Universiteit Utrecht

Corpus Annotation and Exploitation: Contributions by CCL

Jan Odijk

CCL25

Leuven, Feb 9, 2017

UiL OTS
UTRECHT INSTITUTE OF LINGUISTICS

Overview

- Spoken Dutch Corpus (CGN)
- STEVIN
- CLARIN

Spoken Dutch Corpus (CGN)

- FvE chair of the Annotation Working Group
- CGN Tag set (Van Eynde 2004)
  - Has become the *de facto* pos tagging standard for Modern Dutch
    - Used in almost all corpora and treebanks for Dutch
    - Used in the FROG pos-tagger

STEVIN

- Essential LST Resources for Dutch

Evoluon, Eindhoven

Bovendonk, Hoeven


Willem II
Tilburg

STEVIN: D-COI/SoNaR

- CGN tagset extended for written language: CGN/D-COI-tagset
- Pos tag annotation

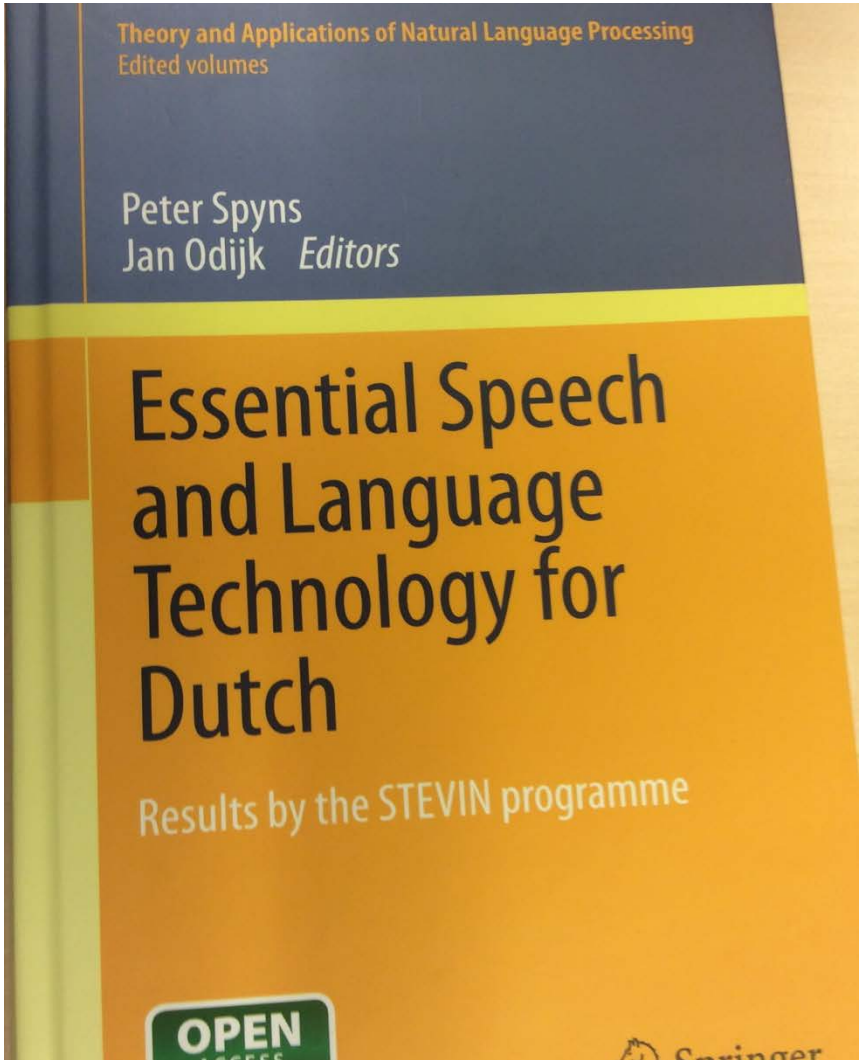
STEVIN: LASSY

- Treebank for Written Dutch
  - LASSY-Small (1 m tokens)
  - LASSY-Large (700 m tokens)

STEVIN: PACO-MT

- Focus on MT
- Parallel Corpora
  - D-E (48 m)
  - D-F (45 m)

STEVIN: DCOI/SONAR



UiL OTS
UTRECHT INSTITUTE OF LINGUISTICS

STEVIN: D-COI/SoNaR

Chapter 13
The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman

13.1 Introduction

Around the turn of the century the Dutch language Union commissioned a survey that aimed to take stock of the availability of basic language resources for the Dutch language. Daelemans and Strik [5] found that Dutch, compared to other

