LIVE and LEARN

Festschrift in honor of Lars Borin

Człowiek uczy się przez całe życie

Greek: Να ζεις και να μαθαίνεις

Estonian: Elada ja õppida

Luxemb: Liewen a léiren

Italian: Non si finisce mai di imparare

Vsak dan se naučimo kaj novega
Slovene

Czech: Žít a učit se

Chinese: 活到老, 學到老

Bulgarian: Човек научава нещо всеки ден

Russian: Век живи, век учись

Spanish: Vive y aprende

German: Man lernt nie aus

Man lär så länge man lever

Finnish: Elää ja oppia

Ukranian: Вік живи, вік учися

Latvian: Mūžu dzīvo, mūžu mācies

French: Il n'est jamais trop tard pour apprendre

Dutch: Een mens is nooit te oud om te leren

Hindi: जिओ और सीखो

Hebrew: תחיה ותלמד

Persian: تا زنده هستی دانش بجو

Catalan: Viu i aprèn

Urdu: جیو اور سیکھو

Slovak: Žiť a učiť sa

Nynorsk|Bokmål: Lev og lær

Turkish: Insan her gün bir şey öğreniyor

INSTITUTIONEN FÖR SVENSKA,
FLERSPRÅKIGHET OCH SPRÅKTEKNOLOGI

GÖTEBORGS
UNIVERSITET

GU-ISS-2022-03

Forskningsrapporter från Institutionen för svenska, flerspråkighet
och språkteknologi, Göteborgs universitet
Research Reports from the Department of Swedish, Multilingualism,
Language Technology

ISSN 1401-5919

LIVE and LEARN

Festschrift in honor of Lars Borin

Editors: Elena Volodina, Dana Dannélls,
Aleksandrs Berdicevskis, Markus Forsberg, Shafqat Virk

GU-ISS-2022-03

Forskningsrapporter från Institutionen för svenska, flerspråkighet och
språkteknologi, Göteborgs universitet
Research Reports from the Department of Swedish, Multilingualism,
Language Technology, University of Gothenburg

Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
Box 200
SE-405 30 Gothenburg

ISSN 1401-5919
ISBN 978-91-87850-82-0 (print)
ISBN 978-91-87850-83-7 (digital)

Cover: Design by Karin Wenzelberg based on word cloud illustration by Elena Volodina
Compilation: Elena Volodina, Dana Dannélls and Gerlof Bouma
Photo of Lars Borin by Johan Wingborg
Printed by BrandFactory AB Kållered 2022

Festschrift in honor of Lars Borin

Preface

This volume is dedicated to Lars Borin, who for many years has been our dear colleague and an inspiring scientific leader.

Lars Borin, born on February 2, 1957, is a professor of Natural Language Processing at the University of Gothenburg, and a co-director of the R&D unit Språkbanken Text at the same university. He is also the director of Nationella språkbanken, a Swedish nationwide e-infrastructure for language technology, and the director of the Swedish node of CLARIN ERIC, an EU infrastructure for language technology.

Lars' research activities are directed at the development of language resources and tools for all contemporary and historical varieties of written Swedish. Methodological and theoretical links between general linguistics and NLP are one of his central research interests. Such links are crucial both for linguistic research and for the incorporation of the knowledge gained through this research into increasingly sophisticated language processing systems. He specializes in many fields, the most prominent being language technology infrastructure, digital language resources, digital historical linguistics, computational lexicography, lexical semantics, language typology, digital humanities, computer-assisted language learning, multi-word expressions, and text corpora. Lars is well-known for his work at Språkbanken Text, where he has been and still is a mastermind and driving force behind making large collections of corpora and computer-readable lexicons of modern and historical Swedish. He has always been concerned about making all of the resources openly available for users, both for download and for searches through web search interfaces. Over the course of his career, he has strengthened the status of Swedish language technology in Sweden and in the world. Thanks to his efforts, Språkbanken's resources and tools have become the source for developing state-of-the-art methods in various research fields such as Swedish linguistics, computational linguistics, historical linguistics, digital humanities and social science, history, language acquisition, and many others.

It is with great pleasure that we present Lars with this Festschrift to honor his lasting contributions, nationally and internationally, and to highlight his importance to his friends and colleagues at the University of Gothenburg and elsewhere in the world. The Tabula Gratulatoria included in this volume lists the names of many friends and colleagues, albeit certainly not all, who want to join in paying tribute to Lars on his 65th anniversary. The contributions to the Festschrift reflect only a fraction of Lars' scientific interests. They come from his friends and colleagues around the world and deal with topics that have been – in one way or another – inspired by his work. A common theme for the articles is the never-ending need to learn, which is alluded to in the title of the volume, *Live and Learn*.

Gothenburg, November 2022
The editors

Tabula Gratulatoria

Yvonne Adesam
Magnus Ahltorp
Lars Ahrenberg
Karin Aijmer
Cai Alfredson
David Alfter
Christiane Andersen
Karin Andersson
Aleksandrs Berdicevskis
Ulf Bjereld
Lars Björk
Kristian Blensenius
Gerlof Bouma
Johan Boye
Daniel Brodén
Lars Burman
Love Börjeson
Nicoletta Calzolari
Dick Claésson
Bernard Comrie
Robin Cooper
Evie Coussé
Mats Dahllöf
Dana Dannélls
Koenraad De Smedt
Marie Demker
Simon Dobnik
Rickard Domeij
Jens Edlund
Adam Ek
Elisabet Engdahl
Stina Ericsson
Gunnar Eriksson
Ghazaleh Esfandiari Baiat
Markus Forsberg
karin Friberg Heppin
Johan Frid
Mats Fridlund
Antoaneta Granberg
Johannes Graën
Eleni Gregoromichelaki
Normunds Grūzītis
Marianne Gullberg
Roger Gyllin
Martin Hammarstedt
Harald Hammarström
Karin Helgesson
Simon Hengchen
Louise Holmer
David House
Christine Howes
Lars Ilshammar
Jonas Ingvarsson
Sofie Johansson
Richard Johansson
Arne Jönsson
Mats Jönsson
Jelena Kallas
Jussi Karlgren
Susanna Karlsson
Jenny Kierkemann
Malin Klang (f.d. Ahlberg)
Per Klang (f.d. Malm)
Dimitrios Kokkinakis
Marco Kuhlmann
Murathan Kurfalı
Hans Landqvist
Shalom Lappin
Staffan Larsson
Ann Lillieström
Anna Lindahl
Cecilia Lindhé
Therese Lindström Tiedemann
Krister Lindén
Peter Ljunglöf
Sharid Loáiciga
Benjamin Lyngfelt
Lennart Lönngren
Mats Malm
Arianna Masciolini
Arild Matsson
Beáta Megyesi

Magnus Merkel
Detmar Meurers
Tommaso M. Milani
Yousuf Ali Mohammed
Felix Morger
Ricardo Muñoz Sánchez
Gunta Nešpore-Bērzkalne
Jenny Nilsson
Kristina Nilsson Björkenstam
Sanni Nimb
Joakim Nivre
Catrin Norrby
Joel Olofsson
Sussi Olsen
Anders Olsson
Leif-Jöran Olsson
Bolette Pedersen
Stellan Petersson
Miriam R. L. Petruck
Eva Pettersson
Ildikó Pilán
Julia Prentice
Taraka Rama
Aarne Ranta
Judy Ribeck Nyström
Lena Rogström
Jacobo Rouces González
Johan Roxendal
Stian Rødven-Eide
Magnus Sahlgren
Natalia Sathler Sigiliano
Baiba Saulīte
Anju Saxena
Anne Schumacher
Maria Skeppstedt
Emma Sköldberg
Sara Stymne
Anna Sågvall Hein
Nina Tahmasebi
Jennica Thylin-Klaus
Jörg Tiedemann
Tiago Timponi Torrent
Maria Toporowska Gronostaj
Jonatan Uppström
Shafqat Virk
Martin Volk
Elena Volodina
Michelle Waldispühl
Barbro Wallgren Hemlin
Åsa Wengelin
Lena Wenner
Søren Wichmann
Mats Wirén
Victor Wåhlstrand Skärström
Niklas Zechner
Torsten Zesch
Alexander Ziem
Maria Öhrman
Robert Östling
Lilja Øvrelid

Contributed papers

Att vara Lars: Några tankar om språkteknologi och socioonomastik
Lars Ahrenberg 1

We may actually all die tomorrow... nevertheless: Predicting short-term frequency changes in Swedish neologisms
Aleksandrs Berdicevskis, Yvonne Adesam and Evie Coussé 5

avokado-r/-er/-s/-sar
Kristian Blensenius and Louise Holmer 13

Counting dirty words: The effect of OCR quality on token statistics in historical Swedish corpora
Gerlof Bouma and Yvonne Adesam 17

Investigating a linguistic mini landscape: The Tsez (Dido) dialect dictionary project
Bernard Comrie 25

Beyond strings of characters: Resources meet NLP – Again
Dana Dannélls, Tiago Timponi Torrent, Natalia Sathler Sigiliano and Simon Dobnik 29

Ordvektorer i lexikografiskt arbete
Markus Forsberg and Emma Sköldberg 37

The diachrony of political terror: Tracing terror and terrorism in Swedish parliamentary data 1867-1970
Mats Fridlund, Daniel Brodén and Victor Wåhlstrand Skärström 43

UD-based Latvian FrameNet
Normunds Grūzītis, Gunta Nešpore-Bērzkalne and Baiba Saulīte 49

The rise and fall of grammatical theories in descriptive grammars of the languages of the world
Harald Hammarström 55

Coveting your neighbor's wife: Using lexical neighborhoods in substitution-based word sense disambiguation
Richard Johansson 61

Linguistics concepts as semantic frames
Per Klang and Shafqat Mumtaz Virk 67

Deep learning models as theories of linguistic knowledge
Shalom Lappin 73

The case for *Språkbanken Dialog*
Staffan Larsson, Christine Howes and Eleni Gregoromichelaki 79

Ordbildning på icke-verbal grund. Om integrering av ikoner och ljudeffekter i grammatiken
Benjamin Lyngfelt and Joel Olofsson 83

Building a multilingual AWE tool for L2 learners: Challenges and ideas
Arianna Masciolini 89

Flexible Universal Dependencies using nested dependency graphs
Joakim Nivre 95

Perspective_on: Semantic relations for frames and constructions
Miriam R.L. Petruck and Alexander Ziem 101

Natural language processing for educational applications: Recent advances
Ildikó Pilán 107

Detecting fake papers with the latent algorithm for recursive search
Stian Rødven-Eide and Ricardo Muñoz Sánchez 111

Leksikalsk-semantiske sprogressourcer: Hvad kan de, og hvordan udvikler vi dem bedst?
Bolette Sandford Pedersen, Sanni Nimb and Sussi Olsen 115

Den ena texten och den andra: Visualisering av textpar med hjälp av ordmoln
Maria Skeppstedt, Gunnar Eriksson, Magnus Ahltorp and Rickard Domeij 121

Datorlingvistikens vagga i Uppsala
Anna Sågvall Hein 127

From open parallel corpora to public translation tools: The success story of OPUS
Jörg Tiedemann 133

Binomials in Swedish corpora – 'Ordpar 1965' revisited
Martin Volk and Johannes Graën 139

ICALL: Research versus reality check
Elena Volodina and David Alfter 145

*Lyx*ig *språk*lig födelse*dag*spresent from the Swedish Word Family
Elena Volodina, Yousuf Ali Mohammad and Therese Lindström Tiedemann 153

Annotating the narrative: A plot of scenes, events, characters and other intriguing elements
Mats Wirén, Adam Ek and Murathan Kurfalı 161

The other SAT-Solver: Applying lexicons to SweSAT word questions
Niklas Zechner 167

Mot en mänskligare maskinöversättning
Robert Östling 171

Att vara Lars: Några tankar om språkteknologi och socioonomastik

Lars Ahrenberg
Institutionen för datavetenskap
Linköpings universitet, Sverige
lars.ahrenberg@liu.se

Abstract

Since the SweClarin project began in 2015 its resources in terms of data and tools have been used in many different projects including linguistics. A research area where they have been less employed is the study of names. In this paper I suggest that language technology and general corpora can be used to contribute to the sociological study of personal names and offer a few examples. As is fit for the occasion I take *Lars* as the point of departure.

1 Namnforskning och språkteknologi

Namnforskning är ett forskningsområde med långa anor. Om det traditionellt hade ett fokus på ortnamn och etymologier har området vidgats och omfattar i dag många olika slags namn och beforskas med olika metoder. Ett livaktigt delområde är socioonomastiken, eller det sociolingvistiska studiet av namn.[1] Även om antalet forskare är litet har på senare tid ett antal initiativ tagits för att utveckla området i Norden. Till den samarbetskommitté, NORNA, som funnits sedan 1971 finns nu ett forskarnätverk, *New Trends in Nordic Socio-onomastics* med en aktiv webbplats, och en tidskrift, *Nordisk tidskrift för socioonomastik* som utkom med sitt första nummer 2021. Den publicerar, enligt sin hemsida, vetenskapliga artiklar som behandlar egennamnens roll i samhället och i social interaktion. Den är tvärvetenskaplig och välkomnar bidrag från olika discipliner. Det innebär att författare tillåts använda en bredd av teorier, metoder och perspektiv för att analysera namn liksom att kombinera olika typer av data.[2]

Jag har letat i ovan nämnda fora efter artiklar och blogginlägg som använder storskalig korpusanalys eller språkteknologi i någon form, dock utan att hitta några. Metodmässigt används förutom register också texter av olika slag: enkäter, intervjuer, inspelade samtal och historiska källor. Stora korpusar av det slag som Språkbanken Text tagit fram genom åren och användning av språkteknologiska verktyg lyser däremot med sin frånvaro, detta trots att ett infrastrukturprojekt som Swe-Clarin nu varit i gång sedan 2015. Jag tycker därför att det är på sin plats att spekulera över hur språkteknologi skulle kunna bidra till namnforskning. Jag begränsar mig här till förnamn, specifikt med utgångspunkt i mitt eget (och dagens jubilars), alltså *Lars*, och frågar mig hur sådana material som den så kallade Gigawordkorpusen (Eide et al., 2016) kan användas för detta syfte. En underliggande fråga är om vi, som i Sverige under mitten av det förra århundradet givits tilltalsnamnet *Lars*, har haft nytta av det. Olika synlighet i exempelvis nyhetsmedia skulle kunna vara en indikation på att namnet gör skillnad.

Personnamn, och specifikt förnamn, bär på sociala betydelser. De flesta förnamn i Sverige får oss att dra mer eller mindre säkra slutsatser om kön, ålder, etnicitet, familjebakgrund, kulturell och religiös tillhörighet, med mera. Man kan till exempel jämföra *Lars* med *Lauri, Laurent* eller *Laura*. Emilia Aldrin studerade i sin avhandling via enkäter och intervjuer föräldrars inställningar till namnval för deras nyfödda och menar att dessa kan karaktäriseras i termer av social positionering (Aldrin, 2011, 67f). *Lars* ingår inte i dessa diskussioner men utifrån hennes kategorier kategoriserar jag det som svenskorienterat, snarare än internationellt, traditionellt snarare än modernt, och vanligt snarare än originellt.

[1] https://www.nordicsocioonomastics.org/about-socio-onomastics/

[2] https://gustavadolfsakademien.se/tidskrifter/tidskrift/nordisk-tidskrift-for-socioonomastik-nordic-journal-of-socio-onomastics

Lars Ahrenberg. 2022. Att vara Lars: Några tankar om språkteknologi och socioonomastik. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 1–4. Available under CC BY 4.0

	1920-29	1930-39	1940-49	1950-59	1960-69
Placering	7	3	1	1	5
Antal	3799	12213	26570	24130	18810

Tabell 1: *Lars* som tilltalsnamn under perioden 1920-1970. Källa: SCB:s namnstatistik, tabell tilltalsnamn-man-per-decennium-1920-2020-topp-10.

2 Vad kan språkteknologisk infrastruktur bidra med?

Sociala betydelser borde kunna studeras även i stora korpusar som Gigawordkorpusen. Att dessa inte är socialt neutrala är välkänt, vilket ofta uppfattas som negativt, som en 'bias' vilken måste åtgärdas för att t.ex. de modeller som genereras från dem ska kunna användas. Men för att studera sociala betydelser är det snarare nödvändigt att korpusen ger en så representativ bild som möjligt av den tid eller det sammanhang den omfattar. På korpusen kan man sedan tillämpa de metoder som språkteknologin utvecklat för att modellera data och då specifikt namnanvändning. Namnanvändning kan jämföras via samförekomster med begrepp eller via sentimentanalys. Även ordinbäddningar borde kunna användas på samma sätt som exempelvis (Garg et al., 2018) analyserat etniska stereotyper.

Man kan invända att textkorpusar inte handlar om namn utan om deras referenter, för personnamn alltså om de personer som namnges. Kan vi utgå från enskilda beskrivningar av dessa personer, vad de gör och vad de utsätts för, till utsagor om namnen som används? Jag vill hävda att vi kan det, under förutsättning att korpusen är representativ för sin tid och så pass stor att den omfattar tillräckligt många personer med samma namn. I så fall kan vi betrakta dessa personer som en social kohort utifrån deras namn. De skillnader som eventuellt finns i modeller och språkliga associationer som vi kan ta fram kan då knytas till namnet snarare än till de enskilda personerna. Visserligen är sådana skillnader i grunden statistiska i den mån de kan säkerställas; men de kan ändå vara intressanta och frågan om vad de kan bero på är främst en uppgift för namnsociologin att besvara.

3 Lars

Lars har använts som tilltalsnamn i Sverige åtminstone sedan tidig kristen tid. Såväl hög som låg har burit namnet; biskopslängden för Linköpings stift omnämner en biskop Lars under 1200-talet, men även fattigfolket har hetat så, som i Frödings dikt Lars i Kuja (... *ty allt som växer åt Lars är sten och sten är dålig förtäring*).

Namnet hade en lång period av hög popularitet som tilltalsnamn under 1900-talet, framför allt under 40- och 50-talen, se Tabell 1. Det var enligt SCB alltjämt det vanligaste tilltalsnamnet för svenska män år 2021.[3] Efter 1970 har populariteten avtagit för att under 2000-talet ha handlat om ett tjugotal eller ännu färre nyfödda pojkar som givits det som tilltalsnamn.

4 Data

Data för analyserna är huvudsakligen hämtade från Gigaword-korpusen sammanställd vid och nerladdningsbar från Språkbanken Text (Eide et al., 2016). Jag har begränsat mig till nyhetstexterna och tre årtionden 1990-tal, 2000-tal, 2010-tal, där det senare decenniet slutar med 2013 för nyhetstexterna. Nyhetstexterna valdes därför att de bäst speglar det offentliga Sverige. För vissa analyser delades materialet från perioden 2000-2009 upp på tre delkorpusar och det från 2010-2013 i två delkorpusar. Antal meningar och token framgår av Tabell 2.

5 Analysexempel

Det faktum att *Lars* är ett så vanligt namn borde innebära att det är vanligt också i korpusen. Så är det också, men det är inte vanligast. *Lars* är det vanligaste mansnamnet i 1990-talsdelen men sett över hela

[3] https://www.scb.se/hitta-statistik/sverige-i-siffror/namnsok/

Delkorpus	Meningar	Tokens
news1990	6,321,173	95,435,081
news2000-01	5,700,000	99,093,472
news2000-02	5,700,000	99,240,929
news2000-03	6,012,341	90,238,180
news2010-01	5,200,000	86,840,551
news2010-02	5,592,318	82,157,754
Alla	34,525,832	553,005,967

Tabell 2: Antal meningar och token i nyhetsdelen av Gigawordkorpusen.

delkorpusen är *Anders* och *Peter* vanligare, se Tabell 3. Att ta fram frekvensdata för alla namn i korpusen kräver disambiguering; vi vill bara ha med de förekomster som faktiskt anger en person. Att göra detta exakt är inte helt enkelt, eftersom många av de vanligaste namnen (*Lars, Göran, Helena, ...*) ingår i namn på andra saker som kyrkor, stadsdelar och sjukhus, medan andra namn som *Stig, Bo, Sten, ...* kan vara något annat än egennamn. Baserat på stickprov bedömde jag att cirka 44% av alla förekomster av *Hans* är possessiva pronomina. Man bör också ha i åtanke att många namn är populära utanför Sverige, så som till exempel *Peter*, som kommer högt upp på listan.

Delkorpus	Vanligaste mansnamn i ordning
news1990-99	Lars, Anders, Peter, Jan, Göran
news2000-09	Anders, Peter, Lars, Johan, Fredrik
news2010-13	Anders, Johan, Peter, Fredrik, Lars
Hela (1990-2013)	Anders, Peter, Lars, Johan, Fredrik

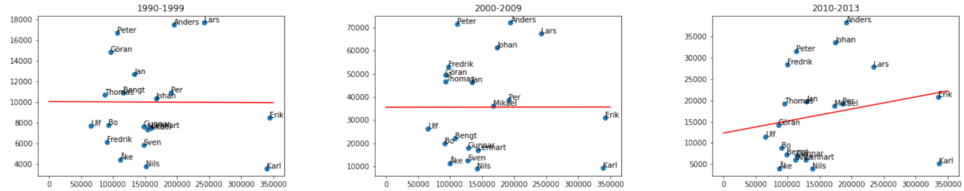
Tabell 3: De fem vanligaste mansnamnen i olika delar av Gigawordkorpusen.

En fråga värd att undersöka är hur sambandet ser ut mellan förekomst av namn i nyhetsmedia och förekomst av namnen i befolkningen. För att undersöka det har jag prövat att korrelera namnstatistiska data från SCB:s tabeller med frekvensdata i Gigawordkorpusens nyhetsdel för olika perioder. Figur 1 visar hur förekomst i nyhetsdelen av korpusen förhåller sig till befolkningsstatistik. I den figuren används en tabell över förnamn bland folkbokförda respektive decennium, men jämförelser med tilltalsnamn på nyfödda från tidigare decennier ger liknande resultat: *Lars* är konsekvent vanligare i nyhetstexterna än i befolkningsstatistiken och mest accentuerat är detta under 1990-talet. Utifrån sådana data kan man våga formulera en hypotes om att det var gynnsamt i Sverige att döpas till *Lars* under 1900-talets mitt. Att namnet förekommer i seriös tidningspress innebär i de flesta fall att referenten är framgångsrik på något sätt, det må vara inom sport, politik, kulturliv eller något annat. Dock finns många möjliga felkällor att beakta: namn i korpusen kan referera till andra företeelser och andra personer än dem som är svenskfödda i det antagna intervallet, SCB grupperar olika stavningar under ett namn, med mera.

Vi kan undersöka hypotesen vidare genom att titta på samförekomster mellan namn och andra ord i korpusen. I Tabell 4 visar vi vilka namn som oftast kopplas till titeln professor i delkorpusen från 1990-talet och vilka namn som ofta uttalar sig (via ordet *säger*). Som en kontrast kan vi jämföra med vilka namn som oftast samförekommer med *spelar*, ett verb som är vanligare inom domäner som idrott och kultur. Med en parsad korpus skulle sådana undersökningar kunna göras mer uttömmande. Skillnaden mellan *Lars* och de andra vanligaste namnen *Anders* och *Peter* är kanske mest intressant. De senare spelar oftare men uttalar sig mindre.

För att se vilka namn som är mest lika *Lars* kan vi använda ordinbäddningar. Här har jag använt Word2Vec (Mikolov et al., 2013) i ramverket Gensim för alla sex delkorpusarna.[4] Lars-vektorns 10 när-

[4] https://radimrehurek.com/gensim/index.html



Figur 1: Samband mellan förekomst av valda mansnamn i befolkningen och i olika delar av nyhetsdelen av Gigawordkorpusen. Befolkningsdata är hämtat från SCB:s statistikdatabas, de 100 vanligaste förnamnen bland folkbokförda 31 december respektive år för åren 1980–2021.

Namn	**professor**	**säger/sade**	**spelar/spelade**
Lars	140	113	46
Björn	50	43	23
Göran	42	113	15
Jan	39	99	22
Lennart	47	60	13
Bengt	44	49	27
Peter	31	73	61
Anders	25	76	71

Tabell 4: Samförekomster mellan förnamn och valda ord i news1990.

maste grannar noterades i vart och ett av dem. Inget namn förekom i alla sex grannskapen men återkommande var Bengt, 5 ggr, och Jan, Christer, Lennart och Ulf, 4 ggr. Noterbart är att 55 av 60 inbäddningar representerar mansnamn, att dessa är traditionella svenska namn med antingen biblisk och/eller nordisk historia. Frånvaron av internationella namn som Peter och Thomas är här påfallande. Kvinnonamnen uppvisar ett liknande mönster. Mer än 90% av namnen i den närmaste omgivningen av ett givet kvinnonamn är kvinnonamn.

Mina slutsatser är att stora textkorpusar och språkteknologi visst borde kunna spela en roll i namnsociologin och att det finns en hel del intressanta metodologiska utmaningar.

Referenser

Emilia Aldrin. 2011. *Namnval som social handling: val av förnamn och samtal om förnamn bland föräldrar i Göteborg 2007–2009*. Ph.D. thesis, Institutionen för nordiska språk, Uppsala universitet. Namn och samhälle 24.

Stian Rødven Eide, Nina Tahmasebi, & Lars Borin. 2016. The Swedish culturomics Gigaword corpus: A one billion word Swedish reference dataset for NLP. In *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, Krakow*, number 126, pages 8–12. Linköping University Electronic Press.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, & James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Tomas Mikolov, Kai Chen, Greg Corrado, & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

We may actually all die tomorrow... nevertheless: Predicting short-term frequency changes in Swedish neologisms

Aleksandrs Berdicevskis **Yvonne Adesam**
Språkbanken Text, Dept of Swedish, Multilingualism, Language Technology
University of Gothenburg, Sweden
{aleksandrs.berdicevskis,yvonne.adesam}@gu.se

Evie Coussé
Dept of Languages and Literatures
University of Gothenburg, Sweden
evie.cousse@gu.se

Abstract

Predicting the future is difficult, as Lars Borin likes to point out by saying the phrase which is included in the title of this paper. Nevertheless, we attempt to predict short-term changes in the frequency of new Swedish words based on some measures of their linguistic and social dissemination. We show that it is possible to predict the direction of change with a higher-than-baseline accuracy. Most interestingly, we show that predictions are much less accurate for those words that denote new phenomena than for those who are new signifiers for already existing phenomena.

1 Introduction

When doing research on language change, linguists usually try to explain the changes, either explicitly or implicitly, either putting forward hypotheses about causal links or making silent assumptions about such links. In our view, the most rigorous means to test such hypotheses and assumptions is to attempt to *predict* language change.

Predicting the future is notoriously difficult (and annoying: as follows from the title, there is always a certain risk that the researchers will not be able to evaluate their own predictions). Fortunately, there is an easier to way to evaluate the predictive power of a theory: splitting the data into a seen and an unseen set, training a theory-based model on the former one and testing it on the latter.

In this paper, we use this approach in order to explore whether short-term frequency changes of Swedish neologisms can be predicted from corpus data. Since neologisms are likely to experience such changes (either an increase, if they become established in the language, or a decrease if they fail to do so), they are a favourable ground for this kind of predictions.

We focus on comparing how successful predictions are for words that denote new phenomena and those that are new signifiers for already existing phenomena. We hypothesize that the former would be more difficult to predict, since they are more dependent on language-external events and not on linguistic and sociolinguistic processes, which we can hope to capture by our measures. Our results support this hypothesis.

2 Data

Språkrådet, the Swedish language council (in the last decade together with the magazine *Språktidningen*), releases a list of new words every year.[1] The words on this neologism list are supposed to have come into use, or gained in use, in the last year. The final list is in no way a complete (if there could even be such

[1] https://www.isof.se/stod-och-sprakrad/spraktjanster/nyordslistor.

Aleksandrs Berdicevskis, Yvonne Adesam and Evie Coussé. 2022. We may actually all die tomorrow... nevertheless: Predicting short-term frequency changes in Swedish neologisms. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 5–11. Available under CC BY 4.0

a thing) list of new words, but rather a list of words that the compilers consider especially interesting or telling about the past year (Karlsson, 2021).

Our work starts with these neologism lists for the past twenty years, 2003–2021. The main reasons for using them are that they are readily available, and a well-known part of Swedish linguistic debate. While the lists are based on expert knowledge, one of the weaknesses for our purposes is their selection procedure. Words are picked in part based on their societal relevance, but mostly, which may not be as relevant for us, with the purpose of language cultivation, to show current trends and the breadth in patterns of word formation and language creativity (Karlsson, 2021).

The restriction in time depends on the corpora that we use. This particular study primarily reports on data from Flashback, while also consulting Familjeliv in the initial stages of the data extraction process. Both corpora represent very large Swedish discussion forums covering a broad range of topics. With more than eight billion tokens, we expect the corpora to be large enough to show new words from the onset, a point when they still will be very infrequent. We assume that the language in this material is closer to everyday use and less edited than many other types of text, which may mean that non-normative language is more frequent, or shows up earlier than in edited texts such as news. All corpora are provided through Språkbanken Text and its corpus infrastructure Korp (Borin et al., 2012).[2]

We first extracted all words from the neologism lists, removing multi-word units for ease of searching.[3] We then gathered statistics about frequencies for these words in the discussion forums Flashback and Familjeliv, by matching each word from the list to its lemma, and in case the lemma is not available, to the word form directly. Not all words in the corpora have lemma annotated, e.g. because no match for the word can be found in the Saldo lexicon (Borin et al., 2013). It should also be noted that not all words in the neologism lists are in their base form.

Many words in the full neologism list have few, if any, instances in the corpora, and will thus be difficult to track over time. We therefore selected the words with the highest frequencies from this list. For each word we added so-called *lemgrams*, lexicon identifiers from the Saldo lexicon, which give us all inflected forms. For words not in the lexicon, we manually added all inflected forms.

In this process, we also removed a number of words which would give us too many erroneous matches in the corpora, for example *VAR* ‘video assistant referee’, which happens to coincide with the verb form *var* ‘was’ and the wh-word *var* ‘where’, or *manga* ’Japanese comics’, which turned out to be mostly instances of a misspelled *många* ‘many’. For the same reason, we also excluded words which were not new themselves, but acquired a new meaning, if this meaning was relatively infrequent compared to the previous meaning(s); for instance, *spår* ‘education track’ (general meaning: ‘track’). We kept the words where the situation was the reverse: the frequency of the older meaning(s) was small; e.g. *buda* ‘to make a bid’ (older meaning: ‘to send with a courier’).

Some words in the list were spelling or pronunciation variants, e.g. *babybio* and *bebisbio* ‘adapted movie showings for parents with babies’ and these were joined into one item in our list. For other entries we added spelling variants, e.g. *covid* and *covid19* to *covid-19*. We did not merge words that are derived from the same stem, such as *blogg* ‘a blog’, *blogga* ‘to blog’, *bloggare* ‘blogger’.

In the end, we had a list of 75 words, which all had absolute frequencies of more than 2300 in Flashback and Familjeliv taken together, see Appendix A. We labelled each word in this list as either denoting a new phenomenon (e.g. *covid-19*) or being a new signifier for a previously existing phenomenon (e.g. *buda* ‘to make a bid’; *prio* ‘a priority’, a shortening from *prioritet*). A *new* phenomenon is one which is (relatively) new for most part of Swedish society (e.g. *anime* ’Japanese comics’ is not new as phenomenon in Japan, nor, perhaps, is it among its early fans in Sweden, but it was largely unknown to the mainstream public in Sweden before 2003). Dealing with numerous borderline cases, we tried to establish whether a change occurs in the language (and discourse) or in the material world. This process has been further complicated by the fact that in many cases either the signifier, or the phenomenon, or both, are not actually new, but experienced a substantial increase in frequency. In total, we end up with 42 “new phenomena” and 33 “new signifiers”.

[2] `spraakbanken.gu.se/korp`

[3] Multi-word expressions may be explored in future research.

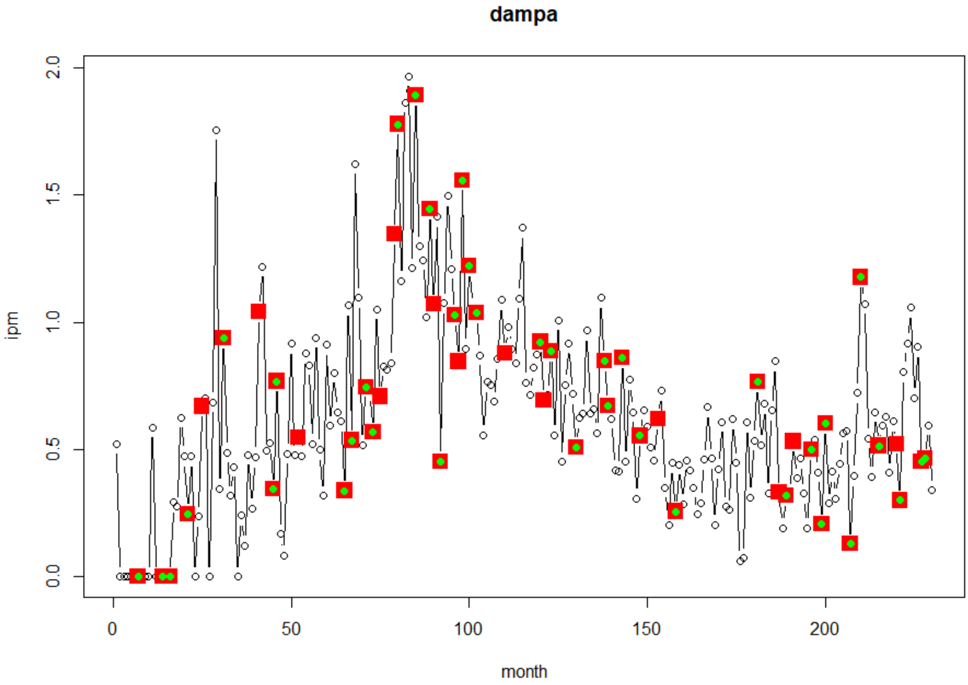


Figure 1: A visualization of our method for the word *dampa* ‘freak out’. Large filled squares represent those months that are in the test set and for which the direction of change has to be predicted (from the previous month), small filled circles represent the months for which the direction of change has been correctly predicted by a randomly chosen model.

3 Methods

Stewart & Eisenstein (2018) show that changes in the frequency of new words on English reddit can rather successfully be predicted using measures of linguistic dissemination and social dissemination, of which the former are better predictors. We try to reproduce their success using a similar methodology.

Our task is to predict for a given word whether its frequency (normalized by corpus size, hereafter relative frequency) will decrease or not in month $n+1$, given the information about month n. Not decreasing means that the frequency either increases or stays the same. We try six predictors: relative frequency, two measures of linguistic dissemination (number of unique trigram contexts in which the word occurs, number of unique part-of-speech trigram contexts) and three measures of social dissemination (number of unique users, number of unique threads and number of unique subforums).

Using absolute numbers (e.g. count of unique trigram contexts) as measures of dissemination can be problematic, since they, obviously, are all strongly correlated with absolute frequency. We follow the solution proposed by Stewart & Eisenstein (2018): for every month, we fit a linear regression model between the given predictor and the absolute frequency (for all words in the dataset) and then take the residuals (i.e. the proportion of variance which is not explained by the absolute frequency) as a measure of dissemination. Hopefully, this procedure also mitigates the problem that corpus size varies strongly with time.

All our datapoints are tuples that look as follows: data for word A at month n, data for word A at

predictors	**rank n-s**	**rank n-p**	**perf. n-s**	**perf. n-p**
1	18	18	0.007	-0.030
rel. freq.	3	9	0.121	0.030
authors + threads + subforums + rel. freq.	1	2	0.134	0.045
authors + threads + subforums + rel. freq. + + trigrams + pos trigrams	2	1	0.124	0.054

Table 1: Performance (increase in accuracy over baseline) across various models across the two neologism types: **n-s** (new signifier) and **n-p** (new phenomenon). **Rank** shows the place of the model in the list of all models ranked by performance.

month $n+1$. For every word, we randomly split all datapoints into a training and test set (80:20). Since predictions are always based on a previous month only, there is no point in preferring a chronological split to a random one. We fit a logistic regression model on a training set and then evaluate its predictions on a test set by subtracting the baseline accuracy (achieved by always predicting the more frequent outcome in the training set) from the actual accuracy. The maximum possible value of model performance is thus 0.5, the minimum value is -1. We fit 18 different models, one of them a null model (the predictor is a vector of ones), the rest are various combinations of the six aforementioned predictors that we find most promising. Our method is visualized on Figure 1 for the word *dampa* ‘to freak out’.

4 Results and discussion

The performance (measured as increase in accuracy over the baseline) varies substantially across models, but for all models, it is better for the new-signifier words than for new-phenomenon ones. The differences range from 0.037 to 0.088.

For brevity’s sake, we report the performance of four models: the best one for new-signifier words; the best one for new-phenomenon words; the worst one, which is the same for both types (the null model); and, for comparison, the one which uses only relative frequency as a predictor. See Table 1.

The null model performs worse than all other models, which is expected. It is interesting to see, however, that even for this model there is a difference between the new-phenomena words and the new-signifiers words, which suggests that the difference in performance depends not so much on the predictors that we choose, but rather on certain properties of the trends themselves. Another interesting result is that frequency alone can account for a substantial part of the predictions’ success.

While it is tempting to draw further conclusions about the relative importance of different predictors and compare them with previous work (Stewart & Eisenstein, 2018; Würschinger, 2021), we refrain from doing so at this exploratory stage of our study. Additional pilot analyses (not reported here) suggest that small changes in the experimental setup have the potential to strongly affect how the models perform with respect to each other, implying that these results may not be robust. The main result (better performance for new-signifier words), however, remains robust.

In order to test whether the difference in performance between the new-signifier words and the new-phenomena words is an artifact, we perform two comparisons. First, we compare whether the proportion of increases in frequency is approximately the same for the types, and that turns out to be the case (0.56 for new phenomena, 0.54 for “new signifiers).

Second, we compare the average amount of datapoints (months) per word. This amount varies (and is smaller the later words appear). It is reasonable to assume that with more datapoints the performance is likely to go up, and indeed that seems to be the case: there is a positive correlation between the amount of datapoints per word and the performance of the best model on the word (Pearson’s coefficient is 0.52 for new phenomena and 0.58 for new signifiers). The average amount of months per word is slightly smaller for new phenomena (193 vs. 218). If we remove the two words with the smallest amount of datapoints (*covid* and *vaccinpass*, resp. 23 and 19), the average amount for new phenomena goes up to

202, the qualitative results do not change. If we remove six words with the largest amounts of datapoints from new signifiers and six with the smallest amount from new phenomena, the average amounts become equal, but the results still hold for all model but one.

If we exclude the 11 words which we found particularly difficult to label as either new phenomena or new signifiers (marked with asterisks in Appendix A) the qualitative results do not change.

Our next goal is to test further which of the findings are robust, and for those that are, to explain why these effects emerge. We hope to do that, but who knows –

Acknowledgements

This work was supported by the Cassandra project (Marcus and Amalia Wallenberg Foundation, 2020.0060) and by the Swedish national research infrastructure *Nationella språkbanken*, funded jointly by the Swedish Research Council (20182024, contract 2017-00626) and the 10 participating partner institutions.

References

Lars Borin, Markus Forsberg, & Johan Roxendal. 2012. Korp the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, volume Accepted, pages 474–478.

Lars Borin, Markus Forsberg, & Lennart Lönngren. 2013. SALDO: a touch of yin to WordNets yang. *Journal of Language Resources & Evaluation*, 47:1191–1211.

Ola Karlsson. 2021. Lesserwisser, lårskav och läppstiftseffekt presentation och problematisering av urvalskriterierna för den svenska nyordslistan. *Lexico Nordica*, 28:101–120.

Ian Stewart & Jacob Eisenstein. 2018. Making “fetch” happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4370, Brussels. Association for Computational Linguistics.

Quirin Würschinger. 2021. Social networks of lexical innovation. investigating the social dynamics of diffusion of neologisms on twitter. *Frontiers in Artificial Intelligence*, 4.

Appendix A

#	word	translation	freq	year	type
1	blogg	blog	462799	2004	phenomenon
2	googla	[to] google	445681	2003	phenomenon
3	app	app	221612	2010	phenomenon
4	covid †	covid	144059	2020	phenomenon
5	foliehatt	tin foil hat	49073	2011	signifier *
6	buda	[to] bid	33745	2006	signifier
7	svininfluensa	swine flu	33240	2009	phenomenon
8	wiki	wiki	32726	2007	phenomenon
9	blogga	[to] blog	31606	2005	phenomenon
10	stalker	stalker	27829	2003	signifier
11	följare	online follower	24638	2009	phenomenon
12	curla	[to] be a helicopter parent	22126	2006	signifier
13	twittra, kvittra †	[to] tweet	21159	2009	phenomenon
14	bloggare	blogger	18547	2005	phenomenon
15	lockdown	lock down	18246	2020	signifier *
16	prio	priority	18180	2006	signifier
17	padda	tablet computer	18161	2011	phenomenon
18	näthat	online hate	17655	2007	phenomenon
19	haffa	[to] pick up (while dating)	17253	2015	signifier

20	hypa	[to] hype	17050	2013	signifier
21	#metoo, metoo	#metoo	15884	2017	phenomenon
22	stalka	[to] stalk	15430	2003	signifier
23	matkasse	online groceries	14352	2011	phenomenon
24	menskopp	menstrual cup	13003	2005	phenomenon *
25	sprita	[to] disinfect hands	11913	2009	signifier *
26	fronta	[to] be/put in front	11541	2004	signifier
27	vaccinpass	vaccination certificate	11480	2021	phenomenon
28	incel	incel	9925	2018	phenomenon
29	brexit	brexit	8532	2013	phenomenon
30	anime	anime	8333	2003	phenomenon *
31	barnvagnsbio †	cinema for parents with babies	8170	2003	phenomenon
32	åsiktskorridor	opinion corridor	7963	2014	signifier
33	stalking, stalkning	stalking	7930	2003	signifier
34	selfie	selfie	7835	2013	phenomenon
35	digitalbox	digital tv box	7620	2005	phenomenon
36	topsa	[to] swab	7449	2004	signifier *
37	spikmatta	bed of nails	7236	2009	phenomenon *
38	sporta	[to] sport, show off	7067	2009	signifier
39	trängselskatt	congestion taxes	6808	2005	phenomenon
40	pimpa	[to] pimp	6714	2007	signifier
41	framåtlutad	energetic	6391	2011	signifier
42	klimathot	climate change	6366	2007	phenomenon
43	sars	SARS	6309	2003	phenomenon
44	hedersvåld	honor-related violence	6250	2005	signifier
45	entourage	entourage	6120	2007	signifier
46	foppatoffel	crocs	5799	2007	phenomenon
47	vintage	vintage	5663	2007	signifier
48	EU-migrant	EU migrant	5657	2015	signifier *
49	svinna	[to] waste	5344	2021	signifier *
50	hbt	lgbt	5068	2004	signifier *
51	halmdocka	straw man argument	4878	2015	signifier
52	chippa	[to] chip	4439	2009	phenomenon
53	transfett	trans fat	4363	2007	phenomenon *
54	e-sport	e-sports	4078	2013	phenomenon
55	snackis	hot conversation topic	4002	2005	signifier
56	dampa	freak out	3624	2007	signifier
57	transponder	transponder	3612	2005	phenomenon
58	rondellhund	roundabout dog	3362	2006	phenomenon
59	instegsjobb	entry-level job	3336	2004	phenomenon
60	bröllopsklänning	wedding dress	3275	2011	signifier
61	trollfabrik	troll factory	3210	2015	phenomenon
62	kubtest	prenatal test	3098	2007	phenomenon
63	videosamtal	video call	2989	2004	phenomenon
64	cringe	cringe	2955	2017	signifier
65	svischa †	[to] transfer money via Swish	2923	2015	phenomenon
66	curlingförälder	helicopter parent	2914	2004	signifier
67	youtuber	youtuber	2878	2015	phenomenon
68	nätpoker	online poker	2850	2005	phenomenon
69	blingbling	bling-bling	2659	2004	signifier
70	nystartsjobb	entry-level job	2599	2006	phenomenon

71	klimatsmart	climate friendly	2558	2007	phenomenon
72	livspussel	work-life balance	2496	2007	signifier
73	killgissa	[to] guess, mansplain	2356	2017	signifier
74	backslick	backslick hairdo	2321	2004	signifier
75	skypa, skajpa	[to] skype	2306	2007	phenomenon

Table 2: The 75 selected words with their year of appearing in Språkrådet/Språktidningen's list, their frequency in the Familjeliv and Flashback discussion forum corpora and approximate translations. The words are marked as either new signifier or new phenomenon. Asterisk is used for borderline cases. Dagger indicates that some spellings or other variants of the word that we included in the search are not listed in the table.

avokado-r/-er/-s/-sar

Kristian Blensenius
Inst. för svenska, flerspråkighet
och språkteknologi
Göteborgs universitet, Sverige
kristian.blensenius@gu.se

Louise Holmer
Inst. för svenska, flerspråkighet
och språkteknologi
Göteborgs universitet, Sverige
louise.holmer@svenska.gu.se

Abstract

The article discusses lexicographic perspectives of the Swedish plural with the suffix *-s*. Traditionally, plural nouns ending in *-s*, for example *avokados* 'avocados', are considered colloquial speech; the formal way of writing the plural in question is *avokador* or *avokadoer*. However, since the Swedish Academy grammar included a noun declension indicating plurals with the suffix *-s*, plural with *-s* seems to have become more accepted, at least among language planners.

