

Dutch Parallel Corpus

A Multifunctional & Multilingual Corpus

Hans Paulussen, Lieve Macken
Piet Desmet, Willy Vandeweghe

Dutch Parallel Corpus

- Annotated sentence aligned corpus
- 10 million words
- Dutch - English / Dutch - French
- Quality control
- Compatible with D-COI

Motivation

- Availability
 - Limited individual or private initiatives
 - E.g. Namur corpus, Scania corpus
- Balanced distribution
 - Big quantities of reduced set of text types
 - Europarl, JRC (*Acquis communautaire*)
- Quality
 - New users requiring higher quality levels

Dutch in multilingual corpora

- OMC: 170,000 Dutch words
- Namur Corpus: 700,000
- MLCC: 7,100,000
- JRC: 7,339,465
- Europarl: > 29,000,000

Corpus evolution

period	corpus sample	compilation duration	input
70's	Brown, LOB 1,000,000	10 years	keying
1993	Namur corpus 2,000,000	1 year	scanning & electronic data
2000	Unesco Courier 1,000,000	1 day	webdownload

Large corpora are useful ...

- Number crunching applications
- Statistical analysis
- Automatic analysis
- No human intervention

... but less adequate in:

- Applications involving quality at all levels
- Applications involving human analysis
- Educational applications

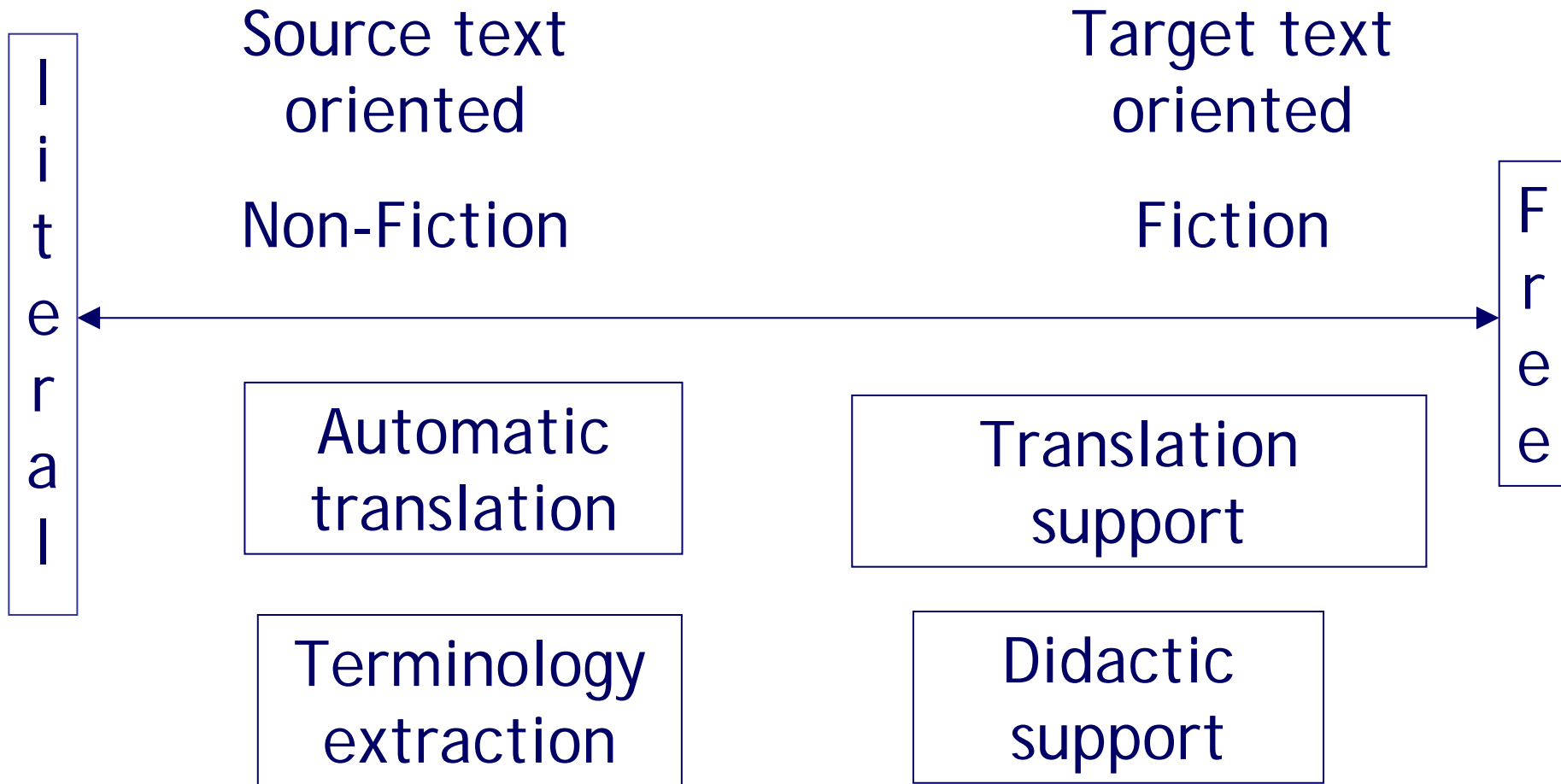
Users and applications

- NLP applications
 - Automatic translations / terminology extraction
 - Training and test data
- Support applications
 - Translation support
 - Didactic support (CALL)
 - Extension to bilingual dictionaries
- Fundamental research
 - Translation science / contrastive linguistics
 - Corpus linguistics

DPC requirements

- 1) Corpus design
- 2) Metadata
- 3) Linguistic annotation
- 4) Quality control
- 5) Corpus exploitation