STEVIN: LASSY

Chapter 9
Large Scale Syntactic Annotation of Written Dutch: Lassy

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste

9.1 Introduction

The construction of a 500-million-word reference corpus of written Dutch has been identified as one of the priorities in the STEVIN programme. The focus is on written language in order to complement the Spoken Dutch Corpus (CGN) [13], completed in 2003. In D-COI (a pilot project funded by STEVIN), a 50 million word pilot corpus has been compiled, parts of which were enriched

STEVIN: PACO-MT

Chapter 17
Parse and Corpus-Based Machine Translation

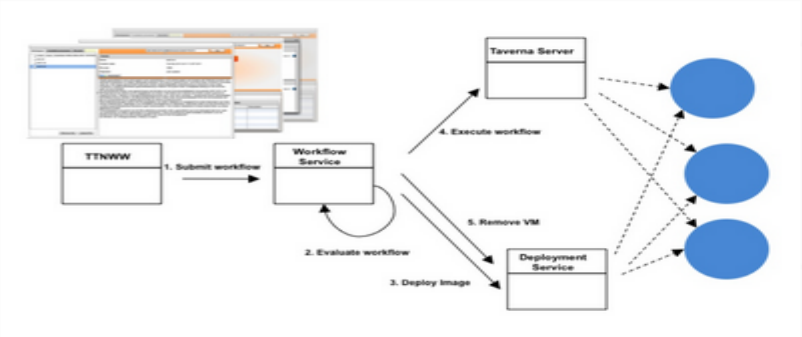
Vincent Vandeghinste, Scott Martens, Gideon Kotzé, Jörg Tiedemann, Joachim Van den Bogaert, Koen De Smet, Frank Van Eynde, and Gertjan van Noord

17.1 Introduction

The current state-of-the-art in machine translation consists of *phrase-based statistical machine translation (PB-SMT)* [23], an approach which has been used since the late 1990s, evolving from word-based SMT proposed by IBM [5]. These *string-based* techniques (which use no linguistic knowledge) seem to have reached their ceiling in terms of translation quality, while there are still a number of limitations