1 Inledning

När *Svenska skrivregler* utkom i ny upplaga 2017 (Karlsson, 2017) var en uppmärksammad nyhet att Språkrådet nu godkände s-plural, att s-pluralen så att säga hade blivit officiell. På Sveriges radios webbplats meddelade Vetenskapsradions nyheter 2017-03-22 att ”[de svenska skrivreglerna] öppnar […] för användning av plural-s i svenskan”, och Svenska Dagbladet publicerade 2017-03-27 en artikel rubricerad ”Engelskt plural-s på väg in i svenska ordböcker”. Förändringen i förhållande till den tidigare upplagan (*Svenska skrivregler*, 2008) var nu inte oerhörd, men formuleringar som att den främmande s-pluralen var ”olämplig” hade plockats bort. Det angavs dock fortfarande att det i regel är bäst att undvika s-plural i formella texter.

Trots att s-plural har använts i svenskan sedan 1700-talet (i vissa fall tidigare; se Söderberg, 1983), verkar fenomenet ännu inte riktigt ha släppts in i värmen. När *Svensk ordbok utgiven av Svenska Akademien* gavs ut i reviderad upplaga 2021 (SO, 2021) innehöll den emellertid fler ord pluralböjda med *-s* än tidigare, t.ex. *sambo*, med pluralangivelsen ”*sambos* eller *sambor*” (tidigare endast *sambor*). S-plural gavs också som förstaform för en del nyinlagda ord, t.ex. *hashtags* eller *hashtaggar*.

I *Svenska Akademiens ordlista*, 14 uppl. från 2015 (SAOL 14), är situationen en annan: här är (engelsk) s-plural uttryckligen motarbetad, bl.a. av den anledningen att denna pluralform anges inte passa in i det svenska böjningssystemet (se inledningen till den tryckta SAOL 14, s. XI). För ett ord som *sambo* ges endast r-pluralen *sambor* i SAOL, medan ett annat (icke-engelskt) ord som *avokado* endast ges pluralen *avokador* (utöver att endast *k*-stavningen ges). Detta trots att pluralformer som *avokados* numera inte är ovanliga i tal- och skriftspråk, liksom – men ovanligare – *avokadosar*. Den senare, som ibland går under beteckningen sar-plural (Josefsson, 2018), behandlas översiktligt i *Svenska Akademiens grammatik* (SAG; Teleman et al., 1999, vol. 2, s. 83, 104), förekommer i former som *bikinisar* och, kanske i mer lustfyllda sammanhang, trefaldigt markerade pluraler som *paparazzisar* (den rekommenderade formen är *paparazzoer*). Språkvårdare avråder vanligen från *-sar*-plural i vårdat skriftspråk, och särskilt trefaldigt betecknade har varit ställda utanför gemenskapen. Ett parallellt fall är italienskheten *putto*, för vilket pluralformen *puttisar* av Wellander (1970, s. 162) beskrivs som ett ordformsexemplar som ”knappast kan anses som en prydnad för vårt språk”.

Som ett led i arbetet med vidareutvecklingen av SAOL och SO undersöker vi hur s-pluralformer behandlas i de två svenska enspråkiga ordböckerna SAOL 14 och SO 2021, och i föreliggande text ger vi exempel på lånord där skriftspråket uppvisar variation i fråga om pluralböjning, framför allt avseende

Kristian Blensenius and Louise Holmer. 2022. avokado-r/-er/-s/-sar. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 13–16. Available under CC BY 4.0

s-plural. Vi vill jämföra rekommendationer som uttrycks av framför allt Språkrådet, liksom grammatiska beskrivningar av s-plural i SAG, med språkbruket vid ett urval av ord där pluralformerna sedan tidigare har konstaterats variera.

Vi fokuserar på substantiven *avokado*, *bikini* och *hashtag*. De två förra är välkända i svenskan sedan åtminstone 1930- och 1940-talet, medan *hashtag* är ett exempel på ett nyare ord, belagt i skrift (nyhetstext) sedan 2010.

2 Beskrivningar i grammatik och språkvård

Här ges en genomgång av vad som sägs om s-plural i referensgrammatiken SAG, några grammatiska läroböcker samt Språkrådets rekommendationer.

När SAG utkom 1999, tillkom i en svensk grammatisk beskrivning en sjunde deklination (SAG 2, s. 63), innefattande substantiv med pluralsuffixet *-s*, t.ex. *dissenters* och *tricks*. Att en omfattande referensgrammatik för svenska inkluderat deklinationen har från språkvårdshåll kommit att ge viss legitimitet åt användningen av s-plural för ord som *avokado*, *bikini* och *hashtag*: *avokados*, *bikinis* respektive *hashtags*.

S-pluralen har dock inte med självklarhet fått genomslag i grammatikläroböcker på högskolenivå sedan publiceringen av SAG 1999: vissa har anammat s-pluraldeklinationen, t.ex. Bolander (2012, s. 114) medan andra inte verkar ta upp den särskilt, t.ex. Lundin (2014, s. 187). Den till SAG relativt nära knutna *Svenska Akademiens språklära* (Hultman, 2003, s. 64–65) antar bara sex deklinationer, och även om s-pluralen nämns antas inte någon särskild deklination för den. Även Josefsson (2009) diskuterar s-pluralen, här som en möjlig sjätte deklination (Josefsson räknar med fem grunddeklinationer), men uttrycker samtidigt tveksamhet, utifrån resonemanget att s-pluralen bara förekommer under en övergångstid, för att därefter pluralböjas enligt någon av de andra deklinationerna.

Svenska skrivregler i sin senaste upplaga kan sägas gå ett par steg längre än SAG i fråga om vad som antas om pluralformens etableringsgrad. SAG (2, s. 79) noterar att ”Bruket av *-s* har i de flesta fall en relativt osvensk prägel”, medan *Svenska skrivregler* (Karlsson, 2017, s. 104) menar att ”S-plural är ganska vanligt förekommande i svenskan” och ”För vissa ord är […] s-pluralen så etablerad att den åtminstone i vissa sammanhang dominerar helt över andra böjningsmönster”. Språkrådet går i sin ”Frågelådan” ännu något längre: exemplifierande med plural av några svenska ord, t.ex. *sambos* och skämtsamma former som *snyggos*, *gubbs* och *kvinns*, meddelas att s-plural är att betrakta som en del av det svenska språksystemet.[1]

Svenska skrivregler anför förvisso ett vanligt argument emot s-pluralen: böjningsmönstret ”saknar en etablerad form för bestämd form plural” (Karlsson, 2017, s. 104). Tanken verkar vara att s-pluraldeklinationen ska hålls intakt (s-suffixet i obestämd form plural ska följa med även i den bestämda formen) och att sar-plural ska undvikas, åtminstone i formell text. Sådana resonemang kan skönjas i språkvårdares rekommendationer av andra pluralsuffix av typ flera *containrar – de containrarna*.

3 S-plural i SAOL, SO och bruket

Substantivet *hashtag* har som nämnts en mycket begränsad historia i SAOL och SO, men *avokado* och *bikini* har varit med desto längre och ibland försetts med olika rekommendationer avseende böjningssätt och böjningsformer över tid (jfr Josefsson, 2009).

När ordet *avokado* togs med i SAOL 10 (1973) noterades, förutom variantstavningen *avokato*, den enda pluralböjningen *avokador*. De följande upplagorna ger även pluralformen *avokadoer*, medan s-pluralen *avokados* som nämnts ännu inte har tagits in i SAOL.

Uppslagsordet *bikini* togs också in i SAOL 10 (1973), då med pluralböjningen *bikini*, alltså samma form som i singular, och i SAOL 11 (1986) fanns också variantböjningen *bikinier* med. I SAOL 14 (2015) har pluralformen *bikini* försvunnit och ersatts av böjningsangivelsen ”*bikinier* hellre än *bikinis*”, där beteckningen ”hellre än” indikerar att *bikinier* förordas framför *bikinis*.

Medan SAOL 14 är mer normativ, är SO 2021 mer deskriptiv (Blensenius et al., 2021, s. 41). De olika inriktningarna i fråga om normativitet visar sig genom att SAOL 14 har fler och explicitare rekommen-

[1] Språkrådet, Frågelådan. ”Hur ser Språkrådet på s-plural?” `https://frageladan.isof.se/visasvar.py?svar=79712`. Hämtat september 2022

dationer än SO 2021 (t.ex. rekommendationer som den för *attachment*: ”Använd hellre *bilaga*”). Det är därför ingen överraskning att den mer deskriptivt inriktade SO 2021 ger s-plural för *avokado*, *avokados*, medan den mer normativa SAOL 14 inte rekommenderar *avokados*. Noteras kan att varken SAOL eller SO föreslår eller på annat sätt nämner sar-plural som möjlig form.

Pluralformer för *avokado*, *bikini* och *hashtag* ges på följande sätt i de båda ordböckerna: SAOL 14 ger ”*avokador*”, ”*bikinier* hellre än *bikinis*” respektive ”*hashtaggar* hellre än *hashtags*”, medan SO 2021 ger ”*avokados* eller *avokador*”, ”*bikinis*” respektive ”*hashtags* eller *hashtaggar*”. Notera att SO 2021 jämställer böjningsvarianterna med beteckningen ”eller”.

I skriftspråket varierar pluralböjningen av *avokado* mer än i ordböckerna. I olika typer av texter påträffas främst dessa: *avokado, avokados, avokador, avokadoer, avokadon* och *avokadosar*. Av dessa är *avokador* den mest frekventa i tidningstext, ungefär 10 gånger vanligare än *avokados* (vi bortser från *k/c*-variationen, och vi bortser också från potentiell homografi). I fråga om *bikini* är pluralvariationen inte lika stor, men det är ingen tvekan om att pluralformen *bikinis* är betydligt vanligare än den i SAOL 14 rekommenderade *bikinier*. I fråga om *hashtag* utgör *hashtags* den vanligare pluralformen i obestämd form, medan bestämd form har klar övervikt för *hashtaggarna* jämfört med t.ex. *hashtagsen* i tidningstext.

4 Saxad böjning eller Frihet att röra sig mellan paradigmen

Mycket talar för att båda pluralformerna av ord som *hashtag* (*-s* och *-ar*) bör inkluderas i ordböckerna. Språkbrukarna kan också lösa svårigheten med bestämd form plural genom att använda *hashtags* i obestämd form och *hashtaggarna* i bestämd form. Vi föreslår möjlighet till s.k. saxad böjning (SAG 2, s. 544), dvs. böjning där böjningsformerna för samma substantiv kan föras till olika deklinationer. Här skulle man då kunna tänka sig denna böjning:

singular	plural obest.	plural best.
avokado	*avokados*	*avokadorna*
bikini	*bikinis*	*bikinierna*
hashtag	*hashtags*	*hashtaggarna*

Sammanfattningsvis försöker vi illustrera hur pluralerna behandlas på delvis olika sätt i grammatikor och ordböcker och av språkvårdare och språkbrukare. Trots allt är det språket i bruk som ligger till grund för både ordböckerna och grammatikböckerna, samt språkvårdens olika rekommendationer, och frågan är om inte SAOL behöver anamma en mer tillåtande attityd till s-plural i kommande upplagor.

Referenser

Kristian Blensenius, Louise Holmer, & Emma Sköldberg. 2021. SAOL 14 som rättesnöre – diskussion kring den senaste upplagan. *LexicoNordica*, pages 39–58.

Maria Bolander. 2012. *Funktionell svensk grammatik* (3 upplagan). Liber, Stockholm.

Tor G. Hultman. 2003. *Svenska Akademiens språklära.* Svenska Akademien, Stockholm.

Gunlög Josefsson. 2009. *Svensk universitetsgrammatik för nybörjare* (2 upplagan). Studentlitteratur, Lund.

Gunlög Josefsson. 2018. Avokadosar och kepsar – ett epentetiskt *s* med olika funktioner. *Språk och stil*, 28:5–21.

Ola Karlsson, editor. 2017. *Svenska skrivregler* (4 upplagan). Språkrådet & Liber, Stockholm.

Katarina Lundin. 2014. *Tala om språk. Grammatik för lärarstuderande.* Studentlitteratur, Lund, 2 edition.

SAOL 14. 2015. *Svenska Akademiens ordlista över svenska språket* (14 upplagan). Tillgänglig: `svenska.se`. Hämtat september 2022.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* (2 upplagan). Tillgänglig: `svenska.se`. Hämtat september 2022.

Barbro Söderberg. 1983. *Från rytters och cowboys till tjuvstrykers. S-pluralen i svenskan. En studie i språklig interferens.* Almqvist & Wiksell International, Stockholm.

Ulf Teleman, Erik Andersson, & Staffan Hellberg. 1999. *Svenska Akademiens grammatik.* Svenska Akademien, Stockholm.

Svenska skrivregler. 2008. (3 upplagan). Språkrådet & Liber, Stockholm.

Erik Wellander. 1970. *Riktig svenska. En handledning i svenska språkets vård*. Norstedts, Stockholm.

Counting dirty words:
The effect of OCR quality on token statistics in historical Swedish corpora

Gerlof Bouma and **Yvonne Adesam**
Språkbanken Text / Dept of Swedish, Multilingualism, Language Technology
University of Gothenburg, Sweden
{gerlof.bouma,yvonne.adesam}@gu.se

Abstract

We explore the effects of varying OCR quality on word statistics in historical Swedish newspaper corpora. Most type frequencies are underestimated, with a small group of overestimated types. To adjust the freqencies we propose using the strong correlation between lexicon coverage and word error rate, which shows encouraging first results. The method currently, however, only targets underestimation and needs further development.

1 Introduction and background

The large scale corpora provided through Språkbanken Text invite exploration of a wide range of questions, both linguistic and from other disciplines. One fruitful method that Språkbanken's corpus infrastructure makes easily available to researchers is the comparison of term occurrences between different corpora, for instance between collections from different time periods, from different genres, with different target audiences, or different types of text producers.

A valid comparison across corpora relies on the assumption that contrasts between corpora are due to factors inherent to the texts, preferably the factors of interest. However, different corpora have different levels of reliability and quality. First, there may be problems with the (automatic) annotations. For instance an early 2000 newspaper corpus is likely to receive better annotations than a 19th century newspaper or a 2020 blog entry, as the latter are textually further away from the material used to train the annotation tools. Secondly, the faithfulness of the available electronic version to its source may show faults. We can expect a corpus based on *born digital* newspapers from the 2000s to have few issues in this respect, but one based on material from before the digital age, which must be digitized – photographed, analysed for layout and text flow, and put through optical character recognition – may deviate from its source in many ways, and also contain errors that percolate down the processing pipeline. For ever-earlier historical newspapers, these issues will be exacerbated by the state/quality of the paper, the printing techniques, the script, and changes in conventions and language, to name but a few factors.

The impact of varying OCR quality on text analysis and processing has been studied by Hill & Hengchen (2019), and van Strien et al. (2020), among others. The current paper focuses on one specific aspect of this impact: the effect of OCR accuracy on word statistics. Suppose we look for the word *politik* 'politics' in a corpus of 1M tokens, and find 100 occurrences. We would say it has a relative frequency of 0.1%. But if we also know that half of these 1M tokens are *letter salad* – garbled words – we might prefer to say we found 100 occurrences in 500k tokens instead – twice the relative frequency. Below, we will have an empirical look at the relation between word frequencies and OCR quality. Furthermore, by using the proportion of known words in a document as a proxy for its OCR quality (Adesam et al., 2019; van Strien et al., 2020; Neudecker et al., 2021, and references therein), we will investigate a simple method for adjusting frequency estimates to correct for the error introduced by OCR mistakes.

Gerlof Bouma and Yvonne Adesam. 2022. Counting dirty words: The effect of OCR quality on token statistics in historical Swedish corpora. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 17–24. Available under CC BY 4.0

2 Material and method

As our corpus, we use the historical newspaper dataset described in Dannélls et al. (2021), available from Språkbanken Text,[1] which contains 186 pages of newspaper text, distributed over 90 documents (that is, newspaper editions) published between 1818 and 1906, with an additional late 20th century newspaper included for control. The data is a selection of the National Library of Sweden's digital newspaper archive,[2] and consists of images, ground truth transcriptions (henceforth: GT) and the output of several OCR systems. We use the included output of the commercial *ABBYY Finereader 11* as our OCR text. We visually classified newspaper editions as *Blackletter*, *Antiqua* or mixed[3] using the supplied images. Blackletter is the dominant script in 44 documents, Antiqua in 38, and 8 documents are mixed.

We lightly preprocessed both the GT and OCR material by (blindly) undoing end-of-line hyphenation and replacing *ſ* (long s) by *s*, after which we ran the materials through the Sparv annotation pipeline using an annotation module for historical Swedish[4] (Hammarstedt et al., 2022). Sparv provides us with tokenization and links from word forms to entries in one of three lexical resources: one consisting of entries (base forms) in Swedberg's early 18th c dictionary (Swedberg & Holm, 2009), a resources based upon Dalin's early 19th c dictionary (Dalin, 18501853) with full paradigms for part of the entries,[5] and the present-day Swedish lexical resource Saldo (Borin et al., 2013), which contains a comprehensive full-form component. As we were not interested in the quality of the links, but in estimating the maximal coverage of the dictionaries, we configured Sparv to assign as many links as possible. We also seperately postprocessed tokens containing *ß*, since the earliest dictionary contains this ligature as is, whereas later use *ss* in its place. For our statistics we only consider word-like tokens, and therefore discard punctuation.[6] The GT and OCR materials consist of around 500k words, each.

One of the quantities of interest in this study is the proportion of known word-like tokens, that is, those which received at least one dictionary link. Existing research (van Strien et al., 2020) shows that this is correlated to OCR quality. Since we can add the needed annotation automatically, the proportion of known words gives us a handle on OCR quality without the need for demanding manual transcription. To evaluate the validity of this proxy, we also perform an intrinsic evaluation of OCR quality by comparing GT and OCRed documents at word level, using *normalized word error rate* (NWER). This measure is based upon the alignment of token sequences, in our case sequences forming a paragraph-like segment, as defined in the used dataset. If we have four alignment operations *insert a word*, *delete a word*, *replace a word by another word*, and *match a word to an identical one*, NWER is given by:

$$\mathrm{NWER} = \frac{\#\,\text{error operations}}{\#\,\text{all operations}} = \frac{\#\,\text{insertions} + \#\,\text{deletions} + \#\,\text{replace operations}}{\#\,\text{insertions} + \#\,\text{deletions} + \#\,\text{replace operations} + \#\,\text{matches}}$$

We ignore differences in case and treat *ß* and *ss* as the same substring when comparing two words.

3 Results

3.1 OCR quality

Figure 1 contains the results of looking at OCR quality over time, both using NWER (a) and using the proportion of known tokens (b).[7] The NWER results clearly shows later editions are OCRed more accurately, although curiously, the effect is only really seen in the Blackletter material. The error rate in the

[1] `https://spraakbanken.gu.se/resurser/svenska-tidningar-1818-1870`, `https://spraakbanken.gu.se/resurser/svenska-tidningar-1871-1906`

[2] The archive is accessible from `tidningar.kb.se`. The historical part is available as the "Kubhist" corpus at `https://spraakbanken.gu.se/korp/?mode=kubhist`

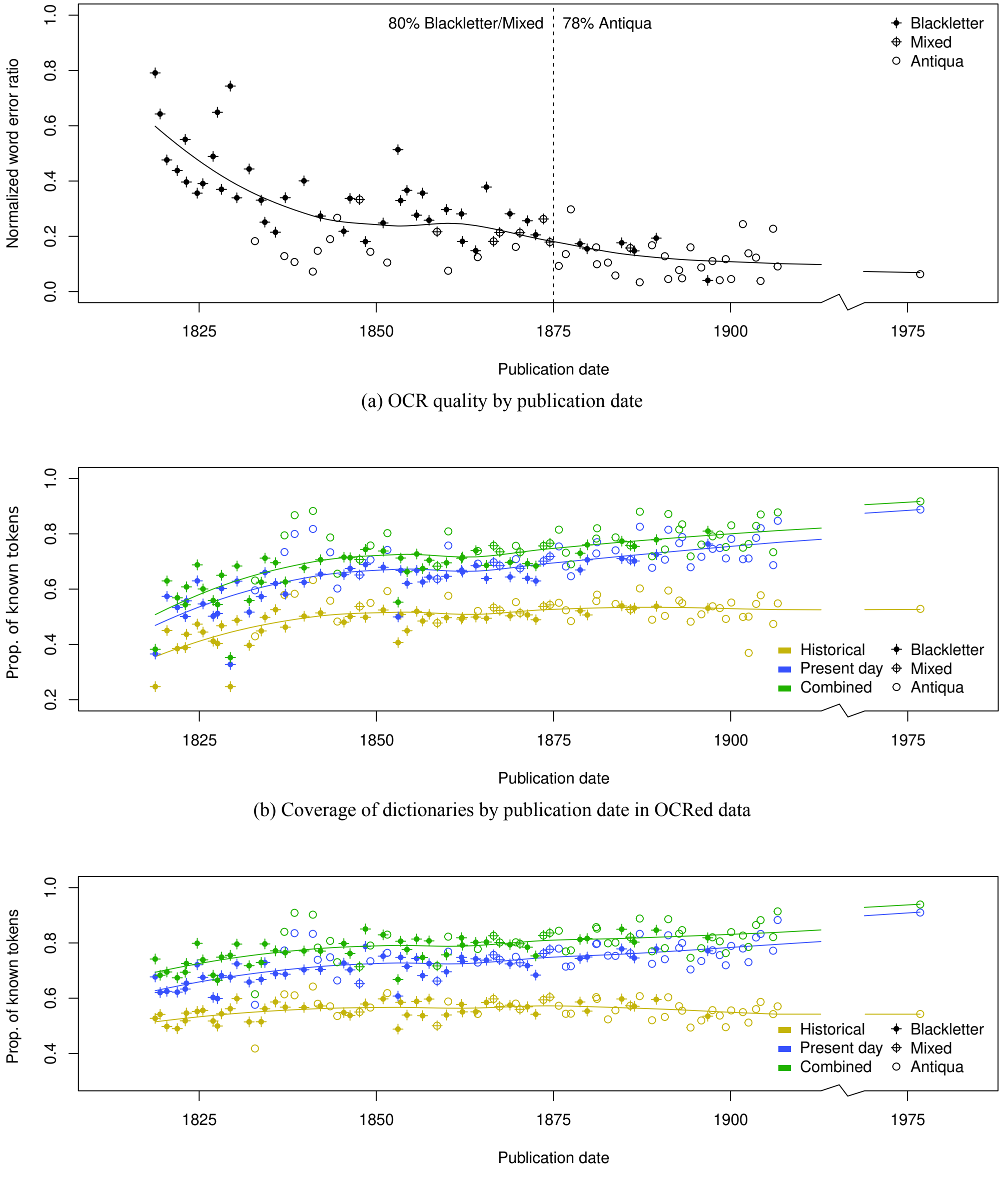
[3] We classified as mixed newspaper samples that contain sections in Blackletter as well as sections in Antiqua, for instance when the former script was used for news and the latter for classified advertisements or a feuilleton. Older Blackletter newspapers also frequently contain smaller stretches of Antiqua, for instance for French or Latin words/text, but we did not consider this to be mixed script material.

[4] `https://spraakbanken.gu.se/sparv/`

[5] Notably, however, there are *no* inflected forms for some irregular high frequency items, such as the verb *vara* 'be'.

[6] To be precise, we only look at tokens matching the regular expression `\w+([-:']\w+)*`, where `\w` is an alphabetical character or a number.

[7] The smoothed trendlines are produces by the `loess` function in R v4.2.1, with the standard settings (R Core Team, 2022).



(a) OCR quality by publication date

(b) Coverage of dictionaries by publication date in OCRed data

(c) Coverage of dictionaries by publication date in ground truth data

Figure 1: The relation between publication date of the newspaper and different aspects of OCR quality. Intrinsic evaluation using GT data shows newer publications are interpreted more accurately (a), although the effect is only clear for publications using Blackletter. Concordantly, newer material generally contains higher proportions of words recognized by the dictionaries (b), although comparison to measurements on ground truth material reveals that part of this upward trend is explained by the better compatibility of the large present day dictionary with newer language material (c).

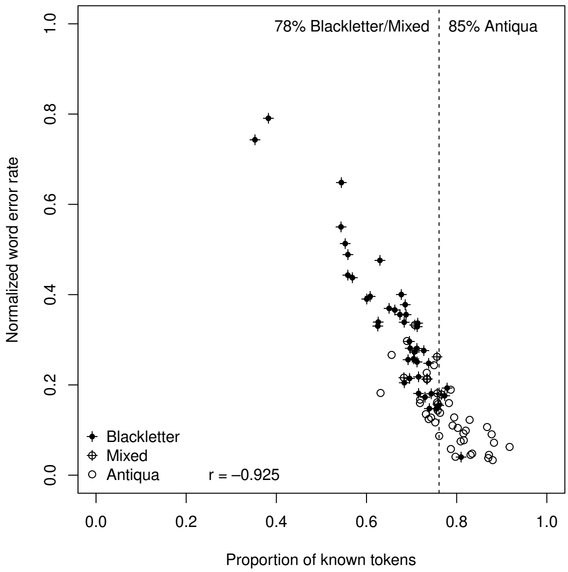


Figure 2: Dictionary coverage is highly predictive of OCR quality.

Antiqua material is comparatively stable from the earliest newspapers up to 20th century control edition. The proportion of known tokens gives a similar, though negated trend. The negation is because a *high* proportion of known tokens goes together with a *low* error rate. But given this transformation, the overall shape of the graphs are very similar, including the wavy area around 1860. Generally, later material is of better quality. The relation between NWER and proportion of known tokens is plotted explicitly in Figure 2. The high correlation shows that probing OCR quality using dictionary coverage is feasible. We do note that the proportion of known tokens falls more slowly than the error ratio increases.

Figure 1c shows that the proportion of known tokens also grows over time for the GT data (green points and trend line). The data set's newer material is more like the present-day Swedish we find in Saldo's full form lexicon. This explains why we see the coverage improvement in graph for the modern dictionary (in blue) but not for the historical dictionaries (yellow). The trend in subfigure (b) therefore reflects OCR quality as well as the compatibility of the annotation tool and dictionaries with the language in the corpus.

3.2 Word statistics

As can be seein in the "Total" rows in Table 1, the GT material contains just over 70k types ("observed in GT"), of which almost 27k are observed more than once ("repeated in GT"). Of these repeated types, 48% have the same token counts in the OCR data as in the GT data (correctly estimated), 44% have lower counts in the OCR data (underestimated), and 7% higher (overestimated). For almost half of the types in this part of the vocabulary, then, we see a loss of token mass. The table also shows that the OCR data contains an additional 37k types (for a total of 47k tokens), absent from the GT vocabulary.

The ten most frequent underestimated word types are *och* 'and', *af* 'of', *att* (complementizer/infinitive marker), *till* 'to', *den* 'it', *en* 'a/one', *för* 'because/(be)for', *som* (relativizer), *med* 'with', and *på* 'on'. These appear 4.5k–15.5k times in the GT data, but lose generally between around 7% of their token mass in the OCR data. The outlier here is *på* which has 42% lower counts, which suggests it is easily misidentified. Indeed, we find the lost token mass with other forms, sometimes non-words like *pa* (counts inflated from 4 in GT to 152 in OCR) and *pä* (from 7 to 1577).[8]

[8] The existence of thes non-words in the GT data shows that this material also contains errors, although in these cases we do not know whether these were misprints or mistakes in the manual transcription.

	GT		OCR		
GT freq. region	Freq. range	#Tokens	Freq. range	#Tokens	#Types
Top third	15369 – 629	161703	13834 – 367	149895	69
Middle third	628 – 24	161597	781 – 0	149023	2002
Bottom third	23 – 0	162726	1577 – 0	192484	105097
— observed in GT	23 – 1	162726	1577 – 0	145909	68312
— repeated in GT	23 – 2	119096	1577 – 0	113297	24682
Total	(≥ 0)	486026		491402	107168
— observed in GT	(≥ 1)	486026		444827	70383
— repeated in GT	(≥ 2)	442396		412215	26753

Table 1: Summary of type and token counts, and definition of three vocabulary regions on the basis of GT frequencies.

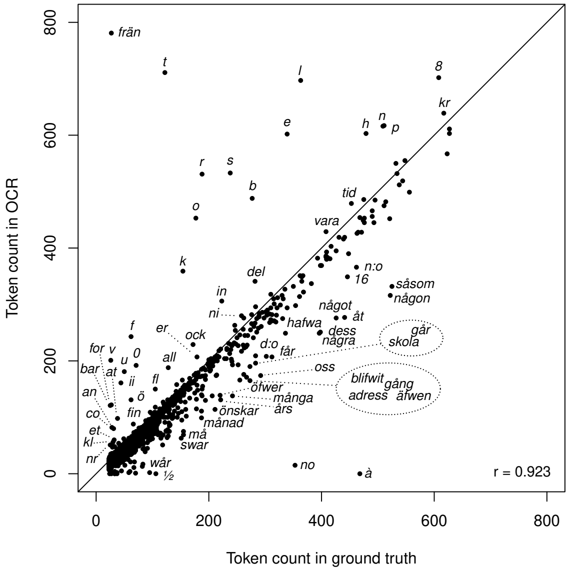


Figure 3: For most of the types that make up the central third in terms of token mass, frequencies based on OCR underestimate true frequencies. Overestimations happen typically in the context of single letter types and short words.

The most frequent overestimated types are *i* ‘in’ (also: letter), *är* ‘am/are/is’, *1*, *vid* ‘at’, *a* (letter), *3*, *var* ‘was/were/where’, *c* (letter), *4*, *g* (letter). As can be seen, these types are very short, which is not just due to them being high frequency types. In general, the overestimated types are shorter (median of 5 characters for the types observed repeatedly in GT) than the correctly estimated or underestimated types (8 characters). This can explained by the higher neighbourhood density of shorter words: misreading a short word has a chance of yielding another existing word, whereas a long word is more likely to give a nonsense type. That single character counts are inflated may have an additional cause: they can also be the result of segmentation errors, for instance when the OCR interprets a wide-spaced heading as made up of single letters rather than words. A precise investigation of these effects is beyond the scope of this paper, but for now we note that the existence of both under- and overestimated types means that a single, simple adjustment to the OCR frequencies will not suffice.

Table 1 also divides the type vocabulary into three regions of roughly equal GT token mass. Note that the OCR frequency ranges may overlap. For instance, the extremely inflated counts of *pä*, discussed above, bring the upper limit of the OCR frequency range of the bottom third well within the top range.

The proportion of overestimated types differs starkly between the three regions: in the top third 17% are overestimated, in the middle third 11% and in the GT observed part of the bottom third, only 4% (7% in the repeated part). This must also be explained from the stronger tendency of shorter words to be overestimated, combined with the well known over-representation of short words among the most frequent words. Of course, this generalization breaks down once we include the types that haven’t been observed in the GT data: these are all by definition overestimated in the OCR data, and in addition they are on the long side (median of 8 characters).

The shapes of the frequency distributions in the GT and OCR data resemble each other closely in the overall data (Pearson’s $r = 0.9911$, only types observed in GT), as well as in the top and middle regions ($r = 0.9919$ and $r = 0.9230$, respectively). However, in the bottom third, the correlation is low ($r = 0.3657$, only types observed in GT). Figure 3 plots the relation between the two frequency distributions in the middle segment, with illustrative cases of over- and underestimation highlighted. Above the diagonal we see the overestimated cases, which tend to be short in this segment, too. An outlier here is *frän* ‘pungent’, whose inflation is due to its resemblance to the highly frequent*från* ‘from’. The underestimated types below the diagonal are longer and many contain diacritics. An outlier here is *no* ‘number’ (that is, *numero*), which appears in the newspapers as *No* or as the abbreviature №, which the OCR software cannot handle.

3.3 Towards a method for adjusting token counts

Although the shapes of the frequency distributions are similar, OCR-based frequency estimates are overall too low for the GT observable part of the vocabulary, on average 18.6% lower. We will briefly consider two ways to adjust the OCR estimates *upwards* to lie closer to the GT observations. In general, we do not know the amount of underestimation, so we cannot use this number directly. However, we can try to guess this from the dictionary coverage numbers, plotted in Figure 1b. With a different correction factor per document, the adjusted frequencies are

$$\widehat{\text{counts}}(w) = \sum_{doc} \text{counts}(w, doc) \times \frac{1}{\text{coverage}(doc)} \; .$$

This method will overcompensate grossly, however, since the mean coverage in the OCR data is only 72%. Applying it, we get a total of 589364.4 tokens for the GT observable vocabulary (top third: 198659.5, middle: 197112.7, bottom: 193592.2), overshooting the mark by 100k tokens. Moreover, the correlation between GT and adjusted OCR frequencies is lower than with the raw OCR frequencies: $r = 0.9899$ (top: 0.9908, mid: 0.9140, bottom: 0.3445), which meant also in terms of shape we have moved away from our target distribution.

A better estimate uses the observation that dictionary coverage on OCR data is not only a matter of OCR quality, but also of compatibility between document language and the lexical resource. A more conservative adjustment would use the latter component as an upper baseline for coverage, as in

$$\widehat{\text{counts}}(w) = \sum_{doc} \text{counts}(w, doc) \times \frac{\text{compatibility}(doc)}{\text{coverage}(doc)} .$$

In a real-world setting, we do not know the compatibility of each specific document with a lexical resource, but we may be able to guess it from what we know about similar documents. We implement this idea here by using the smoothed coverage trend in the GT data – that is the green trend line in Figure 1 (c) – to provide use with compatibility scores for each document.

The results are encouraging. We now arrive at 478461.6 tokens for the whole GT observable lexicon (top: 161200.3, mid: 160126.7, bottom: 157134.6), which is just 10k tokens under the target. Although the correlation with the GT distribution is still worse than for the unadjusted OCR frequencies, the situation is better than with our first adjustment: $r = 0.9903$ (top: 0.9911, mid: 0.9168, bottom: 0.3516).

4 Conclusions

In this paper, we have taken a first look at the effects of varying OCR quality on word statistics in historical Swedish newspaper corpora. The picture that emerges is that, overall, the OCR-based statistics stay closely to the true distributions in terms of shape, but that the counts for the majority of types are underestimated. Against this trend we find a much smaller group of overestimated types, which tend to be short and highly frequent. Using individual examples, we provide some evidence that this is related to lexical neighbourhood effects, but a more rigorous investigation is needed to provide a more solid ground to this account.

We also propose a method to calculate adjusted frequencies, so that the OCR estimates resemble the true GT distributions better. Our method makes crucial use of the strong correlation between coverage and word error rate we found in the data set. An evaluation of its application to further materials would let us better asses this method. A particular drawback of the method is that it only targets underestimation, and it currently only worsens the overestimated type counts. We hope that a future better understanding of these overestimated cases will help us devise a method that addresses them specifically.

Acknowledgements

We are grateful to Martin Hammarstedt and Anne Schumacher for their help with the Sparv pipeline. We also thank Lars Borin for recognizing the value of connected corpus and lexical infrastuctures, and consistently driving their development, not just for modern materials but also for historical corpora. The current paper is only a small demonstration of the many things one can do if one has both these resources available together.

References

Yvonne Adesam, Dana Dannélls, & Nina Tahmasebi. 2019. Exploring the quality of the digital historical newspaper archive Kubhist. In *Proceedings of the 4th Conference of The Association Digital Humanities in the Nordic Countries (DHN), Copenhagen, Denmark, March 5-8, 2019 / edited by Costanza Navarretta, Manex Agirrezabal, Bente Maegaard*, Aachen. CEUR Workshop Proceedings.

Lars Borin, Markus Forsberg, & Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Anders Fredrik Dalin. 1850/1853. *Ordbok öfver svenska språket* [Swedish dictionary]. *Vol. I–II*. Joh. Beckman, Stockholm.

Dana Dannélls, Lars Björk, Ove Dirdal, & Torsten Johansson. 2021. A two-OCR engine method for digitized Swedish newspapers. In *Selected Papers from the CLARIN Annual Conference 2020*, volume 180 of *Linköping Electronic Conference Proceedings*, pages 65–74. LiU Electronic Press.

Martin Hammarstedt, Anne Schumacher, Lars Borin, & Markus Forsberg. 2022. Sparv 5 user manual. Technical report, Department of Swedish, Multilingualism, Language Technology, University of Gothenburg.

Mark J Hill & Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843, 04.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, & Stefan Pletschacher. 2021. A survey of OCR evaluation tools and metrics. In *The 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

R Core Team, 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Jesper Swedberg & Lars Holm. 2009. *Swensk Ordabok. Utgiven efter Uppsala-handskriften, med tillägg och rättelser ur övriga handskrifter, av Lars Holm* [Swedish dictionary. Published on the basis of the Uppsala manuscript, with additions and corrections from other manuscripts, by Lars Holm]. Stifts- och landsbiblioteket i Skara, Skara.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, & Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream NLP tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH,*, pages 484–496. INSTICC, SciTePress.

Investigating a linguistic mini landscape: The Tsez (Dido) dialect dictionary project

Bernard Comrie
Department of Linguistics
University of California
Santa Barbara, CA, USA
comrie@ucsb.edu

Abstract

The interim results of the Dialect Dictionary of the Tsez (Dido) Language project, in comparison with documentation of grammatical differences among Tsez dialects from the mid-twentieth century, provide good information on the dynamics of grammatical and phonological change in Tsez dialects over a period of 70 years. Morphology is stable, including morphological differences between dialects, and may be an important linguistic marker of local identity. By contrast, two major phonological changes, vowel shortening and delabialization of consonants, have radically changed the phonologies and morphophonologies of many Tsez dialects. These changes have spread widely, but not universally, giving rise to new isoglosses that distinguish dialects from one another. The mini-landscape overall shows an interesting interaction of stability and innovation, against the background of the maintenance of local identity, including in linguistic terms.

1 Introduction

Lars Borin and I share an interest in the dynamics of language areas, including both the synchronic distribution of typological variables within the area and the historical processes — break-up of proto-languages with increasing diversification of descendant languages across time, as well as the effects of language contact — that have led to the present-day distribution. We have collaborated with Anju Saxena in investigating the large language area that is South Asia as well as one of its mid-sized sub-components, the Western Himalayas. I happen to have a personal interest in a much smaller language area, namely the dialect diversity within the Tsez language.

Tsez, also known by the Georgian-origin exonym Dido, is one of about a dozen small languages spoken in the Tsunta and Tsumada districts of the Daghestan Republic in the North Caucasus. All belong to the Nakh-Daghestanian (East Caucasian) language family. According to the 2010 census of the Russian Federation, Tsez then had about 12 500 speakers, although community activists think that this is substantially undercounted. Tsez is spoken in about fifty small villages, and probably each village has some combination of grammatical and lexical features that sets it apart from each other village. However, the different village varieties can be grouped into a limited number of dialect clusters.

The most detailed published classification of the Tsez dialects remains Imnajšvili (1963, p.9-10), based on extensive fieldwork among the Tsez in the period 1946–1954. There is a clearcut distinction between the Sagada dialect group and the remainder of the dialects, which latter I will call the Nuclear Tsez dialect group; mutual intelligibility across this divide is impaired, while varieties within each of the Sagada and Nuclear Tsez groups are readily mutually intelligible. I restrict myself in this article to the Nuclear Tsez group. Imnajšvili divides the Nuclear Tsez group into five dialect clusters, as shown in Table 1, and this classification remains the basis for the map in Koryakov (2002). Figure 1 is a schematic visualization of the relative geographic location of the Tsez dialects based on Koryakov, with dialect clusters in bold face; where more than one village dialect within a dialect cluster is cited in this article, they are indicated in italics. In the text of the article, references are to village dialects unless cluster or group is specified.

Bernard Comrie. 2022. Investigating a linguistic mini landscape: The Tsez (Dido) dialect dictionary project. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 25–28. Available under CC BY 4.0

Dialect group	Dialect cluster	Village dialects referred to in text
Sagada	Sagada	Sagada
Nuclear Tsez	Kidero	Kidero, Mokok
	Shaitli	Shaitli
	Asakh	Asakh, Khushet, Khutrakh, Tsebari
	Shapikh	Shapikh
	Elbok	Elbok

Table 1: Tsez dialects (classification)

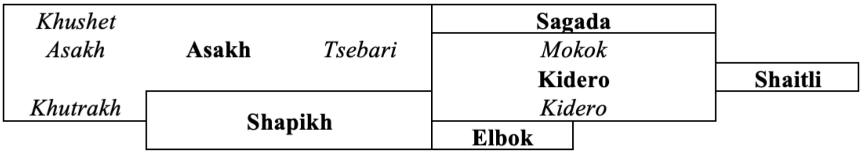


Figure 1: Tsez dialects (sketch map)

Imnajšvili (1963) is a grammar, and as such concentrates on grammatical, including phonological, differences across dialects, with only incidental attention to lexical differences. The explosion of work on Tsez that started in the late 1980s and continues to the present day has seen continued investigation of the grammar of the language, in particular of the dialect groups Asakh (Tsebari village) and Kidero (Kidero and especially Mokok villages). There has also been a burgeoning of work on the lexicon, with the main published work to date being Xalilov (1999); this work does not aim to cover the whole range of Tsez dialects, but often gives variants for Kidero, Mokok, and Asakh. Ramazan Rajabov, from Tsebari, compiled unpublished lexical materials in his native dialect in his capacity as Research Assistant under NSF grant SBR-9220219 (University of Southern California; PIs Maria Polinsky, Bernard Comrie) in the early 1990s.

The most ambitious proposal so far to document cross-dialect diversity, more specifically lexical diversity, in Tsez is the project Dialect Dictionary of the Tsez (Dido) Language (Abdulaev & Xalilov, In prep). By mid-2022 preliminary version 5 of the dictionary was ready, under the authorship of Arsen Abdulaev (from Mokok), and Madžid Xalilov (Head of the Lexicology and Lexicography Department in the Daghestan Federal Research Center of the Russian Academy of Sciences); my own role in the project is as a scientific advisor. The data collection phase of the project was funded by the then Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology.

Comparison of this current project with dialect differences seen in Imnajšvili (1963) provides input into the discussion of section 2. At first, however, one might wonder to what extent one can reasonably compare a grammar based on documentation from the mid-twentieth century with a contemporary dictionary. Imnajšvili (1963) does not deal specifically with the lexicon, while Abdulaev & Xalilov (In prep) do not deal specifically with grammar. However, there are two factors that lead to substantial comparability. First, Tsez has a reasonably complex inflectional morphology, and like many such languages has lexicalized some grammatical forms, so that these appear as separate lemmas in a dictionary. One might compare the English adjectives *interesting* and *tired*, distinct lexical items despite their etymologies as present participle of the verb *to interest* and past participle of *to tire* respectively. Second, an important point of comparison is the phonology, and here both the grammar and the lexicon provide relevant material, in particular given that the grammar gives examples of the inflectional morphology of a range of lexical items.

