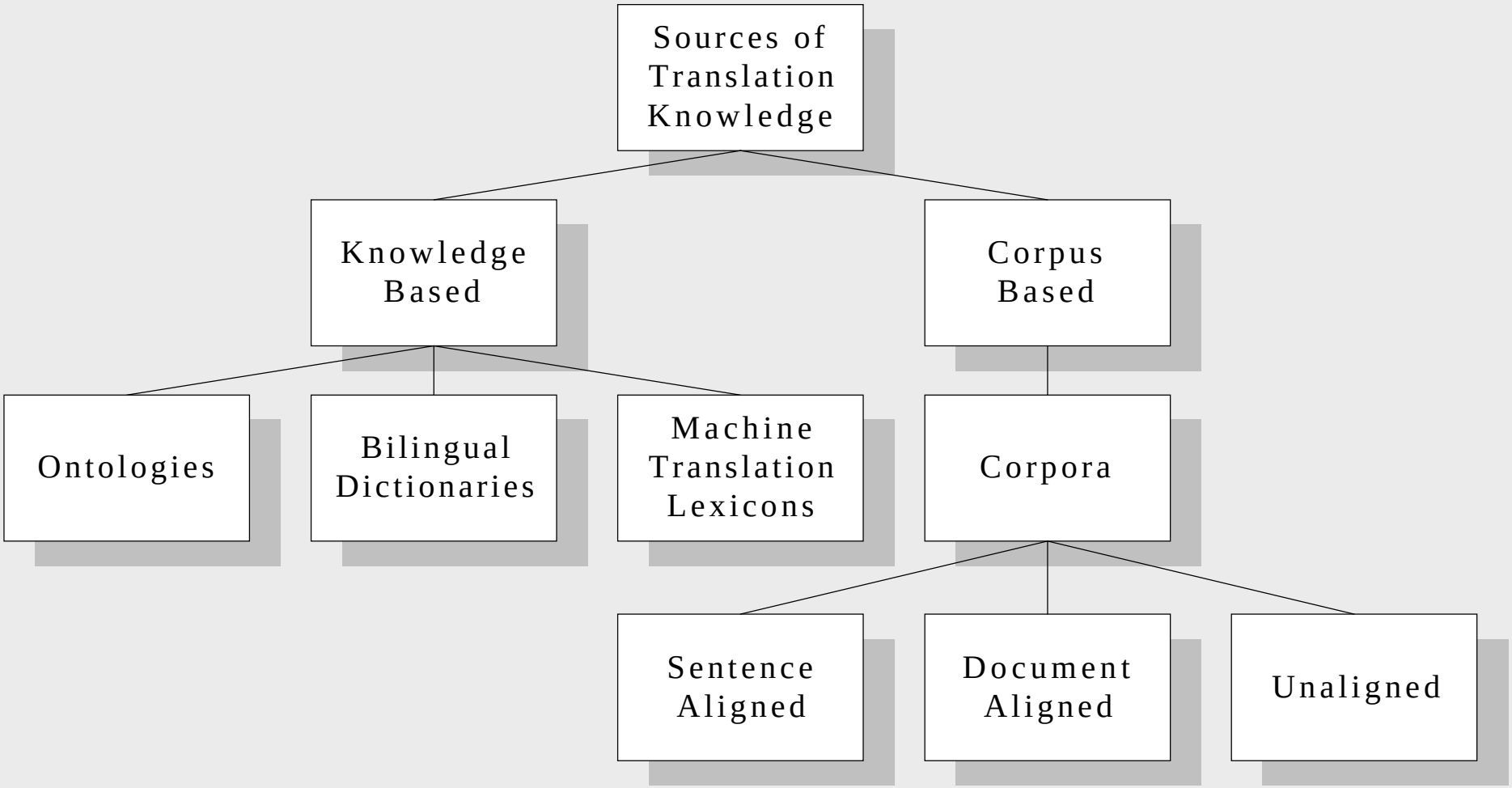
Resource availability

- language resources not always easy to obtain for the average IR researcher because:
- resources can be quite costly
- existing resources often need additional processing
- resources not available for all languages
- creation of resources time consuming and expensive
- free resources possibly unreliable or not comprehensive

Imagine a basic CLIR system

- economic solution: using no resources at all
- cognate matching
- reasonable and economic solution: using translation resources only
- get original through translation as semantically intact as possible

Lexical Translation Resources



Toils of Translation

- lack of a one-to-one mapping of a lexical item and its meaning (lexical ambiguity)
- cultural differences between language communities and the way they lexicalize the world around them
- these two issues create translation problems
- translation problems result in translation errors
- translation errors impact CLIR performance

Translation Problems

- lexical ambiguity
- lexical mismatches
- lexical holes
- figures of speech
- multiword lexemes
- specialized terminology and proper names
- false cognates

Power of good resources

- lower threshold of CLIR performance is no translation
- upper threshold of CLIR performance is manual translation
- clearly, translation is not easy and a good resource is required to do it well

May the best team win

- for a basic CLIR system one needs at least a translation resource
- this translation resource needs to be of good quality
- good quality translation resources costly to acquire

FC Barcelona 1999-2000

http://www.fcbarcelona.es

Van Gaal's Dutch Resources

http:www.fcbarcelona.es

- Frank de Boer
- Ronald de Boer
- Winston Bogarde
- Philip Cocu
- Ruud Hesp
- Patrick Kluivert
- Michael Reiziger
- Boudewijn Zenden

What are we evaluating?

- results in monolingual TREC ad-hoc track are diverging
- systems seem to have reached similar peak performance levels
- basic CLIR is monolingual IR plus translation resources
- if we can say we are evaluating resource quality
- is CLEF evaluation an indication of system quality or financial status?

Too many factors in CLIR system evaluation

- translation
- automatic relevance feedback
- term expansion
- disambiguation
- result merging
- test collection
- need to tone it down to see what happened

The End