CLARIN: TTNWW

TTNWW



TTNWW - TST Tools voor het Nederlands als Webservices in een Workflow

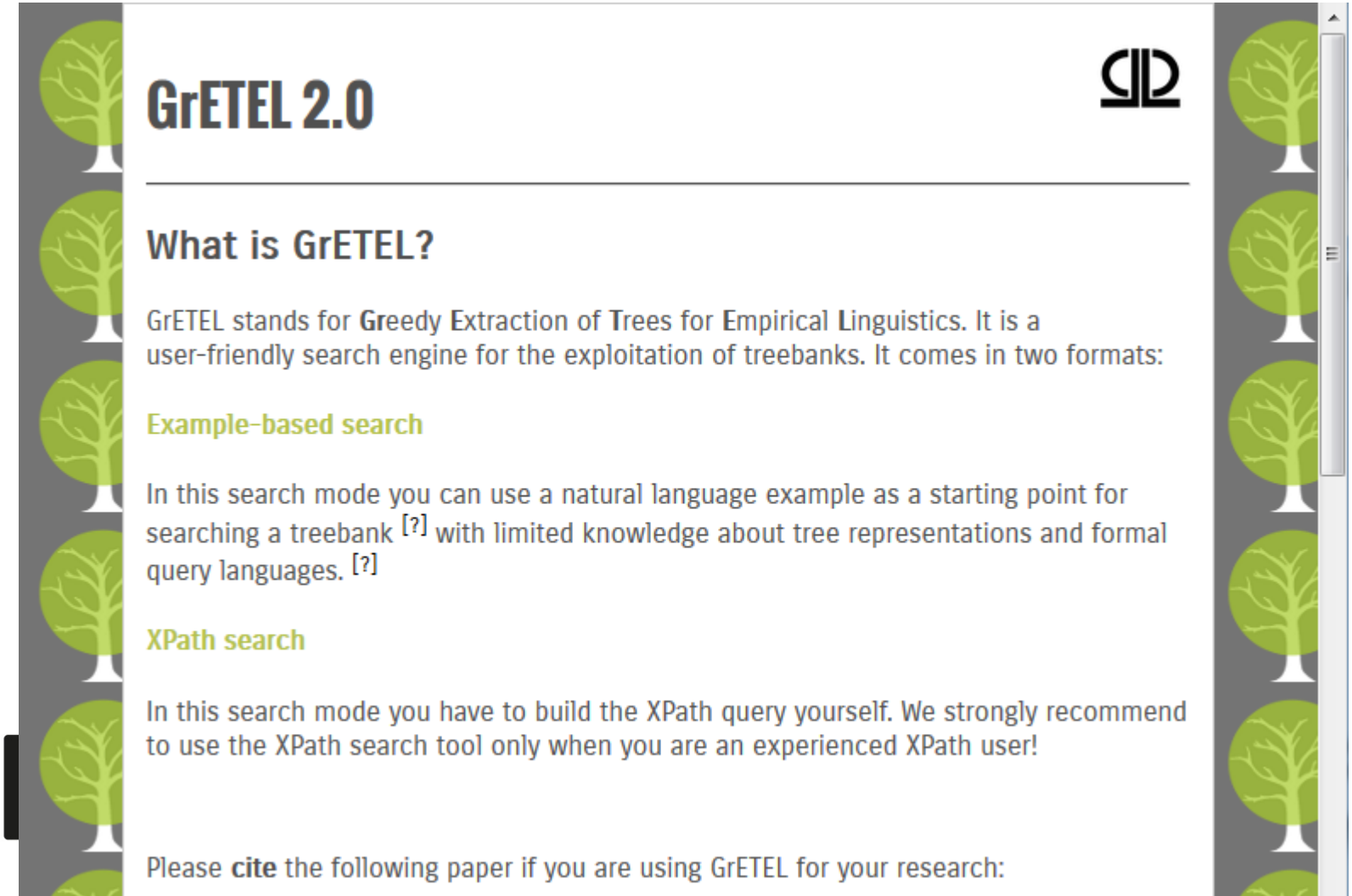
Summary

TTNWW integrates and makes available existing Language Technology (LT) software components for the Dutch language that have been developed in the STEVIN and CGN projects. The LT components (for text and speech) are made available as web-services in a simplified *workflow* system that enables researchers without much technical background to use standard LT *workflow* recipes.