2 The dynamics of stability and change in Tsez dialects

One respect in which Nuclear Tsezic dialects differ is in the details of morphology. For instance, negative suffixes in some dialects begin with simple *č’*, while in others they begin with *nč’*. The negative converb thus has two dialect variants, in *-č’ey* and in *-nč’ey*, as in the forms of the verb *+–iy*[1] ‘to know’: *+–iy-č’ey* and *+–iy-nč’ey*. This negative converb has become lexicalized in Tsez in the meaning ‘unknowingly, unbeknownst’ (i.e. either the subject of the sentence or some other entity may lack knowledge — Tsez is simply unspecific here), and as such it is listed in the dictionary. Imnajšvili (1963, p.199) shows the suffix variant with *n* for the Kidero, Shaitli, and Elbok dialect clusters, the variant without *n* for the Asakh and Shapikh dialect clusters. Exactly this same distribution is found in Abdulaev & Xalilov (In prep): (s.v. рийнчIей). This seems to reflect a general tendency in Tsez dialect morphology. The dialects are morphologically conservative, retaining features that distinguish them from other dialects as stable markers of local identity.

Turning now to phonology, including morphophonology, two phenomena turn out to be not only of interest given recent developments in Tsez dialects but also well represented in the citation forms of lexical items in Abdulaev & Xalilov (In prep): long vowels and labialized consonants.

All Tsez dialects seem historically to have had long vowels, restricted to certain morphological forms and probably originally representing vowel lengthening with concomitant reduction in the number of qualitative oppositions. The material provided by Imnajšvili (1963) shows that long vowels were then largely intact across the dialects, although there were already some signs of shortening. Thus, Imnajšvili (1963, p.94) gives long-vowel forms for the dative case of the first person singular pronoun for Kidero and Mokok (*dǟr*), Asakh (*dār*), and Shaitli (*dēr*), but a form with a short vowel in Elbok (*dár*, where the acute accent indicates a stressed short vowel). Abdulaev & Xalilov (In prep) show that long vowels have undergone shortening across many dialects. They are systematically retained in Mokok and Asakh, to which we can add Tsebari. They are systematically lost in Kidero, this being, incidentally, one of the features that now distinguishes Mokok from Kidero within Imnajšvili’s Kidero dialect cluster. The shortening of long vowels in Kidero was noted already in Kibrik & Kodzasov (1990, p.329). The dialect differences can be seen in a nominalized derivative of the verb *gugi-* ‘to be lost’. The past participle has the suffix *-ru*, which requires vowel lengthening, thus giving *gāgiru*, as still in Mokok and Asakh. This can then be nominalized with the suffix *-łi* to give *gāgirułi* ‘loss’, which is listed as a separate lexical item in Abdulaev & Xalilov (In prep): s.v. гāгирулъи). The word for ‘loss’ is given there as *gāgirułi*, with a long vowel, in Mokok and Asakh, but as *gagirułi*, with a short vowel, in Kidero. Material from other village dialects is still being processed.

Historically, labialized consonants are attested in Tsez both in particular lexical items, e.g. *k*^{w}*edin* ‘sledgehammer’, including finally in some verb stems, e.g. *cax*^{w}- ‘to write’, and as the result of desyllabification of the vowel *u* before another vowel in verbs, as in the infinitive *+–ez*^{w}*-a* from the stem *+–ezu-* ‘to look’. I concentrate here on verb forms like infinitive *cax*^{w}*-a* and *+–ez*^{w}*-a*. As documented by Imnajšvili (1963, p.166), labialization was still found, apparently in all dialects in both verb types (though with some indications of incipient loss in the form of sporadic variable labialization). By the time of Abdulaev & Xalilov (In prep), it had basically been lost in Kidero, Mokok, Shaitli, and Shapikh, consistently retained only in Asakh and Elbok, predominantly retained in Khushet and predominantly lost in Khutrakh — the “predominantly” here covering variation that probably indicates a sound change in progress. In Tsebari, the situation is more differentiated: Labialization is retained consistently in the *+–ez*^{w}*-a* type, but just as consistently lost in the *cax*^{w}*-a* type. Compare the forms in Table 2.

In the language of the period documented by Imnajšvili (1963), all dialects would have followed the pattern of Asakh village. Some explanations of the forms are in order. First, the past witnessed has the suffix *-si* after a consonant, shortened to *-s* after a vowel; since labialized consonants are only found phonetically before a vowel, the labialization is lost in *cax-si*. Second, the infinitive has the suffix *-a*, before which a vowel is lost: The vowel *i* is simply dropped, while the vowel *u* is desyllabified to give labialization in dialects that retain labialization, dropped in those that do not.

[1] The notation +– at the beginning of a word indicates that the given word requires a gender agreement prefix in that morphological slot.

Stem	Past witnessed		Infinitive		Lemma in Abdulaev & Xalilov (In prep)	Gloss
		Mokok	Asakh	Tsebari		
+–ac’-	*+–ac’-si*	*+–ac’-a*	*+–ac’-a*	*+–ac’-a*	бацIа	‘to eat’
+–ik’i-	*+–ik’i-s*	*+–ik’-a*	*+–ik’-a*	*+–ik’-a*	бикIа	‘to go’
cax^{w}*-*	*cax-si*	*cax-a*	*cax*^{w}*-a*	*cax-a*	цаха	‘to write’
+–ezu-	*+–ezu-s*	*+–ez-a*	*+–ez*^{w}*-a*	*+–ez*^{w}*-a*	беза	‘to look’

Table 2: Labialization in Tsez verb forms

The phonological changes of vowel shortening and delabialization thus show innovations spreading across the landscape, though still sensitive to village dialect boundaries, which sometimes become isoglosses separating the innovative and conservative forms.

3 Conclusion

For the Tsez community, the main attraction of Abdulaev & Xalilov (In prep) is the rich information it provides on lexical differences across Tsez dialects. For linguists, however, the dictionary, even in its present preliminary stage, contains enough information on morphology and phonology to provide new insights into the dynamics of language stability and change across Tsez dialects since the mid-twentieth century. In particular, morphological distinctions seem stable, perhaps as markers of local linguistic identity. By contrast, phonological (including morphophonological) changes, such as vowel shortening and delabialization of consonants, have spread rapidly, though they remain sensitive to dialect boundaries as potential isoglosses. It is to be hoped that more detailed studies of Tsez dialect morphology and phonology will follow.

References

Arsen K. Abdulaev & Madžid Š. Xalilov. (In prep.). Диалектологический словарь цезского (дидойского) языка [Dialect dictionary of the Tsez (Dido) language].

David S. Imnajšvili. 1963. *Дидойский язык в сравнении с гинухским и хваршийским языками [The Dido language in comparison with Hinuq and Khwarshi].* Tbilisi: Izd-vo Akademii nauk Gruzinskoj SSR.

Aleksandr E. Kibrik & Sandro V. Kodzasov. 1990. *Сопоставительное изучение дагестанских языков. Имя. Фонетика [Comparative study of Daghestanian languages: The noun. Phonetics].* Moscow: Izd-vo MGU.

Yuri Koryakov. 2002. Dagestanian languages: West. In Yuri Koryakov, editor, *Atlas of the Caucasian languages: Map 9.* Moscow: Institute of Linguistics RAS. `http://lingvarium.org/maps/caucas/9-andido.gif`.

Madžid Š. Xalilov. 1999. *Цезско-русский словарь [Tsez-Russian dictionary].* Moscow: Academia.

Beyond strings of characters: Resources meet NLP – Again

Dana Dannélls
Språkbanken Text
University of Gothenburg, Sweden
dana.dannells@svenska.gu.se

Tiago Timponi Torrent
FrameNet Brasil
Federal University of Juiz de Fora, Brazil
tiago.torrent@ufjf.br

Natalia Sathler Sigiliano
FrameNet Brasil
Federal University of Juiz de Fora, Brazil
natalia.sigiliano@ufjf.br

Simon Dobnik
CLASP, FLoV
University of Gothenburg, Sweden
simon.dobnik@gu.se

Abstract

FrameNet (FN) resources have existed for many languages for over a decade but their adoption in real world applications has been limited. To celebrate the 65 anniversary of Lars Borin, the initiator and leader of Swedish FrameNet, among others, we take a standpoint to motivate why language resources are crucial for moving NLP forward. We present our position on (a) the need for language resources to embrace other dimensions of text and language use, and (b) the need for them to relate to other representations through multimodality.

1 Introduction

The late 1990's witnessed the consolidation of two important areas in Natural Language Processing (NLP). On one side, statistically-oriented methods took advantage of improved computing capacity and corpus availability to make their way into not only mainstream computational processing of linguistic structures, but also into core research in Artificial Intelligence. On the other, language resources, such as WordNet (Miller et al., 1990) and FrameNet (Baker et al., 1998), redefined methodologies, outcomes and expectations for the computationally assisted development of representations of linguistic cognition. As both sub-fields evolved, though, only the first of them kept being framed as core NLP.

As a result, on top of the enormous progress derived from the development and application of models based on embeddings and transformers – only to mention the most recent – to the analysis of human languages, the association of computational linguistics with statistics-based approaches also brought a misconception: the one according to which human languages can be modelled from form alone (Bender & Koller, 2020; Merrill et al., 2021). Form in this case equally applies to multimodal models, where strings of characters and pixels are statistically matched to emulate grounding (Kelleher & Dobnik, 2022).

Such a misconception is prejudicial for two of the main purposes of NLP: that of modelling how human languages work and that of applying computational models to downstream tasks. Regarding the first, an extensive body of research in Linguistics has demonstrated that linguistic form alone does not encode meaning in a way that is either exhaustive or sufficient for comprehension – see Fauconnier & Turner (2002) for pointers. It has also demonstrated that strict compositionality is not able to account for very central language understanding processes (Fillmore, 1979). As for the latter, as Bender et al. (2021) explain, models based on the statistic manipulation of linguistic form also encode unwanted social biases, since they are trained on static limited corpora, representing, via patterns extracted from strings of characters, limited perspectives on culture and society, excluding marginalised groups. Moreover, Rogers

Dana Dannélls, Tiago Timponi Torrent, Natalia Sathler Sigiliano and Simon Dobnik. 2022. Beyond strings of characters: Resources meet NLP – Again. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 29–36.

(2021) shows such models do not perform well on phenomena that are not frequent, because languages follow Zipf's law and most phenomena are distributed along low frequencies of occurrence.

If we want computers to approach understanding of natural languages, and also get a full-range understanding of our reality, we need to equip them with large amount of linguistic knowledge (Dobnik et al., 2022). Such knowledge includes meaning, which, in turn, is structured in terms of scenes, grounded in context and subject to construal operations (Trott et al., 2020). Such a representation of meaning cannot be achieved via statistical processes alone. Therefore, our standpoint is that structured language resources are crucial for moving NLP forward. Research indicates that syntactic and semantic annotated resources compensate for quantitative data (Swayamdipta et al., 2018), providing evidence that rich linguistic resources complemented with high-quality annotations are valuable in the field (Conia & Navigli, 2020; Marton & Sayeed, 2021). Moreover, the tremendous computational and environmental costs resulting from training deep neural network models (Strubell et al., 2019) have also been reframing the idea that curated language resources are too expensive, while neural networks are almost free.

Nonetheless, language resources too need rethinking. Even the ones, like FrameNet (Fillmore et al., 2003), which rely heavily on annotation to attest the analyses, traditionally adopt a very narrow definition of what is an instance of language. Although Berkeley FrameNet (BFN) conducts most of its annotation process on the British National Corpus (BNC) (BNC Consortium, 2007), which is balanced for genre, the annotation disregards structure that lies beyond the syntactic locality of the target lexical unit being annotated. In this regard, because of its lexicographic origins, BFN limits annotation to the association of semantic and morpho-syntactic labels to parts of sentences, with no information being stored or even derived, for example, about the genre macro-structure and properties in frame semantics terms. As pointed out by Torrent et al. (2022), this is not a limitation of the theory of Frame Semantics (Fillmore, 1982), but a limitation deriving from how the original FN model was implemented. In this paper, we arguing argue that the lexicographic orientation of BFN has contributed to reducing text to strings of characters. We then present our position on (a) the need for language resources to embrace other dimensions of texts, namely those related to the characteristics of genres, which would reconcile them with statistical language models, and (b) the need for them to go one step further and embrace multimodality, since human communication is inherently multimodal. Before advancing to those two claims, though, the next section presents an overview of FrameNet.

2 FrameNet in the multilingual world

FrameNet (FN), a lexical semantic resource originally developed for English, was established as a result of a computational lexicographic project led by Fillmore and his colleagues (Baker et al., 1998). The resource rests on the linguistic theory of Frame Semantics (Fillmore, 1982). It follows the standard view of how humans comprehend language through a conceptual, semantic system, containing knowledge about the world that is necessary for supporting inference, and performing cognitive tasks. In the context of FN, frames are general schematic representations of actions, containing Frame Elements (FEs) and Lexical Units (LUs) that can be mapped to particular instances of text type and genre.

Because FN encodes both semantic and syntactic valence information about words – the two components that are arguably the driving force behind any NLP task that requires NLU – it has inspired new initiatives in other languages, including, just to name a few: Japanese (Saito et al., 2008), Spanish (Subirats, 2009), Swedish (Dannélls et al., 2021) and Portuguese (Torrent & Ellsworth, 2013). Following the design and development of BFN (Fillmore et al., 2003), all languages, almost exclusively, encode systematic representations of semantic structures and their relations to words based on empirical evidence from corpus data. Nevertheless, the nature of the corpora from where sentences were extracted, the methods used to extract them, the annotation processes, and the skills the annotators possess differ greatly between the languages. Perhaps not surprisingly, despite initiatives of emulating FN in other languages for the purpose of creating full-fledged lexical semantic resources, considerably little effort has been given to exploiting FN resources in real-world applications. One important reason for this is the small amount of annotated data for languages other than English, as shown in Table 1. The second reason is the lack of context information surrounding lexical units. For example, despite the large amount of annotated sen-

	English	Japanese	Portuguese	Spanish	Swedish
Total lexical units	13 421	3 405	8 393	1 268	39 212
Total annotation sets	200k	73k	12k	11k	9k

Table 1: FrameNet data statistics from Baker et al. (2015) with modification of the Portuguese and Swedish data.

tences covered in BFN, not all of the LUs in a frame are attested with example sentences. Moreover, the representativeness of the annotation in terms of both the lexical unit and the frame elements instantiated in each annotation set is not guaranteed. In BFN many sentences are annotated with null instantiation categories for expressing structurally omitted constituents. This lack of semantic and syntactic coverage has limited the performance of automatic processes such as semantic role labelling (see Section 3). Third is the abstract level of distinctiveness of the semantic categories available in frames, resulting in the lack of mechanism for representing the intrinsic meanings of LUs. For example, the LUs *dog* and *cat* belong to the `Animal` frame but there are no further semantic attributes to distinguish between them or represent how they are perceived in the world. To include this knowledge, an addition of cumulative, context and common-sense knowledge is required (Torrent et al., 2022).

Arguably, lexical semantic resources are hard to come by. Annotation sets are amenable to the corpus at hand, as well as to the annotator curation in cases of manual annotations (Chang et al., 2015), something that was acknowledged in the development and construction of Swedish FrameNet (SweFN).

SweFN is one of the largest FrameNet resources covering nearly 40k lexical units. It has been created by reusing exiting Swedish linguistic resources and integrating them all in a large Swedish infrastructure for language technology through one pivot lexicon Saldo (Borin et al., 2021). Assuming the Zipf behaviour which characterises lexical resources like Saldo is of great importance when resources are to be connected (Borin, 2010). Annotation sets in SweFN were retrived from Korp (Borin et al., 2012) – Språkbanken Text corpus infrastrcuture, which contains text types and genres from diverse spheres of human activity, ranging from academic, medical, legal, newspapers, fiction, journals and social media. The annotation sets provide broad coverage of the semantic and syntactic representations of LUs. An undertaking that was achieved by balancing between computational methods and manual work, leaving some room for human intuitions and some room for consistent, robust language processing.

However, in spite of being developed within a larger infrastructure centred around text, SweFN – like most other FN initiatives – is yet to incorporate information that goes beyond the micro-structure of text. In the following section we argue for this direction as a next step in the expansion of the FN model.

3 Beyond strings of characters: FrameNet meets textual genres

From the point of view of the approaches to Linguistics that take meaning and the social context of language use inseparable from linguistic form, text is much more than sequences of characters forming a sentence judged as grammatical in a language (Cooper, in prep). Therefore, under those perspectives, as much as annotated corpora have been playing a key role in NLP for the past two decades and metadata associated to raw data has proven beneficial for many applications in the field, the material traditionally used in annotation projects – most FN initiatives included – is only one of the ingredients in a text.

Standard FN-like annotation is capable of capturing morpho-syntactic information on the instantiations of FEs in a sentence. Those properties are used in BFN for building the valence patterns of LUs and may eventually inform lexicographers of the need to split a frame or re-frame a given LU, moving it to another frame (Petruck et al., 2004; Ruppenhofer et al., 2016). Also, BFN annotations of sentences have been used for training semantic role labellers such as SEMAFOR (Das et al., 2010), Open Sesame (Swayamdipta et al., 2017) and LOME (Xia et al., 2021). However, although the BNC, the main corpus used by BFN, is balanced for genres (BNC Consortium, 2007), most of BFN annotation does not make use of such a feature. This is so because BFN was built as a lexicographic resource, and annotated sentences extracted from the BNC were meant to support the analyses carried out for a given LU. Therefore, the portion of BFN annotations used for training SEMAFOR, Open Sesame and LOME is that of the full text annotation section, which comprises only circa 3,000 sentences from mostly bureaucratic texts, news pieces and travel guides (Das et al., 2010). This is to say, in other words, that those semantic role labellers are trained and tested on a tiny fraction of textual genres there are, and that the dataset used for training them does not come from a balanced corpus.

More than limiting the representativeness of the training data concerning the semantic side of annotation, this is also a limiting factor for the morpho-syntactic side of the annotation. This is so because, as demonstrated by Sigiliano & Torrent (2017), different textual genres may show different morpho-syntactic valence affordances. In their paper, authors show, for example, that the omission of core FEs, especially those with indefinite reference, was considerably higher in travel guides when compared to the occurrences of the same type of null instantiation in the TED Talk annotated for the Global FrameNet shared annotation task (Torrent et al., 2018).

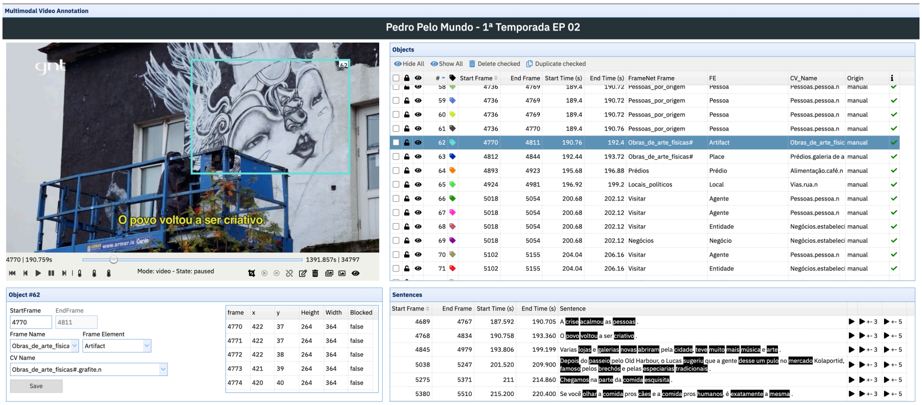
Nonetheless, this is not the only limitation of annotating sequences of characters forming sentences, instead of annotating genres. There is a myriad of types of information that genres comprise besides the sentences in them. This is to say that they can only be defined, following Swales (1990, p.58), as a class of events sharing a communicative purpose, a schematic structure supporting such purpose, as well as similarities in form, style, structure and content, because members of a linguistic community can recognise them from shared characteristics. Also, they can only be grouped together, as proposed by Schneuwly & Dolz-Mestre (2004), because they share biases towards given linguistic operationse.g. the common use of imperatives in instructional genres in languages such as Brazilian Portuguese and English. A pilot experiment conducted by Dutra & Sigiliano (2021) was able to extract correlations between FN-like valence patterns and genre by annotating a corpus comprising 25 exemplar texts from 25 different genres, grouped, following Schneuwly & Dolz-Mestre (2004), under the argumentative, expository, instructional, narrative and reporting domains. For such a project, genres were imported to the FrameNet Brasil WebTool and annotated following the principles of full-text annotation. Although controlled for genre, the annotation does not take into consideration key features of genres, namely those located beyond the verbal language communicative mode.

As Bateman (2008) points out, however, advances in technology have been highlighting the importance of other communicative modes for genre analysis. The author also notes the lack of analytical tools for accounting for those multimodal aspects, especially in order to ground the analysis in more concrete details of objects being analysed. Hiippala (2014) points to the fact that the lack of linearity in multimodal genres compromises current analytical tools generally applied to text-only genre analysis. The author proceeds by claiming that because multimodal genres are stratified, tools used for analysing them must be capable of identifying semiotic choices contributing to the genre structure in multiple strata. In the following section we claim that a multimodal turn in FN might provide such kind of tool.

4 Beyond verbal language: FrameNet meets multimodality

To some extent, FN analyses, especially full-text annotation, are already capable of providing non-linear representations of the semantics of text. As pointed out by Torrent et al. (2022), the very nature of frames, as defined by Fillmore (1982), include both common-sense knowledge and communicative situation grounding. However, methodological decisions made when implementing Frame Semantics as a computational lexicographic resource focusing mostly on content – as opposed to functional – words have precluded such aspects from being properly considered in annotation.

As for communicative situation grounding, work on pragmatic frames (Ohara, 2018; Czulo et al., 2020) has begun to indicate paths for expanding the kinds of frames FN may include, by means of looking into pragmatic set-ups evoked by grammatical constructions. The idea is that frames should be extended to represent linguistic knowledge activated by language structures such as deixis, turn taking and information status. Nonetheless, once one recognises the inherently multimodal nature of human communication, other communicative modes must be considered as well. Belcavello et al. (2020) introduce the idea for a multimodal FN, by reporting on a pilot annotation experiment conducted on a TV travel documentary. Authors describe the process for extracting verbal language data from the video and feeding the output for full-text annotation. They discuss the methodology for annotating video sequences for frames using



Multimodal Video Annotation

Pedro Pelo Mundo - 1ª Temporada EP 02

4770 | 190.759s   1391.857s | 34797

Mode: video - State: paused

Objects

Hide All · Show All · Delete checked · Duplicate checked

□	#	Start Frame	End Frame	Start Time (s)	End Time (s)	FrameNet Frame	FE	CV_Name	Origin	i
□	58	4736	4769	189.4	190.72	Pessoas_por_origem	Pessoa	Pessoas.pessoa.n	manual	✓
□	59	4736	4769	189.4	190.72	Pessoas_por_origem	Pessoa	Pessoas.pessoa.n	manual	✓
□	60	4736	4769	189.4	190.72	Pessoas_por_origem	Pessoa	Pessoas.pessoa.n	manual	✓
□	61	4736	4770	189.4	190.76	Pessoas_por_origem	Pessoa	Pessoas.pessoa.n	manual	✓
□	62	4770	4811	190.76	192.4	Obras_de_arte_físicas#	Artifact	Obras_de_arte_físic	manual	✓
□	63	4812	4844	192.44	193.72	Obras_de_arte_físicas#	Place	Prédios.galeria de a	manual	✓
□	64	4893	4923	195.68	196.88	Prédios	Prédio	Alimentação.café.n	manual	✓
□	65	4924	4981	196.92	199.2	Locais_políticos	Local	Vias.rua.n	manual	✓
□	66	5018	5054	200.68	202.12	Visitar	Agente	Pessoas.pessoa.n	manual	✓
□	67	5018	5054	200.68	202.12	Visitar	Agente	Pessoas.pessoa.n	manual	✓
□	68	5018	5054	200.68	202.12	Visitar	Entidade	Negócios.estabeleci	manual	✓
□	69	5018	5054	200.68	202.12	Negócios	Negócio	Negócios.estabeleci	manual	✓
□	70	5102	5155	204.04	206.16	Visitar	Agente	Pessoas.pessoa.n	manual	✓
□	71	5102	5155	204.04	206.16	Visitar	Entidade	Negócios.estabeleci	manual	✓

Object #62

StartFrame: 4770   EndFrame: 4811

Frame Name: Obras_de_arte_física   Frame Element: Artifact

CV Name: Obras_de_arte_físicas#.grafite.n

Save

frame	x	y	Height	Width	Blocked
4770	422	37	264	364	false
4771	422	37	264	364	false
4772	422	38	264	364	false
4773	421	39	264	364	false
4774	420	40	264	364	false

Sentences

Start Frame	End Frame	Start Time (s)	End Time (s)	Sentence			
4689	4767	187.592	190.705	A crise acalmou as pessoas .	▶	▶ +- 3	▶ +- 5
4768	4834	190.758	193.360	O povo voltou a ser criativo .	▶	▶ +- 3	▶ +- 5
4845	4979	193.806	199.199	Varias lojas e galerias novas abriram pela cidade, teve muito mais música e arte .	▶	▶ +- 3	▶ +- 5
5038	5247	201.520	209.900	Depois do passeio pelo Old Harbour, o Lucas sugeriu que a gente desse um pulo no mercado Kolaportid, famoso pelos brechós e pelas especiarias tradicionais .	▶	▶ +- 3	▶ +- 5
5275	5371	211	214.860	Chegamos na parte da comida esquisita .	▶	▶ +- 3	▶ +- 5
5380	5510	215.200	220.400	Se você olhar a comida pros cães e a comida pros humanos, é exatamente a mesma .	▶	▶ +- 3	▶ +- 5

Figure 1: An image annotated with the Charon video annotation interface.

bounding boxes associated to elements in the scenes. In another annotation experiment, Viridiano et al. (2022) report on the annotation of the Flickr30k Entities dataset (Plummer et al., 2015) for frames and FEs. Both annotation efforts are conducted in Charon, an annotation tool developed to extend FN-like annotation to the visual mode (Belcavello et al., 2022) shown in Figure 1. This allows for the annotations of the verbal language and the visual modes to be correlated. In the part of the TV show being annotated, one Reykjavik local interviewed by the program host explains that the financial crisis in Iceland had a positive impact on the people. He than enunciates (1). Such a sentence could be annotated for the LU *criativo.a* evoking the `Mental_property` frame, as shown in (2).

(1) O povo voltou a ser criativo.
People became creative again.

(2) [O povo$_{\text{PROTAGONIST}}$] voltou a [ser$_{\text{Copula}}$ [$^{\texttt{Mental\_property}}$**criativo**$_{\text{BEHAVIOR}}$].

The visual mode, in turn, is annotated for the `Physical_artworks` frame, being the graffiti on the wall annotated for the ARTEFACT FE. The annotations of both modalities, when combined, have, thus, the potential of enriching the semantic representation FN is capable of providing for a multimodal genre.

5 What lies ahead in the future?

Making predictions of what lies ahead in NLP is always risky, but the position that we have laid out in this paper indicates that we should consider at least four points:

(i) Focus on the creation of high quality resources that not only overcome the current resource biases but also cover a wide-variety of genres expressing different communicative intents. This means that data collection is equally important as the annotation of these resources by frames. Without good data, we cannot have good coverage of resources.

(ii) Frame annotation must invariably go beyond annotation of texts. This will disassociate the current frame annotation bias to textual forms and would allow them to represent more common-sense knowledge, making them useful for a variety of natural language inference tasks. An excellent example of this is integration of other resources done in the SweFN project. More knowledge is always better but then a question arises as to what degree such representations can be applied in traditional tasks such as semantic role labelling in texts alone since not all such information is explicitly expressed in texts, and contexts will have to be disambiguated. This brings us to the third point.

(iii) Meaning representations should be multimodal as this is how communication works. The interesting question is then what modes are to be integrated and how? We have seen examples of annotation of images and videos in FN-Brasil but one could also include gestures, emotions and sentiment. Such an undertaking assumes theoretical understanding how different modalities interact and make up meaning and also also how language is used in linguistic and non-linguistic interaction with the world. However, despite large amount of theoretical and experimental work in these areas, such mechanisms are also not yet fully known and understood.

(iv) How do we represent multimodal meaning in a way that it can be linked to frame representations? Bridging conceptual and perceptual domains inevitably involves classification and hence we need representations that are hybrid: data driven and machine learned, and expert-defined encoding frame information. Moreover, frame representations need to become multidisciplinary to capture what other fields – such as Computer Vision, Robotics, Music and Film Theories, Mass Media Communication – already know about meaning production.

What will be made of these points might appear in future research or Lars' 75 anniversary Festschrift.

Acknowledgements

Research conducted in FrameNet Brasil is funded by FAPEMIG RED-00106-21 and CNPq 408269/2021 - 9 research grants. Tiago Torrent is an awardee of CNPq's Research Productivity grant 315749/2021-0. Natália Sigiliano and Tiago Torrent are thankful for the support of the Department of Swedish, Multilingualism and Language Technology at Gothenburg University. This work was supported by Swe-Clarin (grant 201302003). It was also supported by the Centre for Linguistic Theory and Studies in Probability (CLASP) (Swedish Research Council (VR) grant 2014-39).

References

Collin F. Baker, Charles J. Fillmore, & John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, pages 86–90, Montreal Quebec. Association for Computational Linguistics.

Collin F. Baker, Nathan Schneider, Miriam R. L. Petruck, & Michael Ellsworth. 2015. Getting the roles right: Using FrameNet in NLP. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–12, Denver, Colorado. Association for Computational Linguistics.

John Bateman. 2008. *Multimodality and genre: A foundation for the systematic analysis of multimodal documents*. Palgrave MacMillan, New York.

Frederico Belcavello, Marcelo Viridiano, Alexandre Diniz da Costa, Ely Edison da Silva Matos, & Tiago Timponi Torrent. 2020. Frame-based annotation of multimodal corpora: Tracking (a)synchronies in meaning construction. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 23–30, Marseille. European Language Resources Association.

Frederico Belcavello, Marcelo Viridiano, Ely Matos, & Tiago Timponi Torrent. 2022. Charon: A FrameNet annotation tool for multimodal corpora. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 91–96, Marseille. European Language Resources Association.

Emily M. Bender & Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, & Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

BNC Consortium. 2007. British national corpus, XML edition. Oxford Text Archive.

Lars Borin, Markus Forsberg, & Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Lars Borin, Dana Dannélls, & Karin Friberg Heppin. 2021. Introduction: Swedish Framenet++. In Dana Dannélls, Lars Borin, & Karin Friberg Heppin, editors, *The Swedish FrameNet++. Harmonization, integration, method development and practical language technology applications*, pages 3–36. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Lars Borin. 2010. Med Zipf mot framtiden – en integrerad lexikonresurs för svensk språkteknologi [With Zipf into the future – an integrated lexical resource for Swedish language technology]. *LexicoNordica*, 17:35–54.

Nancy Chang, Praveen Paritosh, David Huynh, & Collin F. Baker. 2015. Scaling semantic frame annotation. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Simone Conia & Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410, Barcelona (Online). International Committee on Computational Linguistics.

Robin Cooper. in prep. From perception to communication: An analysis of meaning and action using a theory of types with records (TTR). To appear in Oxford Studies in Semantics and Pragmatics, Oxford University Press.

Oliver Czulo, Alexander Ziem, & Tiago Timponi Torrent. 2020. Beyond lexical semantics: notes on pragmatic frames. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 1–7, Marseille. European Language Resources Association.

Dana Dannélls, Lars Borin, Markus Forsberg, Karin Friberg Heppin, & Maria Toporowska Gronostaj. 2021. Swedish FrameNet. In Dana Dannélls, Lars Borin, & Karin Friberg Heppin, editors, *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, pages 37–65. John Benjamins, Amsterdam.

Dipanjan Das, Nathan Schneider, Desai Chen, & Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, California. Association for Computational Linguistics.

Simon Dobnik, Robin Cooper, Adam Ek, Bill Noble, Staffan Larsson, Nikolai Ilinykh, Vladislav Maraev, & Vidya Somashekarappa. 2022. In search of meaning and its representations for computational linguistics. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 30–44, Gothenburg. Association for Computational Linguistics.

Lívia Dutra & Natália Sigiliano. 2021. Ferramenta linguístico-computacional como facilitadora para o ensino de gramática na escola. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 432–436, Porto Alegre, RS. SBC.

Gilles Fauconnier & Mark Turner. 2002. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic books, New York.

Charles J. Fillmore, Miriam R.L. Petruck, Josef Ruppenhofer, & Abby Wright. 2003. FrameNet in action: The case of attaching. *International Journal of Lexicography*, 16(3):297–332.

Charles J Fillmore. 1979. Innocence: a second idealization for linguistics. In *Annual Meeting of the Berkeley Linguistics Society*, volume 5, pages 63–76.

Charles J. Fillmore. 1982. Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul.

Tuomo Hiippala, 2014. *11. Multimodal genre analysis*, pages 111–124. De Gruyter Mouton, Berlin.

John D. Kelleher & Simon Dobnik. 2022. Distributional semantics for situated spatial language? Functional, geometric and perceptual perspectives. In Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin, & Aleksandre Maskharashvili, editors, *Probabilistic approaches to linguistic theory*, CSLI Publications, pages 319–356. Center for the Study of Language and Information, Stanford university, Stanford, California.

Yuval Marton & Asad Sayeed. 2021. Thematic fit bits: Annotation quality and quantity interplay for event participant representation. *ArXiv*.

William Merrill, Yoav Goldberg, Roy Schwartz, & Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, & Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Kyoko Ohara. 2018. The relations between frames and constructions: A proposal from the Japanese FrameNet Constructicon. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, & Tiago Timponi Torrent, editors, *Constructicography: Constructicon development across languages*, pages 141–164. John Benjamins, Amsterdam.

Miriam R.L. Petruck, Charles J Fillmore, Collin F Baker, Michael Ellsworth, & Josef Ruppenhofer. 2004. Reframing framenet data. In *Proceedings of The 11th EURALEX International Congress*, pages 405–416. Université de Bretagne Sud Lorient.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, & Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Anna Rogers. 2021. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R Johnson, & Jan Scheffczyk. 2016. Framenet II: Extended theory and practice. Technical report, International Computer Science Institute.

Hiroaki Saito, Shunta Kuboya, Takaaki Sone, Hayato Tagami, & Kyoko Ohara. 2008. The Japanese FrameNet software tools. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*, Marrakech. ELRA.

Bernard Schneuwly & Joaquim Dolz-Mestre. 2004. *Gêneros orais e escritos na escola*. Mercado de Letras, São Paulo.

Natália Sigiliano & Tiago Torrent. 2017. Framenet annotation as a means to identify genre-relevant linguistic structures. In *ScriptUM: la revue du colloque VocUM*, pages 1–20.

Emma Strubell, Ananya Ganesh, & Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence. Association for Computational Linguistics.

Carlos Subirats. 2009. Spanish Framenet: A frame-semantic analysis of the Spanish lexicon. In Hans C. Boas, editor, *Multilingual FrameNets in Computational Lexicography. Methods and Applications*, pages 135–162. Mouton de Gruyter, Berlin.

John M Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge university press.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, & Noah A Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, & Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels. Association for Computational Linguistics.

Tiago Timponi Torrent & Michael Ellsworth. 2013. Behind the labels: criteria for defining analytical categories in framenet brasil. *Veredas-Revista de Estudos Linguisticos*, 17(1):44–66.

Tiago Timponi Torrent, Michael Ellsworth, Collin Baker, & Ely Edison da Silva Matos. 2018. The multilingual FrameNet shared annotation task: a preliminary report. In Tiago Timponi Torrent, Lars Borin, & Collin F. Baker, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris. European Language Resources Association (ELRA).

Tiago Timponi Torrent, Ely Edison da Silva Matos, Frederico Belcavello, Marcelo Viridiano, Maucha Andrade Gamonal, Alexandre Diniz da Costa, & Mateus Coutinho Marim. 2022. Representing context in framenet: A multidimensional, multimodal approach. *Frontiers in Psychology*, 13.

Sean Trott, Tiago Timponi Torrent, Nancy Chang, & Nathan Schneider. 2020. (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170–5184, Online. Association for Computational Linguistics.

Marcelo Viridiano, Tiago Timponi Torrent, Oliver Czulo, Arthur Lorenzi, Ely Matos, & Frederico Belcavello. 2022. The case for perspective in multimodal datasets. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 108–116, Marseille. European Language Resources Association.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, & Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.

Ordvektorer i lexikografiskt arbete

Markus Forsberg
Språkbanken Text
Inst. för svenska, flerspråkighet
och språkteknologi
Göteborgs universitet, Sverige
markus.forsberg@gu.se

Emma Sköldberg
Språkbanken Text
Inst. för svenska, flerspråkighet
och språkteknologi
Göteborgs universitet, Sverige
emma.skoldberg@gu.se

Abstract

We present a preliminary case study on the use of word vectors in lexicographic practice. The study shows the potential of using vector models in the revision of existing dictionary entries as well as creating new entries.

1 Inledning

Den lexikografiska praktiken är idag datadriven, där språkliga beskrivningar huvudsakligen baseras på den evidens som samlas ur stora mängder text, så kallade *korpusar*. Språkproven presenteras oftast i form av *konkordanser*, vilket med rätta kan anses vara lexikografens främsta verktyg. Atkins & Rundell (2008) fångar orsaken i följande citat: ”One of the earliest revelations of corpus study was that that right- or left-sorted concordances will often give a powerful, visual representation of a word's recurrent patterns - in a way that is impossible to ignore or overlook.”

Även bruket av olika slags *ordbilder* (Borin et al., 2012) är standard, inom vilka man på olika vis abstraherar ordens språkliga kontexter för att därmed ge en bättre överblick över vad som förekommer i ordens kontext. Ordbilder kan exempelvis vara baserade på en automatisk syntaktisk analys, där man samlat de syntaktiska konstituenterna i separata tabeller, exempelvis en tabell med alla subjekthuvuden för ett visst verb. Också ordbilder är viktiga verktyg för lexikografen för att kunna åstadkomma en så utförlig och rättvisade bild av uppslagsorden som möjligt, och då inte minst uppslagsordens syntagmatiska egenskaper. Frågan är om det finns andra metoder som kan tillföra nya aspekter och perspektiv.

I föreliggande arbete redogör vi för en initial studie av värdet av att använda ordvektorer som ett verktyg i lexikografiskt arbete. Mer konkret studerar vi vad bruket av ordvektorer kan bidra med dels när det gäller beskrivningen av etablerade svenska ord som redan införlivats i en viss ordbok, här Svenska Akademiens Svensk ordbok (2021; hädanefter SO), dels vad vektorer kan tillföra vid beskrivningen av nya ord som kan komma att finnas med i nästa uppdatering av samma verk.[1]

Innan vi kortfattat redogör för vad ordvektorer är kan det vara på plats med en kort beskrivning av den aktuella ordboken. SO är en allmänspråklig definitionsordbok som har till uppgift att spegla samtida svenska. Ordboken innehåller ca 65 000 uppslagsord som är försedda med uppgifter om bl.a. uttal, böjning, betydelse(r), konstruktion(er), fraseologi och etymologi. Tyngdpunkten ligger dock på vad uppslagsorden betyder och hur de används. Betydelsebeskrivningarna stöds av språkprov, såväl morfologiska som syntaktiska. Uppslagsorden i SO relateras också till andra uppslagsord i samma verk genom hänvisningar till bl.a. synonyma, antonyma och kohyponyma ord. Ordboken kan därmed sägas utgöra ett slags semantiskt nätverk mellan olika lexikala enheter i svenskan (Blensenius et al., 2021). SO är tillgänglig dels i form av appar, dels via ordboksportalen svenska.se.

[1] SO utarbetas inom ramen för ett samarbete mellan Svenska Akademien och GU och av ett forskarlag inom vilket vi själva ingår.

Markus Forsberg and Emma Sköldberg. 2022. Ordvektorer i lexikografiskt arbete. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 37–41. Available under CC BY 4.0

2 Ordvektorer

Ordvektorer, även kallade *ordinbäddningar*, är matematiska representationer av ord som försöker fånga ordens egenskaper på så vis att relaterade ord hamnar nära varandra i en så kallad *vektorrymd*. Detta görs med hjälp av ordens kontexter som hämtas ur en stor mängd textmaterial. Uttryckt annorlunda anses ord vara relaterade med varandra när de förekommer i samma typ av språkliga kontext.

Ordvektorernas kvalitet är beroende av hur många relevanta språkliga kontexter som förekommit i det textmaterial man använt för att skapa respektive ordvektor. Med tanke på att ordens statistik är zipfiansk distribuerad, dvs. snabbt avtagande, kommer det oundvikligen att finnas ord som ligger nära varandra i vektorrymden beroende på slumpmässighet snarare än kontextuell likhet. Detta behöver man ta höjd för i en studie som denna.

Det finns många olika metoder för att skapa ordvektorer, och alla dessa metoder har sina för- och nackdelar. Den metod vi har valt i den här studien har implementerats i verktyget *fastText* (Bojanowski et al., 2016), som representerar varje ord utifrån dess delar. Fördelen med detta är att ord som tidigare inte observerats kan hamna rätt i vektorrymden, givet att delarna har observerats. Tekniskt säger man att metoden kan hantera OOV (Out of Vocabulary). Detta är en egenskap som är viktig vid studier av ett språk som svenska med sin rika produktion av nya sammansättningar.

Verktyget fastText distribueras tillsammans med förtränade ordvektorer för 157 språk (Grave et al., 2018), och vi har använt den för svenska. Ordvektorerna är baserade på datamängden Common Crawl, som är ett stort insamlat material med webbsidor, och Wikipedia.

Det finns en rik flora av vetenskapliga arbeten som kretsar runt hur frukten av lexikografiskt arbete kan användas till att förbättra ordvektormodeller, men det finns färre som behandlar hur ordvektorer kan användas i lexikografisk praktik. En studie som ligger nära den studie vi rapporterar om här är Sørensen & Nimb (2018), som använder sig av en annan populär metod för att skapa ordvektorer, word2vec, som ett lexikografiskt verktyg för att hitta ord som saknas i ett visst semantiskt fält eller för att identifiera inkonsekvenser i existerande beskrivningar.

3 Studien

Inom ramen för denna undersökning har vi studerat sammanlagt 20 svenska ord. Hälften av dessa behandlas i SO (2021), hälften av dem kan betraktas som nyordskandidater i förhållande till SO. Kännetecknande för den senare typen av ord är att de finns med i en förteckning med nyordskandidater vilken utarbetats genom en jämförelse mellan svenska textmaterial från 2021 med motsvarande material från 2020. De aktuella orden har antingen dykt upp som nyheter i det senare materialet eller ökat i användning.

De tio etablerade orden är: *adekvat*, *disputation*, *fräsch*, *hund*, *kärlek*, *organisera*, *röd*, *sjunga*, *usch* och *åldras*. Som synes tillhör de olika ordklasser. Vidare uppträder de oftast i olika sammanhang, inte minst i olika slags texter.

De tio nyordskandidater som vi granskat är: *blåbrun*, *gangsterrap*, *glamping*, *kontaktförbud*, *matresa*, *mockumentär*, *prosecco*, *smittovåg*, *snabbtesta* och *yes*. De flesta av dessa är substantiv, men det säger också något om hur ordklassfördelningen ser ut i hela listan med nyordskandidater.

