TICCLops

Martin REYNAERT

Tilburg centre for Creative Computing

CLARIN-NL, Utrecht. 19 February, 2010

TEXT-INDUCED CORPUS CLEAN-UP: INTRODUCTION

TICCL for TYPOS and OCR-errors

- Tool to perform large scale, unsupervised spelling correction of corpora
- Spelling correction = reduction of lexical variation caused by typos, OCR-errors, historical orthographical changes...
- Currently no direct text editing, but linking of variants to their most likely canonical form
- Prototype developed during a pilot project by invitation of the National Library, The Hague
- Production version developed according to KB specifications, second half 2008

EVALUATION: Het Volk 1918

				Overall Score		
	TP	FN	FP	R	P	F
types	1414	1671	134	0.458	0.913	0.610
tokens	11334	3163	2971	0.782	0.792	0.787
types				Score per LD		
LD 1	312	8	10	0.975	0.969	0.972
LD 2	1039	60	108	0.945	0.906	0.925
LD 3	62	1569	16	0.038	0.795	0.073
types				Cumulated score per LD		
LD 1	312	8	10	0.975	0.969	0.972
LD 2	1351	68	118	0.952	0.920	0.936
LD 3	1413	1637	134	0.463	0.913	0.615

TICCL online processing system

- A demonstration project which will allow CLARIN users to submit their corpora for fully automatic spelling correction and normalization by TICCLops, the online processing version of our core component TICCL. This system should be widely applicable in all manner of curation projects and lexicographical work.
- Start Date: 15 January 2010
- End Date: 15 July 2010

TICCLops: PARTNERS

- Coordination & Technology Provider:
  Tilburg centre for Creative Computing (TiCC) - Tilburg
  - Martin Reynaert: Researcher
  - Maarten van Gompel - Scientific Programmer
- User & Data Provider:
  Koninklijke Biblotheek (KB) - The Hague
  - Astrid Verheusen - Head Digitisation Department
- CLARIN Center & Data Provider:
  Instituut voor Nederlandse Lexicologie (INL) - Leiden
  - Remco van Veenendaal - Head TST-Centrale

Research Data

- Staten-Generaal Digitaal: OCR-ed Acts of Parliament (http://www.statengeneraaldigitaal.nl) covering 1920 to 1995, i.e. in historical and contemporary spelling.
- Databank Digitale Dagbladen: Digitized newspapers: we also have a gold standard for Het Volk-1918 (http://www.kb.nl/hrd/digi/ddd/index-en.html).
- New collections of recently digitized copyright-free books and magazines

Demonstration Scenario

- Digitized copyright-free book and its gold standard
- All aspects of using the system will be demonstrated to prospective users on the basis of this book
- Aspects include pre-packaging the corpus (i.e. the book) as a compressed archive, choosing and setting the system parameters and receiving and interpreting the results
- Actual evaluation results returned will then teach the prospective user what to expect from the system and to choose which output formats best suit his own needs

Demonstrator Book

44950.

KORT BEGRIP

DER

WAERELD-HISTORIE

VOOR DE JEUGD,

DOOR

KORT BEGRIP

DER

WAERELD-HISTORIE

VOOR DE JEUGD,

DOOR

J. F. MARTINET,

PREDIKANT TE ZUTPHEN, &c.

MET KAARTEN.

TE AMSTERDAM, BIJ
JOHANNES ALLART.
MDCCLXXXIX.

Thanks!!

Thanks for your attention!

Papers about TICCL are available at:
`http://ilk.uvt.nl/`

TICCLops

Martin REYNAERT

Tilburg centre for Creative Computing

CLARIN-NL, Utrecht. 19 February, 2010