Towards the automatic detection of syntactic differences

Kroon, M.S.

Towards the Automatic Detection of Syntactic Differences

Published by

LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461
e-mail: lot@uva.nl
http://www.lotschool.nl

Cover illustration: Kalle Wolters

ISBN: 978-94-6093-414-8
DOI: https://dx.medra.org/10.48273/LOT0629
NUR: 616

Towards the Automatic Detection of Syntactic Differences

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 10 november 2022
klokke 10.00 uur

door

Martin Siebren Kroon

geboren 6 november 1993
te Groningen, Nederland

Promotores:	Prof. dr. L.C.J. Barbiers
	Prof. dr. J.E.J.M. Odijk (Universiteit Utrecht)
Co-promotor:	Dr. S.L. van der Pas (Amsterdam UMC)
Promotiecommissie:	Prof. dr. S.A. Raaijmakers
	Prof. dr. ir. J. Nerbonne (Rijksuniversiteit Groningen)
	Prof. dr. P.D. Grünwald
	Prof. dr. G.J.M. van Noord (Rijksuniversiteit Groningen)
	Dr. J. Prokic

The research reported in this thesis was funded by the Data Science Research Programme and the Faculty of Humanities of Leiden University.

Dedicated to my dear parents.

Contents

Acknowledgements

There are many people I owe thanks to, either for their direct involvement in my research endeavours during my PhD, for supporting me outside of my work, or even by simply existing. In fact, everyone that I had the pleasure of talking to ever since I moved to Leiden in 2017, I am to some extent indebted to. However, a few people I am particularly thankful to.

First and foremost I want to thank my three brilliant supervisors Sjef, Jan and Stéphanie, who have advised and guided me throughout this sometimes difficult project. You have taught me many things, and I could not have been luckier with you as my supervisors. Our collaboration was very enjoyable to me and I hope to be working together again on a new project at some point in the future.

Then, I want to thank my sweet parents, who have supported and loved me unconditionally – not only during my time in Leiden, but all my life. I count myself lucky with you as my parents and I treasure the way you have raised me. As my great-grandfather would have said in his native Bolserters: *kist op un minder like*.

I also want to thank my big brother Peter, who has often had the time to distract me with fun games as well as to stimulate and motivate me to perform research and to always stay curious. You have often been an example to me, and I am proud to be your little brother.

A great deal of thanks I owe to my Pascale. Your love and support seem unending and make me a very happy man indeed, even during the dark days of the pandemic.

A big thank you also goes out to my friends, who have proven to be fantastic people by keeping me sane during my time off. In particular Max, with whom I have learned to check whether people actually left before taking their cheese platter; Laura odB., for introducing me to Leiden and making me feel at home in a city I didn't know; Xander, Lotte, Laura D., and Vera, with whom I have developed a nearly cultist fascination with a type of Cypriot cheese; Hanjo and Froos Bommee, who have led my linguistic creativity into a wanton subversion

of reality; and Timo, Susan and Janoël, who even after all those years remain among my closest of friends.

I was also kept relatively sane during my time at work when the country wasn't in lockdown, so I am also very grateful to all of my colleagues at LUCL, my colleagues of the DSO and my colleagues of the LUCDH, chief among which are Xander and Lis (who I am also honoured and grateful to call my paranymphs), Laura odB., Manolis and all the data-science PhDs. Special thanks go out to Astrid and Anikó, who helped me analyse some of my data, and to the staff of ÚFAL of the Charles University in Prague, who have so graciously hosted me in the autumn of 2019.

Finally, I want to thank Toneelgroep Imperium – especially those with whom I've worked together closely during productions. Thank you for accepting me into your theatrical family with open arms. I thoroughly enjoy spending time and sharing the spotlight with you, and hope to continue to do so for many years to come.

Glossary of abbreviations, acronyms and notations

AAA	Affix-Attribute Associator
ADJ	adjective (UD tag)
ADP	adposition (UD tag)
ADV	adverb (UD tag)
`advmod`	adverbial modifier (UD relation)
AL list	the list containing characteristic differences between Hungarian and English as compiled by dr. Lipták
`amod`	adjectival modifier (UD relation)
Art	article (UD feature value)
AUC	area under the (ROC) curve
`aux`	auxiliary verb (UD relation)
AUX	auxiliary verb (UD tag)
`cc`	coordinating conjunction (UD relation)
`ccomp`	clausal complement (UD relation)
CCONJ	coordinating conjunction (UD tag)
`conj`	conjunct (UD relation)
CoNLL-U	a specific annotation format; short for *Conference on Computational Natural Language Learning–Universal Dependencies*
CS	Czech
DE	German
Def	definite (UD feature value)
Definite	definiteness (UD feature)
deprel	dependency relation

DeSDA	Detecting Syntactic Differences Automatically
`det`	determiner (UD relation)
DET	determiner (UD tag)
DGAE	Data Grouper for Attribute Exploration
EN	English
EU	European Union
Fin	finite (UD feature value)
GED	graph-edit distance
Gen	genitive (UD feature value)
GTI	Generalization Tree Inducer
Hn	hypothesis number n, referring to Table 4.3
HU	Hungarian
Ind	1. indefinite (UD feature value); 2. indicative (UD feature value)
Inf	infinite (UD feature value)
Intrans	intransitive
KJB	King James Bible
Lev.	Levenshtein distance
MDL	Minimum Description Length
Mn	missed difference number n, referring to Table 4.4
N	noun
n-gram	a contiguous sequence of n items
NL	Dutch
NLP	natural language processing
`nmod`	nominal modifier (UD relation)
NP	noun phrase
`nsubj`	nominal subject (UD relation)
`nsubj:pass`	passive nominal subject (UD relation)
Num	number (UD feature)
NUM	numeral (UD tag)
`obj`	object (UD relation)
`obl`	oblique nominal (UD relation)
OSV	object-subject-verb order
OVS	object-verb-subject order
Part	participle (UD feature value)
PART	particle (UD tag)

Plur	plural number (UD feature value)
pmi	pointwise mutual information
POS	part of speech
PRON	pronoun (UD tag)
PronType	pronoun type (UD feature)
PROPN	proper noun (UD tag)
Prs	personal or possessive personal pronoun or determiner (UD feature value)
PRS	present tense
PST	past tense
PUNCT	punctuation (UD tag)
ROC curve	receiver operating characteristic curve
S	sentence
SCONJ	subordination conjunction (UD tag)
Sent. length	sentence length
Sing	singular number (UD feature value)
sg.	singular number
SOV	subject-object-verb order
SQS	'Summarising event seQuenceS'; Tatti and Vreeken (2012)
SV	subject-verb order
SVO	subject-verb-object order
Trans	transitive
UD	Universal Dependencies; Nivre et al. (2016)
V	verb
V2	verb-second word order
VB	Vizsoly Bible
VBP	verb, non-3rd person singular present (Penn Treebank tag)
VO	verb-object order
VP	verb phrase
WEB	World English Bible
wh	question-, interrogative
X	other (UD tag)
<…>	used to identify an individual grapheme or character
$x\rangle\ldots\langle y$	used to mark off circumfixes