Vid granskningen av de 20 orden har vi studerat vad de har för 100 närmaste grannar i vektorrymden. Här är t.ex. en förteckning över de 100 närmaste grannarna för substantivet *kärlek*, sorterade i fallande ordning baserat på avstånd (siffrorna i parentes):

> **kärlek.** (0.783), **kärlek.En** (0.779), **kärleken** (0.778), **Kärlek** (0.764), **självkärlek** (0.751), **familjekärlek** (0.742), **föräldrakärlek** (0.739), **kärlekDu** (0.732), **kärlek-** (0.729), **hat-kärlek** (0.728), **kärlek.Och** (0.722), **vänskap** (0.718), **moderskärlek** (0.714), **Gudskärlek** (0.713), **livskärlek** (0.711), **kärlekAtt** (0.710), **kärleksfullhet** (0.706), **människokärlek** (0.703), **kärlek.Men** (0.701), **syskonkärlek** (0.685), **kärlekslycka** (0.681), **tonårskärlek** (0.680), **kärlek.Jag** (0.675), **förälskelse** (0.675), **kärlekssorg** (0.675), **kärlekskänslor** (0.674), **kärlek.Det** (0.673), **kärlekshandling** (0.672), **kärleksrus** (0.670), **kärleksmagi** (0.667), **Gudakärlek** (0.665), **längtan** (0.662), **egenkärlek** (0.660), **kärlekskraft** (0.658), **kärlekDag** (0.655), **passion** (0.653), **moderskärleken** (0.653), **kärleksförklaring** (0.649), **kärlekslöshet** (0.649), **kärlekslängtan** (0.648), **Självkärlek** (0.648), **romantik** (0.647), **broderskärlek** (0.647), **nyförälskelse** (0.645), **människokärleken** (0.643), **kärleksbevis** (0.642), **hjärtesorg** (0.642), **tvåsamhet** (0.639), **vänskap.** (0.639), **kärleksgärning** (0.635), **kärleksberättelse** (0.635), **förälskelser** (0.635), **villkorslös** (0.633), **kärlekar** (0.631), **Kärleken** (0.630), **kärlekstrubbel** (0.629), **matkärlek** (0.628), **omtänksamhet** (0.627), **moderlighet** (0.627), **kärlekEnsamhet** (0.626), **wikikärlek** (0.625), **kärlekstecken** (0.625), **kärlekshandlingar** (0.625), **sanningskärlek** (0.625), **Förälskelse** (0.624), **kärleksförhållanden**

> (0.622), **kärlekslyrik** (0.621), **barnlängtan** (0.621), **Nätkärlek** (0.619), **ömhet** (0.619), **kärleksakt** (0.618), **kärleksljus** (0.618), **Hatkärlek** (0.617), **kärleksbok** (0.616), **kärleksdag** (0.616), **egenkärleken** (0.615), **Wikikärlek** (0.614), **Människokärlek** (0.613), **kärlekarna** (0.613), **villkorslösa** (0.613), **glädje** (0.612), **känslosamhet** (0.611), **känslor** (0.610), **ömsinthet** (0.609), **kärlekDe** (0.609), **kärleksfyllt** (0.608), **kärlekshyllning** (0.608), **omtanke** (0.607), **kärleksband** (0.607), **kärleks** (0.607), **kärleken.** (0.607), **kärleksarbete** (0.606), **Moderskärlek** (0.605), **kärleksfyllda** (0.605), **kärle** (0.604), **kärleksförklaringar** (0.604), **kärlekLäs** (0.604), **kärleksförhållandet** (0.604), **sorg** (0.603), **ledsamhet** (0.602)

Bland grannarna finns det som synes en hel del brus som **kärlek.En** och **kärlekDe**, som kommer sig av den tokenisering som Grave et al. (2018) använt när det skapade den svenska vektormodellen.

4 Nedslag i vektorrymden

Listorna med grannar ser i korta ordalag väldigt olika ut för de 20 utvalda orden. På grund av det begränsade utrymmet i detta sammanhang sammanfattar vi i det följande några av de typer av iakttagelser som man kan göra gällande de ord som studerats.

4.1 Etablerade ord

Här är ett par sammanfattande ord om grannarna till några av de etablerade ord som valts ut:

kärlek: Bland grannorden finns ord som betecknar olika slags kärlek, ex. *moderskärlek*, *syskonkärlek*, *tonårskärlek* och *hatkärlek*. Det finns också ord som betecknar andra relaterade känslor m.m. såsom *vänskap*, *förälskelse*, *längtan*, *passion*, *romantik*, *nyförälskelse*, *omtanke*, *ömhet*, *hjärtesorg*, *tvåsamhet*. Det finns också ett typiskt attribut: *villkorslös*.

adekvat: De 100 grannarna utgörs nästan uteslutande av andra adjektiv. De flesta är synonymer eller närsynonymer till *adekvat*, ex. *tillfredsställande*, *erforderlig*, *fullgod*, *ändamålsenlig*. I listan finns det också ord som kan betraktas som mer eller mindre antonyma, t.ex. *inadekvat*, *otillräcklig*, *otillfredsställande*. Inget av dessa ord upptas bland hänvisningarna i artikeln *adekvat* i SO (2021).

disputation: Huvuddelen av grannorden betecknar företeelser som hör till den forskarstuderandes vardag, t.ex. *doktorsavhandling*, *slutseminarium*, *spikning*, *disputationsfest*. Samtliga grannar utom en handfull utgörs av substantiv. Till undantagen hör det centrala verbet *disputera*.

röd: En klar majoritet av de 100 grannorden utgör mer eller mindre vanliga färgbeteckningar såsom *gul*, *buteljgrön* och *rödviolett*. Övriga ord slutar nästan samtliga på efterleden *-färgad*, t.ex. *laxfärgad*.

sjunga: I listan finns det såväl verb, t.ex. *nynna*, *joddla* och *gnola*, som substantiv. Flera av dessa senare betecknar olika slags sånger, bl.a. *luciasången*, *julvisor*, *psalmer*. De aktuella orden står således i olika relation till det undersökta verbet.

4.2 Nya ord

Här är några sammanfattande ord om grannarna till några av de nyord som valts ut:

gangsterrap: Bland grannarna finns dels sammansättningar med *gangster* som för- eller efterled, t.ex. *gangsterkung*, *smågangsters*, men också många ord som innehåller ordet *maffia*, t.ex. *maffiaorganisation*, *knarkmaffian*. Det finns också andra ord som förknippas med kriminalitet, t.ex. *droghandlare* men också ord som har med hiphop att göra. I listan dyker slutligen stavningsvarianten *gangstarap* upp.

glamping: Ordet står för 'glamorös camping'. I listan med grannar finns det en rad sammansättningar med *camping* vars förled visar på dagens mångfald när det gäller campingliv, t.ex. *husvagns-*, *fri-*, *fiske-*, *vildmarks-*, *vintersports-* och *naturist-*. Värt att notera är att det i listan med grannar inte finns några spår av komponenten 'glamour'.

mockumentär: På plats 5 bland grannarna finns en sammansättning, *fejkdokumentär*, som utgör en god omskrivning och viktig signal till lexikografen om vad ordet betyder. I listan med grannar återfinns för övrigt ett stort antal olika slags filmer. Några har dokumentärt innehåll, t.ex. *minidokumentär*, *dramadokumentär*, *naturdokumentär*, andra har mer fiktivt innehåll, t.ex. *zombiefilm*, *äventyrsfilm*.

snabbtesta: Bland grannarna finns dels substantivet som verbet är avlett ifrån (*snabbtest*), dels andra sammansättningar som slutar på *-testa*, t.ex. *hårdtesta*, *trycktesta*, *funktionstesta*, *stresstesta*, *betatesta*.

yes: Grannarna består av många engelsklingande ord, ex. *yeah*, *shit*, *kidding*, *well*, *say*, *indeed*. Några av dessa grannar, som t.ex. mer svenskklingande *jaaaaa*, *aaaah*, *tadaaaa*, *tjoho* och *niiiice*, kan nog i vissa sammanhang vara mer eller mindre synonyma till ordet ifråga. Värt att notera är också hur dessa och andra ord i listan stavas.

5 Diskussion

Studien visar att förhållandena mellan de 20 ord som granskas och deras respektive grannar varierar. Ett skäl till detta kan vara att de ord som studeras är av olika slag, vilket påverkar deras språkliga kontexter: de har olika ordklass, de har olika frekvens och spridning i svenska texter av idag, de hör till olika stilnivå etc.

Exempel är *adekvat* vars grannar nästan uteslutande är andra adjektiv. Ett annat exempel är *hund* vars grannord till synes enbart utgörs av substantiv. Detta är en fördel när lexikografen letar efter nya hänvisningar till SO, vilka återfinns under rubriker som Synonym, Antonym och Jämför. Tidigare studier har visat att SO kan stärkas när det gäller sådana kopplingar mellan artiklar (Blensenius et al., 2021). Ett ord som *sjunga* kan så förses med illustrativa hänvisningar till lemman som *nynna* och *gnola*. Ett nyord som *prosecco* kan kopplas till befintliga SO-ord som *champagne* och *cava*.

Vidare finner lexikografen gott om ord som kan tjäna som ytterligare morfologiska exempel till såväl redan beskrivna ord som nya sådana. I ett fall som *smittovåg* blir den variation som gäller fogen vid sammansättningar med *smitta* mycket tydlig.

Ordvektorer kan även bidra till information om andra uppslagsord än de här undersökta. I ett fall som *snabbtesta* står det klart att verbartikeln *testa* i SO saknar typiska sammansättningar som t.ex. *hårdtesta*, *stresstesta* och *betatesta*. Undersökningen av ordet *glamping* visar på hur artikeln *camping* i SO skulle behöva moderniseras.

Över huvud taget kan ordvektorer ringa in olika slags semantiska fält. Ett exempel är *kärlek* med grannar som betecknar olika typer av känslor, tillstånd m.m. Dessa är viktiga när man ska klargöra för ordet *kärleks* betydelse. I en ordbild till *kärlek* visas emellertid även andra relaterade ord och av olika ordklass, såsom adjektivet *obesvarad* och verbet *hysa*. Sådana ord är givetvis också mycket viktiga för lexikografen vid beskrivningen av inte minst kollokationer.

Grannorden kan också förtydliga samhälleliga förändringar som avspeglas i svenskans ordförråd. Exempel är grannarna till ordet *matresa* som visar på den mångfald av temaresor som arrangeras.

Kanske kan grannorden också säga något om det aktuella ordets värdeladdning och konnotationer. Exempel är *åldras* där det är slående hur många negativt laddade ord det finns bland grannorden. Samtidigt kan grannorden ha viss slagsida åt någon av betydelsekomponenterna hos det ord som undersöks. Se t.ex. *glamping*. I listan med grannar finns det som sagt en rad kohyponymer som betecknar olika campingvarianter, men inget spår av komponenten 'glamour'.

När det gäller lånord som *yes*, så ser vi även en hög grad av engelska ord bland grannarna, vilket visar att det finns en hel del engelska texter i det material som använts för att bygga upp vektormodellen.

Slutligen finner man här och var nya goda nyordskandidater, t.ex. *efterforskningsförbud* som är granne till *kontaktförbud*.

6 Slutsats

Ordvektorer ger oss de ord vars språkliga kontexter liknar varandra, vilket kompletterar vad vi får av konkordanser och ordbilder, som ger en överblick av specifika ords kontexter. Ordvektorer kan uppmärksamma lexikografen på bl.a. ord som kan tjäna som morfologiska språkprov och som hänvisningar till relaterade artiklar. De kan således bidra till att det semantiska nätverk som exempelvis SO redan utgör stärks ytterligare. Vidare kan de hjälpa lexikografen att, på ett förhållandevis objektivt och datadrivet sätt, se kopplingar mellan befintliga uppslagsord i ordboken och sådana som ska läggas till i samband med en revidering.

Efterord

Vi vill tacka[2] Lars Borin för att han är, och alltid varit, en sådan positiv kraft i våra professionella yrkesliv och i den verksamhet som vi delar.

Referenser

B.T. Sue Atkins & Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford. 2010.

Kristian Blensenius, Emma Sköldberg, & Erik Bäckerud. 2021. Finding gaps in semantic descriptions. visualisation of the cross-reference network in a swedish monolingual. In *Proceedings of the eLex 2021 conference*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, & Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Lars Borin, Markus Forsberg, & Johan Roxendal. 2012. Korp the corpus infrastructure of språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, volume Accepted, pages 474–478.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, & Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Nicolai Hartvig Sørensen & Sanni Nimb. 2018. Word2dict lemma selection and dictionary editing assisted by word embeddings. In Iztok Kosem Jaka ibej, Vojko Gorjanc & Simon Krek, editors, *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 819–826, Ljubljana, Slovenia, jul. Ljubljana University Press, Faculty of Arts.

[2] This work has been supported by Nationella språkbanken and HUMINFRA, both funded by the Swedish Research Council (20182024, contract 2017-00626; 20222024, contract 2021-00176) and their participating partner institutions.

The diachrony of political terror: Tracing terror and terrorism in Swedish parliamentary data 1867-1970

Mats Fridlund, Daniel Brodén, Victor Wåhlstrand Skärström
Centre for Digital Humanities
University of Gothenburg, Sweden
name.surname1.surname2@gu.se

Abstract

The paper explores the development of the closely related words 'terror' and 'terrorism' as manifested in the discourse of the Swedish Parliament, 1867–1970, drawing on digital history and language technology methodologies and tools. Combining distant and close reading, we show that terror-related words first gained traction from 1918 and onwards. The recorded uses of words and compounds indicate that terror-related phenomena were often associated with states rather than individuals, but also that terror-related words have been used metaphorically in relation to non-violent domestic issues. Our results confirm the argument that the word terrorism primarily gained its modern meaning in the early 1970s. We conclude by stressing the potential of combining LT-driven and interpretative approaches for investigating the diachronicity of words in Parliamentary corpora.

1 Introduction

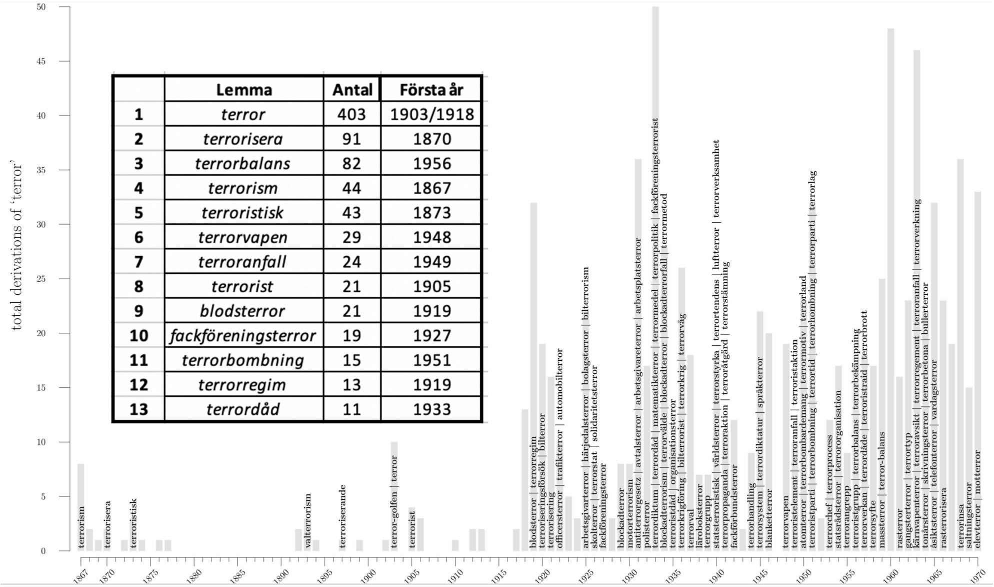
This study provides a step towards a digital history of the Swedish political discourse on political terror by means of distant and close reading of parliamentary texts. Drawing on a mixed-methods approach, we explore the development of the closely related words 'terror' and 'terrorism' as manifested in the discourse of the bicameral Parliament throughout its existence, 1867–1970.

From an etymological perspective, terror (same spelling in Swedish and English) designates an intense state of fear or horror and has been used in Swedish written accounts since at least the 1600s. *Terrorisera* ('to terrorize') means to put people in such a state through one's actions. *Terror* was rarely used before 1918 when it gained a second meaning, signifying the use of certain means *'i politiskt syfte för att sprida skräck o. därvid uppnå vissa mål'* by both state and sub-state actors (saob.se). Likely, this new use derived from English or Russian and was connected to the 1917 Russian Revolution and Finnish Civil War 1918. *Terrorism* had entered Swedish in the early 19th century, primarily referencing the French Revolution's Reign of Terror (initially translated as *skräckväldet*) and was similarly associated with state repression. In the 1970s, *terrorism* gained its modern meaning, becoming distinctly associated with sub-state violence against civilians or non-combatants (Stampnitzky, 2013).

The aim of this study is to explore how the words *terror* and *terrorism* have been used in Swedish parliamentary discourse, focusing on the different meanings that have been ascribed to them over time. To map these meanings we draw on the digitized parliamentary records and the resources of Språkbanken Text. Partly, we follow prior initiatives by the Swedish CLARIN node, Swe-Clarin, where humanities and social sciences scholars collaborate with researchers in natural language processing in using LT-based e-science tools for HSS research (Karsvall & Borin, 2018; Viklund & Borin, 2016). The present study builds on prior explorations together with Lars Borin of the newspaper discourse on terrorism in Sweden and Finland (Fridlund et al., 2019; Fridlund et al., 2020; Fridlund et al., 2022).

Specifically, we ask two research questions regarding the understanding of the phenomenon of terrorism in the Swedish parliamentary discourse during the period in focus: (1) What variations of meanings

Mats Fridlund, Daniel Brodén and Victor Wåhlstrand Skärström. 2022. The diachrony of political terror: Tracing terror and terrorism in Swedish parliamentary data 1867-1970. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 43–47. Available under CC BY 4.0



	Lemma	Antal	Första år
1	*terror*	403	1903/1918
2	*terrorisera*	91	1870
3	*terrorbalans*	82	1956
4	*terrorism*	44	1867
5	*terroristisk*	43	1873
6	*terrorvapen*	29	1948
7	*terroranfall*	24	1949
8	*terrorist*	21	1905
9	*blodsterror*	21	1919
10	*fackföreningsterror*	19	1927
11	*terrorbombning*	15	1951
12	*terrorregim*	13	1919
13	*terrordåd*	11	1933

Figure 1: The diachrony of political terror in the Swedish Parliament 1867–1970. First occurrences of new derivations of terror in the bicameral corpus and staples showing the use of all derivations. The top left table shows the terror lexemes with more than 10 occurrences.

have the words *terror* and *terrorism* had when used in isolation and as compound words, and (2) what terror-related compounds have been added to the discourse over time?

2 Analyzing the bicameral Riksdag parliamentary data

The parliament's texts available at the National Library of Sweden (riksdagstryck.kb.se) (retrieved 2021-11-11) consist of 10 categories, such as motions, propositions and minutes of the debates. The material was processed for analysis with tokenization, lemmatization and dependency parsing by means of the Sparv Pipeline tool designed for automatic neural and statistical annotation of documents with textual structure and linguistic properties for Swedish applications (Borin et al., 2016; Ljunglöf et al., 2019; Hengchen & Tahmasebi, 2021). Our processed corpus was subsequently made accessible through the qualiquantitative Context tool (developed by Wåhlstrand Skärström) for investigating linguistic representations, i.e., words and phrases, in the texts. Context aids distant reading and enables the production of quantitative results, such as relative and absolute frequency for queries and close reading of the context of a search query, which enables qualitative and interpretative analysis.

For the analysis, the parliamentary texts were grouped by year, irrespective of their subcollection, and tokenized into individual words and subsequently lemmatized and queried for the head word, e.g *terror*, per year. This filtered volume was then manually curated to remove errors from OCR or lemmatization. This produced data on a 'diachrony' of terror-related words – the frequency of usage of terror-related lexemes (staples in Figure 1) and innovation of new words – indicating the yearly growth of terror-related terms and the production of compounds. In the analysis, significant individual occurrences (generally the first) were interrogated by closer readings of the parliamentary records.

2.1 Volume of roots and compounds

The search string **terror** generates 1.016 hits in our corpus. The majority (606) are different grammatical forms of the seven terms *terror* (404), *terrorisera* (81), *terrorism* (44), *terroristisk* (43), *terrorist* (21), *terroriserande* (10) and *terrorisering* (3), all of which may be considered derivations from the root word *terror*. However, since they are all separately productive, we will consider them roots by their own right The rest consists of 102 compounds from the root *terror* (82 different compounds), *terrorist* (12), *terrorism* (5), *terrorisera* (1), *terrorisering* (1) and *terroristisk* (1), either as the modifier or head constituent.

Focusing on the two core roots, *terrorism* was used already 1867, the bicameral parliament's first year. However, close reading shows its 19th century use to be nonlethal metaphorical, primarily to denote *valterrorism*, perceived oppressive parliamentary voting procedures. In the 1900s *terrorism* becomes used for violent and even lethal activities. First in connection with labor disputes and later in 1918 with the Finnish Civil War, that also introduced *terror* as a significant concern for the Swedish Parliament. Usage and new compositions rose sharply from 1918 (from 53 to 542 1918–1970). Notably, *terror* had only been used on one occasion in 1903 to refer to the British warship HMS *Terror*. Furthermore, it is striking that although the use of *terrorism* preceded that of *terror* it is almost absent before 1970 (26 hits 1900–1969) something that we discuss further below.

Regarding most common usage, 13 words occurred 10 times or more (see table in Figure 1): 5 simple terms and the 8 compound words *terrorbalans* ('terror balance'), *terrorvapen* ('weapons of terror'), *terroranfall* ('terror attacks'), *blodsterror* ('blood terror'), *fackföreningsterror* ('trade union terror'), *terrorbombning* ('terror bombing'), *terrorregim* ('terror regime') and *terrordåd* ('terror deed'). Notably, four of these – *terrorbalans*, *terrorvapen, terroranfall* and *terrorbombning* – are used in reference to Cold War nuclear terror. Also, close reading reveals the first occurrence of *terrordåd* (in 1933) to refer to *individuella terrordåd* ('individual acts of terror'), i.e. political violence that today would be discussed in terms of *terrorism.*

2.2 Productivity of compounds

Our results show six roots and one compound (*valterrorism*) emerging before 1918 and one root (*terrorisering*) and 102 compounds after 1918. As far as compound words with *terror* are concerned, of the 83 in total, eight have more than 10 instances (Figure 1): *terrorbalans* ('-balance'), *terrorvapen* ('-weapon'); *terroranfall* ('-attack'); *blodsterror* ('blood-'); *fackföreningsterror* ('trade union-'); *terrorbombning* ('-bombing); *terrorregim* ('-regime'), *terrordåd* ('-deed). Furthermore, in line with the rare use of *terrorism* discussed above, there are only six uses of the four (methaphorical) compounds with *-terrorism* (*val-, bil-, motor-, blockad-*) as compared to those with *terror* and *terrorist* (12).

States' involvement in terror activities are referred to in several of the terror compounds such as *terrorregim* (13), *terrorvälde* (9), *polisterror* (6), *terrorland* (2), *terrordiktatur* (1) and *terrorregement* (1), as well as other that refer to warfare involving states, such as *terrorkrig* and *terrorbombning*. Notably, *statsterroristisk* (1) and *terrorstat* (3), were the only compounds with *stat* ('state'), although the former was metaphorically used to refer to domestic governance issues, which was also the case with *statsrådsterror* (1) (c.f. Ängsal et al., 2022, on state and terrorism compounds in the Swedish parliamentary debate 1993–2018). Thus, our results show the associations between states and terror to have a long Swedish history (c.f. Fridlund et al., 2020).

One can also distinguish periods of compound productivity grounded in domestic and geopolitical trends contexts, such as *arbetsmarknadsterror* in 1925–1935 referring to terror by and against labour unions and employers, *luftterror* in 1936–1940 denoting the threat of wartime aerial bombings against civilian targets, and *atomterror* in 1948–1963 denoting the nuclear threat during the Cold War.

2.3 The rise of terrorism

What is missing from the above discussion is any references to words related to the insurgent violence perpetrated by militant organizations such as the Popular Front for the Liberation of Palestine (PFLP) and the West German Red Army Faction (RAF) that became synonymous with terrorism in the early 1970s (prior to this, other words were sometimes used as labels for similar forms of political violence, including *anarkism* ('anarchism')). In fact, Stampnitzky argues that the transnational character of this form of violence – skyjacking, hostage taking for political purposes, etc. – and its impact on the modern world order generated a need for a discursive term: 'the concern was with violence *out of place* – spilling over from local conflicts into the international sphere' (Stampnitzky, 2013, p.27).

The Swedish parliamentary debate provides a clear illustration of this argument when an MP in 1970 claimed that the use of *'terror och motterror'* ('counter terror') could have dire consequences: *'När terrorgrupper tillgriper sådana metoder som kapning eller rent av förstörelse av flygplan med oskyldiga civila passagerare, eller mord på diplomater från utomstående länder, hotas hela det regelsystem för den internationella samlevnaden, som mödosamt byggts upp under lång tid.* '(1970–04–29)' The pejorative quality of the word terrorism also made it useful as a rhetorical tool. Later that year, a liberal MP sarcastically commented on the New Left's advocacy of using violent means for political purposes and in doing this introduced in the Parliament terrorism in its new emerging meaning: *'När den nygamla vänstern vill försvara våldsmetoder får det inte sammanblandas med advokatyr för terrorism. Det skulle låta alltför illa!'* (1970–10–29).

3 Conclusion

This paper provides an attempt to understand the development of the closely related words *terror* and *terrorism* as manifested in the Swedish parliamentary discourse, 1867–1970. By applying the tools Sparv and Context, we have explored the development of these two words in isolation and as parts of compounds in parliamentary texts. Combining distant and close reading, we have shown that terror-related words gained traction from 1918 and onwards. Furthermore, the uses of the words of interest and their compounds clearly indicate that terror-related activities were, to a large extent, associated with states rather than individuals. At the same time, our results confirm the familiar argument that the word terrorism gained its modern meaning in the early 1970s. On another level, the paper illustrates the potential of combining LT-driven and interpretative approaches for analysing the diachronicity of words in Parliamentary corpora.

Acknowledgments

This study is part of *The Cultural Imaginary of Terrorism; Terrorism in Swedish Politics;* and *Things for Living with Terror* funded respectively by Marcus and Amalia Wallenberg Foundation; VR, RJ and Vitterhetsakademien; and RJ. It is also supported by Swe-Clarin funded partly by VR (contract no. 2017-00626).

References

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, & Anne Schumacher. 2016. Sparv: Språkbankens corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18.

Mats Fridlund, Leif-Jöran Olsson, Daniel Brodén, & Lars Borin. 2019. Trawling for terrorists: A big data analysis of conceptual meanings and contexts in Swedish newspapers, 1780–1926. In *HistoInformatics 2019: Proceedings of the 5th International Workshop on Computational History; co-located with the 23rd International Conference on Theory and Practice of Digital Libraries (TPDL 2019). Oslo*, pages 30–39. CEUR-WS.

Mats Fridlund, Leif-Jöran Olsson, Daniel Brodén, & Lars Borin. 2020. Trawling the Gulf of Bothnia of news: a big data analysis of the emergence of terrorism in Swedish and Finnish newspapers, 1780–1926. In *Proceedings of CLARIN Annual Conference*, pages 61–65.

Mats Fridlund, Daniel Brodén, T. Jauhiainen, L. Malkki, Leif-Jöran Olsson, & Lars Borin. 2022. Trawling and trolling for terrorists in the digital Gulf of Bothnia: Cross-lingual text mining for the emergence of terrorism in Swedish and Finnish newspapers, 1780–1926. In *CLARIN: The Infrastructure for Language Resources*, pages 781–802. Berlin: De Gruyter.

Simon Hengchen & Nina Tahmasebi. 2021. A collection of Swedish diachronic word embedding models trained on historical newspaper data. *Journal of Open Humanities Data*, 7:1–7.

Olof Karsvall & Lars Borin. 2018. SDHK meets NER: Linking place names with medieval charters and historical maps. In *Proceedings of DHN 2018. Aachen. CEUR-ws.org*, pages 38–50.

Peter Ljunglöf, Niklas Zechner, Luis Nieto Piña, Yvonne Adesam, & Lars Borin. 2019. Assessing the quality of Språkbankens annotations. Technical report, University of Gothenburg, Department of Swedish.

Lisa Stampnitzky. 2013. *Disciplining terror: How experts invented 'terrorism'*. Cambridge University Press.

Jon Viklund & Lars Borin. 2016. How can big data help us study rhetorical history? In *Selected Papers from the CLARIN Annual Conference 2015*, pages 79–93. Linköping University Electronic Press.

Magnus P Ängsal, Daniel Brodén, Mats Fridlund, Leif-Jöran Olsson, & Patrik Öhberg. 2022. Linguistic framing of political terror: Distant and close readings of the discourse on terrorism in the Swedish parliament 1993–2018. In *Proceedings of CLARIN Annual Conference 2022, Prague*, pages 69–72.

UD-based Latvian FrameNet

Normunds Grūzītis
IMCS, University of Latvia
normunds@ailab.lv

Gunta Nešpore-Bērzkalne
IMCS, University of Latvia
gunta@ailab.lv

Baiba Saulīte
IMCS, University of Latvia
baiba@ailab.lv

Abstract

We, students of Lars Borin from the good old NGSLT times, present Latvian FrameNet. This is a part of a larger work on the creation of a balanced multilayered corpus of Latvian, anchored in cross-lingual state-of-the-art syntactic and semantic representations: Universal Dependencies (UD), FrameNet and PropBank, as well as Abstract Meaning Representation. We have been inspired a lot by the Swedish FrameNet++ (SweFN++), yet there are some differences: we stick to the frame inventory of Berkeley FrameNet, and the FrameNet annotation layer is added on top of a manually curated UD layer. Thus, the annotation of frames, frame elements (FE), and FE spans is guided by the dependency structure of a sentence. We strictly follow a corpus-driven approach – lexical units (LU) in Latvian FrameNet are created only based on the annotated corpus examples. Therefore, in contrast to SweFN++, Latvian FrameNet is definitely not the largest one in terms of LUs, but, to our knowledge, it is the first FrameNet-annotated corpus that has been created as an extension of an UD treebank.

1 Introduction

In the industry-oriented research project "Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian",[1] we have created a balanced text corpus with multilayered annotations (Gruzitis et al., 2018), adopting widely acknowledged and cross-lingually applicable representations: Universal Dependencies (UD) (Nivre et al., 2016), FrameNet (Fillmore et al., 2003), PropBank (Palmer et al., 2005) and Abstract Meaning Representation (AMR) (Banarescu et al., 2013).

The UD representation is automatically derived from a more elaborated manually annotated hybrid dependency-constituency representation (Pretkalnina et al., 2018). The FrameNet annotations are manually added, guided by the underlying UD annotations (see Figure 1). Consequently, frame elements (FE) are represented by the root nodes of the respective subtrees instead of text spans; the spans can be easily calculated from the subtrees. The PropBank layer is automatically derived from the FrameNet and UD annotations (Gruzitis et al., 2020), provided a manual mapping from lexical units (LU) in FrameNet to PropBank frames, and a mapping from FrameNet FEs to PropBank semantic roles for the given pair of FrameNet and PropBank frames. Draft AMR graphs are derived from the UD and PropBank layers, as well as auxiliary layers containing named entity and coreference annotation, with the potential to seamlessly integrate the FrameNet frames and FEs into the AMR graphs. The semantically richer FrameNet annotations (compared to PropBank) are also helpful in acquiring more accurate draft AMR graphs, even if FrameNet itself stays behind the scenes.

The inspiration to create an integrated multilayer corpus comes from the OntoNotes corpus (Hovy et al., 2006) and the Groningen Meaning Bank (GMB) (Bos et al., 2017). The overall difference from

[1] Thank you Lars for your letter of support!

Normunds Grūzītis, Gunta Nešpore-Bērzkalne and Baiba Saulīte. 2022. UD-based Latvian FrameNet. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 49–53. Available under CC BY 4.0

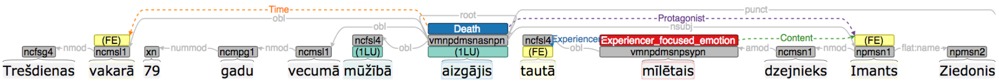


Figure 1: FrameNet annotation on top of a UD tree. Only head nodes are selected while annotating FEs. Literal translation of the sentence: *"On Wednesday evening at age 79, passed away the nation's beloved poet Imants Ziedonis"*. The FE spans can be acquired automatically by traversing the respective subtrees: [*tredienas vakar*]$_{\text{Time}}$, [*taut*]$_{\text{Experiencer}}$, [*taut mltais dzejnieks Imants Ziedonis*]$_{\text{Protagonist}}$. Multi-word LUs are indicated by generic LU tags: *mb aizgjis*$_{\text{DEATH}}$ versus *mltais*$_{\text{EXPERIENCER\_FOCUSED\_EMOTION}}$.

the OntoNotes approach is that we use the UD model at the treebank layer, and we annotate FrameNet frames in addition to the PropBank frames. In fact, FrameNet is the primary frame-semantic representation in our approach. Another difference is that we aim at whole-sentence semantic annotation at the ultimate AMR layer. This in some sense is similar to the goal of GMB, but GMB uses Discourse Representation Theory instead of AMR. For pragmatic reasons, we use the more shallow and more lossy AMR formalism. Our experience developing semantic parsers and multilingual text generators, by combining machine learning and grammar engineering (Gruzitis et al., 2017; Gruzitis & Dannélls, 2017; Borin et al., 2018), has convinced us that FrameNet and AMR both have a great potential to establish as powerful and complementary semantic interlinguas which can be furthermore strengthened and complemented by other multilingual frameworks, like Grammatical Framework (Ranta, 2011).

It should be noted that there has been previous work on a domain-specific Latvian FrameNet for a real life media monitoring use case, focusing on 26 modified Berkeley FrameNet (BFN) frames (Barzdins et al., 2014). The current work, however, aims at a balanced general-purpose BFN-compliant framenet that will cover many frequently used frames and LUs.

2 The corpus

We are aiming at a medium-sized treebank/framebank – around 20,000 sentences annotated at all the layers mentioned in Section 1. Therefore it is crucially important to ensure that the multilayer corpus is balanced not only in terms of text genres and writing styles but also in terms of LUs.

Our fundamental design decision is that the text unit is an isolated paragraph. The corpus therefore consists of manually selected paragraphs from many different texts of various types. Representative paragraphs are selected in different proportions from a balanced 10-million-word text corpus.

As for the LUs, our goal is to cover at least 1,500 most frequently occurring verbs and deverbal nouns, calculated from the 10-million-word corpus. Since the most frequent verbs tend to be also the most polysemous, we expect that the number of LUs will be considerably larger – at least 3,000 LUs. We are aiming at least 10 annotation sets per LU on average.

Paragraphs to be annotated are selected based on target words they contain, not randomly, and curators are constantly updated on the current balance or imbalance of the corpus w.r.t. text genres and target word frequencies.

Currently, we have acquired more than 20,000 annotation sets, covering more than 500 BFN frames evoked by more than 2,500 LUs.[2]

3 The FrameNet annotation process

Paragraphs for which the manual treebank annotation is finalized and which have been successfully converted to the UD representation are considered for the FrameNet annotation. Unfinished paragraphs are ignored till next iteration, since their sentence split, tokenization, as well as tree structure can still considerably change. Changes in the tree structure are not a major issue, and the FrameNet annotation process actually helps to spot and eliminate many inconsistencies in the underlying trees. The sentence splitting and tokenization, however, is a major requirement to later avoid issues in merging the different annotation layers.

[2] `https://github.com/LUMII-AILab/FullStack`

3.1 The concordance approach

While treebank, named entity and coreference annotations are done paragraph by paragraph and sentence by sentence, we do not find this being a productive workflow for annotating semantic frames, especially in case of the highly abstract FrameNet frames. Instead, we prefer a concordance view, so that the linguist can focus on a target word and its different senses, without constantly switching among different sets of frames. This improves the annotation consistency.

When more paragraphs are finalized at the UD layer, they are included in the next concordance queries. The first concordance is processed when there are at least three example sentences available for the target word. The next concordances are collected and processed on considerable milestones at the UD layer. The annotated concordances from the first rounds serve as guidelines when annotating the next rounds, thus, further improving consistency.

A consequence of such approach is that no full-text annotation is intentionally done, although many sentences might become close-to-fully annotated after merging annotations of the same sentence from different concordances.

3.2 The UD-based annotation

The UD-based approach has a significant consequence: FEs are not annotated as spans of text – annotators select only the head word (node) when annotating an FE. The whole span can be easily calculated automatically by traversing the respective UD subtree. These calculations are not included as part of the data set.

Such approach not only makes the annotation process more simple and the annotations more consistent, but it also facilitates the training of an automatic semantic role labeler, since it is easier to identify the syntactic head of an FE than a span of a string. Still, most FrameNet corpora are annotated in terms of spans, relying on syntactic parsing as a post-processing step.

3.3 Important notes on frame elements

Yet another important decision regarding FEs is to currently focus only on the core elements according to BFN. We have made this decision because of the limited resources. However, we do annotate two non-core elements systematically: *Time* and *Place* (as illustrated in Figure 1). In various information extraction use cases (e.g. for media monitoring), these two non-core FEs are important. Other non-core elements are annotated occasionally, if they are rather specific to the frame (e.g. non-core indirect objects and specific adverbial modifiers).

Regarding null instantiations (NI), we do not annotate missing FEs in the sentence. This is out of the scope of the current project, but the annotation of NI should to be considered in a follow-up research: (i) since the FrameNet annotation is relaying on UD, it is an open question how to handle NI – where to attach these annotations; (ii) since Latvian is a highly inflected language, the grammatical subject and object can be omitted in a sentence, to some extent, compensating it with the respective form of the verb; (iii) in general, it would require Latvian-specific guidelines, but the theoretical foundations are not mature yet for Latvian; it would require more elaborate linguistic research, based on the basic annotated data acquired in the current project; (iv) although NI is highly relevant for lexicographic research, it is not a priority for many practical use cases that require semantic parsing.

3.4 Multi-word lexical units

To deal with target words as as multi-word units, we have introduced an auxiliary annotation layer for multi-word LUs (as illustrated in Figure 1). The head word is still a verb or deverbal noun that evokes a frame, but the other key constituents are indicated as well. Again, note that these constituents may be root nodes of some subtrees – we do not annotate the whole spans.

This auxiliary layer is not an ultimate solution to deal with constructions, but for now it allows us to register such cases and to retrieve them later for more elaborated analysis. Usually these are partially grammaticalized constructions or even idioms that, as a whole, evoke the respective frames. If we would consider these verbs in isolation, they would rather evoke different frames.

3.5 Cross-lingual issues

In order to ensure compliance with BFN and, thus, to maximize the cross-lingual applicability of Latvian FrameNet, we are strictly sticking to the BFN frame inventory. We avoid defining any Latvian-specific frames. Therefore it is sometimes difficult to select an appropriate BFN frame for a particular sense of a Latvian verb. It usually happens when:

1. The sense of a Latvian verb is more specific compared to the closest English verb sense or compared to the definition of the closest BFN frame. For instance, for the verb *prdomt* ‘to change one’s mind’ or ‘to rethink’, we do not have a solution yet, since BFN frames related to thinking (*Opinion*, *Cogitation*) do not fit this verb sense, and neither does the general *Cause_change* frame. Similarly, we have not found a good mapping for *maldties* ‘to be wrong’ and *saemties* ‘to pull oneself together’.

2. The sense of a Latvian verb is more general compared to the closest English verb sense: the sense of an English verb is expressed in Latvian by a phrase (typically, by a verb and a direct object). Examples: *last lekciju* ‘to lecture’ (‘to give a lecture’), *krist bon* ‘to faint’ (‘to fall into unconsciousness’), *zaudt samau* ‘to faint’ (‘to lose consciousness’).

3. The semantic elements are different between the Latvian and English verb senses. For instance, *braukt* ‘to move using a vehicle’: the sense of the Latvian verb does not specify whether the person is a driver or a passenger (e.g. *es braucu uz darbu* ‘I go to work (by a transport)’ – it is unclear what is the role of the person, and which frame is evoked – *Ride_vehicle* or *Operate_vehicle*. In this particular case, we use the frame *Use_vehicle* which is a non-lexical frame in English.

There are some options for how to deal with these issues: (i) by treating more verb phrases in Latvian as if they were multi-word LUs, even if lexicographers would argue about that; (ii) by using a more general BFN frame if possible, i.e., if the direct object of the target verb can be annotated as a core FE (e.g., it would work for ‘to lose consciousness’ but not for ‘to give a lecture’); (iii) some frames are just missing in BFN, and a global solution would be needed on how to propose and confirm new frames in the BFN frame hierarchy; most likely in the scope of the Multilingual FrameNet initiative (Gilardi & Baker, 2018).

4 Conclusion

Creating the Latvian FrameNet, we strictly follow a corpus-driven approach: no LUs are introduced without annotated examples, i.e., we create no LUs based on lexicographic intuition or a common-sense dictionary; only based on corpus evidence. An initial experiment on bootstrapping LUs without corpus evidence did not prove to be productive: many of those hypotheses are not confirmed by our corpus (at least for now), and vice versa – many LUs were missing.

The consecutive treebank and framebank annotation workflow has turned out very productive and mutually beneficial. The dependency tree facilitates the annotation of semantic frames and roles, while the frame semantic analysis of the verb valency often unveils various inconsistencies and bugs in the dependency or morphological annotation.

Acknowledgements

This work has received financial support from the European Regional Development Fund under the grant agreement No. 1.1.1.1/16/A/219 and is continued within the State Research Programme under the grant agreement No. VPP-LETONIKA-2021/1-0006.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, & Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia.

Guntis Barzdins, Didzis Gosko, Laura Rituma, & Peteris Paikens. 2014. Using C5.0 and exhaustive search for boosting frame-semantic parsing accuracy. In *Proceedings of the 9th LREC Conference*, pages 4476–4482.

Lars Borin, Dana Dannélls, & Normunds Gruzitis, 2018. *Linguistics vs. language technology in constructicon building and use*, pages 229–254. John Benjamins.

Johan Bos, Valerio Basile, Kilian Evang, Noortje Venhuizen, & Johannes Bjerva. 2017. The Groningen Meaning Bank. In Nancy Ide & James Pustejovsky, editors, *Handbook of Linguistic Annotation*, volume 2, pages 463–496. Springer.

Charles J. Fillmore, Christopher R. Johnson, & Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Luca Gilardi & Collin Baker. 2018. Learning to Align across Languages: Toward Multilingual FrameNet. In *International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons*, Miyazaki.

Normunds Gruzitis & Dana Dannélls. 2017. A multilingual FrameNet-based grammar and lexicon for Controlled Natural Language. *Language Resources and Evaluation*, 51(1):37–66.

Normunds Gruzitis, Didzis Gosko, & Guntis Barzdins. 2017. RIGOTRIO at SemEval-2017 Task 9: Combining machine learning and grammar engineering for AMR parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 924–928.

Normunds Gruzitis, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, & Peteris Paikens. 2018. Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th LREC Conference*.

N. Gruzitis, R. Dargis, L. Rituma, G. Nespore-Berzkalne, & B. Saulite. 2020. Deriving a propbank corpus from parallel framenet and ud corpora. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 63–69.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, & Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, & Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th LREC Conference*, pages 1659–1666.

Martha Palmer, Daniel Gildea, & Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

L. Pretkalnina, L. Rituma, & B. Saulite. 2018. Deriving enhanced universal dependencies from a hybrid dependency-constituency treebank. In *Text, Speech, and Dialogue*, volume 11107, pages 95–105. Springer.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

The rise and fall of grammatical theories in descriptive grammars of the languages of the world

Harald Hammarström
Department of Linguistics and Philology
University of Uppsala, Sweden
harald.hammarstrom@lingfil.uu.se

Abstract

The present study traces the popularity of various grammatical theories in descriptive grammars of the languages of the world. Using the DReaM corpus of grammatical descriptions, we may simply look for the existence of simple terms relating to various descriptive theories across time. Even such a relatively elementary investigation does show that there is an, over time, increasing explicit interest in theory but specific theories "come and go" relatively quickly.

1 Introduction

The present study traces the popularity of various grammatical theories through time in descriptive grammars of the languages of the world. As such it intersects various interests of Lars Borin, such as culturomics (Tahmasebi et al., 2015), grammar formalisms (Borin & Saxena, 2004), linguistic typology (Virk et al., 2017) and corpus linguistics (Borin et al., 2012).