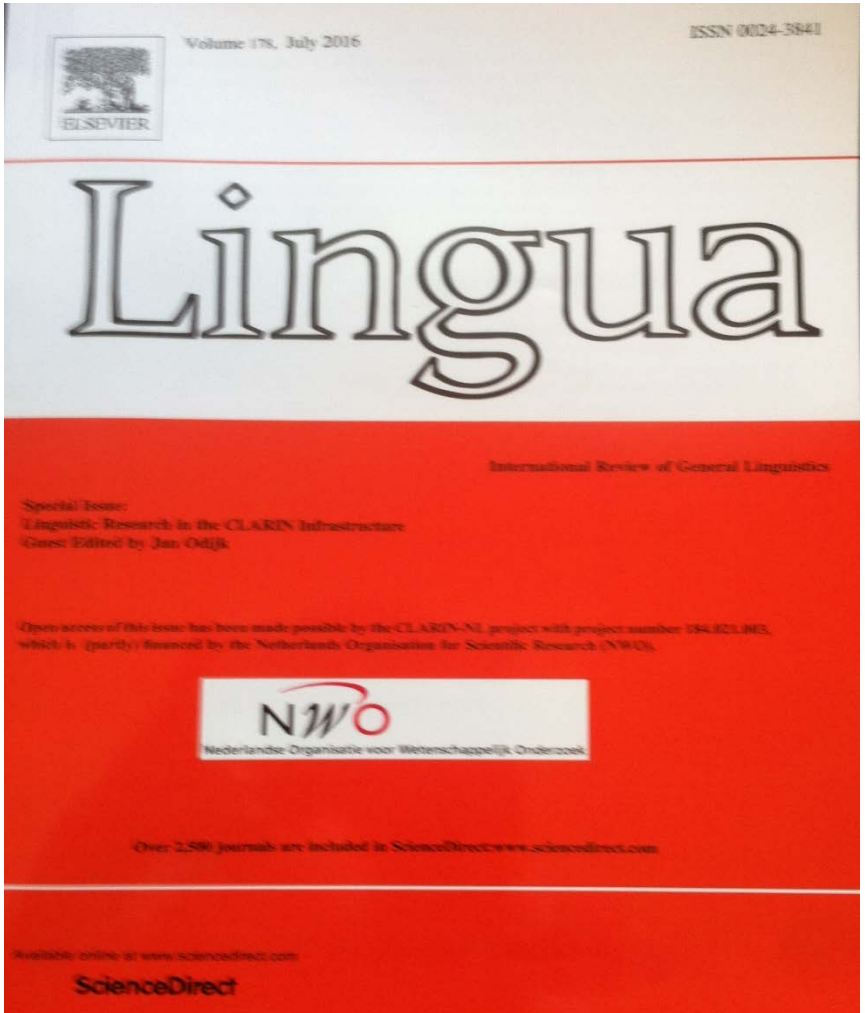
Background

UIL OTS
UTRECHT INSTITUTE OF LINGUISTICS

CLARIN: GrETEL

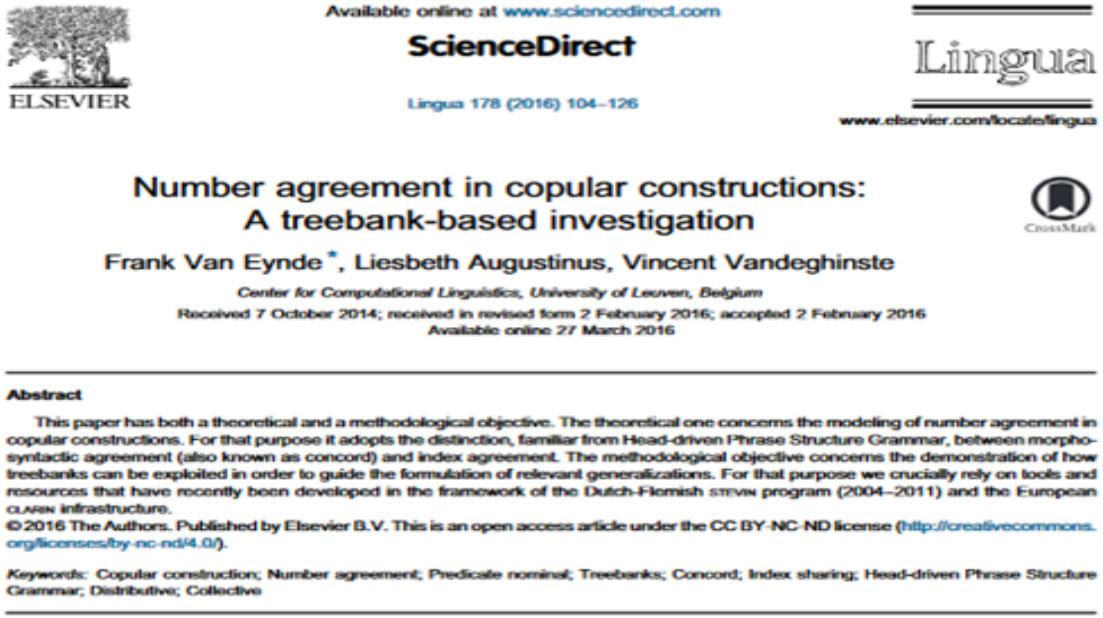


CLARIN: GrETEL



UiL OTS
UTRECHT INSTITUTE OF LINGUISTICS

CLARIN: GrETEL



Available online at www.sciencedirect.com

ScienceDirect

Lingua 178 (2016) 104–126

Lingua

www.elsevier.com/locate/lingua

Number agreement in copular constructions: A treebank-based investigation

Frank Van Eynde *, Liesbeth Augustinus, Vincent Vandeghinste

Center for Computational Linguistics, University of Leuven, Belgium

Received 7 October 2014; received in revised form 2 February 2016; accepted 2 February 2016
Available online 27 March 2016

Abstract

This paper has both a theoretical and a methodological objective. The theoretical one concerns the modeling of number agreement in copular constructions. For that purpose it adopts the distinction, familiar from Head-driven Phrase Structure Grammar, between morpho-syntactic agreement (also known as concord) and index agreement. The methodological objective concerns the demonstration of how treebanks can be exploited in order to guide the formulation of relevant generalizations. For that purpose we crucially rely on tools and resources that have recently been developed in the framework of the Dutch-Flemish STEVIN program (2004–2011) and the European CLARIN infrastructure.

Keywords: Copular construction; Number agreement; Predicate nominal; Treebanks; Concord; Index sharing; Head-driven Phrase Structure Grammar; Distributive; Collective

1. Introduction

UiL OTS
UTRECHT INSTITUTE OF LINGUISTICS

CLARIN: GrETEL

References

[Van Eynde 2004] Frank Van Eynde. Part of Speech Tagging en Lemmatisering van het Corpus Gesproken Nederlands. CCL, KU Leuven, 2004. [url]