Corpus design



Qualitative resource providers

Resource type	Sector
Journalistic	Newspapers & magazines
Essayistic & fiction	Book publishers
Business	Banking & insurance
Technical	Software houses & medical sector
Administrative	Government

Metadata

- Translation direction
 - EN → NL vs. NL → EN
- Translation modality
 - Human translation, CAT, MT
- Direct vs. indirect translations
 - Indirect via English (e.g. Europarl)

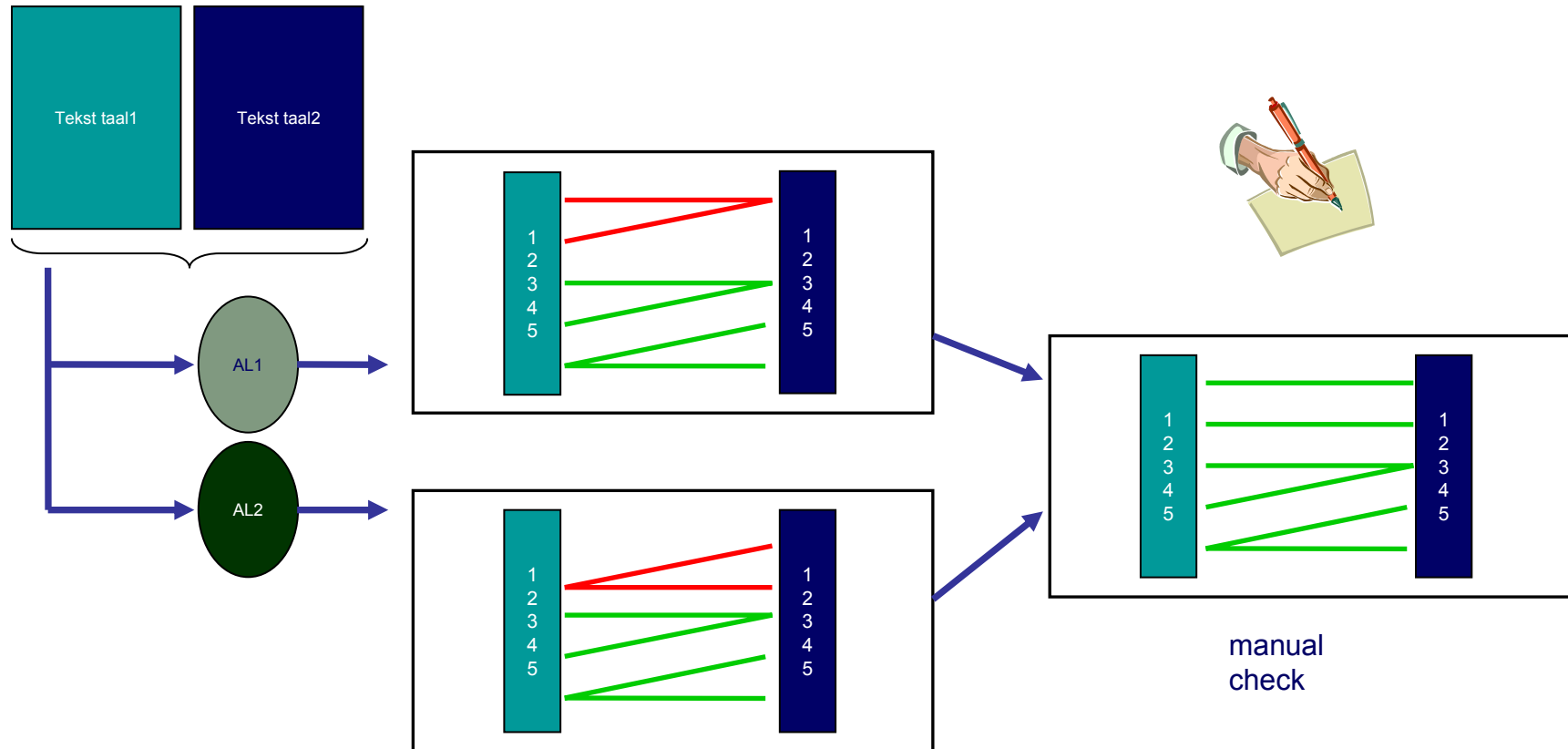
Annotation

- Tokenisation
 - Paragraphs, sentences, words
- Alignment
 - Sentence
- Linguistic annotation
 - Lemma
 - PoS
 - Syntactic structures

Quality control

- Manually checked
- Spot checking
- Automatic control procedures
 - e.g. automatic comparison of output from different alignment programs

Alignment merge



Corpus exploitation

- Web search interface
 - Simple queries
 - Extended queries
 - Pattern matching & annotation labels
- Text exploitation
 - Data-driven automatic learning
 - Statistical MT

DPC organisation

- DPC core team
 - K.U.Leuven - Campus Kortrijk
 - HoGent
- DPC research team
- DPC user group

DPC core research team

- K.U.Leuven – Campus Kortrijk
 - Prof. Dr. Piet Desmet
 - Dr. Hans Paulussen
 - Dr. Yulia Trushkina
 - Lic. Antoine Besnehard
- HoGent – School of Translation Studies
 - Prof. Dr. Willy Vandeweghe
 - Dra. Lieve Macken
 - Lic. Lidia Rura

DPC Research partners

- University of Groningen
- Radboud University Nijmegen
- Tilburg University
- K.U. Leuven
- University of Antwerp
- University of Gent

DPC User group

- Consulting in important design decisions
- Industrial partners
 - CALL, translation services, terminology extraction, information extraction
- Academic partners
 - Language technology
 - Translation studies
 - Contrastive linguistics

Thank you

www.kuleuven-kortrijk.be/dpc