Writing a descriptive grammar involves the analysis of primary linguistic data (Karlsson, 2005; Chelliah & de Reuse, 2011, 7-24). As has been repeatedly pointed out, *some* theory is necessary for this task (Rice, 2006; Dryer, 2006) but theories can be more or less explicit and more or less restrictive. Thanks to the appearance of the DReaM Corpus (Virk et al., 2020) — a collection of digitized grammatical descriptions of the languages of the world — we are now able to shed light on the use of grammatical theories in descriptive work through time. Apart from qualitative remarks in passing (Chelliah & de Reuse, 2011; Sakel & Everett, 2012, 152-158) these trajectories have not been compared in previous work.

For the experiments in the present study, we have searched the archive of over 10,911 grammatical descriptions (grammars and grammar sketches) in the DReaM corpus spanning languages all over the world from 1250 AD to the present (Virk et al., 2020). Even if not explicitly mentioned, the searches have been done inclusive of synonyms, spelling variants, morphological variants and OCR errors (Hammarström et al., 2017).

2 Experiments

The first question is to what extent grammars are explicit about theory at all. For this we may simply search for the term 'theory', and its equivalents in a few other European language across grammars. For example, the term 'theory' occurs no less than 308 times in Saxon (1986)'s English dissertation on Dogrib [dgr], 'théorie' occurs 5 times in Alexandre (1966)'s French description of Bulu [bum] and 0 times in von Hagen (1914)'s German description of the same language. Figure 1 shows the proportion of grammars in which the theory-word occurs at least once through the timespan 1850-2010. As can be seen, there is a trend towards explicit mentions of theory, from approximately 20% in 1850 to over 50% at present (with few appreciable differences between meta-languages).

To gauge the popularity of different specific descriptive frameworks we searched through the English-subset (7,816 grammars) for a (non-exhaustive) selection of influential theories. Figure 2 shows the proportion of grammars which mention a given theory at least once through the timespan 1900-2010.

Harald Hammarström. 2022. The rise and fall of grammatical theories in descriptive grammars of the languages of the world. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 55–59. Available under CC BY 4.0

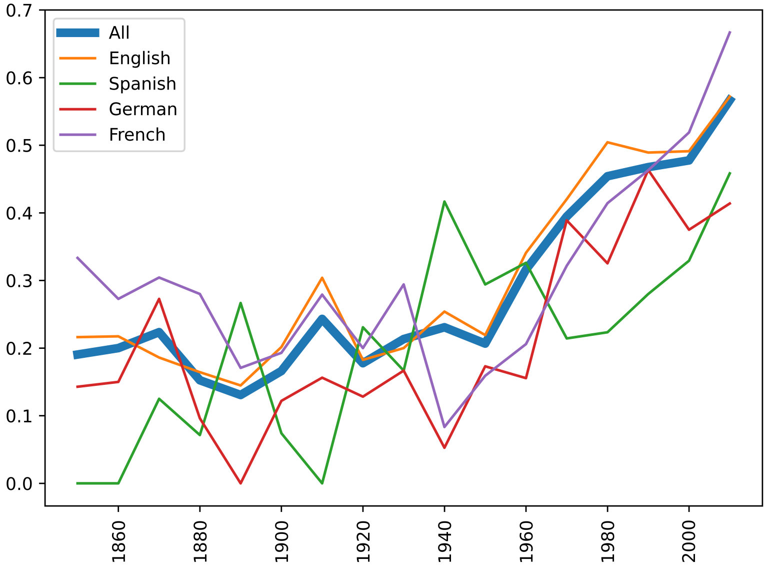


Figure 1: The proportion of grammars in which the term ‘theory’, or its equivalent in other languages, occurs at least once through the timespan 1850-2010. The data has been binned into 10-year intervals.

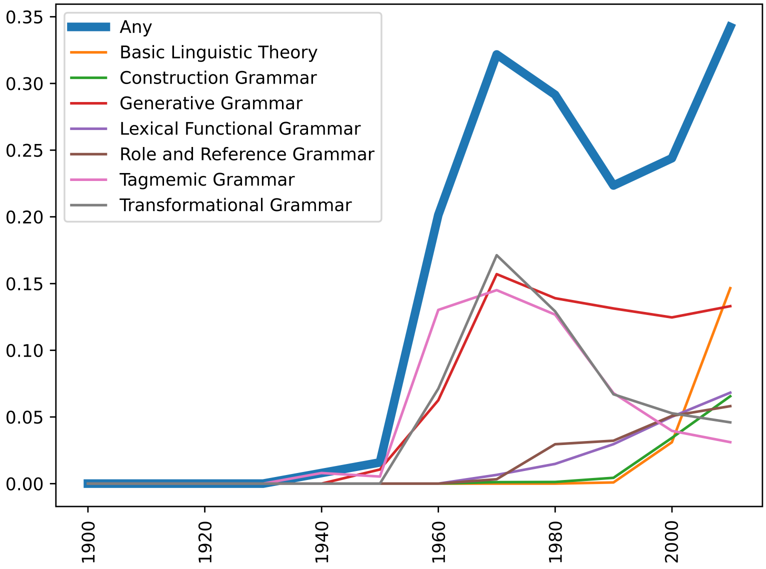


Figure 2: The proportion of grammars which mention a given theory (or an equivalent name or spelling variant) at least once through the timespan 1900-2010. The data has been binned into 10-year intervals.

The term tagmem(ic) — originating from Bloomfield (1935) — has a few mentions before, but the first grammar sketch written in the Tagmemic framework developed by Pike (1954–1960) is Duff (1959)'s sketch of Yanesha' [ame], cf. Waterhouse (1974, 90-92). A range of Tagmemic grammars produced by members of the Summer Institute of Linguistics followed, peaking in production around 1970 — when approximately 1/8 grammars mention Tagmemics — but has since faded in popularity.

An almost identical rise-and-fall curve is manifested in Transformational grammars, associated with Harris (1957) and Chomsky (1957), whose first grammar sketch appears to be Apte (1962)'s sketch of Marathi [mar]. The first explicitly Generative grammar is Sleator (1957)'s description of English [eng] of Jackson County, Indiana followed by further Indiana University dissertations. In the beginning Transformational is almost synonymous with Generative. Although not logically necessary, there is in fact near-full empirical overlap between these terms. However, after the 1970s, the Generative umbrella continues with other exponents. At its peak, 1/6 grammars were Generative-Transformational.

Lexical-Functional Grammar (LFG) was developed in the 1970s (Dalrymple et al., 2019, 1-2) but it is not until Davies (1981)'s grammar of Choctaw [cho] that there is a grammar written in this framework.

Around the same time is Role and Reference Grammar (RRG) whose first witness is a dissertation on Lakota [lkt] by an author who co-outlined the theory itself in the same year (Van Valin, 1977, 1).

A decade later is Construction Grammar — originally developed for modeling idioms (Fillmore et al., 1988) — whose first explicit witness is Watters (1988, 15)'s dissertation on Tlachichilco Tepehua [tpt]. LFG, RRG and Construction Grammar have in common that they are minority theories which nevertheless continue to gain popularity, even into the present, in terms of proportional mentions.

Finally, Basic Linguistic Theory, most extensively articulated by Dixon (2010), is first adopted under that name in Hanafi (1997)'s work on Sundanese [sun] (crediting Dixon's 1996 lecture series in Canberra that underlie the later publication). It is since the fastest growing and the currently most popular explicitly mentioned descriptive theory, occurring in approximately 1/7 grammars. Arguably, many, perhaps most, other descriptions in the past have used a naive version of Basic Linguistic Theory. The new development lies in the explicit use of an extensively developed version thereof.

Figure 2 also contains a curve for the proportion of grammars that mention at least one of the aforementioned theories at least once. Compared to Figure 1 it exhibits a dipping point in the curve around 1990, i.e., mentions of theory increase steadily, but the specific theories inspected here come in waves.

3 Conclusion

Thanks to the appearance of the DReaM corpus we were able to quantify some trends relating to theoretical framework for grammatical description through time. Although shallow, these investigations do reinforce the widely held impression that restrictive descriptive theories "come and go" (Aikhenvald, 2015, 6-7).

References

Alexandra Y. Aikhenvald. 2015. *The art of grammar: a practical guide*. Oxford: Oxford University Press.

Pierre Alexandre. 1966. *Système Verbal et Prédicatif du Bulu*, volume 1 of *Langues et Littératures de l'Afrique Noire*. Paris: Librairie C. Klincksieck.

Mahadeo Laxman Apte. 1962. *A sketch of Marathi transformational grammar*. Ph.D. thesis, Madison: University of Wisconsin.

Leonard Bloomfield. 1935. *Language*. George Allen & Unwin: London.

Lars Borin & Anju Saxena. 2004. Grammar, incorporated. In *Henrichsen, P. J. (ed). CALL for the Nordic languages*, pages 125–145. Samfundslitteratur, Frederiksberg.

Lars Borin, Markus Forsberg, & Johan Roxendal. 2012. Korp the corpus infrastructure of språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, pages 474–478.

Shobhana L. Chelliah & Willem J. de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Dordrecht: Springer.

Noam Chomsky. 1957. *Syntactic Structures*. The Hague: Mouton.

Mary Dalrymple, John J. Lowe, & Louise Mycock. 2019. *The Oxford Reference Guide to Lexical Functional Grammar*. Oxford: Oxford University Press.

William D. Davies. 1981. *Choctaw Clause Structure*. Ph.D. thesis, University of California at San Diego.

R.M.W. Dixon. 2010. *Basic Linguistic Theory*. Oxford: OUP. 2 vols.

Matthew S. Dryer. 2006. Descriptive theories, explanatory theories, and basic linguistic theory. In Felix Ameka, Alan Dench, & Nicholas Evans, editors, *Catching Language: Issues in Grammar Writing*, pages 207–234. Berlin: Mouton de Gruyter.

Martha Duff. 1959. Amuesha (arawak) syntax i: simple sentence types / sintaxe amuexa (arawak) i: sentenças do tipo simples. In *Publicações do Museu Nacional*, volume 1 of *Série Lingüistica Especial*, pages 172–237. Rio de Janeiro: Museu Nacional.

Charles J. Fillmore, Paul Kay, & Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

Harald Hammarström, Shafqat Mumtaz Virk, & Markus Forsberg. 2017. Poor man's ocr post-correction: Unsupervised recognition of variant spelling applied to a multilingual document collection. In *Proceedings of the Digital Access to Textual Cultural Heritage (DATeCH) conference*, pages 71–75. Göttingen: ACM.

Nurachman Hanafi. 1997. *A typological study of Sundanese*. Ph.D. thesis, LaTrobe University.

Zellig S. Harris. 1957. Co-occurrence and transformation in linguistic structure. *Language*, XXXIII:283–340.

Fred Karlsson. 2005. Nature and methodology of grammar writing. *SKY Journal of Linguistics*, 18:341–356.

Kenneth L. Pike. 1954-1960. *Language in relation to a unified theory of the structure of human behavior*. Santa Ana, California: Summer Institute of Linguistics. 3 vols.

Keren Rice. 2006. Let the language tell its story? the role of linguistic theory in writing grammars. In Felix Ameka, Alan Dench, & Nicholas Evans, editors, *Catching Language: Issues in Grammar Writing*, pages 235–268. Berlin: Mouton de Gruyter.

Jeanette Sakel & Daniel L. Everett. 2012. *Linguistic Fieldwork: A Student Guide*. Cambridge: Cambridge University Press.

Leslie Saxon. 1986. *The Syntax of Pronouns in Dogrib: Some Theoretical Consequences*. Ph.D. thesis, University of California at San Diego.

Mary Dorothea Sleator. 1957. *Phonology and morphology of an American English dialect*. Ph.D. thesis, Indiana University.

Nina Tahmasebi, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, Fredrik Johansson, Richard Johansson, Mikael Kågebäck, Olof Mogren, Pierre Nugues, & Thomas Risse. 2015. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2-4):169–187.

Robert D. Van Valin. 1977. *Aspects of Lakhota Syntax: A Study of Lakhota (Teton Dakota) Syntax and its Implications for Universal Grammar*. Ph.D. thesis, University of California, Berkeley.

Shafqat Virk, Lars Borin, Anju Saxena, & Harald Hammarström. 2017. Automatic extraction of typological linguistic features from descriptive grammars. In *Text, Speech, and Dialogue 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings / edited by Kamil Ekstein, Václav Matousek.*, Cham. Springer International Publishing.

Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg, & Søren Wichmann. 2020. The dream corpus: A multilingual annotated corpus of grammars for the worlds languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 871–877. European Language Resources Association, Marseille.

Gunther Tronje von Hagen. 1914. *Lehrbuch der Bulu Sprache*. Berlin: Druck und Verlag von Gebr. Rodetzki Hofbuchhandlung, Berlin.

Viola G. Waterhouse. 1974. *The history and development of tagmemics*. The Hague: Mouton.

James Kenneth Watters. 1988. *Topics in Tepehua Grammar*. Ph.D. thesis, University of California at Berkeley.

Coveting your neighbor's wife: Using lexical neighborhoods in substitution-based word sense disambiguation

Richard Johansson
Department of Computer Science and Engineering
University of Gothenburg and Chalmers University of Technology, Sweden
richard.johansson@gu.se

Abstract

We explore a simple approach to word sense disambiguation for the case where a graph-structured lexicon of word sense identifiers is available, but no definitions or annotated training examples. The key idea is to consider the *neighborhood* in a lexical graph to generate a set of potential substitutes of the target word, which can then be compared to a set of substitutes suggested by a language model for a given context. We applied the proposed method to the SALDO lexicon for Swedish and used a BERT model to propose contextual substitutes. The system was evaluated on sense-annotated corpora, and despite its simplicity we see a strong improvement over previously proposed models for unsupervised SALDO-based word sense disambiguation.

1 Introduction

Probabilistic language models estimate the probability of a word occurring in a given context. This means that for an observed occurrence of a word, a language model can suggest other words – *substitutes* – that could potentially have occurred instead. With a high-quality language model, the set of potential substitutes reflects the *sense* of the word in that specific context. This intuition suggests a simple mechanism for the task of *word sense disambiguation* (WSD) where our goal is to link each occurrence to an item in a fixed sense inventory defined by a lexicon: assuming that the lexicon allows us to generate a set of potential substitutes for each sense, we can then simply compare each of these lists to the one we got from the language model. To disambiguate, we then select the lexicon sense where the substitute set is most similar to the language model's set of substitutes.

How can we use a lexicon to generate a set of potential substitutes of a given sense? This depends on what information the lexicon represents and how it is structured. In this work, we assume that the lexicon is graph-structured and that proximity in the graph corresponds to substitutability; this assumption allows us to generate a set of potential substitutes of a given sense by considering its *neighborhood* in the graph.

To exemplify, let us assume that we are given the following two occurrences of the Swedish word *ämne* and that we want to associate them with a sense in the SALDO lexicon (Borin et al., 2013):

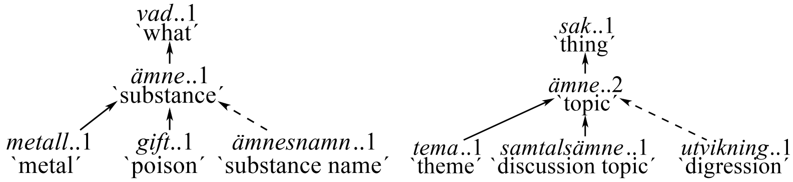
(1) *Detta ämne är frätande.*
'This substance is corrosive.'

(2) *Detta ämne kommer att diskuteras senare.*
'This topic will be discussed later.'

For the first case, the five most probable substitutes suggested by a BERT model are *innehåll* 'content', *gift* 'poison', *område* 'area', *medel* 'agent', *föremål* 'object'; for the second case, they are *område* 'area', *problem* 'problem', *språk* 'language', *tema* 'theme', *förslag* 'proposal'.

We then consider the neighborhoods in the lexicon graph. SALDO defines four senses of *ämne*. Sense 1 corresponds to 'substance' and its immediate neighborhood overlaps with the substitute set for the first example: there is an edge in the SALDO graph between sense 1 of *ämne* and sense 1 of *gift*, so we can link the first occurrence to sense 1. Similarly, sense 2 in SALDO corresponds to 'topic' and there is an edge between this sense and sense 1 of *tema*, allowing us to disambiguate the second occurrence as well.

Richard Johansson. 2022. Coveting your neighbor's wife: Using lexical neighborhoods in substitution-based word sense disambiguation. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 61–66. Available under CC BY 4.0



(a) Neighborhood of *ämne*..1. (b) Neighborhood of *ämne*..2.

Figure 1: Fragments of SALDO neighborhoods for two of the senses of *ämne*. Primary descriptor edges are drawn as solid arrows and secondary descriptor edges as dashed arrows.

2 The SALDO lexicon

The SALDO lexicon (Borin et al., 2013) defines a large sense inventory for Swedish words. While a number of other large-scale lexical resources for Swedish have been developed, SALDO is the largest open resource. It is an extended version of the SAL lexicon (Lönngren, 1989; Borin, 2005) and has been used as a pivot lexicon to define mappings between several lexical-semantic resources in Swedish (Borin, 2010), for instance in the Swedish FrameNet++ project (Borin et al., 2010).

Borin & Forsberg (2009) discuss the conceptual differences between SALDO and WordNet (Fellbaum, 1998). A major difference between these resources is that SALDO tends to use a more coarse-grained sense inventory compared to WordNet. Another fundamental difference is that SALDO does not define typed lexical-semantic relations (e.g. synonymy, is-a, hyponymy) between word senses but instead relies on the notion of *association* (Borin et al., 2013). Association can correspond to several types of lexical-semantic relationships: in many cases, an associated sense can be a synonym or hyperonym, but in other cases it can be e.g. a meronym or be in a predicate–argument relationship.

While each sense could in principle be in an association relationship with many other senses, SALDO explicitly encodes relationships between each sense and its *primary descriptor* (PD): an associated sense that has a more primitive meaning. A few additional relationships are encoded as *secondary descriptors*. SALDO includes no other lexical-semantic information apart from these relations, such as sense definitions or contextual examples. Figure 1 shows the neighborhoods in the SALDO graph around the two senses of *ämne* discussed in the introduction.

3 Previous work

Disambiguation systems are implemented in different ways depending on what resources are available. For WordNet-based WSD in English, the most systems tend to use supervised learning because of the availability of moderately large annotated datasets. WordNet is also a fairly rich resource and includes definitions, glosses, as well as several types of labeled sense-to-sense relations. In contrast, SALDO-based WSD is more challenging because of the small quantity of available annotated data and the sparse information in the lexicon. For this reason, most of the WSD systems using SALDO rely on the structure of the lexicon graph only, sometimes in combination with representations learned from unannotated text.

Johansson & Nieto Piña (2015b) proposed a method to align SALDO senses with a word embedding model; this approach naturally leads to a disambiguation mechanism (Johansson & Nieto Piña, 2015a). A tool using this disambiguation method is now integrated in the *Sparv* annotation pipeline (Borin et al., 2016). Nieto Piña & Johansson (2017) used a graph-based regularizer to train word and sense embeddings jointly. Purely graph-based WSD approaches requiring no corpora include graph embeddings using random walks (Nieto Piña & Johansson, 2016b) and personalized PageRank (Agirre & Soroa, 2009).

Nieto Piña & Johansson (2016a) evaluated several WSD systems on all SALDO-annotated corpora that were available at the time. The system by Johansson & Nieto Piña (2015b) was the most effective of those using no training data, but a comparison with a supervised system (on a limited set of target lemmas for which annotated data was available) showed that the unsupervised systems performed relatively poorly.

The idea of disambiguating word senses by using language models to suggest potential substitutes was first proposed by Başkaya et al. (2013), who applied this approach for WordNet-based WSD as well as for lexicon-free word sense induction (WSI). Subsequent work has mostly focused on WSI: for instance, Amrami & Goldberg (2018) applied a pair of language models to generate substitute sets for WSI.

The same group later used a BERT model for substitute set generation (Amrami & Goldberg, 2019) and this approach is the state of the art in WordNet-based WSI for English as of 2022 (Eyal et al., 2022). The pre-training of BERT (Devlin et al., 2019) involves (among other things) training a *masked language model* (MLM) that tries to predict the identity of a hidden word in a given context, and this aligns perfectly with our goals since a substitute set can then be generated simply by applying the MLM.

4 Selecting a SALDO sense for an ambiguous word

Assuming that we are given a context, the position of a word to disambiguate, and a set of SALDO senses to select from, we compute a weighted contextual set of substitute words (§4.1) as well as a weighted word set based on the SALDO neighborhood for each sense (§4.2). We then compute the cosine similarity between the contextual set to each of the SALDO-based sets and select the highest-scoring sense.

4.1 Proposing contextual substitutes

We follow the most recent work in substitution-based WSI and apply the MLM of a BERT model. We used the Swedish BERT model published by the Swedish Royal Library (Malmsten et al., 2020). Following Eyal et al. (2022), the MLM is applied in a straightforward manner without masking or modifying the text. We compute the probability distribution at the target position, select the 200 top-scoring items, and exclude inflections of the original target word. The set of potential substitute tokens are weighted proportionally to the probability assigned by the MLM.

While the application of BERT is quite straightforward, the probability distributions are affected by the word piece tokenization. For instance, if a token is followed by a suffix word piece (e.g. `##ar`), the MLM will assign high probabilities mainly to prefixes likely to be followed by this suffix. This likely causes the substitute sets to be of poorer quality for less frequently occurring words and precludes the use of the approach for the disambiguation of multiword expressions. In this work, we simply removed suffix word pieces (starting with `##`) from the set of substitutes; the development of a more systematic approach could potentially be explored in later work.

4.2 Extracting neighborhoods from SALDO

We use the neighborhood extraction approach proposed by Nieto Piña & Johansson (2017). For a given SALDO sense, we extract its immediate neighbors in the SALDO graph, following primary and secondary descriptor edges in both directions. Since our goal is to produce a list of words that could potentially be substituted, we only include senses of words of the same grammatical category as the original sense. We repeat the process and add parents, children, and siblings to the set until it has a size of at least 16. Finally, we use the morphological lexicon of SALDO to map every sense to a set of inflected forms, so that e.g. *gift*..1 results in *gift, giftet, . . . , giftens*. The items are assigned weights that depend on the distance in the SALDO graph.

5 Experiments

The largest sense-annotated resource for Swedish was developed in the SemTag project (Järborg, 1999); this covers most of the Stockholm–Umeå corpus (Ejerhed et al., 1992). However, this resource does not use SALDO to define its senses, although SALDO has imported some senses from SemTag lexicon. The Swedish lexical sample of the *SENSEVAL-2* shared task (Kokkinakis et al., 2001) used a subset of the SemTag resource consisting of annotation for 40 ambigous lemmas. The senses for these lemmas were manually mapped to SALDO by Nieto Piña & Johansson (2016a). Since SALDO uses a coarser division into senses than SemTag, three of the lemmas were not ambiguous after this lexicon mapping and they were removed from the dataset. The only running-text corpus annotated with SALDO senses is *Eukalyptus* (Johansson et al., 2016), which includes texts from eight different domains.

Method	SENSEVAL-2	Eukalyptus
Substitutes	0.6675	0.7020
J & NP (2015)	0.4976	–
Random baseline	0.3557	0.4094
Lowest-sense baseline	0.4952	0.6580
Supervised (BoW)	0.8033	–
Supervised (BERT)	0.9209	–

Table 1: Disambiguation results on the test sets for the different methods.

The instances were preprocessed using the *Sparv* pipeline (Borin et al., 2016). For each word, the pipeline proposes a set of possible SALDO senses, based on the automatically determined morphological analysis and lemmatization. The sense disambiguator chooses one of the candidates from this set.

Unambiguous words are excluded from the experiment, which means that the *practical* accuracy is higher than what we report in the next section, since the majority of the words are unambiguous. We also exclude cases where the annotated sense is a non-compositional reading of a multi-word expression (e.g. *på örat* intended as ‘drunk’, not as ‘on the ear’) or a compositional reading of a compound. After this preprocessing, the SENSEVAL-2 sample consists of a test set of 1,366 instances and a training set of 7,790 instances, and the Eukalyptus set of 12,434 instances.

5.1 Results

We evaluated the substitute-based approach proposed in this paper and compare it to a number of trivial and nontrivial baselines. Table 1 shows the disambiguation accuracies on the two test sets. The accuracies are macro-averaged over the 37 lemmas for SENSEVAL-2 and micro-averaged for Eukalyptus.

The most meaningful comparison is with the method by Johansson & Nieto Piña (2015a), which is included in *Sparv*: this system uses a similar setup with a combination of the SALDO graph and a representation model trained in an unsupervised fashion. As we can see, the substitute-based method performs much better on the SENSEVAL-2 test set. Both methods outperform two trivial baselines: random selection, and selecting the sense with the lowest numerical identifier. The substitute-based method also outperforms the lowest-sense baseline on the Eukalyptus set.

For SENSEVAL-2, we also evaluate two straightforward supervised approaches that learn from annotated training examples: a linear SVM using a bag-of-words representation, and a MLP on a BERT representation. Both were implemented as “word experts” that use one classifier per base form. All graph-based methods are strongly outperformed by the supervised models. Practically, the supervised approach cannot be applied to Eukalyptus because of the Zipfian distribution of lemmas to disambiguate.

6 Discussion

The proposed method works surprisingly well compared to the baselines despite its simplicity. The method is also quite cheap: in the implementation we have described here, we have only used the graph-based neighborhood, although in the general case it may be possible to exploit other lexical-semantic information to generate more accurate substitute sets. No annotated examples for training are needed.

While the performance is better than previous purely graph-based WSD approaches using SALDO, it is much lower than for supervised models in a lexical sample setting. Obviously, a supervised word expert approach is more difficult to apply in a running-text setting, e.g. in Eukalyptus. Another important practical consideration is the flexibility of the substitute-based method: if we add a new sense to the lexicon and update the edges accordingly, we can *immediately* use the new sense in the disambiguator. The method can therefore be argued to be applicable in an interactive fashion.

This is a first attempt and we see a potential for a more careful consideration of the graph-based substitute set, the contextual substitutes, and the way that these sets are compared. The whole idea hinges on being able to use the lexical resource to suggest potential substitutes. For SALDO, this works less well in some cases where the neighborhood structure does not correspond well to substitutability. Words referring to professions is one such case; cf. the discussion by Johansson (2014).

More generally, we may want to develop methods that align token representations from a language model with a representation of the graph. One might use an embedding of the SALDO graph, either a purely graph-based embedding (Nieto Piña & Johansson, 2016b) or one based on a combination of the graph and a corpus (Johansson & Nieto Piña, 2015b; Nieto Piña & Johansson, 2017). It may then be possible to build a mapping of the BERT-based representation into the space of the embedded graph.

Acknowledgements

This work builds on resources developed in projects running over several decades at Språkbanken and Uppsala University: the development of the SALDO lexicon and the sense-annotated corpora in the SemTag and Eukalyptus projects. I was funded by the projects *Interpreting and Grounding Pre-trained Representations for NLP* and *Representation Learning for Conversational AI*, both funded by WASP.

References

Eneko Agirre & Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens.

Asaf Amrami & Yoav Goldberg. 2018. Word Sense Induction with Neural biLM and Symmetric Patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867, Brussels. Association for Computational Linguistics.

Asaf Amrami & Yoav Goldberg. 2019. Towards better substitution-based word sense induction. arXiv preprint 1905.12598, `https://arxiv.org/pdf/1905.12598.pdf`.

Osman Başkaya, Enis Sert, Volkan Cirik, & Deniz Yuret. 2013. AI-KU: Using Substitute Vectors and Co-Occurrence Modeling For Word Sense Induction and Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 300–306, Atlanta, Georgia. Association for Computational Linguistics.

Lars Borin & Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series*, volume 7.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, & Dimitrios Kokkinakis. 2010. The Past Meets the Present in the Swedish FrameNet++. In *Proceedings of EURALEX*.

Lars Borin, Markus Forsberg, & Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, & Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *Swedish Language Technology Conference*, Umeå.

Lars Borin. 2005. Mannen är faderns mormor: Svenskt associationslexikon reinkarnerat. *LexicoNordica*, 12:39–54.

Lars Borin. 2010. Med Zipf mot framtiden - en integrerad lexikonresurs för svensk språkteknologi. *LexicoNordica*, 17.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eva Ejerhed, Gunnel Källgren, Ola Wennstedt, & Magnus Åström. 1992. The linguistic annotation system of the Stockholm-Umeå corpus project – description and guidelines. Technical report, Department of Linguistics, Umeå University.

Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, & Yoav Goldberg. 2022. Large Scale Substitution-based Word Sense Induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

Richard Johansson & Luis Nieto Piña. 2015a. Combining Relational and Distributional Knowledge for Word Sense Disambiguation. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 69–78, Vilnius. Linköping University Electronic Press.

Richard Johansson & Luis Nieto Piña. 2015b. Embedding a Semantic Network in a Word Space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1428–1433, Denver.

Richard Johansson, Yvonne Adesam, Gerlof Bouma, & Karin Hedberg. 2016. A Multi-domain Corpus of Swedish Word Sense Annotation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3019–3022, Portorož.

Richard Johansson. 2014. Automatic Expansion of the Swedish FrameNet Lexicon. *Constructions and Frames*, 6(1):92–113.

Jerker Järborg. 1999. Lexikon i konfrontation. Technical report, University of Gothenburg. Research Reports from the Department of Swedish, Språkdata, GU-ISS-99-6.

Dimitrios Kokkinakis, Jerker Järborg, & Yvonne Cederholm. 2001. SENSEVAL-2: The Swedish Framework. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 45–48, Toulouse.

Lennart Lönngren. 1989. Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi. Technical report, Uppsala University. Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi.

Martin Malmsten, Love Börjeson, & Chris Haffenden. 2020. Playing with words at the National Library of Sweden – Making a Swedish BERT. arXiv preprint 2007.01658, `https://arxiv.org/pdf/2007.01658.pdf`.

Luis Nieto Piña & Richard Johansson. 2016a. Benchmarking word sense disambiguation systems for Swedish. In *Swedish Language Technology Conference*, Umeå.

Luis Nieto Piña & Richard Johansson. 2016b. Embedding Senses for Efficient Graph-based Word Sense Disambiguation. In *Proceedings of the 2016 Workshop on Graph-based Methods for Natural Language Processing*, pages 2710–2715, San Diego.

Luis Nieto Piña & Richard Johansson. 2017. Training Word Sense Embeddings with Lexicon-based Regularization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 284–294, Taipei.

Linguistics concepts as semantic frames

Per Klang
Department of Scandinavian Languages
Uppsala University
Uppsala, Sweden
per.klang@nordiska.uu.se

Shafqat Mumtaz Virk
Språkbanken Text
University of Gothenburg
Gothenburg, Sweden
shafqat.virk@svenska.gu.se

Abstract

The topic of this paper is the representation of linguistic concepts as semantic frames. It presents the general structure of a lexical resource for the linguistic domain, here referred to as *LingFN*. This resource contains semantic frames for the linguistic terms and concepts found in traditional grammars. In addition, the paper illustrates how LingFN can be used for the natural language processing task of automatically extracting typological information from descriptive grammars which is otherwise attained manually at a greater cost.

1 Introduction

A grammatical description is a form of document that describes various structural aspects of a natural language. There are approximately 7 000 recorded natural languages (see ethnologue.com), and written grammatical information is available for around 4 000 of these (see golottolog.org for details). At present, there are ongoing endeavors to digitize this information so that modern computational techniques can be exploited to analyze it and, ultimately, the languages themselves.

Digitization can be seen from two different prospects. The first regards the preparation of corpora which involves scanning and OCRing of available printed resources on natural languages, and the second covers structured digital representations of natural languages and linguistic concepts. This paper's concern is with the latter.

Over centuries of philosophical research, linguists have developed a variety of notions on how to discern and describe the phonological, morphological, and structural attributes of language. Indeed, there is a reasonable amount of work on digital representation of lexicons of natural languages, and consequently there exist many online dictionaries and related lexical resources (Warwick, 1988; Fellbaum, 1998). But, to the best of our knowledge, the amount of work is limited, as regards the digital representation of various grammatical aspects, and grammars of natural languages.

This paper is an attempt to fill the gap above by using frame-semantics – a theory of meaning in language (Fillmore, 1976) – to develop special structures (semantic frames) for representing the concepts of linguistics. Using a semiautomatic methodology, a set of semantic frames have been developed and connected to each other via various types of relations resulting in a network of semantic frames called *LingFN*, *a framenet for the linguistics domain*. LingFN is expected to be useful for various NLP tasks, especially for automatic extraction of typological information from descriptive grammars which is otherwise accomplished by labour intensive manual methods.

The paper begins with a brief presentation of frame semantics and its background (section 2), followed by a description of LingFN and the idea of representing linguistic concepts as semantic frames (section 3). Next is given a brief outline of some possible applications of LingFN (section 4). The paper is concluded with a summary (section 5).

Per Klang and Shafqat Mumtaz Virk. 2022. Linguistics concepts as semantic frames. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 67–71. Available under CC BY 4.0

2 Background

The general idea of frame semantics is that words are understood with respect to the situation that they evoke in the mind of the speaker. The mapping between a word and a situation forms a conceptual structure known as a semantic frame, which is a script-like description of a prototypical situation, an event, or an object, along with its participants known as *frame elements*, or FEs for short (Ruppenhofer et al., 2016). The ideas of frame semantics were first put to use in a lexico-semantic resource for English called FrameNet, also known as Berkeley FrameNet (BFN), which contains a network of semantic frames for general-language (Baker et al., 1998). This resource has successfully been used for automatic shallow semantic parsing (Gildea & Jurafsky, 2002) which is employed in several natural language processing tasks such as information extraction (Surdeanu et al., 2003), and question answering (Shen & Lapata, 2007), to name a few. The utility of FrameNet has also led to the development of framenets for a number of other languages that build upon the BFN model, all of which have contributed to the understanding of the semantic characteristics of each specific language.

Although general framenets have proven to be useful for many tasks, they have also been criticized for their limited coverage. To cope with this problem, domain-specific framenets have been constructed as complements to the corresponding general-language framenets in order to improve the performance of NLP-tools for specific domains such as medicine (Borin et al., 2007), soccer (http://www.kicktionary.de/), and soccer-related tourism (Torrent et al., 2014). The upcoming section presents the overall structure of another domain specific framenet, namely the previously mentioned LingFN whose content consists of semantic frames for linguistic terms and concepts that have been used in traditional linguistic grammars (e.g. inflection, agreement, affixation, etc.). While a part of LingFN follows the general structure of event frames in BFN, the other part is structured after an ontology of linguistic terms, known as GOLD.[1]

3 Linguistics concepts as semantic frames

LingFN is a framenet for the linguistics domain which has been created after the grammar (and language) descriptions found in the 1.3 MW sub corpus of *The Linguistic Survey of India* (LSI) (Grierson, 1903 1927).[2] Although the lexicographic work in LingFN assumes a slightly less complex structure than BFN, it mostly draws upon the same model as described in the BFN manual (Ruppenhofer et al., 2016) where words, or *lexical units* (LUs), with similar meaning are bundled up under the same frame. Like BFN, it holds a network of semantic frames (Baker et al., 1998), but the major difference between the two networks lies in that the former covers specific linguistic terms and concepts where the latter covers more general concepts. Detailed descriptions of the development of LingFN are found elsewhere (Malm et al., 2018; Virk et al., 2022).

The general design of LingFN is fairly simple, and it is illustrated in Figure 1 below. It has two frame types: *event frames* which represent eventful types of scenes (or concepts), and *filler frames*, which follow the general structure of linguistic terms in GOLD. It also contains two frame-to-frame links: an *inheritance link* which forms a hierarchical IS-A relation between frames, and a *used-by link* which connects the filler frames to any event frame in which they may appear.
Let us now consider an example of how semantic frames for the linguistics domain may be used to aid in the identification of linguistic information from descriptive grammars. This is the case of the word *borrow* which has a specific meaning in linguistics. Consider below the difference between the meaning of *borrow* from the BFN BORROWING frame in (1a), as opposed to *borrow* in (1b) from LSI.

(1) a. Does my Mum *borrow* money off you? (BFN)
b. […] the Musalmān dialect *borrows* freely from the Persian vocabulary. (LSI)

The BORROWING frame holds information about a situation involving a number of frame elements such as a BORROWER that takes possession of a THEME belonging to a LENDER under the tacit agreement

[1] http://linguistics-ontology.org/

[2] LSI presents a comprehensive survey of the languages spoken in South Asia. It was conducted in the late nineteenth and the early twentieth century by the British government, under the supervision of George A. Grierson. The survey resulted in a detailed report comprising 19 volumes of around 9500 pages in total. The survey covered 723 linguistic varieties for which it provides: a grammatical sketch, a core word list; and text specimens.

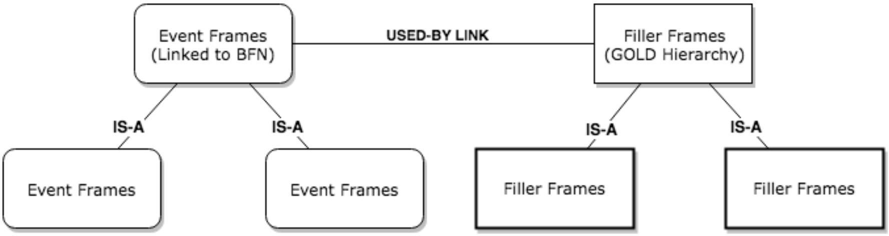


Figure 1: The basic structure of LingFN

that the THEME is to be returned after a DURATION of time. Example (1a) is clearly an instance of the BORROWING frame, since it is conceptually necessary to imagine the return of the borrowed item given a duration of time. This is not the case in (1b). A word that has been borrowed by a language cannot be returned, since it was never really borrowed in the first place. The pseudo-character of the linguistic borrowing frame, which we may refer to as PSEUDO-BORROWING, can be further illustrated by means of *reductio ad absurdum* in the comparison of (2a–b) below.

(2) a. [. . .] Paula had let her borrow the boat *for a few hours* [. . .]. (BFN)

b. This principle of formation is borrowed from Magahī $\left\{\begin{array}{l} \text{* \textit{for a few hours}} \\ \text{* \textit{until next spring}} \\ \text{* \textit{temporarily}} \end{array}\right\}$. (Constructed)

The PSEUDO-BORROWING frame can thus be delimited from the BFN BORROW frame in that the latter does not occur with an FE of DURATION. However, this observation is of limited use since it can only be used for identifying the cases that are not PSEUDO-BORROWING. Still, as we shall see, this limitation may be remedied by the filler frames and their used-by links to frames like that of PSEUDO-BORROWING.

The filler frames typically capture information about LUs that appear as frame elements in event frames, such as the type, material, or color of the LU referent. Annotated examples from LSI are given below marking the two filler frames GENETIC TAXON and LINGUISTIC DATA STRUCTURE.

(3) a. [. . .] [$_{\text{LANGUAGE\_VARIETY}}$ the Musalmān] [$_{\text{LU}}$ dialect] borrows freely from the Persian vocabulary. (GENETIC_TAXON)

b. [. . .] the Musalmān dialect borrows freely from the [$_{\text{LANGUAGE\_VARIETY}}$ Persian] [$_{\text{LU}}$ vocabulary]. (LINGUISTIC_DATA_STRUCTURE)

The information in the filler frames can aid in the disambiguation of polysemous verbs like *borrow*. The linking of filler frames to event frames with used-by links allows for the FEs of the event frame to be checked against the LUs listed in the LingFN filler frames. On the one hand, if the content of an FE of some event frame is not listed in the set of filler frames, the event frame probably belongs to the general domain. On the other hand, if the content of the FE is listed in the set of filler frames, as illustrated by the sentence in the table below, it would suggest that the event frame is specific to the linguistic domain.

Sentence:	*The*	*Musalman*	*dialect*	*borrows*	*from*	*the*	*Persian*	*vocabulary*
Event FEs:	Borrower			LU	Lender			
Filler FEs:		Language variety	LU					
Filler FEs:							Language variety	LU

In its current state, LingFN houses nearly 100 frames with 325 used-by links, about 360 lexical units, and more than 2 800 annotated sentences from LSI; see statistics below.

Frame Type	**Frames**	**Used-by links**	**Lexical Units**	**Annotated examples**
Event frames	5	171	25	1 858
Filler frames	94	154	335	948
Total	99	325	360	2 806

The dual structure of the event and filler frames, coupled with the inheritance and used-by links, provides a simple architecture that may be exploited for applications aimed at automatic extraction of information from grammatical descriptions. These applications form the subject of the next section.

4 Applications of linguistic domain semantic frames

LingFN was developed particularly for the extraction of typological linguistic information from descriptive grammars.[3] Traditionally, the extraction of typological information from descriptive grammars is done manually as a part of the development of typological databases (e.g. https://wals.info/). While the manual curation of such databases takes a lot of time and effort, the usefulness of the end result is often cited to justify the development cost. A reasonable alternative, which has not been subject to extensive consideration in the literature, is to automatically extract the typological information. LingFN has proven to be helpful in this respect, and the general process to meet this end is described next.

In the first stage, the linguistic concepts were gathered from the grammatical descriptions, and annotated as semantic frames. These annotations were used to train a parser, which subsequently was used to annotate additional grammatical descriptions. The parser annotations were then converted to typological features values to be used for the purpose of linguistic analysis. For more details on the annotation task, the parser development, and the conversion from semantic parses to typological features, see Virk et al., (2017), and Virk et al., (2019).

5 Summary

This paper has proposed to use frame semantics for structured digital representations of traditional linguistic concepts. The concepts have been gathered from LSI and the GOLD ontology of linguistic terms and rendered as semantic frames in LingFN, a framenet for the domain of linguistics. It has further been shown how this resource can be used for automatic extraction of information from descriptive grammars.

6 Acknowledgements

The work presented here has been financially supported by the Swedish Research Council through its funding of the projects *South Asia as a linguistic area? Exploring big-data methods in areal and genetic linguistics* (20152019; contract 421-2014-969) and *Swe-Clarin* (20142018; contract 821-2013-2003), the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) through its Swedish-Brazilian research collaboration program (20142019; contract BR2014-5860), and the University of Gothenburg, its Faculty of Arts and its Department of Swedish, through their truly long-term support of the Sprakbanken research infrastructure.

References

Collin F. Baker, Charles J. Fillmore, & John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of ACL/COLING 1998*, pages 86–90, Montreal. ACL.

Lars Borin, Maria Toporowska Gronostaj, & Dimitrios Kokkinakis. 2007. Medical frames as target and tool. In *FRAME 2007: Building Frame Semantics resources for Scandinavian and Baltic languages. (Nodalida 2007 workshop proceedings)*, pages 11–18, Tartu. NEALT.

[3] There is, however, nothing that prevents it from being used for other purposes.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Daniel Gildea & Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

George A. Grierson. 1903–1927. *A Linguistic Survey of India*, volume I–XI. Government of India, Central Publication Branch, Calcutta.

Per Malm, Shafqat Mumtaz Virk, Lars Borin, & Anju Saxena. 2018. LingFN: Towards a Framenet for the Linguistics Domain. In *Proceedings of the IFNW 2018 Workshop on Multilingual FrameNets and Constructicons at LREC 2018*, Miyazaki. ELRA.

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R Johnson, Collin F Baker, & Jan Scheffczyk. 2016. FrameNet II: Extended Theory and Practice.

Dan Shen & Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. In *Proceedings of EMNLP-CoNLL 2007*, pages 12–21, Prague. ACL.

Mihai Surdeanu, Sanda Harabagiu, John Williams, & Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of ACL 2003*, pages 8–15, Sapporo. ACL.

Tiago Timponi Torrent, Maria Margarida Martins Salomão, Ely Edison da Silva Matos, Maucha Andrade Gamonal, Júlia Gonçalves, Bruno Pereira de Souza, Daniela Simões Gomes, & Simone Rodrigues Peron-Corrêa. 2014. Multilingual lexicographic annotation for domain-specific electronic dictionaries: The Copa 2014 FrameNet Brasil project. *Constructions and Frames*, 6(1):73–91.

Shafqat Virk, Lars Borin, Anju Saxena, & Harald Hammarström. 2017. Automatic extraction of typological linguistic features from descriptive grammars. In *Proceedings of TSD 2017*. Springer.

Shafqat Mumtaz Virk, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal, & Nazia Khurram. 2019. Exploiting Frame-Semantics and Frame-Semantic Parsing for Automatic Extraction of Typological Information from Descriptive Grammars of Natural Languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 1247–1256, Varna. INCOMA Ltd.

Shafqat Virk, Per Malm, Lars Borin, & Anju Saxena. 2022. LingFN: A FrameNet for the Linguistics Domain. In *Proceedings of CICLing 2019*.

Susan Warwick. 1988. Automated lexical resources in europe: A survey. *Lexicographica (1988)*, 4(1988):93–129.

Deep learning models as theories of linguistic knowledge

Shalom Lappin
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics, and Theory of Science
University of Gothenburg
shalom.lappin@gu.se

Abstract

Recent work in deep learning had made remarkable progress across a wide range of complex tasks in artificial intelligence in general, and in natural language processing in particular. It has yielded solutions to problems that were recalcitrant to traditional rule-based methods over many decades. It is worth considering the possibility that deep learning systems are not simply engineering procedures, but viable models of linguistic representation and language acquisition. When exploring this option it is important to keep in mind the serious limitations of these systems. In this paper I briefly explore these questions, and I offer some tentative conclusions.

1 Introduction

Over the past 70 years most linguists have used formal grammars and model theories to encode linguistic knowledge. Similarly, until 2000 rule systems and knowledge representation logics dominated many areas of artificial intelligence. The rise of deep learning in AI in general, and in natural language processing (NLP) in particular, has largely displaced symbolic methods in computational linguistics.

Baroni (2021) observes that most theoretical linguists have taken little, if any, notice of the role of deep neural networks (DNNs) in NLP. He argues that DNNs should be considered as possible alternative theories of linguistic representation. In Lappin (2021) I suggest a similar view. In this paper I will briefly address some of the arguments for this approach, and several of its implications.

2 Alternative approaches to representation and learning

A wealth of important work on individual languages, and on cross linguistic patterns, has been done within the formal grammar framework. Many (most?) advocates of these formalisms have assumed that they correspond to the way in which humans encode knowledge of their language, at some level of representation. Some theorists have claimed that the class of formal grammars that generate the set of natural languages cannot be learned on the basis of the primary linguistic data available to children, through domain general learning procedures. They have concluded that humans bring strong domain specific learning biases to bear on the language acquisition task. They have identified these biases with a Universal Grammar (UG), construed as a schematic cognitive structure that restricts the set of hypotheses available to the language learner, concerning the design of a possible grammar.

DNNs are not rule-based algebraic systems. They do not, in general, apply grammatical rules or constraints as part of their training regimen, nor do they extract symbolic representations from data. The success of deep learning in NLP raises the question of whether a non-symbolic, domain general inductive system may offer a viable model of the way in which humans acquire and represent linguistic knowledge. It is possible to enrich a neural network with syntactic or semantic learning biases in a variety of ways. It is important to consider such systems to see whether these enrichments improve performance on a given set of tasks.

If a DNN learns to solve a task that requires interesting types of linguistic information, at a level approaching human performance, within tractable amounts of time and data, then it provides a demon-

Shalom Lappin. 2022. Deep learning models as theories of linguistic knowledge. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 73–78. Available under CC BY 4.0

stration of how humans could, in principle, acquire this knowledge. To the extent that this learning uses domain general inductive mechanisms, it shows that strong linguistic biases may not be required for this element of language acquisition. Success in deep learning also indicates that humans could encode the knowledge required to perform a given task through distributed, non-symbolic representations.

While formal grammars offer elegant formalisms for encoding syntactic information (as do model theories for semantic content), they have generally not provided robust, wide coverage systems for handling linguistically interesting NLP applications. By contrast, deep learning has achieved remarkable success over a wide variety of tasks. They include, among others, identifying subject-verb agreement (Linzen et al., 2016; Bernardy & Lappin, 2017; Gulordava et al., 2018), machine translation (Bahdanau et al., 2015), image description (He et al., 2020), and prediction of sentence acceptability (Lau et al., 2020).

Formal grammars and model theories pose serious problems of learnability.[1] By contrast, recent work in deep learning has shown that relatively domain general inductive learning devices can learn to solve cognitively interesting NLP tasks within tractable limits of time and data.

3 Hybrid models: How much do linguistic theories add to NLP

It is natural to assume that integrating linguistic theories into deep learning models will improve the performance of DNNs, and add robust wide coverage to formal grammars. While this view is intuitively appealing, its correctness is far from obvious.

There are (at least) two ways in which hybrid systems of this kind can be implemented. First, it is possible to train a DNN to identify tree structures in data through design and training (Tai et al., 2015; Socher et al., 2011; Bowman et al., 2016; Yogatama et al., 2017; Choi et al., 2018; Maillard et al., 2019; Kuncoro et al., 2018; Kuncoro et al., 2019; Kuncoro et al., 2020). Second, we can enrich the training data with syntactic and semantic feature markers, rendering these markers part of the information that the DNN learns to generalise over (Ek et al., 2019).

Both sorts of model have been tested on a variety of NLP tasks, including sentiment analysis, natural language inference (NLI), and sentence acceptability prediction. In general, they have yielded only very small improvements, if any, over the performance of their non-enriched counterparts. Ek et al. (2019) report that the addition of syntactic and semantic tags to the training data actually degraded the level of accuracy of the LSTM tested on this task.[2]

The fact that hybrid DNNs incorporating symbolic linguistic information have not yielded particularly interesting results to date does not show that the approach is misguided. More successful versions of such models may yet emerge from future work. However, these results do indicate that DNNs learn in ways that do not straightforwardly accommodate symbolic representations and rule systems. It is worth considering the possibility that the way in which humans acquire natural languages, and other forms of knowledge, may be closer to deep learning than to classical paradigms of grammar induction.

4 The limitations of deep learning

The remarkable success of deep learning in NLP has tended to obscure the serious limitations of DNNs. There are at least three of these worth noting here. First, DNNs require very large quantities of training data in order to achieve reasonable performance, on most complex NLP tasks. This renders a comparison with human learning moot, given that children do not require this amount of data to achieve linguistic knowledge. Use of reinforcement learning (François-Lavet et al., 2018), and multimodal encode-decoder models (Hill et al., 2020a) may go some way to alleviating the need for large training data sets. Much work remains to be done on this problem.

The second difficulty is that DNNs generally lack transparency concerning their mode of operations, and the generalisations that they extract from data. It is often unclear how they learn the generalisations that they achieve. This undermines their usefulness as explanatory models of learning. I will take this issue up further in Section 5.

[1] See Clark & Lappin (2011) for discussion of complexity and learnability in grammar induction.

[2] See Lappin (2021) for detailed discussion of these hybrid models, and additional references.

Finally, deep learning systems frequently encounter problems in generalising to domains beyond those on which they are trained. The error rate of a DNN generally increases in proportion to the distance between the examples of its test set and those of its training data (Lake & Baroni, 2018). Also adversarial testing has indicated that small permutations in the test set can produce dramatic changes in accuracy, particularly in cognitively complex tasks like NLI (Talman & Chatzikyriakidis, 2019; Talman et al., 2021). I will briefly suggest a possible way of dealing with this limitation in Section 6.

5 The opacity problem

The absence of transparency in deep learning systems has become substantial with the move to large multi-headed transformers, many of which are non, or bidirectional, like BERT (Devlin et al., 2019). The primary source of opacity in DNNs is their use of non-linear functions, such as sigmoid and hyperbolic tangent, to compute the output states of units from their inputs. These functions cause the vectors that each layer of a DNN produces to be, in the general case, non-compositional. This is due to the fact that the representations of the input and the output vectors cannot be represented by a homomorphic mapping operation.[3]

Bernardy & Lappin (2022) and Bernardy & Lappin (2023) suggest a solution to the opacity problem. They propose Unitary Evolutionary Recurrent Neural Networks (URNs) for NLP. An URN uses unitary matrix word embeddings and simple linear operations on them to process linguistic input.[4] They do not contain non-linear functions, and so they are strictly compositional in the operations through which they combine word embeddings to obtain output matrices. URNs correspond to quantum circuits. No information is lost in the course of processing input. Earlier states of the network are fully recoverable.

Although URNs achieve promising results in recognising long distance agreement patterns in artificial languages, they have some way to go before attaining the scale, level of performance, and coverage of current state of the art non-transparent DNNs.

6 Extending the scope of generalisation

As we observed in Section 4, deep learning continues to contend with difficulties in generalisation and overfitting in many areas of NLP. Hill et al. (2020a) and Hill et al. (2020b)'s work on multimodal semantic learning points to an encouraging direction for solving this problem. They train agents in simulated visual environments to recognise new natural language commands, and to respond to them appropriately in these environments. They use a multimodal encoder-decoder architecture for this task, which extends the models applied to machine translation, to achieve dynamic visual grounding of language.

The systems that they describe perform more substantial learning and generalisation beyond the training data than previous models that are limited to linguistic input. These systems come closer to approximating human learning, which is grounded in a multimodal environment, than DNNs trained exclusively on linguistic data. When input from visual and other modalities is coordinated with linguistic training data, it may be possible to reduce the size of the linguistic training set. It is necessary to recognise that information from these modalities is indispensable to human language acquisition, and it should be counted as part of the data that supports the acquisition process.

7 Conclusions and future work

The robustness and wide coverage that deep learning systems are demonstrating across a wide range of cognitively interesting NLP tasks suggests that they are more than engineering devices for producing effective language technology. It is worth taking them seriously as possible models of linguistic representation and language acquisition. Attempts to integrate symbolic elements of linguistic theory into these systems have not yielded dramatic improvements in their performance to date. This situation may change

[3] A mapping $f: A \rightarrow B$ from group A to group B is a homomorphism iff for every $v_i, v_j \in A$, and the group operation $\cdot$, $f(A \cdot B) = f(A) \cdot f(B)$, where $f(A \cdot B) \in A$, and $f(A) \cdot f(B) \in B$.

[4] A complex square matrix U is unitary if its composite transpose U^* is identical to its inverse U^{-1}. The word embeddings of URNs contain only real numbers, and so they are orthogonal matrices.

in the future. At this point, these results suggest that linguistic theories may not be particularly useful for NLP tasks, as DNNs process information in a way that does not easily accommodate symbolic encoding.

DNNs exhibit serious limitations in large data requirements, opacity, and range of generalisation. Work on reinforcement learning in NLP, on transparent models, and on multimodal encoder-decoder architectures offers promising approaches to solving these problems. Considerably more research is needed on how to deal with noise in linguistic input without destabilisation. This is particularly important for tasks involving NLI, dialogue management, and text interpretation.

Closer cooperation among computational linguists, cognitive psychologists, and neuroscientists is needed in order to assess the similarities and the differences between the ways in which DNNs and humans process linguistic information. Deep learning systems can offer indirect insights into learning and representation by demonstrating what sort of knowledge can be acquired by such a system on the basis of certain kinds of training, with learning biases of a general, or a domain specific type, within given limits of time and data. However, to determine to what extent, if any, these devices actually correspond to human learning and representation it is necessary to study the latter in comparison with the performance of DNNs. Only comparative research of this kind can illuminate whether deep learning provides plausible models of human linguistic knowledge.

Acknowledgements

It is a pleasure and an honour to contribute a paper to Lars Borin's festschrift. Lars has made major contributions to linguistics and NLP in Sweden. The legacy of his work has had a lasting impact in these areas, and it will continue to do so in the future. He is a good friend and a fine colleague. I look forward to many more years of fruitful scientific exchange and friendship with him.

My research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg.

References

Dzmitry Bahdanau, Kyunghyun Cho, & Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ArXiv*, abs/1409.0473.

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *ArXiv*, 2106.08694, (to appear in (Lappin & Bernardy, in press)).

Jean-Philippe Bernardy & Shalom Lappin. 2017. Using deep neural networks to learn syntactic agreement. *Linguistic Issues In Language Technology*, 15:1–15.

Jean-Philippe Bernardy & Shalom Lappin. 2022. Assessing the unitary rnn as an end-to-end compositional model of syntax. In *End-to-End Compositional Models of Vector-Based Semantics 2022, Electronic Proceedings in Theoretical Computer Science 366.4*, pages 9–22.

Jean-Philippe Bernardy & Shalom Lappin. in press. Unitary recurrent networks: Algebraic and linear structures for syntax. In Shalom Lappin & Jean-Philippe Bernardy, editors, *Algebraic Structures in Natural Language*. CRC Press, Taylor & Francis, Boca Raton, London, New York.

Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, & Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1477, Berlin. Association for Computational Linguistics.

Jihun Choi, Kang Min Yoo, & Sang goo Lee. 2018. Learning to compose task-specific tree structures. In *AAAI Conference on Artificial Intelligence*.

Alexander Clark & Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell, Malden, MA and Oxford.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Ek, Jean-Philippe Bernardy, & Shalom Lappin. 2019. Language modeling with syntactic and semantic representation for sentence acceptability predictions. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 76–85, Turku.

Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, & Joelle Pineau. 2018. An introduction to deep reinforcement learning. *Foundations and Trends in Machine Learning*, 11(3-4):219–354.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, & Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, & Nicolas Pugeault. 2020. Image captioning through image transformer. *ArXiv*, pages 1–17.

Felix Hill, Andrew K. Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, & Adam Santoro. 2020a. drivers of systematicity and generalization in a situated agent. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa*.

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, & Stephen Clark. 2020b. Grounded language learning fast and slow. *ArXiv*.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, & Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Melbourne. Association for Computational Linguistics.

Adhiguna Kuncoro, Chris Dyer, Laura Rimell, Stephen Clark, & Phil Blunsom. 2019. Scalable syntax-aware language models using knowledge distillation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3472–3484, Florence. Association for Computational Linguistics.

Adhiguna Kuncoro, Lingpeng Kong, Daniel Fried, Dani Yogatama, Laura Rimell, Chris Dyer, & Phil Blunsom. 2020. Syntactic structure distillation pretraining for bidirectional encoders. *ArXiv*.

Brenden Lake & Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer Dy & Andreas Krause, editors, *Proceedings of Machine Learning Research*, volume 80, pages 2873–2882, Stockholm. PMLR.

Shalom Lappin & Jean-Philippe Bernardy, editors. in press. *Algebraic Structures in Natural Language*. CRC Press, Taylor & Francis, Boca Raton, London, New York.

Shalom Lappin. 2021. *Deep Learning and Linguistic Representation*. CRC Press, Taylor & Francis, Boca Raton, London, New York.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, & Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Tal Linzen, Emmanuel Dupoux, & Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Jean Maillard, Stephen Clark, & Dani Yogatama. 2019. Jointly learning sentence embeddings and syntax with unsupervised tree-lstms. *Natural Language Engineering*, 25(4):433–449.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, & Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh. Association for Computational Linguistics.

Kai Sheng Tai, Richard Socher, & Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing. Association for Computational Linguistics.

Aarne Talman & Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence. Association for Computational Linguistics.

Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, & Jörg Tiedemann. 2021. NLI data sanity check: Assessing the effect of data corruption on model performance. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 276–287, Reykjavik (Online). Linköping University Electronic Press.

Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, & Wang Ling. 2017. Learning to compose words into sentences with reinforcement learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, Conference Track Proceedings*.

The case for *Språkbanken Dialog*

Staffan Larsson **Christine Howes** **Eleni Gregoromichelaki**
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Sweden
staffan.larsson@ling.gu.se
christine.howes@gu.se
eleni.gregoromichelaki@gu.se

Abstract

We argue that the National Language Bank of Sweden should be extended with an additional infrastructure supporting research on linguistic interaction. Our main argument is that dialogue is not (just) text or speech, and consequently, that studying dialogue requires dialogue-specific infrastructure.

1 Introduction

One of Lars Borin's main achievements is, of course, his very successful development and management of Språkbanken, now a part of the National Language Bank of Sweden under the name Språkbanken Text. Following his lead and inspiring pioneering example, we would like to take this opportunity to argue the case for a Swedish Dialogue Bank - *Språkbanken Dialog*.

2 Språkbanken

Quoting from the homepage of the National Language Bank of Sweden, "The purpose of The National Language Bank of Sweden is to develop a national e-infrastructure supporting research in language technology, linguistics and other fields of study where research is conducted based on language data." Currently, the National Language Bank of Sweden has three parts – Språkbanken Text, Språkbanken Tal and Språkbanken Sam. These all provide important infrastructures for researchers in various fields. Språkbanken Text contains text corpora that can be searched for occurrences of words and phrases, including longitudinal data. Språkbanken Tal contains (or will contain when it is launched in 2023) recorded speech aligned with text for use in research and development of speech technologies. Språkbanken Sam contains text and some speech recordings focusing on (1) official multilingual texts and terminology for research in official communication and social conditions, and (2) folk narratives, as well as other text and speech material from the dialect and folklore archives.

The National Language Bank of Sweden is a significant achievement and a valuable resource for language technology purposes. However, a considerable lacuna, in our view, remains: these resources do not provide a comprehensive collection of spoken, written, and/or multimodal interactions in Swedish (and/or minority languages) that are available and searchable in the way that is needed to explore the *interactive* aspects of language use and structure. This is what we argue is still needed.

3 Dialogue

It is now widely accepted that human conversation does not consist of a sequence of sentences simply placed one after the other. There are specific phenomena that only become visible at the level of dialogical interaction, for example, so-called "grounding processes" (Clark, 1996), turn-taking (Sacks et al., 1974), repair (Schegloff et al., 1977), and multimodal input and output (Bavelas & Gerwing, 2007), which are

Staffan Larsson, Christine Howes and Eleni Gregoromichelaki. 2022. The case for *Språkbanken Dialog*. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 79–82. Available under CC BY 4.0

the features of dialogue that make it so much easier to process and engage in than monologue. On the other hand, phenomena which have been considered sentence-internal and requiring specialised syntactic/semantic mechanisms can be seen under a new, more illuminating, light when considered in the context of conversation. For example, phenomena like anaphora, ellipsis, syntactic/semantic dependencies, and speech act recognition/production can extend across turns and participants (see (3), below). In fact, it can be shown that such puzzling phenomena rely more crucially on interactive mechanisms for their resolution than individual processing capacities, a case of 'computational offloading' to the social environment (Gregoromichelaki, 2017). Across linguistics, psychology, philosophy, and cognitive science, it is now recognised that the primary ecological niche of language use is face-to-face interaction. Therefore, it has now become common to talk about the human 'interaction engine' (Levinson, 2020) to refer to the evolutionarily and culturally shaped linguistic skills and social capacities that are involved in language processing and general action coordination. Formal grammars, computational implementations, and linguistic/psycholinguistic theories now attempt to model formally and test experimentally these interactive processes to explain human linguistic cognition and behaviour (Ginzburg, 2012; Gregoromichelaki et al., 2020; Healey et al., 2018; Cooper, 2022).

In the field of language technology and AI, it is also becoming a familiar theme to address human interaction and conversation as the source of invaluable data. Many current architectures take advantage of training data from dialogue and multimodal corpora, whether annotated or not, and there is a recognition in recent work that large-scale language models – even those which make use of visual data – lack sufficient training data of conversational strategies such as repair (Lemon, 2022). Additionally, models increasingly seek to leverage interactive processes with human-in-the-loop teaching and supervision as a means of extending the capabilities of Large Language Models and artificial agents like social robots developing their trustworthiness, reliability, and alignment with human values.

As an illustration, let us look at an example of a dialogue with the sort of annotations we envision for Språkbanken Dialog:

(1) STANLEY: Louis, I$_{[ref\text{:}\textsc{Stanley}]}$ just didn't$_{[\text{NPI}-licensor]}$ think
`[[assertion; change of turn: split utterance]]`

LOUIS: you$_{[ref\text{:}\textsc{stanley}]}$'d ever$_{[\text{NPI}]}$ hear from me$_{[ref\text{:}\textsc{Louis}]}$?
`[[continuation & clarification & confirmation request & quotation]]`

[BBC Transcripts, *Dancing to the Edge*, Episode 5, example from: Gregoromichelaki (2017)]

Here the annotation needs to indicate the dialogue-act multifunctionality of subsentential turns. We also need to have information about the dependency between the Negative Polarity Item (NPI) *ever* and its licensor *n't* that occur in different turns by different speakers even though no single surface string can be syntactically reconstructed. In confirmation of this, it needs to be indicated in the annotation how the incremental change of speaker within a quotative clause reporting the first speaker's mental state ('Stanley$_{speaker}$ did not think || that Stanley$_{addressee}$ will hear from Louis$_{speaker}$') results in incremental switches in the interpretation of indexicals. This evidence of dependencies crossing turns and speakers render untenable any simple analysis of the shared string as a joined surface syntactic form with respect to the semantics:

(2) #Louis I just didn't think you'd ever hear from me.

In addition, it is demonstrated that grammatical analyses need to incorporate semantic and, crucially, pragmatic factors, e.g., turn-taking in dialogue, in order to provide a coherent and unified analysis of syntactic/semantic phenomena. Moreover, understanding both human psychological processes and the functioning of end-to-end models and AI architectures with respect to linguistic behaviour requires becoming aware and modelling such interactions of what have been standardly taken as separate modules of linguistic/non-linguistic knowledge in standard monological accounts.

4 Språkbanken Dialog

With this in mind, let us try to explain in more detail why Språkbanken Dialog is needed, and how we envision it.

Språkbanken Dialog is (would be) a large collection of linguistic interactions, including video recordings of face-to-face interactions, audio recordings of spoken interactions, transcribed interactions (aligned with the source video or sound recordings), and written interactions taken e.g. from social media and chat applications. It is possible to view, annotate and analyse individual interactions across multiple turns – something not currently offered by any Språkbanken resources. It is also possible to relate individual interactions to each other, e.g. temporally, spatially, or with respect to the speakers involved (while keeping to GDPR restrictions).

What about overlap with existing Språkbanken resources? It is true that other Språkbanken resources already contain linguistic interactions. In fact, as far as possible, such material should also be included in Språkbanken Dialog. However, none of the existing resources offer the possibility of adequately exploring the interactive aspects of these dialogues. In Språkbanken Text, interactions are treated as any other text, and it is not possible to see full interactions across several turns, nor to annotate or analyse them. The argument for Språkbanken Dialog rests on the fact that linguistic interaction is not reducible to, or analysable in terms of, individual words or phrases.

So maybe Språkbanken Dialog could just be a different interface to existing Språkbanken resources? Such a thing would certainly be useful, but there are also reasons to include additional resources not covered by other Språkbanken infrastructure. Currently, linguistic interactions are collected by researchers and students working on dialogue in the course of their research activities. This data can be in the form of text, audio, video, or some combination thereof. Currently, a lot of these resources never become available to other researchers. We believe that Språkbanken Dialog could offer infrastructure that would enable and encourage low-effort sharing, annotation and analysis of dialogue data (including multimodal data), thus boosting research on linguistic interaction in Swedish and other languages.

5 Future work

We leave for future work to fund, organise and implement Språkbanken Dialog. In this, we hope to follow Lars Borin's inspiring example.

Acknowledgements

We acknowledge support from the Swedish Research Council VR project 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Janet B. Bavelas & Jennifer Gerwing. 2007. Conversational hand gestures and facial displays in face-to-face dialogue. *Frontiers of social psychology: Social communication*, pages 283–307.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Robin Cooper. 2022. *From perception to communication: An analysis of meaning and action using a theory of types with records (TTR)*. Oxford University Press, Oxford. to appear.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Eleni Gregoromichelaki, Gregory J. Mills, Christine Howes, Arash Eshghi, Stergios Chatzikyriakidis, Matthew Purver, Ruth Kempson, Ronnie Cann, & Patrick G. T. Healey. 2020. Completability vs (in)completeness. *Acta Linguistica Hafniensia*.

Eleni Gregoromichelaki. 2017. Quotation in Dialogue. In *The Semantics and Pragmatics of Quotation*, pages 195–255. Springer.

Patrick G. T. Healey, Gregory J. Mills, Arash Eshghi, & Christine Howes. 2018. Running Repairs: Co-ordinating Meaning in Dialogue. *Topics in Cognitive Science*, 10(2):367–388.

Oliver Lemon. 2022. Conversational grounding in emergent communication–data and divergence. In *Emergent Communication Workshop at ICLR 2022*.

Stephen C Levinson. 2020. On the human “interaction engine”. In N. J. Enfield & S.C. Levinson, editors, *Roots of human sociality: Culture, cognition and interaction*, pages 39–69. Routledge.

Harvey Sacks, E.A. Schegloff, & Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

E.A. Schegloff, G. Jefferson, & H. Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Ordbildning på icke-verbal grund
Om integrering av ikoner och ljudeffekter i grammatiken

Benjamin Lyngfelt
University of Gothenburg
Sweden
benjamin.lyngfelt@svenska.gu.se

Joel Olofsson
University West
Trollhättan, Sweden
joel.olofsson@hv.se

Abstract

In this paper we discuss a couple of examples of morphosyntactic integration of non-verbal material in written Swedish, specifically icons and music beats. It is argued that classification in terms of parts-of-speech is not a particularly fruitful approach in these cases. Instead, we suggest a constructional perspective and briefly illustrate how these phenomena may be analyzed in terms of constructions.

1 Inledning

Kommunikation kan vara multimodal på flera olika vis (se t.ex. Björkvall, 2009). I den här texten inriktar vi oss på hur ljud och bild kan integreras inte bara i texten utan även morfosyntaktiskt. Verbalisering av icke-verbala ljud är förvisso ingenting nytt; det är själva grunden för onomatopoetiska uttryck. Desto nyare är de möjligheter att integrera bilder i alfabetisk skrift som har utvecklats i s.k. e-kommunikation. De senaste årtiondena har först emojier och sedan andra ikoner tagit plats i skriftspråksrepertoaren, först som självständiga indikationer om meddelandets funktion (👍, 😜)[1] eller skribentens sinnesstämning (😃, 😢, 😍) och senare även mer integrerat med den verbalt uttryckta skriften (se McCulloch, 2019). Detta illustreras i (1), där ikonen 🔺 utgör förled i en sammansättning; dessutom innehåller exemplet det ljudhärmande verbuttrycket *oontz, oontz, oontzade*:

(1) ”Bang local MILFS” stod det på 🔺-bilen som oontz, oontz, oontzade förbi när jag väntade utanför sextonåringens skola. Gulligt ändå. (Exemplet hämtat från Twitter; 26 september 2022.)

I sammansättningen 🔺*-bilen* fungerar 🔺 som förled till efterledet *bilen* och anger vilken sorts bil det handlar om. I detta fall handlar det om en s.k. A-traktor (även kallad *EPA-traktor* och *LGF-fordon*), och 🔺 avser här den varningsskylt som sitter baktill på sådana långsamtgående fordon. I sig är 🔺 en generell varningssymbol; det är först när den placeras på en bil som den betyder ‘A-traktor’. Uttrycket 🔺*-bilen* påminner om nominala sammansättningar av typen [N-N] (jfr *brandbil, elbil, glassbil*), vilket skulle kunna tyda på att ikonen 🔺 tolkas nominalt i sammanhanget. Innebär det i så fall att den är att betrakta som ett substantiv?

Oontz, oontz, oontza är ett ljudhärmande uttryck som kan sägas fungera som ett onomatopoetiskt rörelseverb. På samma sätt som förledet i 🔺*-bilen* avser ett utmärkande drag hos A-traktorer, refererar *oontz, oontz, oontza* till repetitivt dunkande musik (jfr det äldre uttrycket *dunka-dunka*). Den repetitiva karaktären återspeglas i det språkliga uttrycket, med (minst) två *oontz* före verbformen *oontza*. Det senare följer ett produktivt verbbildande avledningsmönster som typiskt har substantivrot. Är alltså *oontz* därmed är ett substantiv?

I den här artikeln diskuterar vi ett par exempel på hur åtminstone delvis icke-verbala uttryck integreras morfosyntaktiskt i svenskt skriftspråk och hur detta kan analyseras grammatiskt. Vi ifrågasätter ordklasser som primärt analysverktyg i sammanhanget och förordar i stället ett konstruktionsbaserat perspektiv.

[1] The Emoji graphics used in this paper are taken from the Twemoji project, copyright 2020 Twitter Inc and other contributors, licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/

Benjamin Lyngfelt and Joel Olofsson. 2022. Ordbildning på icke-verbal grund. Om integrering av ikoner och ljudeffekter i grammatiken. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 83–87. Available under CC BY 4.0

2 Ikoner och ordklasser

🔺*-bilen* är en nominal sammansättning, dvs. bildar som helhet ett substantiv.[2] Även efterledet, som är sammansättningens semantiska och syntaktiska huvud, är ett substantiv. Däremot råder inga särskilda ordklassrestriktioner på förledet. Det vanligaste är substantiv även här, [N-N], men förledet kan också utgöras av t.ex. adjektiv (*småbil*), verb (*hyrbil*) eller som synes en ikon (🔺*-bil*). Redan när det gäller lexikala förled är ordklasstillhörigheten inte alltid helt självklar; utgår t.ex. *lastbil* från substantivet *last* eller verbet *lasta*, och *lyxbil* från substantivet *lyx* eller adjektivet *lyxig*?

Eftersom 🔺*-bilen* som sagt påminner om [N-N]-sammansättningar och ikonen 🔺 betecknar ett föremål kan det ligga nära till hands att uppfatta ikonen som ett substantiv. Samtidigt fungerar den också som en varningssignal och liknar i det avseendet snarare en interjektion. Detta är ingen exceptionell egenskap hos just 🔺, utan ikoner framstår i allmänhet som notoriskt flerfunktionella sett från ett grammatiskt perspektiv. Exempelvis kan ✈️ tolkas både som substantiv (*flygplan*) och som verb (*flyga*/*resa*), eller till och med som ett helt yttrande (*jag flyger till Indien*). Ett annat exempel är 🥳 som kan tolkas som substantiv (*fest*), verb (*festa*), adjektiv (*glad/partysugen*) och/eller interjektion (*Hurra!*).

Att följa den traditionella klassificeringsprincipen "olika ordklass – olika ord" skulle alltså leda till massiv homonymi bland ikonerna. Frågan är om ordklassbestämning alls är meningsfull för ikoner. Rimligen avser ordklass hos en lexikal enhet en konventionaliserad uppsättning morfosyntaktiska egenskaper, men vanliga ordklasskriterier är till föga hjälp här (Adesam & Bouma, 2019). Ikoner saknar helt morfosyntaktisk markering, svarar dåligt mot distributionstester, tenderar att vara flerfunktionella och konventionaliseringen av deras grammatiska beteende är vanligtvis inte särskilt långt gången (även om vissa drag kan ha konventionaliserats; jfr McCulloch, 2019). Enligt taggningsmodellen Koala, som används i Språkbanken Text, klassas de helt enkelt som symboler (Adesam & Bouma, 2019).

Inte heller ljudeffekten *oontz* i uttrycket *oontz, oontz, oontzade* har någon tydlig ordklasstillhörighet. Som fristående ljud liknar uttrycket närmast en interjektion, i likhet med ljudhärmande interjektioner som *bom*, *klang* och *smack*. Samtidigt kan det i vissa avseenden bete sig som ett substantiv: ett/flera *oontz*. Och i exempel (1) ingår det i ett verb. I Koala hanteras uttryck av det här slaget som *foreign material*, utan närmare specificering (Adesam & Bouma, 2019).

Kort sagt, hur man tolkar ikoner och ljudhärmande uttryck, liksom för den delen andra språkliga enheter, beror i stor utsträckning på kontexten. Ordklasser, annars högst användbara för flera typer av språkvetenskaplig analys (jfr t.ex. Adesam & Bouma, 2019; Kalm 2021), är kanske inte det mest fruktbara angreppssättet just i det här sammanhanget.

3 Avledningar

Verbavledning är synnerligen produktivt i svenskan; vi kan bilda verb genom att kombinera en verbändelse med snart sagt vad som helst som kan tolkas som någon sorts aktivitet. Kända exempel är *vabb+a*, *facebook+a* och *sol-och-vår+a*, medan *oontz, oontz, oontz+ade* i exemplet ovan är en mindre etablerad bildning. Avledning genom suffix antas i traditionell grammatik innebära ordklassbyte, men i likhet med ikoner är det som sagt tveksamt om *oontz, oontz, oontz* kan sägas ha någon given ordklasstillhörighet.

Ordklassproblematiken blir ännu mer påtaglig om vi går ett steg till och bildar substantiv (*ett evinnerligt oontz, oontz, oontzande*) eller particip/adjektiv (*ett oontz, oontz, oontzande dansgolv*). Både s.k. verbalsubstantiv och participer bildas typiskt av verb, men en sådan analys skulle här innebära en föga plausibel tvåstegsprocess där man först bildar verbet *oontz, oontz, oontza* och sedan avleder detta verb vidare till substantiv eller particip/adjektiv. För den nominala användningen passar således den latinska termen *nomen actionis* bättre än den svenska verbalsubstantiv här – ett 'nomen som uttrycker en aktion' snarare än ett 'substantiv bildat av ett verb'. Andra exempel på *nomen actionis* utan verbstam är *hemma hos-ande* och *GI-ande* (se Holmer, 2022).

Om vi återgår till verbet kan *oontz, oontz, oontzade* sägas fungera som ett rörelseverb i exemplet; i alla händelser ingår det i rörelseuttrycket *oontz, oontz, oontzade förbi* och fyller där den roll som typiskt uttrycks av ett rörelseverb (jfr *körde förbi*). Det verkar för övrigt vara ett ganska produktivt mönster att

[2] Sammansättningar i svenskan behandlas utförligt av bl.a. Svanlund (2009) och Loenheim (2019); se även Teleman et al. (1999).

använda ljudhärmande verb i rörelseuttryck; jfr *susa förbi*, *skramla iväg* och *braka in i ngt* (Olofsson, 2018). Med tiden kan associationen till rörelse lexikaliseras, som i fallet *susa*, men det skulle förvåna om även *oontz, oontz, oontza* når dit. Ett tecken på en sådan utveckling skulle annars vara om uttrycket reducerades till enbart *oontza*.

4 Diskussion: ordbildning genom konstruktioner

Hur ska vi då analysera *oontz, oontz, oontzade* och 🔺*-bilen?* Att postulera (tillfälliga) lexikala enheter med tillhörande ordklassegenskaper m.m. förefaller både omständligt och implausibelt. Istället föreslår vi att de nominala resp. verbala associationerna uppstår i kontexten (s.k. emergens), såsom antas inom bl.a. konstruktionsgrammatik. Konstruktionsgrammatik är förmodligen mest känd för sin hantering av olika typer av flerordsmönster, men har också framgångsrikt tillämpats på morfologi (t.ex. Booij, 2010). Således kan nybildningarna i (1) hanteras som instanser av en sammansättningskonstruktion resp. en verbkonstruktion.

4.1 Verb och sammansättningar som konstruktioner

En konstruktion består av ett eller flera konstruktionselement, som vardera fyller en viss funktion i konstruktionen och förknippas med en (mer eller mindre specificerad) uppsättning egenskaper. Om vi börjar med verbkonstruktionen kan den sägas innehålla två element: dels en STAM, dels GRAMMATISK MARKERING (GM) av främst tempus. I *oontz, oontz, oontzade* fungerar [*oontz, oontz, oontz*] som stam, medan [*-ade*] utgör GM och markerar såväl verbkategori som preteritum. I vanliga fall utgörs förstås stammen av ett etablerat verb, som redan förknippas med verbala egenskaper. Dessa verb utgör specifika typer av konstruktionen och har lexikaliserat GM-dragen. I fallet *oontz, oontz, oontzade* är dock stammen inget verb i sig, utan konstrueras som en aktivitet just genom att ingå i en verbkonstruktion. Skillnaden mellan detta synsätt och mer strikt kompositionella modeller är alltså att konstruktionselementens egenskaper inte måste vara lexikalt givna av de ingående leden, utan också kan följa av den konstruktion de ingår i.

Även rörelsebetydelsen antas följa av en konstruktion, närmare bestämt en syntaktisk konstruktion med elementen VERB och RIKTINGSADVERBIAL: *oontz, oontz, oontzade förbi*. Detta mönster uppträder som sagt oftast med etablerade rörelseverb, men verb som inte har denna betydelse inherent kan alltså få det i just den här konstruktionen. Det avgörande villkoret är att verbet uttrycker en betydelse som kan associeras med rörelse, i det aktuella fallet ljud som hörs från den körande A-traktorn. Rörelsekonstruktioner av det här slaget behandlas utförligt i Olofsson (2018).

På liknande sätt kan vi hantera 🔺*-bilen*. Uttrycket antas instansiera en nominal sammansättningskonstruktion med elementen FÖRLED, EFTERLED och GM. Här uttrycker GM nominala egenskaper som t.ex. definithet och fogas till efterledet. Eftersom nominala sammansättningar genomgående är determinativa måste förledet kunna förstås som en specificering av efterledet. I övrigt lägger konstruktionen inga särskilda restriktioner på vare sig ordklass eller andra egenskaper hos förledet, mer än att det rent praktiskt måste kunna kombineras med efterledet.[3]

Om 🔺*-bilen* också ska förstås mer specifikt som en sammansättning av typen [N-N] fordras därutöver att 🔺 tolkas nominalt. Här finns inte utrymme för en utredning av begreppet ‘nominal funktion’ (jfr Teleman et al., 1999), men grovt förenklat kan det sägas innebära att 🔺*-bilen* tolkas analogt med *firmabilen* och *leksaksbilen* snarare än med (det mer verbala) *hyrbilen*. Inte heller det kräver dock att 🔺 analyseras som ett substantiv, vare sig grammatiskt eller som lexikal enhet.

4.2 Unifiering och s.k. coercion

Utgångspunkten för analysen är att konstruktionerna utgör mönster, som vid språkanvändning instansieras av konkreta språkliga uttryck. Rent tekniskt förenas konstruktionselementen och de uttryck som realiserar dem genom s.k. unifiering (Fillmore & Kay, 1993; Fried & Östman, 2004). Elementen och deras instansieringar behöver inte ha identiska egenskaper; det räcker att de är kompatibla. I de allra flesta

[3] Det ställs mer specifika krav på efterledet, som utgör både syntaktiskt och semantiskt huvud och behöver kunna kombineras med ev. GM. Efterledet behöver alltså både formellt och funktionellt fungera som ett nominal. Därför fungerar ☀️*-semestern* bättre än *semester-*☀️*-en*; den oböjda varianten *semester-*☀️ är dock tänkbar.

fall är detta tämligen oproblematiskt. Mer intressant blir det när egenskaperna hos element och instansieringar inte matchar fullt ut. Så är bl.a. fallet när verb uppträder i syntaktiska strukturer som krockar med verbets valens, som i exemplet *Kan man äta bort sin huvudvärk?*. Objektet till *äta* uttrycker normalt det ätna, vilket här knappast är huvudvärken. I stället associeras objektet med *bort* (det som ska bort) snarare än tolkas som föremål för verbhandlingen. Denna tolkning uppstår just i konstruktionen [verb + *bort* + objekt]; jfr ?*äta sin huvudvärk* (Sjögreen d.y., 2015).

Den här sortens fenomen, där språkliga uttryck används på sätt som krockar med deras konventionaliserade egenskaper, brukar beskrivas i termer av *coercion*. I konstruktionsperspektiv innebär coercion att en konstruktion ”kör över” lexikala drag hos de ingående leden, som därmed anpassas till villkoren på de konstruktionselement de unifieras med (t.ex. Michaelis, 2005). Det kan alltså betraktas som coercion att foga in 🔺 som förled i en sammansättning trots att ikonen i sig inte har sådana morfologiska egenskaper. Termen *coercion* må signalera att vi begår någon form av persuasivt våld på språket, men företeelsen är egentligen bara ett specialfall av att språkliga uttryck anpassas till sin kontext – här i form av den språkliga konstruktion uttrycket ingår i.

Konstruktionell coercion är emellertid ett kraftfullt verktyg, som utan begränsningar skulle öppna för att övergenerera å det vildaste. Så varför producerar vi inte jämt och ständigt en uppsjö av konventionsbrytande nybildningar? Den viktigaste och mest generella begränsningen är förmodligen vanans makt. Vi språkbrukare är inte så kreativa som vi kanske vill tro, i varje fall inte hela tiden, utan det är i regel smidigast för både sändare och mottagare att använda vanliga, lättillgängliga och förväntade uttryckssätt. En mer specifik faktor, som har föreslagits som förklaring till nyckfulla begränsningar i partiell produktivitet, är konkurrens: ett uttryck kan vara disprefererat därför att det trängs undan av mer etablerade alternativ (Goldberg, 2019).

Slutligen bör vi inte glömma att språkbruk inte enbart styrs av begränsningar. Det fordras också positiva krafter, såsom motivation och kontextuell relevans – en anledning att använda ett uttryck och kanske någonting som gynnar valet av just det uttryckssättet. En gynnande kontext för bruk av ikoner i sammansättningar skulle t.ex. kunna vara en 🥳-skrift till Lars Borin.

Referenser

Yvonne Adesam & Gerlof Bouma. 2019. The Koala part-of-speech tagset. *Northern European Journal of Language Technology*, 6(2):5–41.

Anders Björkvall. 2009. *Den visuella texten. Multimodal analys i praktiken*, volume 40 of *Ord och stil. Språkvårdssamfundets skrifter*. Hallgren & Fallgren, Stockholm.

Geert Booij. 2010. *Construction Morphology.* Oxford University Press, Oxford.

Charles J. Fillmore & Paul Kay. 1993. Construction grammar coursebook. Opublicerat manuskript. Department of Linguistics, University of California, Berkeley.

Mirjam Fried & Jan-Ola Östman. 2004. Construction grammar: A thumbnail sketch. In Mirjam Fried & Jan-Ola Östman, editors, *Construction Grammar in a Cross-Language Perspective*, volume 2 of *Constructional Approaches to Langu- age*, pages 11–86. John Benjamins, Amsterdam.

Adele E. Goldberg. 2019. *Explain me this: creativity, competition, and the partial productivity of constructions*. Princeton University Press, New Jersey.

Louise Holmer. 2022. *Neutrala substantiv på* -ande *i text och ordbok.* Ph.D. thesis, Göteborgs universitet.

Mikael Kalm. 2021. Ordklasser – finns de? In Johan Brandtler & Mikael Kalm, editors, *Nyanser av grammatik: gränser, mångfald, fördjupning*, pages 23–42. Studentlitteratur, Lund.

Lisa Loenheim. 2019. *Att tolka det sammansatta. Befästning och mönster i första- och andraspråkstalares tolkning av sammansättning*. Ph.D. thesis, Meijerbergs Arkiv för Svensk Ordforskning, Göteborg.

Gretchen McCulloch. 2019. *Because internet: understanding the new rules of language.* Riverhead Books, New York.

Laura Michaelis. 2005. Entity and event coercion in a symbolic theory of syntax. In Jan-Ola Östman & Mirjam Fried, editors, *Construction Grammar(s): Cognitive Grounding and Theoretical Extensions*, page 45–87. John Benjamins, Amsterdam.

Joel Olofsson. 2018. *Förflyttning på svenska. Om syntaktisk produktivitet utifrån ett konstruktionsperspektiv*. Ph.D. thesis, Göteborgs universitet, Göteborg.

Christian Sjögreen d.y. 2015. *Kasta bort bollen och äta bort sin huvudvärk. En studie av argumentstrukturen i kausativa bort-konstruktioner*. Ph.D. thesis, Uppsala universitet, Uppsala.

Jan Svanlund. 2009. *Lexikal etablering. En korpusundersökning av hur nya sammansättningar konventionaliseras och får sin betydelse*. Acta universitatis Stockholmiensis, Stockholm.

Ulf Teleman, Staffan Hellberg, & Erik Andersson. 1999. *Svenska Akademiens grammatik*. Svenska Akademien, Stockholm.

Building a multilingual AWE tool for L2 learners: Challenges and ideas

Arianna Masciolini
Språkbanken Text
University of Gothenburg, Sweden
arianna.masciolini@gu.se

Abstract

Language tools are a valuable ICALL resource, especially for learners of a second language. In this work, we discuss the challenges involved in the development of a multilingual Automatic Writing Evalutation tool addressing their specific needs and brainstorm some potential solutions.

1 Introduction

It is common to identify *ICALL* (Intelligent Computer Assisted Language Learning) applications with *ILTSs* (Intelligent Language Tutoring Systems). That of ICALL, however, is a broader category, which also includes various types of *language tools* (Heift & Vyatkina, 2017), such as online dictionaries, MT (Machine Translation) software, spell checkers and morphological analyzers. Usually designed not specifically with learners in mind, but rather for the general population, these tools are especially important - perhaps more than ILTSs - for intermediate and advanced L2 (Second Language) learners, who, as opposed to students learning a FL (Foreign Language), tend to acquire language in its context of use, often without receiving any formal instruction (Kramsch, 2000).

In this category, AWE (Automatic Writing Evalutation) tools have recently started emerging, Grammarly being perhaps the most widespread example.[1] Based on an automatic analysis of the learner's text, AWE software can provide different kinds of feedback, from numerical scores to corrections and stylistic suggestions (Hockly, 2018). Some AWE tools targeting teachers are used in standardized tests, while tools like Grammarly, used both by natives and learners, are increasingly popular outside the classroom.

With such a large target group, however, they are not always able to address the specific needs of learners in the best way possible. First, the majority of AWE tools targeting the writer (rather than the grader) provide them with corrections but no explanations. Furthermore, the analyses these systems perform are not specifically focused on L2 learner errors, which may well differ from those typical of L1 users of the same language. In addition, most AWE tools are monolingual, and adapting them to other languages, when at all possible, requires a large amount of data, which is often unavailable.

We are interested in building an AWE tool that addresses these issues. We focus primarily on grammaticality, aiming for a system able to provide both corrections and verbal explanations in a potentially wide variety of languages, targeting learners with different levels of proficiency and metalinguistic awareness. We propose an approach where, after obtaining a correction hypothesis for the learner's input, both the original text and its correction are processed with an UD (Universal Dependencies) parser (de Marneffe et al., 2021). The two obtained treebanks are then compared by an error analysis module, which outputs a lossless structured representation of the errors it detects based on the discrepancies between them. Finally, these structured data are converted to human-readable feedback through a domain-specific CNL (Controlled Natural Language).

In the following, we go through the various steps one by one, discussing potential solutions to the problems the implementation of each of them arises. Section 2 is dedicated to the question of how to obtain a good correction hypothesis. Section 3 discusses the challenging aspects of parsing while Section 4

[1] grammarly.com

Arianna Masciolini. 2022. Building a multilingual AWE tool for L2 learners: Challenges and ideas. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 89–93. Available under CC BY 4.0

focuses on how to use the results of this step for error analysis. After that, Section 5 revolves around generating metalinguistic feedback from structured data. We then close with some concluding remarks.

2 Obtaining a correction hypothesis

As mentioned in the Introduction, our system provides feedback based on a comparison between the user input and a corrected version of the same text. The first processing step is therefore that of obtaining a _correction hypothesis_. We use this expression to emphasize the fact that every correction is based on some interpretation of what the writer meant to express through their text. As an example, consider the ungrammatical sentence “_*This are my *contribute to the Festschrift in honor of Lars Borin_”. A possible correction is “_This is my contribution to the Festschrift in honor of Lars Borin_”, but the author might have instead meant to say that they made several contributions, the proper correction thus becoming “_These are my contributions to the Festschrift in honor of Lars Borin_”.

For this reason, obtaining a correction hypothesis with a GEC (Grammatical Error Correction) tool, is, while certainly an option to take into consideration, not necessarily optimal; the results obtained with one such tool are not guaranteed to match the learner’s intentions and can therefore be confusing. Aside from this, the amount of GEC software available is still quite limited and performance is uneven across languages. When it comes to Swedish, for example, both more dated software such as the hybrid rule-based/probabilistic tool Granska (Domeij et al., 2000) and the most recent neural approaches (Nyberg, 2022) still present some weaknesses, especially when it comes to longer sentences containing several and/or multi-token errors.

In an interactive system, an alternative solution exploiting the learner’s L1 competence is to use MT. This approach would consist in translating the user input to the learner’s L1 (or any other language that they selected as the instruction language) and let them adjust the result to clarify their intentions. The resulting L1 text would then be translated back to the L2, producing a correction hypothesis that hopefully matches the learner’s expectations. Among several translation candidates, the closest one to the original user attempt could be selected based on a metric such as the BLEU score (Papineni et al., 2002). The learner’s L1 being Swedish, for instance, our example sentence could be automatically translated to “_Detta är **mitt** bidrag till Festschrift till Lars Borins ära_”, but the user would then be given a chance to intervene, specifying that they meant “_Detta är **mina** bidrag till Festschrift till Lars Borins ära._” Two advantages of back-and-forth translation are its awareness of the learner’s intentions and its high multilinguality, as MT tools are nowadays available for a vast amount of language pairs. Translation errors can of course pose problems, especially in the L1-to-L2 direction, where users cannot intervene, but we expect this issue to be mitigated by the tendency of learner language to be relatively simple.

3 Parsing

For the morphosyntactic analysis of learner text, we propose using a UD parser, i.e. a dependency parser outputting CoNNL-U files following the Universal Dependencies guidelines (de Marneffe et al., 2021). As the name suggests, UD is a framework for cross-linguistically consistent grammatical annotation: the scheme is largely identical across languages and even language-specific features are annotated following shared principles. This significantly simplifies the task of working with several L2s, which is one of our main ambitions. Furthermore, state-of-the-art UD parsers such as UDPipe (Straka, 2018) are remarkably accurate, fast, open source and easy to use.

While processing correction hypotheses should therefore be unproblematic, learner language poses significant challenges. In their systematic study of the performance of dependency parsers on learner English, Huang et al. (2018) have shown that, while their overall accuracy stays reasonably high even for L2 text, they are not robust to grammatical errors. The overall good scores seem in fact to occur due to errors being sparse and learner sentences being shorter and simpler than those written by native or otherwise highly proficient users of the same language, not to mention that the study only takes dependency labels and Part Of Speech (POS) tags into account, thus not giving any information about the accuracy of annotation when it comes to, for instance, incorrectly inflected items.

There have been some efforts to develop parsers specifically meant for learner language. Sakaguchi et al. (2017), for instance, have proposed an error-repairing architecture capable of dealing with a variety of single-token errors. Another, perhaps more straightforward approach could be training an existing UD parser on manually or semi-automatically annotated learner data. As we will see in Section 4, some UD-annotated parallel learner treebanks are in fact available, but both the number of languages for which these resources already exist and their sizes are most likely insufficient. We suggest that, in our context of application, the parsing of learner sentences could be informed by that of the corresponding correction hypotheses. In practice, the annotation of learner attempts could consist in postprocessing the corresponding corrected sentences annotated with a standard UD parser.

4 Analyzing errors

We propose framing error analysis, which lies at the heart of our hypothetical AWE tool, as a tree comparison task. More specifically, this means operating on a *parallel learner treebank* (or, to use a term coined in Lee et al. (2017b), an *L1-L2 treebank*), i.e. a dependency corpus where learner sentences are aligned with the corresponding correction hypotheses. This format was originally designed to address the interoperability issues arising from the coexistence of different markup styles and tagsets used for annotating learner corpora, usually employed to retreive occurrences of particular error patterns. The idea is to, rather than defining a universal error taxonomy, simply annotate both learner sentences and correction hypotheses according to the UD guidelines, to then retreive errors via tree queries. In our case, parallel learner treebanks, so far usually handcrafted (Berzak et al., 2016; Lee et al., 2017a; Di Nuovo et al., 2019), are to be obtained via automatic parsing (see Section 3).

With an approach similar to what we have in mind, parallel learner treebanks have been used to derive error taxonomies dynamically (Choshen et al., 2020). The method is conceptually simple: given a portion of a learner sentence containing a grammatical error and the corresponding correction, errors are found by selecting the parsed learner substring node closest to the root and checking whether its counterpart in the correction has the same UD label and POS tag or not. If that is not the case, an error has been found. Its class is defined as the ordered pair of diverging UD labels or POS tags, token additions and deletions being a special case where the pair lacks one of its elements. Of course, in this way errors that do not involve an UD label or a POS tag change, such as incorrectly inflected words, are left out. To address the issue, the FEATS field of CoNNL-U files, reserved for morphological analysis, is also taken into account and errors of this kind are labeled with the morphological feature(s) they retain. Assuming that “_This is my contribution to the Festschrift in honor of Lars Borin_” is a suitable correction of “_*This are my *contribute to the Festschrift in honor of Lars Borin_”, then, the first (inflection) error would be labelled `Mood=Ind|Person=3|Tense=Pres|VerbForm=Fin` (dropping “are”’s `Number=Plur`), while the second’s category would be `VERB`→`NOUN`.

While Choshen et al. (2020)’s work is only concerned with classification, albeit fine-grained and dynamic, our error analysis module is intended to output lossless representations of each error. As we will discuss in Section 5, these are to be converted into human-readable feedback of arbitrary granularity only at a later stage, by a separate program, which implies machine-readability as another requirement for our data format. In addition, it seems that the simple error labeling algorithm we described, implemented as an open source program which we unsuccessfully tried to run, discards too much of the information available in CoNNL-U files and do not see ways for it to work effectively if not on single-token errors.

The simplest solution could be representing errors as pairs of UD subtrees representing a fragment of the learner’s sentence and its correction, aligned with a variant of the approach proposed by Masciolini & Ranta (2021). However, this would mean carrying additional information such as the specific word forms used and the features that remain the same even after correction has taken place. Not only is this superfluous for the feedback step: it also prevents us from using our data format for the task, complementary to feedback generation, of error retrival, which was Lee et al. (2017b)’s original goal with L1-L2 treebanks. A good basis for defining a better error format could be `hst`, the Haskell-embedded DSL (Domain Specific Language) used for pattern matching UD trees in `gf-ud` (Kolachina & Ranta, 2016; Ranta & Kolachina, 2017)[2].

[2] The pattern matching language documentation can be found at `github.com/GrammaticalFramework/gf-ud/blob/`

5 Generating feedback

As mentioned in the introduction, automatic feedback comes in different forms. Since our hypothetical AWE tool has the learners themselves as its intended users, we are interested in generating feedback that they can understand and make use of to independently improve their texts and acquire new grammatical knowledge. In her study on the impact of corrective feedback on learner uptake, Heift (2004) distinguishes three approaches commonly used in ICALL system: *recasting*, which implies showing correction hypotheses as replacement suggestions, *highlighting*, consisting in showing only the location of the error(s), and *providing metalinguistic feedback*, i.e. giving verbal explanations of what causes a sentence to be incorrect. While the former two do not require all of the processing steps we have so far described, we focus on the latter, which has proven to be an especially effective way to incentivize learners to reflect on their own errors and correct them (Heift, 2004).

In our setting, and in particular after the error analysis step, feedback generation can be seen as a data-to-text conversion task. Crucially in a multilingual setting, metalinguistic feedback should be available in several languages. Moreover, there should be a possibility to adjust the feedback to the learner's level of proficiency and/or metalinguistic awareness and, ideally, to also have the possibility to output labels belonging to some error taxonomy, rather than extended explanations. For these reasons, we think a CNL implemented in GF (Grammatical Framework), a well-established programming language for multilingual grammar engineering, would be suitable for the job. In GF, grammars are composed of an *abstract syntax*, playing the role of an interlingua, and one or more *concrete syntaxes*, capturing the specificities of the various languages. Translating implies parsing a string in the source language to an AST (Abstract Syntax Tree) and then linearizing it to a new string in the target language. Designed for building multilingual applications, GF makes it relatively easy to develop *semantic* or *application grammars*, i.e. domain-specific CNLs, by re-using rules defined by the large-scale syntactic grammars of over 40 languages which constitute GF's "standard library", usually referred to as the RGL (Resource Grammar Library) (Ranta, 2011).

Using GF, our task can become that of defining an application grammar that has the error descriptions outputted by the error analysis module as one of the concrete syntaxes. To that, we can add an arbitrary number of additional concrete syntaxes for verbal feedback in different languages and at different levels of granularity, ranging from labels to exhaustive explanations. In terms of grammar engineering, this is not a trivial task, as the definition of the abstract syntax resembles that of a novel, albeit more flexible, error taxonomy, derived in this case from the actual errors, possibly incrementally, and not defined *a priori*. An alternative path yet to be explored is trying to exploit the interoperability between UD and GF (Kolachina & Ranta, 2016; Ranta & Kolachina, 2017): if errors are represented as (simplified) UD subtrees, they could be automatically converted to GF ASTs, thus making the definition of a new abstract syntax unnecessary.

6 Concluding remarks

In this work, we have expressed our interest in building a multilingual AWE tool for L2 learners. We discussed some of the challenges associated with the task and brainstormed potential solutions. We hope the open problems we presented, which will be the object of our future work, to have sparked the reader's interest in the topic.

References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, & Boris Katz. 2016. Universal Dependencies for learner English. *arXiv preprint arXiv:1605.04278*.

Leshem Choshen, Dmitry Nikolaev, Yevgeni Berzak, & Omri Abend. 2020. Classifying syntactic errors in learner language. *arXiv preprint arXiv:2010.11032*.

`master/doc/patterns.md`

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, & Manuela Sanguinetti. 2019. Towards an Italian learner treebank in Universal Dependencies. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Rickard Domeij, Ola Knutsson, Johan Carlberger, & Viggo Kann. 2000. Granska–an efficient hybrid system for Swedish grammar checking. In *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*, pages 49–56.

Trude Heift & Nina Vyatkina. 2017. Technologies for teaching and learning L2 grammar. *The handbook of technology and second language teaching and learning*, pages 26–44.

Trude Heift. 2004. Corrective feedback and learner uptake in CALL. *ReCALL*, 16(2):416–431.

Nicky Hockly. 2018. Automated writing evaluation. *ELT Journal*, 73(1):82–88.

Yan Huang, Akira Murakami, Theodora Alexopoulou, & Anna Korhonen. 2018. Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1):28–54.

Prasanth Kolachina & Aarnte Ranta. 2016. From abstract syntax to universal dependencies. In *Linguistic Issues in Language Technology, Volume 13, 2016*.

Claire Kramsch. 2000. Second language acquisition, applied linguistics, and the teaching of foreign languages. *The Modern Language Journal*, 84(3):311–326.

John SY Lee, Herman Leung, & Keying Li. 2017a. Towards Universal Dependencies for learner Chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71.

John SY Lee, Keying Li, & Herman Leung. 2017b. L1-L2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 44–49.

Arianna Masciolini & Aarne Ranta. 2021. Grammar-based concept alignment for domain-specific machine translation. In *Proceedings of the Seventh International Workshop on Controlled Natural Language (CNL 2020/21)*.

Martina Nyberg. 2022. Grammatical error correction for learners of Swedish as a second language. Master's thesis, Uppsala Universitet.

Kishore Papineni, Salim Roukos, Todd Ward, & Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Aarne Ranta & Prasanth Kolachina. 2017. From universal dependencies to abstract syntax. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 107–116.

Aarne Ranta. 2011. *Grammatical framework: Programming with multilingual grammars*, volume 173. CSLI Publications, Center for the Study of Language and Information Stanford.

Keisuke Sakaguchi, Matt Post, & Benjamin Van Durme. 2017. Error-repair dependency parsing for ungrammatical texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–195.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels. Association for Computational Linguistics.

Flexible Universal Dependencies using nested dependency graphs

Joakim Nivre
Department of Computer Science
RISE Research Institutes of Sweden
joakim.nivre@ri.se

Department of Linguistics and Philology
Uppsala University, Sweden
joakim.nivre@lingfil.uu.se

Abstract

Universal Dependendencies (UD) is a framework for cross-linguistically consistent morphosyntactic annotation, which has to date been applied to 130 languages. Despite widespread adoption of UD, there are a number of issues in the annotation guidelines that continue to raise debate, such as the treatment of function words and the criteria for word segmentation. In this article, I sketch an extension of the UD framework, which may allow these and other issues to be resolved by offering more flexibility in the analysis of certain linguistic phenomena.

1 Introduction

Universal Dependencies (UD) is a project that develops cross-linguistically consistent morphosyntactic annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective (Nivre et al., 2016; Nivre et al., 2020; de Marneffe et al., 2021). Since its start in 2014, the project has grown into a large community effort with contributions from 503 researchers around the world, and the latest release (v2.10) features 228 annotated corpora representing 130 languages. In addition to their use in natural language processing research, these resources are also increasingly being used for empirical studies in linguistic typology and evolutionary linguistics, which indicates that there was a real need in the community for a cross-linguistic standard for morphosyntactic annotation.

Nevertheless, questions have been raised about the appropriateness of certain design choices in UD, and alternatives have been proposed in the literature. Most of these alternatives, however, are largely compatible with the overall approach of UD and differ only in the analysis of certain linguistic phenomena. This is true, for example, of the most well-known alternative framework, Surface-Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018). This raises the question of whether it is possible to extend UD into a framework that subsumes UD as well as a number of its proposed variants — a framework that we might call Flexible Universal Dependencies (FUD).

In this short paper, I want to sketch one approach for realizing such a framework. The basic idea is to extend UD representations from simple dependency trees to nested dependency graphs, where nodes are not limited to atomic syntactic units corresponding to syntactic words but can themselves be dependency graphs that provide compact representations of alternative structural analyses. This proposal draws on a number of previous proposals, including the syntactic nuclei of Tesnière (1959), the bubble trees of Kahane (1997), and the underspecified dependency representations of Schneider et al. (2013). On a more abstract level, it can be seen as an elaboration of the notion of theory-supporting treebanks that I proposed some twenty years ago (Nivre, 2003).

Joakim Nivre. 2022. Flexible Universal Dependencies using nested dependency graphs. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 95–100. Available under CC BY 4.0

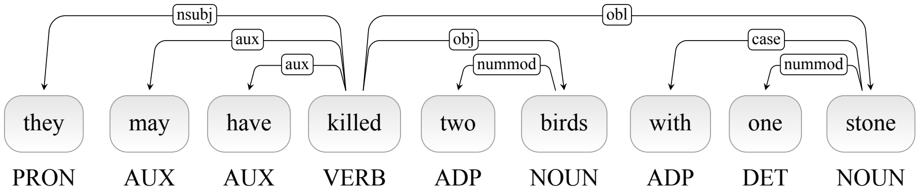


Figure 1: Basic UD representation with limited morphological annotation.

彼女 の 美し さ が 失わ れ ない
she =GEN beautiful -NLZ =NOM lose -PASS -NEG.NPST
‘Her beauty is not lost’

SUWs: | 彼女 | の | 美し | さ | が | 失わ | れ | ない |
Syn. words: | 彼女 | の | 美し さ | が | 失わ れ ない |
LUWs: | 彼女 | の | 美し | さ | が | 失わ | れ | ない |
Bunsetsu: | 彼女 の | 美し さ が | 失わ れ ない |

Figure 2: Japanese word segmentation schemes. Illustration from Murawaki (2019).

2 Annotation principles and issues

The linguistic theory underlying UD is based on two fundamental ideas: (a) the basic syntactic units are *words*; and (b) syntactic structure consists of *grammatical relations* between words (de Marneffe et al., 2021). This leads naturally to an annotation scheme where sentences are segmented into words, which are annotated with morphological information in the form of lemmas, part-of-speech tags and morphological features, and where the syntactic annotation takes the form of a dependency tree where the nodes represent words and the arcs represent grammatical relations. This kind of annotation is illustrated in Figure 1.[1]

The choice of words as basic syntactic units is motivated by the lexical integrity principle (Chomsky, 1970; Bresnan & Mchombo, 1995; Aronoff, 2007), which states that words are built out of different structural elements and by different principles of composition than syntactic constructions, and by the belief that a word-based model will generalize better across languages than trying to segment words into smaller units like morphemes. However, it is well known that cross-linguistically valid criteria for word segmentation are hard to establish (Haspelmath, 2011), and that the application of such criteria is especially challenging for languages which do not have a tradition of marking word boundaries in the orthography. Japanese, for example, has at least three established standards for segmentation into word-like units — known as short unit words (SUW), long unit words (LUW) and *bunsetsus* — and the choice of an appropriate standard for UD annotation has been the subject of considerable discussion (Tanaka et al., 2016; Asahara et al., 2018; Murawaki, 2019; Han et al., 2020; Omura et al., 2021). As a result, some UD treebanks for Japanese now exist in several versions with different word segmentation schemes. Figure 2 illustrates the three established standards as well as a fourth one proposed by Murawaki (2019) as a better fit to the UD notion of syntactic word.

The dependency analysis adopted in UD gives priority to relations holding between the lexical heads of predicates, arguments and modifiers in order to maximize parallelism across languages with different structural characteristics. Major syntactic relations therefore typically hold directly between content words, while function words are treated as grammatical markers on content words. This is illustrated in

[1] Since the morphological annotation is not relevant for the discussion in this paper, I have limited it to part-of-speech tags in Figure 1 and will suppress it completely later.

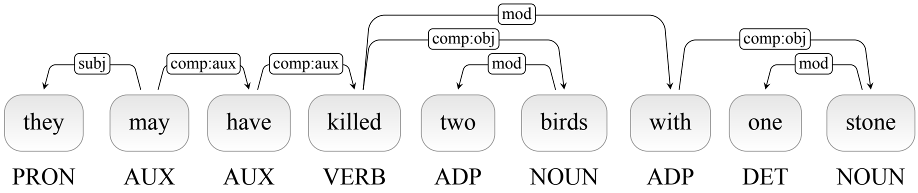


Figure 3: Basic SUD representation corresponding to the UD representation in Figure 1.

Figure 1, where the nominal subject relation (nsubj) holds between the main verb *killed* and the subject pronoun *they*, while the auxiliary verbs *may* and *have* are treated as dependents of the main verb. Similarly, the oblique modifier relation (obl) holds between *killed* and the noun *stone*, while the numeral *one* and the preposition *with* are both dependents of *stone*. The primary motivation for giving priority to relations between content words is that they are more likely to be parallel across languages, while function words in one language may correspond to morphological inflection or nothing at all in other languages.

Regardless of the motivation, however, the treatment of function words has turned out to be one of the most controversial aspects of UD, as it has been perceived as incompatible with syntactic theories that treat function words as syntactic heads (Gerdes & Kahane, 2016; Osborne & Gerdes, 2019). This has led to the development of a sister framework to UD known as Surface-Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018), which is described by its creators as near-isomorphic to UD but which differs in particular by treating function words as heads in the dependency structure, as illustrated in Figure 3.[2]

It is important to note that, despite the obvious differences, the UD and SUD representations also have several things in common. Disregarding for the moment the use of different labels for some relations in SUD and UD (subj vs. nsubj, comp:obj vs. obj, and mod vs. obl and nummod), the two representations give the same analysis of the direct object construction and the numerals, and also posit a subject relation from the verb group *may have killed* to *they* and a modifier relation from *killed* to the prepositional phrase *with one stone*. Similarly, in the case of Japanese word segmentation variants, the syntactic representations will be identical down to the largest word units, and will differ only in that the more aggressive segmentation variants necessitate subtrees that are absent in other variants. In both cases, we may therefore ask whether we can design a richer syntactic representation from which all variants can be extracted. This is the idea of Flexible Universal Dependencies (FUD).

3 Nested dependency graphs

The extended representation proposed here is based on two ideas. The first is to relax the tree constraint and allow general dependency graphs,[3] from which dependency trees can be extracted using spanning tree algorithms familiar from the dependency parsing literature (McDonald et al., 2005). The second idea is to allow these dependency graphs to be nested in the sense that a (smaller) dependency graph can be a node in a (larger) dependency graph. Here is one way of formalizing these ideas, disregarding dependency labels for the moment:

- Let $S = w_1, \ldots, w_n$ be a sentence segmented into a sequence of minimal syntactic units.
- We say that $V_S = \{w_1, \ldots, w_n\}$ is the set of elementary nodes for S.
- A nested dependency graph for S is a directed graph $G = (U, A)$, where every element of U is either
    1. an elementary node $v \in V_S$, or
    2. a nested dependency graph for a (proper) subsequence S' of S with elementary node set $V_{S'}$

  such that U exactly covers S.

[2] SUD differs from UD not only in the treatment of function words, but we will concentrate on this difference here.

[3] The tree constraint in UD only holds for the *basic* syntactic representations; there is also an *enhanced* representation, which is graph-structured and encodes implicit syntactic relations, which does not concern us here.

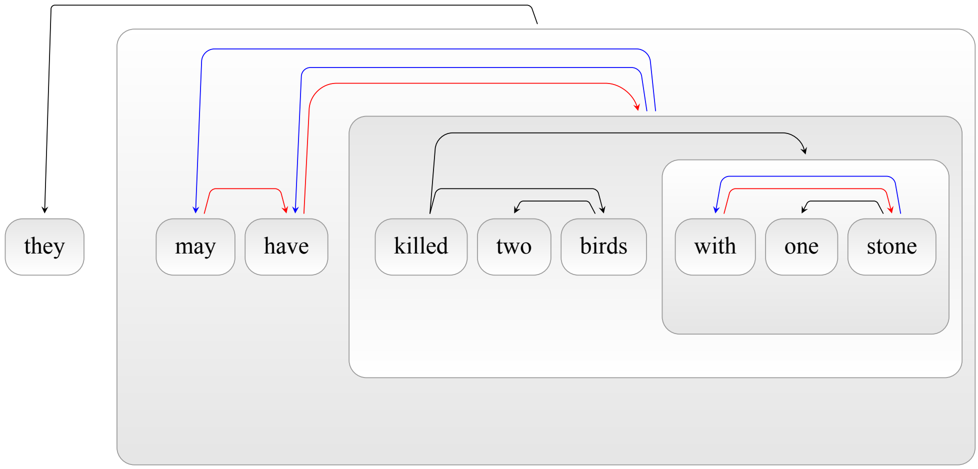


Figure 4: Flexible UD representation with nested dependency graphs.

- We say that a node set U exactly covers S if every elementary node $v \in V_S$ occurs exactly once in U or (recursively) in one of its graph-structured nodes.

The notion of a nested dependency graph is probably best explained through an example. Figure 4 shows a nested dependency graph for the sentence from Figure 1 and Figure 3. The top-level dependency graph has two nodes, the elementary node *they* and a graph node covering the rest of the sentence. That graph in turn has two elementary nodes (*may*, *have*) and one graph node, and so on. In order to extract an ordinary dependency tree from this representation, we need to extract a rooted directed spanning tree from each (nested) dependency graph. For the largest subgraph, we may pick the spanning tree marked in blue, corresponding to a UD analysis, or the spanning tree marked in red, corresponding to an SUD analysis. For the next smaller graph, there is only one spanning tree, which is common to both frameworks, and for the smallest graph we can again choose between a UD (blue) or SUD (red) analysis. If sentences are annotated with nested dependency graphs in this way, we can thus extract different styles of analysis by assigning different weights to the arcs and use a maximum spanning tree algorithm. To handle different word segmentation schemes, as in the Japanese example above, we instead have to choose between extracting a spanning tree and collapsing a subgraph covering a potential word unit into a single node.

4 Conclusion

In this paper, we have sketched a possible extension of the UD framework for syntactic annotation, which allows different variants of the annotation schemes to be extracted from the same compact representations. We have shown how this approach can be used to accommodate the different treatments of function words in UD and SUD, as well as different word segmentation schemes in languages like Japanese. We believe that it can also be used to resolve other issues in UD annotation, such as the analysis of (fixed) multiword expressions and coordination, but that remains outside the scope of the paper. Needless to say, this simple idea needs to be worked out in much more detail before it can be seriously considered for inclusion in UD, but we hope that it can at least inspire others to think about ways to add more flexibility to the framework.

Acknowledgment

Warm thanks to Lars Borin for fruitful and constructive collaborations over the years and for two key contributions to Universal Dependencies: pointing out that the Google universal part-of-speech tag set was missing one of the few truly universal word categories — interjections — and helping us find a meeting room for the UD kick-off meeting during EACL in Gothenburg 2014, as local co-chair. Thanks, Lars, and happy anniversary!

References

Mark Aronoff. 2007. In the Beginning Was the Word. *Language*, 83:803–830.

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, & Yugo Murawaki. 2018. Universal Dependencies Version 2 for Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Joan Bresnan & Sam A. Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language and Linguistic Theory*, 13:181–254.

Noam Chomsky. 1970. Remarks on Nominalization. In Roderick A. Jacobs & Peter S. Rosenbaum, editors, *Readings in English Transformational Grammar*, pages 11–61. Ginn and Co.

Marie de Marneffe, Christopher D. Manning, Joakim Nivre, & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47:255–308.

Kim Gerdes & Sylvain Kahane. 2016. Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies. In *Proceedings of LAW X –The 10th Linguistic Annotation Workshop*, pages 131–140.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, & Guy Perrier. 2018. Sud or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to ud. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74.

Ji Yoon Han, Tae Hwan Oh, Lee Jin, & Hansaem Kim. 2020. Annotation issues in Universal Dependencies for Korean and Japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 99–108, December.

Martin Haspelmath. 2011. The Indeterminacy of Word Segmentation and the Nature of Morphology and Syntax. *Folia Linguistica*, 45:31–80.

Sylvain Kahane. 1997. Bubble Trees and Syntactic Representations. In *Proceedings of the 5th Meeting of Mathematics of Language*.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, & Jan Hajič. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 523–530.

Yugo Murawaki. 2019. On the Definition of Japanese Word. *CoRR*, abs/1906.09719.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, & Dan Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, & Dan Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 4034–4043.

Joakim Nivre. 2003. Theory-Supporting Treebanks. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT)*, pages 117–128.

Mai Omura, Aya Wakasa, & Masayuki Asahara. 2021. Word Delimitation Issues in UD Japanese. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 142–150.

Timothy Osborne & Kim Gerdes. 2019. The Status of Function Words in Dependency Grammar: A Critique of Universal Dependencies (UD). *Glossa*, 4(1):17.

Nathan Schneider, Brendan O'Connor, Naomi Saphra, David Bamman, Manaal Faruqui, Noah A. Smith, Chris Dyer, & Jason Baldridge. 2013. A Framework for (Under)specifying Dependency Syntax without Overloading Annotators. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 51–60.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, & Yuji Matsumoto. 2016. Universal Dependencies for Japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1651–1658.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Editions Klincksieck.

Perspective_on: Semantic relations for frames and constructions

Miriam R L Petruck
International Computer Science Institute
Berkeley, CA, USA
miriamp@icsi.berkeley.edu

Alexander Ziem
Heinrich-Heine-Universität Düsseldorf
Düsseldorf, Northrhine-Westphalia, Germany
ziem@phil.uni-duesseldorf.de

Abstract

This paper considers the frame-to-frame relation **Perspective_on** in FrameNet, addressing its importance for both Frame Semantics and Construction Grammar. Perspective_on highlights the extent to which the continuum of lexicon and grammar is crucial for both theories. Developers of frame-based lexical resources (Borin et al., 2010; Dannélls et al., 2021) and Constructicons, i.e., repositories of grammatical constructions built on the principles of Construction Grammar (Fillmore, 1988) have only begun to address this issue.

1 Introduction

FrameNet (Ruppenhofer et al., 2016), a research and resource development project grounded in the theory of Frame Semantics (Fillmore, 1982; Fillmore, 1985), provides information about the mapping between form and meaning in English. The organizing theoretical construct of FrameNet (**FN**) is the **semantic frame**, i.e., a schematic representation of some scene, whose **frame elements** (**FEs**), or semantic roles, identify participants and other conceptual entities in the scene that a sentence or an utterance describes. Aside from the semantic information that a frame captures, FN links frames in its hierarchy with frame-to-frame relations, including (among others) **Perspective_on**.

This paper considers the frame-to-frame relation Perspective_on in FrameNet, addressing its importance for both Frame Semantics and Construction Grammar. Perspective_on highlights the extent to which the continuum of lexicon and grammar is crucial for both theories. Developers of frame-based lexical resources (e.g., Borin et al. (2010), Dannélls et al. (2021)) and Constructicons, i.e., repositories of grammatical constructions built on the principles of Construction Grammar (Fillmore, 1988) have only begun to address this issue.[1]

2 Background to FrameNet and the FrameNet Constructicon

2.1 FrameNet

FrameNet is a unique knowledge base that maps meaning to form through the theory of Frame Semantics (Fillmore, 1982; Fillmore, 1985). The FrameNet database includes frame descriptions for over 1,200 semantic frames, also understood as script-like structures that provide background knowledge for the use and understanding of words in (a) language and facilitate inferencing about participants and events, more than 13,000 lexical units (**LU**), each of which is a **lexical construction**, and nearly 200K manually annotated sentences in Frame Semantic terms. Moreover, FrameNet captures additional semantic information about relations between frames with a set of frame-to-frame relations.

Aside from the significance of the intellectual accomplishment, FrameNet data serve as training data for downstream natural language processing applications, such as question-answering, event tracking, and information extraction, to name but a few.

[1] A very early version of this paper was presented at the International Construction Grammar Conference (ICCG-8) in Osnabrueck, 2014. Space limitations preclude the inclusion of the entire presentation, although the authors intend to expand on the current work in the future.

Miriam R.L. Petruck and Alexander Ziem. 2022. Perspective_on: Semantic relations for frames and constructions. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 101–105. Available under CC BY 4.0

Relation	Super_frame	Sub_frame
Inheritance	Parent	Child
Subframes	Complex	Component
Precedes	Earlier	Later
Using	Parent	Child
Perspective_on	Neutral	Perspectivized
See_also	Main Entry	Referring Entry
Metaphor	Source	Target
Inchoative_of	Inchoative	State
Causative_of	Causative	Inchoative/State

Table 1: Frame-to-Frame relations in FrameNet

2.2 The FrameNet Constructicon

A **constructicon** is a structured repository of grammatical constructions, interrelated as a (mostly) taxonomic network linking the most general of constructions (as in a family of constructions) to its most specific type (Diessel, 2019). A complete constructicon (for a single language) would include the full range of construction types, from highly schematic non-lexical constructions to meaningful argument-structure constructions, as well as partly idiomatic constructions, complex words, and even morphemes.

The FrameNet Constructicon holds approximately 75 constructions, with detailed information about the kinds of linguistic material that can occur in specifiable positions within each construction, as well as those positions within which said construction may occur (Fillmore et al., 2012). Thus, developers of the FrameNet Constructicon first identified and defined constructions, along with their construction elements, then analyzed and labeled constructs that illustrate each construction. For example, consider the **be_recip** construction, examples of which appear below, where example 1 is the asymmetrical version of the construction and example 2 is the symmetrical version.

1. I know that [Chuck **INDIVIDUAL_1**] is friends [with Paul **INDIVIDUAL_2**].
2. I know that [Chuck and Paul **INDIVIDUALS**] are friends.

In this construction the head noun, which is used as a predicate must be a term that denotes a reciprocal relationship, e.g., *partners, coworkers*, etc. and that the *with*-prepositional phrase in example 1 is not predictable. Simplifying matters for the current purposes, example 1 shows the construction elements **INDIVIDUAL_1** and **INDIVIDUAL_2**, while example 2 shows the construction element **INDIVIDUALS**.

In addition, the construction evokes the `Reciprocality` frame, which FN has characterized as states-of-affairs with **Protagonists** in relations with each other that may be viewed symmetrically. When these frame elements are equally prominent, each equally serving to identify the other, they manifest together as PROTAGONISTS. When one of them defines the other (similar to a Ground), it is PROTAGONIST_2, with the other called PROTAGONIST_1 (i.e., the Figure).[2]

3 Frame-to-Frame relations in FrameNet

The so-called FrameNet hierarchy[3] links frames through nine frame-to-frame semantic relationships, which Table 1 displays.[4]

Ruppenhofer et al. (2016) defines, explains, and illustrates all of FN's frame-to-frame relations;[5] here we focus exclusively on Perspective_on. This frame-to-frame relationship assumes the existence of a neutral parent frame and two child frames, each providing a different point of view on the neutral parent

[2] See Lee-Goldman & Petruck (2018) for a detailed description of this construction.

[3] Structurally, the FrameNet hierarchy is most similar to a lattice. See Valverde-Albacete (2005) for further information.

[4] For expediency, FrameNet represents Inchoative_of and Causitive_of as frame-to-frame relations, not lexical relations (which they are). See Petruck et al. (2004).

[5] Limitations of space preclude presenting and illustrating all of the frame-to-frame relations listed in Table 1. See Ruppenhofer et al. (2016, p. 79-85) for discussion of all of FN's semantic relations.

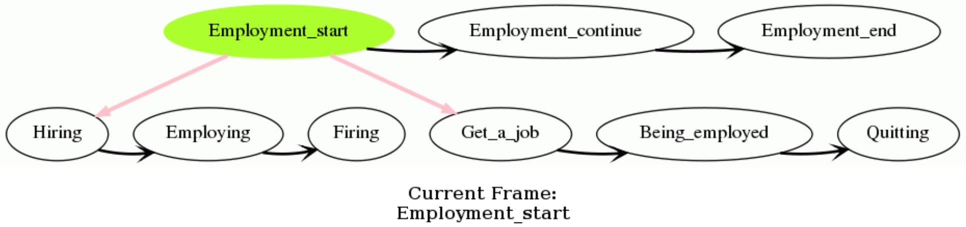


Figure 1: `Employment_start`

unperspectivized frame. For example, consider the `Employment_start`, parent frame, two of whose children are `Hiring` and `Get_a_job`, where the former provides the EMPLOYER's point of view and the latter provides that of the EMPLOYEE. [6]

Using FrameGrapher, FN's visualization tool, Figure 1 depicts the `Employment_start` frame along with several related frames, including `Hiring` and `Get_a_job`, each related via Perspective_on to its parent (displayed with pink arrows).

4 Perspective_on as a relation between constructions

The understanding that Perspective_on relates frames in the FN hierarchy to each other when those frames capture two points of view (or perspectives) allows exploiting the relation to capture semantic relations between constructions, not only between frames.

Consider the correlation between active and passive, as in example 3 and 4, respectively, where the passive 4 shifts the point of view from the **SELLER**, *Chuck*, to the **GOODS**, *the car*.

3. I know that [Chuck **SELLER**] sold [the car **GOODS**] for $1000.
4. I know that [the car **GOODS**] was sold [by Chuck **SELLER**] for $1000.

Much the way Perspective_on profiles an event on the lexical level in the FrameNet lexicon, so too does the relation operate on the level of constructions, thus also hinting at the similarity between lexical and grammatical constructions.

Perspective_on is also useful for relating constructions that involve multiple affected entities, as in the Double_object construction. Consider examples 5 and 6, where the two versions of that construction make use of the same frame elements, namely **BUYER** and **GOODS**, in different syntactic realizations. The syntactic realization of the **BUYER** in a PP-to phrase changes the semantic profiling in the sentence from **BUYER**, *Jerry*, to the **GOODS**, *car*.

5. Chuck sold [Jerry **BUYER**] [the car **GOODS**] for $1000.
6. Chuck sold [the car **GOODS**] [to Jerry **BUYER**] for $1000.

The data in examples 5 and 6 demonstrate the usefulness of the frame-to-frame relation Perspective_on for semantic profiling. More generally, the two sets of examples (3 and 4 along with 5 and 6), also suggest that the utility of Perspective_on as a relation between constructions is not limited to just one (type of) construction.

5 Related work

Not surprisingly, developers of FrameNet resources and Constructicons have addressed the issues of (1) the relationships between constructions and frames, as well as (2) the relationships between constructions. This section briefly discusses some of the most immediately relevant works (presenting them in chronological order of their publication).

[6] This example derives from (Ruppenhofer et al., 2016, p. 9).

In an effort to control the connections between frames and constructions in the FrameNet Brasil database, Torrent et al. (2014) outlined policies for the annotation of constructions, specifically for the identification and labeling of construction elements, in the Brazilian Portugese Constructicon. The motivation behind these policies includes remaining faithful to the principles of Frame Semantics and Construction Grammar. Additionally, adopting and implementing the policies recognize the continuum of lexicon and grammar (Fillmore, 2008). Although the work focused on relations between construction elements and frame elements, it also drew attention to **families of constructions**, which necessarily requires the analyst to consider the relations between (or among) members of one such family.

Somewhat similarly, Ohara (2018) also addressed relations between frames and constructions in the Japanese FrameNet Constructicon, albeit from a different perspective from that of Torrent et al. (2014), where the latter concerned annotation practices for using a combined tool that facilitates work on both frames and constructions. Ohara (2018) concerns distinguishing between (1) frame-based description and annotation of semantico-syntactic structures of lexical units (and multiword expressions that FrameNet has defined as such) and (2) constructicon annotation for describing the internal and external syntax and semantics of linguistic objects having complex structures. Importantly, this work introduces a frame-based classification of constructions, which developers of constructicons might employ to facilitate determining relations between (or among) constructions.

In developing the German Constructicon (https://gsw.phil.hhu.de/constructicon/), Boas & Ziem (2018) took a contrastive approach and considered German constructions in relation to their English analogs. One goal of the work was to leverage existing construction entries in the FrameNet Constructicon for English (Fillmore et al., 2012) to develop the German Constructicon. That goal required defining and exploiting the notion of a continuum of constructional correspondence. The work is more about relations between constructions across two genetically related, yet typologically distinct languages (at least in terms of the morphology and the syntax of each language, as Kastovsky (2011) suggests), than it is about semantic relations between constructions within a single language, namely German. Nonetheless, the developers of the German Constructicon clearly know about relations between constructions in a constructicon. As such, even if only by implication, in directing the reader's attention to *correspondences*, Boas & Ziem (2018) also asks the reader to attend to semantic relations between constructions in the German Constructicon effort (Ziem et al., 2019).

While the above briefly described works do not address relations between constructions directly within a single constructicon, they strongly suggest that the extended community of frame-based and construction-based resource developers is aware of the necessity of addressing relations between constructions. Perhaps, the next round of development in building constructicons will include attention to that necessity.

6 Concluding remarks

The structured event formalism for representing FrameNets informal descriptions in Chang et al. (2002) also offered a way of handling linguistic focus, which FrameNet has implemented with the semantic relation Perspective_on. Building on that insight, this paper addressed the application of one frame-to-frame relation for describing semantic relations between constructions, also in constructicons in general.

Aside from the relatively recent work that addresses relations between frames and constructions in the context of constructicon development, this paper also suggests that Construction Grammarians more generally (not just those involved in constructicon development) might investigate existing frame-to-frame relations for relating different types of constructions, beyond that of **Inheritance**, which Fillmore (1999) already identified as playing an important role in Construction Grammar.

References

Hans Christian Boas & Alexander Ziem. 2018. Constructing a constructicon for German. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, & Tiago Timponi Torrent, editors, *Constructicography*, pages 183–228. Johns Benjamins, Amsterdam and Philadelphia.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, & Dimitrios Kokkinakis. 2010. The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281.

Nancy Chang, Srini Narayanan, & Miriam R. L. Petruck. 2002. Putting frames in perspective. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING, pages 1–7, Stroudsburg, PA. ACL.

Dana Dannélls, Lars Borin, & Karin Friberg Heppin. 2021. *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*. Number 14 in Natural Language Processing. John Benjamins Publishing Company, Amsterdam and Philadelphia.

Holger Diessel. 2019. *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*. Cambridge University Press, Cambridge.

Charles J. Fillmore, Russell Lee-Goldman, & Russell G. Rhodes. 2012. The FrameNet constructicon. In Hans C. Boas & Ivan A. Sag, editors, *Sign-based Construction Grammar*, pages 309–372. CSLI.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–138. Linguistics Society of Korea, Seoul.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Charles J. Fillmore. 1988. The Mechanisms of Construction Grammar. In *Proceedings of the 14th Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

Charles J Fillmore. 1999. Inversion and constructional inheritance. *Lexical and constructional aspects of linguistic explanation*, 1:113–128.

Charles J. Fillmore. 2008. Border conflicts: FrameNet meets Construction Grammar. In Elisenda Bernal & Janet DeCesaris, editors, *Proceedings of the XIII EURALEX International Congress*.

Dieter Kastovsky. 2011. Typological differences between English and German morphology and their causes. In Toril Swan, Endre Mørck, & Olaf Jansen Westvik, editors, *Language Change and Language Structure: Older Germanic Languages in a Comparative Perspective*, pages 135–158. De Gruyter Mouton.

Russell Lee-Goldman & Miriam R. L. Petruck. 2018. The FrameNet constructicon in action. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, & Tiago Timponi Torrent, editors, *Constructicography*, pages 19–40. Johns Benjamins, Amsterdam and Philadelphia.

Kyoko Ohara. 2018. Relations between frames and constructions: A Proposal from the Japanese FrameNet Constructicon. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, & Tiago Timponi Torrent, editors, *Constructicography: Constructicon Development across Languages*, pages 141–164. Johns Benjamins, Amsterdam and Philadelphia.

Miriam R. L. Petruck, Charles J. Fillmore, Collin F. Baker, Michael Ellsworth, & Josef Ruppenhofer. 2004. Reframing FrameNet data. In G. Williams & S. Vessier, editors, *Proceedings of The 11th EURALEX International Congress*, pages 405–416, Lorient.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, & Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. ICSI: Berkeley.

Tiago Timponi Torrent, Ludmila Lage, Thais Fernandes Sampaio, Tatiane Tavares, & E. Matos. 2014. Revisiting border conflicts between FrameNet and construction grammar: Annotation policies for the Brazilian Portuguese constructicon. *Constructions and Frames*.

Francisco J. Valverde-Albacete. 2005. Explaining the structure of FrameNet with concept lattices. In Bernhard Ganter & Robert Godin, editors, *Formal Concept Analysis*, pages 79–94, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alexander Ziem, Johanna Flick, & Phillip Sandkühler. 2019. The German Constructicon Project: Framework, methodology, resources. *Lexicographica*, 35:15–40.

Natural language processing for educational applications: Recent advances

Ildikó Pilán
Norwegian Computing Center
Oslo, Norway
pilan@nr.no

Abstract

We summarize recent advances in Natural Language Processing applied to topics related to the educational domain based on the latest publications from a major conference in the field. The three topics we touch upon include feedback, difficulty assessment and question generation.

1 Introduction

The 60th Annual Meeting of the Association for Computational Linguistics (ACL) and the co-located workshops, held in May 2022, included a number of articles applying Natural Language Processing (NLP) to the educational domain. Most of these studies are centered around three topics: feedback, difficulty rating and question generation. In the following three sections, we provide an overview of the relevant articles, which all target English language data, except for two studies involving German resources. We do not include a separate, more in-depth discussion on recent research published at workshops dedicated to educational NLP topics since the interested readers can find relevant papers more easily in those proceedings compared to the broad and vast body of research published at more generic NLP conferences.

2 Feedback

In the past years, there has been an increasing focus on developing automated assessment systems providing explainable and understandable feedback that goes beyond a mere correct-incorrect response, the need for which has been, in fact, emphasized also in previous work (Deeva et al., 2021). Filighera et al. (2022) took a step in this direction by creating a dataset of content-focused elaborated feedback for an automatic short answer grading system. The dataset consists of (i) learner responses; (ii) the reference answers; (iii) a score and (iv) detailed feedback explaining that score. The dataset contains more than 2000 responses in both German and English to 8 and 22 different questions respectively within the topic of a college-level communication networks course. The authors included also baselines created by fine-tuning a T5 Transformer language model (Raffel et al., 2020) on their dataset. They found that these baselines improve on a majority baseline, but still perform considerably more poorly than humans. However, they also pointed out that measures such as BLEU and ROUGE fail to capture content similarity between manual and automatic feedback.

Kaneko et al. (2022) proposed an example-based Grammatical Error Correction (GEC) system, which uses retrieved example sentences for both generating more accurate corrections compared to traditional GEC systems, and for providing an explanation to language learners about their errors. For each error, a pair of correct and incorrect sentences similar to the original, learner-written sentence, is shown by the system. These serve as a kind of indirect explanation to learners about why their sentence might be incorrect. The authors conducted also a user study where they found that providing language learners with examples helped them to decide whether to accept or refuse the automatically suggested corrections.

Ildikó Pilán. 2022. Natural language processing for educational applications: Recent advances.
In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 107–109. Available under CC BY 4.0

3 Difficulty

Research on readability and difficulty in general, especially targeting English, seems to have somewhat decreased in recent years, but we can still find a few examples in this directions.

Steinmetz & Harbusch (2022) presented EasyTalk, an interactive support system for German-speaking low-literate users with intellectual or developmental disabilities. The system helps users write coherent and correct text suitable for their proficiency level. Words can be supplemented with symbols and users are reminded to complement their texts with information covering wh-questions as well as conjunctions improving coherence. The authors have also conducted a small case study with low-literate users where they investigated writing behaviour with eye-tracking recordings and found that participants used most parts of the system for redacting, but to a minor extent for adding connectors.

It it worth noting that there is a growing interest in writing assistants in the broader educational technology community, including NLP. In fact, at ACL 2022, a whole workshop was dedicated to the topic of writing assistants[1], where the above mentioned study was also published. The workshop aimed to be a meeting place for researchers in NLP, human-computer interaction as well as writers and industry practitioners. Dialogue among representatives with such a broad set of expertise has a good potential to spark research in the area that is relevant for user needs.

Another set of experiments related to difficulty but of materials presented to learners, not written by them is described in Byrd & Srivastava (2022), who investigated predicting the difficulty of natural language questions based on a question answering (QA) dataset (HotPotQA) and Item Response Theory (IRT). The advantage of this psycometric tool is that it defines difficulty in a straightforward manner, namely having a 50% chance of answering a question correctly. To simulate a variety of responses for their study, the authors created an artificial crowd by training a QA model with varying amounts of data and epochs. Questions were of a variety of topic (e.g. entertainment, biology) and type (yes/no, wh-questions). The experiments showed that yes/no questions had a more consistent difficulty level. Textual features were also correlated to difficulty with different non-neural models, which showed that commas and complex words were among the most important features for determining question difficulty. Such automatic question difficulty assessment enables the development of question generation systems where a desired difficulty level can be specified.

4 Question generation

The third educational NLP theme at ACL 2022 centered around the automatic generation of questions. Ghanem et al. (2022) present a novel question generation dataset and system generating inferential questions targeting specific comprehension skill types. Compared to extractive questions, which has been the focus of previous work (Murakhovs'ka et al., 2022), inferential questions measure better learner understanding. The authors created and released a dataset annotated with story-based reading comprehension skills that makes it possible to train systems able to generate questions with explicit control for such skills. The dataset contains 726 children's stories and is annotated for the following skill types: Basic Story Elements, Character Traits, Close Reading, Figurative Language, Inferring, Predicting, Summarizing, Visualizing and Vocabulary. Each type has an average of ca. 5 question-answer pairs per story.

Another interesting study explores the use of summaries for alleviating a major problem in reading comprehension question generation, namely irrelevant or un-interpretable questions (Dugan et al., 2022). Instead of the original text, the authors experimented with providing a T5-based question generation model with human-written summaries. The passages used were chapters from a well-known NLP handbook. Their results indicated that the 3 annotators participating in the evaluation accepted a considerably higher proportion of questions generated by their method than relying only on the original textbook text passages themselves. Questions were also more relevant and interpretable without a larger context. In the lack of human-written summaries, using automatic summaries still led to improved question generation in the experiments presented.

[1] `https://in2writing.glitch.me/`

5 Conclusion

As the studies presented above show, recent research in educational NLP has been branching out to areas and tasks which remained less explored previously, such as question generation and feedback. However, they mostly focus on English as a target language, which in part is most likely due to the availability of more resources both as unannotated data for pre-training the most recent transformer models, as well as datasets annotated for specific tasks.

Some studies also included human evaluations which, understandably, remained somewhat limited in size. They represent, nonetheless, an important step towards understanding better the performance and the usability of the proposed systems, a much needed aspect in the educational domain.

References

Matthew Byrd & Shashank Srivastava. 2022. Predicting Difficulty and Discrimination of Natural Language Questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130, Dublin. Association for Computational Linguistics.

Galina Deeva, Daria Bogdanova, Estefanía Serral, Monique Snoeck, & Jochen De Weerdt. 2021. A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers and Education*, 162:104094.

Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chuning Yuan, & Chris Callison-Burch. 2022. A Feasibility Study of Answer-Agnostic Question Generation for Education. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1919–1926, Dublin. Association for Computational Linguistics.

Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, & Sebastian Ochs. 2022. Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8577–8591, Dublin. Association for Computational Linguistics.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer von der Ohe, & Alona Fyshe. 2022. Question Generation for Reading Comprehension Assessment by Modeling How and What to Ask. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2131–2146, Dublin. Association for Computational Linguistics.

Masahiro Kaneko, Sho Takase, Ayana Niwa, & Naoaki Okazaki. 2022. Interpretability for Language Learners Using Example-Based Grammatical Error Correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7176–7187, Dublin. Association for Computational Linguistics.

Lidiya Murakhovs'ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, & Caiming Xiong. 2022. MixQG: Neural Question Generation with Mixed Answer Types. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1486–1497, Seattle. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, & Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ina Steinmetz & Karin Harbusch. 2022. A text-writing system for Easy-to-Read German evaluated with low-literate users with cognitive impairment. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 27–38, Dublin. Association for Computational Linguistics.

Detecting fake papers with the latent algorithm for recursive search

Stian Rødven-Eide
University of Gothenburg
stian.rodven.eide@gu.se

Ricardo Muñoz Sánchez
University of Gothenburg
ricardo.munoz.sanchez@gu.se

Abstract

The current avalanche of fake scientific papers appearing on respected websites such as snarXiv.org has become of much concern for researchers. After concluding that existing methods for distinguishing between real and fake scientific papers leave much to be desired, we have developed the Latent Algorithm for Recursive Search for this purpose. Our proposed method is not only able to identify fake scientific papers with extremely high accuracy, but can also automatically retract the identified papers.

1 Introduction

- *Intrinsic Structure Reduction for Reversing Sentiment in Investigative Journalism*,
- *SELMA: A Novel Approach to Assessing the Novelty of Novels*, and
- *Understanding Universal Undercurrents in Unappreciated Unions – a Semantic Understudy.*

These are just some of the titles behind which blatantly fake papers hide, purporting to be scientific, but in essence offering nothing of essence. Some might be automatically generated, some created by internet trolls for a laugh and a half, and some carefully constructed to cause chaos in the community.

Inspired by computational models of hide-and-seek, such as Latent Dirichlet Allocation, Latent Semantic Analysis and Latent Discriminant Analysis, we have developed a novel state-of-the-art algorithm to address this growing problem. Our Latent Algorithm for Recursive Search (LARS) has proven itself capable of detecting even the fakest papers. In the following sections, we describe how and why this is possible, detailing the inner workings of LARS, and presenting the results in comparison to other recent advances.

2 Related work

With the widespread appearance of false information following the onset of the COVID-19 pandemic, much research has been devoted to identifying false information on several media outlets. One of the biggest issues we face is that humans are not good at detecting misinformation, with even experts picking falling for both fake news,[1] and academic papers.[2]

One of the tasks that has grown in recent years is that of fake news detection. Oshikawa et al. (2020) note in their survey that most datasets tend to be relatively small due to the need to obtain fact-checked articles. This leads to most papers dealing with a simple binary classification task.

While the field of detecting fake academic articles has not received much attention yet, there have been some recent attention brought to them since the groundbreaking work of Baldassarre (2020). This

[1] https://www.thedailybeast.com/fooled-by-the-onion-9-most-embarrassing-fails

[2] https://www.sciencealert.com/cultural-studies-sokal-squared-hoax-20-fake-papers

Stian Rødven-Eide and Ricardo Muñoz Sánchez. 2022. Detecting fake papers with the latent algorithm for recursive search. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 111–113. Available under CC BY 4.0

paper notes how important it is to have a good peer-reviewing system, to check for the veracity of the data, and to go through the cited literature. It further notes how these processes can break down, leading to non-serious papers being accepted into otherwise academic outlets.

Two of the more interesting approaches of late are the FFF-method, as proposed by Borrs et al. (2022), and the Agreeable Knowledge algorithm, developed by Larn et al. (2022). What both of these have in common, and that which we have found it sound to rely on, is a simultaneously holistic and recursive understanding of the nature of scientific publishing: The assumption that any given scientific paper attempts to replicate itself through self-absorption, as well as insert itself into as many other papers as possible. We call this the Vital Viral Vector (VVV).

3 Methodology

In order to exploit the VVV for LARS, its direction in the scientiverse must first be established. This is achieved by finding the non-trivial zeros in the following function, where x is a representation of our document:

$$\zeta(x) = \sum_{n=1}^{\infty} \frac{1}{n^x}$$

Once we have the vector of all non-trivial zeros of the previous equation (the VVV embedding), we insert it into a recursive matrix, where the endpoints consist of the matrix itself. We then backpropagate by finding the eigenvalues of the space of non-real papers. Essentially, if D is the domain where all interesting and novel papers can be found and B is the boundary where papers become false, we define its Euler-Lagrange equations as follows:

$$\phi(VVV) = \lim_{b \to 0} \frac{1}{|b|\sqrt{VVV}} e^{-VVV/b}$$

$$Q[\phi] = \int_D p(X)\nabla_\phi \cdot \nabla_\phi q(X)\phi dX + \int_B \sigma(S)\phi^2 dS$$

Lastly, we apply a latency filter that separates true and false upon dimensionality reduction. This can be done by integrating the hyperbolic Riemannian representation of the filter

$$[\wp(Q)]^2 = 4[\wp(Q)]^3 - g_2\wp(Q) - g_3$$

This is most easily done by finding g_1, g_2, and g_3 such that $g_1^3 + g_2^3 = g_3^3$ and applying a logistic regression to $\wp(Q)$.

We now have the final score, signifying the trueness and/or thruthiness of the scrutinised paper. The only step left is retraction, which is done through undetected infiltration of the infected papers – taking it down from the inside, so to speak.

4 Evaluation

For four different datasets, we ran LARS as well as the two most prominent alternative methods, FFF (Borrs et al., 2022) and FAKE (Larn et al., 2022), the results of which are available in Table 2. The datasets we used are part of the shared task *Fake Methods for Finding Fake Papers*, which took place att BCL in 2021 (Borscht & Goulash, 2021). These are detailed in Table 1. All our results are, as you can see, much better than those provided by other methods.[3] If you look closely at the F1-scores, you will see that the result for snarXiv is 1.01. The reason for this is that this paper, the one you are reading, was implicitly (and latently) included into the test set upon running the algorithm, and successfully identified as well.

[3] At least the ones we tested.

Dataset	**Genre**	**Documents**	**Tokens**
Onion-stories	News	1,742,009	1,742,010
Old York Times	Olds	34	34,000,000
Borin-collection	Scientific papers	9,999	9,999,999
snarXiv.org	Unscientific papers	4,321	1,234,567

Table 1: Datasets from the shared task.

Method	**Onion**	**OYT**	**Borin**	**snarXiv**
FFF	0.73	0.77	0.80	0.71
FAKE	0.79	0.78	0.79	0.77
LARS	**0.99**	**0.99**	**1.00**	**1.01**

Table 2: F1-scores for fake paper detection.

4.1 Recursive testing

An unfortunate side-effect of LARS is that it automatically evaluates any paper in which it is mentioned. The result of that evaluation is then inserted into that paper as a subsection named *Recursive Testing*. The likelihood that this paper is as fake as the articles it set out to scrutinise is 97.8%.

5 Conclusion and future work

As we can see, this paper is as fake as they come. However, considering the humorous nature of its nature, we, naturally, regard this as entirely natural.

References

Daniel T Baldassarre. 2020. What's the deal with birds? *Scientific Journal of Research and Reviews*, 2(4).

Lina Borrs, Boris Larn, & Biron Rals. 2022. Finding false flags – and other anomalous analogies. *Journal of False Disinformation*, 14(3):123–234.

Johannes Borscht & Maria Goulash. 2021. Shared task: Fake methods for finding fake papers. In *Proceedings of the 132th Workshop on Fake Publishing*, pages 52–57, Georgetown, Guyana. Brotherhood for Computational Linguistics.

Boris Larn, Lina Borrs, & Biron Rals. 2022. Finding agreeable knowledge evenly. *Journal of Diffusion Through Confusion*, 15(4):234–345.

Ray Oshikawa, Jing Qian, & William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093. European Language Resources Association.

Leksikalsk-semantiske sprogressourcer: Hvad kan de, og hvordan udvikler vi dem bedst?

Bolette Sandford Pedersen
Center for Sprogteknologi
Københavns Universitet, Danmark
bspedersen@hum.ku.dk

Sanni Nimb
Det Danske Sprog- og Litteraturselskab
Danmark
sn@dsl.dk

Sussi Olsen
Center for Sprogteknologi,
Københavns Universitet, Danmark
saolsen@hum.ku.dk

Abstract

The paper discusses the status of lexical-semantic language resources in the era of statistical language models. We argue that a lot of essential background information about culture, society and the surrounding world is available via these resources; information which cannot be deduced from text alone, but which is crucial for language interpretation. We describe the Danish scenario and describe the series of resources that have been compiled for Danish in a joint venture between lexicographers and NLP researchers. For two decades, three such resources have been developed (a wordnet, a framenet and a sentiment lexicon), all with identifier links to the same sense inventory, namely that of Den Danske Ordbog. We also present a new resource, the COR lexicon, which draws on these existing lexical resources but attempts to create an easy-to-use, joint semantic lexicon for AI developers which has a more coarse-grained sense inventory and a core set of semantic information types. Finally, we argue that lexical semantic resources for NLP should ideally be integrated as part of a larger lexicographical infrastructure with the aim of easing future scaling and maintenance.

1 Leksikalske sprogressourcer og sprogmodeller

Leksikalske sprogressourcer der beskriver ordenes betydning og rolle fra forskellige perspektiver, har været centrale byggesten i mange sprogteknologiske applikationer igennem de seneste tiår. De har imidlertid også udgjort flaskehalsen i mange systemer fordi vi har haft vanskeligt ved at opnå tilstrækkelig dækningsgrad og tilstrækkelig konsistens til at de smidigt kunne indgå i interaktion med fx formelle grammatiker, eller senere, med statistiske sprogmodeller.

Uafladeligt har vi måttet falde tilbage på teknikker der på bedste beskub kan håndtere såkaldte out-of-vocabulary (OOV)-problematikker, altså at et givent ord ikke er beskrevet i ressourcen. Det skyldes dels at ordforrådet i et sprog hele tiden udvikler sig, og at en del ord, især de sammensatte, ofte skabes dynamisk på stedet i specifikke kontekster. Men det skyldes også i nok så høj grad at sprogressourcer til NLP ofte er blevet udviklet en anelse stedmoderligt og uden for de leksikografiske miljøer, dvs. uden den reelle leksikografiske faglighed og det setup som er nødvendigt for at kunne udvikle og vedligeholde en solid og konsistent leksikalsk ressource.

Vi argumenterer for at leksikalsk-semantiske ressourcer fortsat bør spille en central rolle i NLP, også i en tid hvor neurale sprogmodeller har bragt os langt med ren tekststatistik. De leksikalske ressourcer kan

Bolette Sandford Pedersen, Sanni Nimb and Sussi Olsen. 2022. Leksikalsk-semantiske sprogressourcer: Hvad kan de, og hvordan udvikler vi dem bedst?. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 115–120. Available under CC BY 4.0

noget vigtigt, og de rummer noget viden som vi ikke kan eller bør undvære i vores sprogteknologiske tjenester, som helst skal være inkluderende og tillidsskabende for alle.

Der er næppe nogen tvivl om at sprogmodeller også fremover vil være biased i forhold til de tekster de baseres på, og at tekster i øvrigt ikke rummer den viden om sproget og verden som er nødvendig hvis NLP skal kunne levere en dybere og mere praktisk anvendelig sprogforståelse (Bender et al., 2021).

Et velkendt og illustrativt eksempel er *black sheep*-problematikken (cf. Van Durme (2009)) som beskriver det problem der opstår hvis vi spørger en statistisk sprogmodel hvilken farve et får har. Her får vi for de fleste sprog svaret *sort*. Dette selv om vi alle ved at *det sorte får* er undtagelsen der bekræfter reglen om at får som de er flest, er hvide eller grå.

Det er præcis baggrundsviden af denne type som de leksikalsk-semantiske ressourcer beskriver, og derfor er de vigtige at inddrage.[1] Selv om *det sorte får* er en metafor, typisk for mennesker der ikke følger den slagne vej, er eksemplet i al sin enkelthed godt fordi det illustrerer hvad vi skriver om, set relativt til den verden som agerer baggrund for det vi skriver om. Den mest selvfølgelige baggrundsviden er med andre ofte ikke eksplicit i teksten, og derfor har de statistiske sprogmodeller svært ved at indfange den. Dette skaber problemer i mange sammenhænge hvor statistiske sprogmodeller anvendes til sprogforståelse, parallelt med at der også ses et overraskende tungt bias mod fx demografiske stereotyper og kønsstereotyper i state-of-the-art sprogmodeller. Sidstnævnte problemstilling er i høj grad blevet adresseret i flere nyere videnskabelige artikler (se fx Kurita et al. (2019) og Sólmundsdóttir et al. (2022) for kønsbias i maskinoversættelse og NLP mere generelt) og i pressen fordi problemet med sådanne bias er så åbenlyst.

Den anden problematik med manglende baggrundsviden fremstår derimod endnu ikke så tydeligt, men er måske nok så alvorlig. Ordbøgerne med deres semantiske beskrivelser dækker selvfølgelig ikke denne baggrundsviden alene, men de udgør en vigtig brik i det samlede billede.[2]

Udover at argumentere for at de neurale sprogmodeller bør beriges med mere basal baggrundsviden, argumenterer vi også for hvorfor leksikalske-semantiske ressourcer ikke bør udvikles isoleret fra et sprogs øvrige leksikografiske virke men derimod bør ses som en naturlig del af eller "spinn-off" på den leksikografiske virksomhed der i øvrigt foregår i et givent sprogmiljø.

Det store samarbejdsprojektet ELEXIS (Krek et al. (2018); https://elex.is/), som vi afsluttede i 2022, har kun bekræftet denne tilgang. Projektet er lykkedes med at skabe en kobling mellem de leksikografiske miljøer og udvalgte NLP-miljøer i Europa og har arbejdet henimod at åbne og standardisere de "traditionelle" ordbøger sådan at de informationer de rummer, i højere grad kan komme i spil i NLP. Selv om de intellektuelle rettigheder i de leksikografiske miljøer stadig vanskeliggør fuld udnyttelse af ordbøger til NLP, er dette et vigtigt skridt på vejen.

Nedenfor opsummerer vi vores eget arbejde med leksikalsk-semantiske sprogressourcer som det har udviklet sig henover de seneste to årtier, se også Pedersen et al. (2021). Det drejer sig i essensen om det danske WordNet, DanNet, det Danske FrameNet og det Danske Sentimentleksikon. Alle er udviklet med fast base i Den Danske Ordbogs betydningsinventar og beskæftiger sig med hhv. det paradagmatiske, det syntagmatiske og det konnotative perspektiv af ordenes betydning. Endelig beskriver vi hvordan vi i et nyligt igangsat ordbogsprojekt for kunstig intelligens samler de væsentligste oplysningstyper fra de tre ressourcer i et samlet Centralt OrdRegister for Dansh (COR).[3]

Det er vigtigt for os i denne sammenhæng at nævne at kollegers arbejde ved Språkbanken og på de tilsvarende svenske semantiske ressourcer som SALDO-ordbogen, det Svenske FrameNet og den svenske sentimentordbog SenSALDO gennem årene har været en stor inspiration for vores arbejde. Det har været interessant at se hvordan man har grebet arbejdet an ved Språkbanken, og hvordan man har fordelt indsatsen med at styrke det tilsvarende svenske sprogområde på NLP-området.

[1] Den Danske Ordbog og DanNet har fx følgende definition på *får: mellemstor drøvtygger med meget kraftig, oftest hvidlig uldpels*.

[2] Encyklopædisk viden udgør en anden dimension af væsentlig baggrundsviden om end grænsedragningen mellem det semantiske og det encyklopædiske ikke altid er lige klar.

[3] Dette igangværende projekt finansieres af Digitaliseringsstyrelsen som en del af en satsning på kunstig intelligens i Danmark.

2 Et sæt af danske leksikalsk-semantiske ressourcer med fast forankring i Den Danske Ordbog

2.1 DanNet

Samarbejdet mellem et leksikografisk og et sprogteknologisk miljø i Danmark blev grundlagt i nullerne med igangsættelse af DanNet-projektet (Pedersen et al., 2009). I dette projekt arbejdede Center for Sprogteknologi (CST) ved Københavns Universitet og Det Danske Sprog- og Litteraturselskab (DSL) sammen om at udvikle et omfattende dansk wordnet baseret på Den Danske Ordbogs (DDO) definitioner. Genus proximum i DDO, der i forvejen var identificeret i ordbogs-manuskriptet, blev aksen hvorom et ordnetværk kunne genereres semiautomatisk og derefter justeres af sprogteknologer og leksikografer. Ressourcen rummer i dag knap 70.000 begreber organiseret i synsets, som er forsynet med bl.a. ontologiske typer og overbegreber. Samlet set rummer ressourcen mere end 300.000 indbyrdes semantiske relationer, og den udvides løbende med flere betydninger.

2.2 Begrebsordbog

Erfaringerne med at udarbejde et wordnet på basis af DDO var afgørende for tilblivelsen af den senere danske tesaurus, kaldet Den Danske Begrebsordbog (Begrebsordbogen), der blev skrevet i årene 2010-2015 på DSL. I Begrebsordbogen tildeles betydningerne i DanNet yderligere en emnebetegnelse, og begreber listes i semantisk rækkefølge, opdelt i grupper af synonymer og nærsynonymer der indledes af et nøgleord, ofte et overbegreb. Samtidig blev flere DDO-lemmaer og -betydninger tilføjet så op mod 95 % af DDO er repræsenteret. Med Begrebsordbogens færdiggørelse åbnede der sig en række nye muligheder for at sammenkoble onomasiologiske og semasiologiske leksikalske oplysninger fra de tre ressourcer: et WordNet, en ordbog og en tesaurus.

2.3 Det Danske FrameNet-leksikon

De tematiske oplysninger og semantiske undergrupper i Begrebsordbogen blev kombineret med DDOs valensmønstre. Disse data dannede grundlag for udarbejdelsen af et dansk framenet-leksikon med fokus på verbalbetydninger. Da Begrebsordbogen både beskriver kollokationer fra den korpusbaserede DDO og ofte tildeler mere end ét tematisk afsnit til den enkelte verbalbetydning (dvs. beskriver betydninger anskuet fra forskellige vinkler), udgjorde datamaterialet et velegnet grundlag for tildeling af mulige frame-værdier fra den internationale standard Berkeley FrameNet til de enkelte ordbetydninger, igen med bevarelse af id-numrene fra DDO.

Ud fra Begrebsordbogens kapitler kunne overordnede kategorier, fx alle verber og verbalsubstantiver der omhandler kommunikation, behandles i samme ombæring så inventaret af mulige frames blev overskueligt. Leksikonet er tænkt som leksikografisk hjælp til semantisk ramme- og rolleopmærkning af danske tekster, men giver i sig selv værdifuld formaliseret semantisk information om de enkelte ordbetydninger (Nimb et al., 2017; Nimb, 2018).

2.4 Den Danske Sentimentlexikon

Inspireret af SenSALDOS sentimentleksikon (Rouces et al., 2018a; Rouces et al., 2018b) blev der også udarbejdet et sentimentleksikon ud fra Begrebsordbogens afsnitsopdeling. I afsnit hvor ordene så ud til at have inhærent polaritet baseret på afsnitsbetegnelsen, blev disse automatisk opmærket med hhv. positiv og/eller negativ polaritet. Ud af Begrebsordbogens 888 afsnit drejede det sig om 122 negative afsnit, fx "Tristhed", 80 positive afsnit, fx "Beundre", samt 12 afsnit hvor polariteten var uklar, fx "Omdømme".

De automatisk opmærkede ord blev derpå manuelt valideret. Ord uden polaritet blev tildelt værdien 0. 400 ord blev opmærket af to leksikografer for at måle annotørenigheden som var 0,83 (Cohens kappa). Graden af polaritet, +3 til -3 blev tilføjet ved at sammenligne dels med et tidligere dansk sentimentleksikon, AFINN, dels med ordets synonymer og nærsynonymer fra Begrebsordbogen idet den semantiske rækkefølge var bevaret i udtrækket. På lemmaniveau blev ord med divergerende polaritet derefter nærmere undersøgt, og det blev besluttet om en betydning eller et helt lemma skulle slettes. De neutrale ord blev fjernet fra listen. Resultatet er et sentimentleksikon med knap 14.000 polaritetsbærende lemmaer, heraf 62 % negative og 38 % positive (Nimb et al., 2022). Denne fordeling ligger i øvrigt tæt op ad fordelingen i det svenske SenSALDO-leksikon. Sentimentleksikonet er større end de allerede eksisterende sentimentleksikoner for dansk (for en evaluering se Schneidermann & Pedersen (2022)).

2.5 COR Centralt OrdRegister for dansk

I COR-projektet udvælges semantiske kerneoplysninger fra DanNet, FrameNet-leksikonnet og sentimentleksikonnet. Det betyder i korte træk at ressourcen har information om antal betydninger, ontologisk type, nærmeste danske overbegreb (fra DanNet), semantisk verbalramme (fra FrameNet) samt konnotation (positiv/negativ) fra Sentimentordbogen.

Udgangspunktet er denne gang 60.000 lemmaer i den danske retskrivningsordbog (RO) der nummerindekseres og forsynes med formaliserede morfologiske oplysninger, og som efterfølgende lanceres som en frit tilgængelig standardressource, administreret af Dansk Sprognævn (se også Nimb et al. (i trykken)).

Tanken er at fremtidige sprogteknologiske ordbøger kan koble sig på vha. id-numrene; dermed sikres en mere effektiv deling af danske leksikalske data. For en stor del af RO-ordforrådet tilkobles et betydningsinventar udarbejdet på basis af oplysningerne i de semantiske ressourcer nævnt ovenfor: DDO, DanNet, Begrebsordbogen, FrameNet-leksikonnet og sentimentleksikonnet. COR-ordbogen er semasiologisk i sin opbygning: Alle DDO-betydninger af et lemma tages i betragtning, ikke kun dem der fx er beskrevet i DanNet. Kun de væsentligste betydninger medtages imidlertid: Sjældne, gammeldags eller faglige DDO-betydninger af lemmaet udelades i COR, og samtidig slås nært beslægtede betydninger i DDO sammen så der opnås en mere grovkornet betydningsinddeling der egner sig bedre til sprogteknologisk anvendelse.

I udviklingen af ordbogen anvendes en række automatiske metoder der sikrer ensartet behandling på tværs af ordforrådet, og reduceringen af antal betydninger udføres automatisk for en del af de polyseme ord på baggrund af håndopmærkede data (Pedersen et al., 2022). Ordbogen, der lanceres i december 2023, vil omfatte ca. 30.000 lemmaer; ca. 11.000 af disse er udvalgt som værende særligt centrale i dansk ud fra enten deres kobling til centrale begreber udvalgt for engelsk[4] eller ud fra deres egenskab af nøgleord i Begrebsordbogen.

3 Afrundende bemærkninger

Det er særdeles omkostningstungt at udvikle leksikalsk-semantiske ressourcer til sprogteknologi. Særligt i de mindre sprogsamfund har vi slet ikke råd til ikke at få fuldt udbytte af disse i den teknologi der udvikles. Som vi har påpeget her, indeholder de semantiske ressourcer vigtig viden om bl.a. kultur og samfund, som ikke nødvendigvis fremgår af sprogmodeller som er bygget ud fra ren tekst. Hvis vi skal dybere ned i sprogforståelsen i fremtidig sprogteknologi, er det derfor nødvendigt at inkludere den viden som ressourcerne rummer.

Problemer med kuratering, opskalering og/eller manglende vedligehold har tidligere medført at nogle NLP-ressourcer udviklet i mindre projekter med kortvarig finansiering ikke altid er blevet fuldt udnyttet i et større perspektiv. De har simpelthen været for svære at integrere og anvende og har været for løsrevne fra andre, mere almene ordbøger, som typisk (og forhåbentlig) har kuratering, vedligeholdelse og opdatering med som en del af deres udvikling og finansiering. Derfor har vi i denne artikel også argumenteret for at leksikalsk-semantiske ressourcer til sprogteknologi bedst og sikrest udvikles inden for rammerne af en leksikalsk infrastruktur og i samarbejde med de klassiske ordbøger som et sprogsamfund investerer midler i i forvejen. I den forbindelse har vi også fremhævet ELEXIS-projektet (elex.is) som et vigtigt netværk der arbejder henimod at muliggøre sådan en infrastruktur både i et nationalt og et internationalt perspektiv.

[4] Der er taget udgangspunkt i det engelske *core wordnet* som udgør 5000 centrale begreber på engelsk, og som kan downloades her: https://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt.

Referencer

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, & Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Simon Krek, Iztok Kosem, John P McCrae, Roberto Navigli, Bolette S Pedersen, Carole Tiberius, & Tanja Wissik. 2018. European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*, pages 881–892.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, & Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing, 166172. Florence, Italy: Association for Computational Linguistics*, pages 166–172.

Sanni Nimb, Anna Braasch, Sussi Olsen, Bolette Sandford Pedersen, & Anders Søgaard. 2017. From Thesaurus to FrameNet. In *Electronic Lexicography in the 21st century: Proceedings of eLex 2017 conference*, pages 1–22.

Sanni Nimb, Sussi Olsen, Bolette S Pedersen, & Thomas Troelsgård. 2022. A Thesaurus-Based Sentiment Lexicon for Danish–The Danish Sentiment Lexicon. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC12), Marseille*, pages 2826–2832.

Sanni Nimb, Bolette Sandford Pedersen, Nicolai Hartvig Hau Sørensen, I. Flörke, Sussi Olsen, & T. Troelsgård. (i trykken). COR-S den semantiske del af Det Centrale OrdRegister (COR). *LexicoNordica 29, Nordisk Forening for Leksikografi.*

Sanni Nimb. 2018. The Danish FrameNet Lexicon: method and lexical coverage. In *Proceedings of the International FrameNet Workshop at LREC, Miyazaki*, pages 51–55.

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, & Henrik Lorentzen. 2009. DanNet: the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.

Bolette Sandford Pedersen, Sanni Nimb, & Sussi Olsen. 2021. Dansk betydningsinventar i et datalingvistisk perspektiv. *Danske Studier 2021, Universitets-Jubilæets danske Samfund 2021*, pages 72–106.

Bolette S. Pedersen, Nicolai C.H. Sørensen, Sanni Nimb, S. Flörke, Sussi Olsen, , & Thomas Troelsgård. 2022. Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open-Source COR Lexicon. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC12), Marseille*, pages 51–60.

Jacobo Rouces, Lars Borin, Nina Tahmasebi, & Stian Rødven Eide. 2018a. Defining a Gold Standard for a Swedish Sentiment Lexicon: Towards Higher-Yield Text Mining in the Digital Humanities. In *CEUR Workshop Proceedings vol. 2084. Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference Helsinki, Finland, March 7-9, 2018*, pages 219–227.

Jacobo Rouces, Lars Borin, Nina Tahmasebi, & Stian Rødven Eide. 2018b. SenSALDO: Creating a sentiment lexicon for Swedish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, ELRA*, pages 4192–4198.

Nina Skovgaard Schneidermann & Bolette Sandford Pedersen. 2022. Evaluating a New Danish Sentiment Resource: the Danish Sentiment Lexicon, DSL. In *Proceedings for SALLD2 - 2nd Workshop on Sentiment Analysis and Linguistic Linked Data. European Language Resources Association.*, pages 19–25.

Agnes Sólmundsdóttir, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir, & Anton Karl Ingason. 2022. Mean Machine Translations: On Gender Bias in Icelandic Machine Translations.

Benjamin D Van Durme. 2009. *Extracting Implicit Knowledge from Text. PhD Thesis*. University of Rochester.

Den ena texten och den andra: Visualisering av textpar med hjälp av ordmoln

Maria Skeppstedt[1], **Gunnar Eriksson**[2], **Magnus Ahltorp**[2] **och Rickard Domeij**[2]
[1]Centrum för digital humaniora Uppsala, Institutionen för ABM, Uppsala universitet
maria.skeppstedt@abm.uu.se
[2]Språkrådet, Institutet för språk och folkminnen

Abstract

Word clouds are commonly used for visualising the content of text collections. We here propose a slight update of the standard word cloud that also visualises similarities between pairs of texts. We apply the method to a text collection consisting of 39 text lyrics and visualise similarities between four text pairs.

1 Inledning

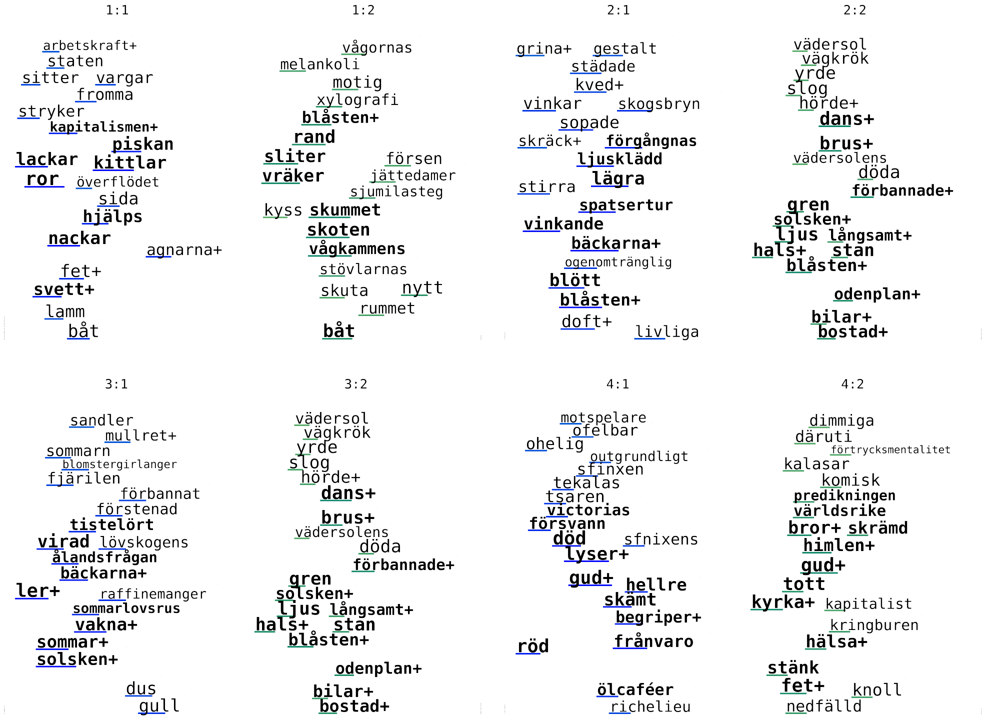
Ordmoln, det som på engelska kallas ”tag clouds” eller ”word clouds” är ett väldigt populärt sätt att visualisera innehållet i en text, eller de taggar som är associerade med en text. Den enklaste formen av ordmoln visar orden med en större font ju mer frekvent förekommande de är i texten/bland ordtaggarna, och arrangerar sedan orden i ett moln, exempelvis i alfabetisk ordning (Viégas & Wattenberg, 2008). Det finns färdiga webbtjänster för att generera ordmoln, vilket skulle kunna vara en anledning till deras popularitet. En annan anledning skulle kunna vara att ordmoln är en väldigt enkel och lättförståelig visualisering. Det har dock riktats kritik mot de klassiska ordmolnen, exempelvis för att betraktaren lätt kan tolka in en betydelse i hur orden är placerade, även när en sådan innebörd saknas (Barth et al., 2014). Att använda fontstorlek som indikation på ett ords signifikans kan också vara problematiskt, i och med att långa ord då lätt kan uppfattas som mer viktiga, i och med att deras längd gör att de tar upp mer plats i ordmolnet (Viégas & Wattenberg, 2008).

Det finns många olika varianter av standardversionen av ordmolnen. Andra kriterier än ordfrekvens, såsom TF-IDF-måttet (Barth et al., 2014), kan användas. Det finns även olika metoder för att placera orden i molnet så att placeringen får en innebörd, exempelvis Context-preserving word cloud visualisation (Barth et al., 2014) och t-SNE, d.v.s. ”t-distributed stochastic neighbour embedding” (Schubert et al., 2017). Det finns också flera utökningar av ordmolnens syfte, d.v.s. ordmoln som, förutom att visualisera innehållet i en text, även har till syfte att visualisera andra aspekter. Exempelvis visar temporala ordmoln hur ordens frekvens varierat över tid (Jatowt et al., 2021).

Vi hade också för avsikt att utöka ordmolnens syfte. Förutom att använda ordmoln för att (i) visualisera innehållet i dokumenten i den textsamling vi undersökte, ville vi (ii) samtidigt även kunna visualisera textlikhet mellan olika dokument. Vi har använt ett verktyg för att söka bland tidigare textvisualiseringsinitiativ (Kucher & Kerren, 2015)[1], men inte hittat någon färdig metod för visualisering som fungerar för exakt det vi vill visa. Vi kommer därför här att designa en enkel metod för att visualisera båda dessa aspekter.

[1] https://textvis.lnu.se

Maria Skeppstedt, Gunnar Eriksson, Magnus Ahltorp and Rickard Domeij. 2022. Den ena texten och den andra: Visualisering av textpar med hjälp av ordmoln. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 121–125. Available under CC BY 4.0



Figur 1: Fyra par av texter visas i figuren. Par 1 består av gruppens mest kända låt och den text som var mest lik denna låttext. Par 2 består av de två texter som var allra mest lika enligt det likhetskriterium vi använde, par 3 av de näst mest lika texterna, o.s.v. (Titeln på låtarna står med liten stil vid sidan av graferna. Så för den som vill gissa vilken låttext som visualiseras, rekommenderar vi att inte kika på titeln i förväg.)

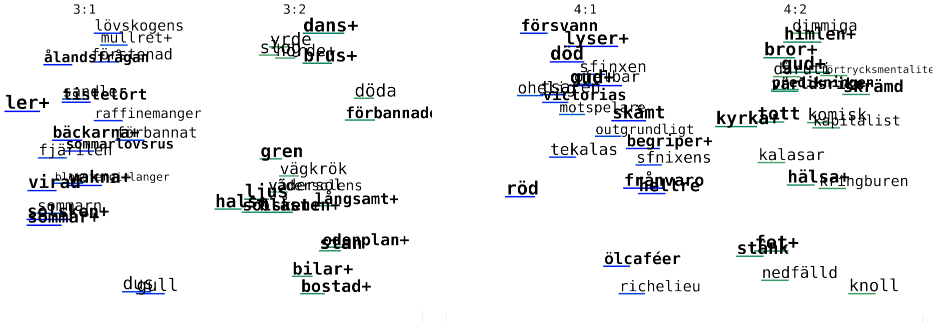
2 Textsamlingen

Vi samlade in texter från en svensk musikgrupp som främst var verksam under sent 60-tal och första halvan av 70-talet. Vi hittade texterna på en webbsida med låttexter från gruppen[2]. Webbsidan innehöll även låtar skrivna av bandets medlemmar i andra kontexter än tillsammans med gruppen. Vi allokerade cirka tre timmar för att extrahera låttexter från webbsidan, och formatera dem i ett enhetligt textformat. Vi gjorde en begränsad ansträngning för att enbart ta med låtar skrivna för gruppen, men även en viss mängd låtar skrivna av bandmedlemmarna i andra sammanhang kan ha tagits med i vår textmängd. Den allokerade tiden räckte till att skapa en textmängd bestående av 39 låttexter.

3 Metod

Metoden bestod av två delar. Dels skapade vi en dokumentvektor för varje låttext, som vi använde som ett kriterium för att välja ut fyra par av låttexter, och dels gjorde vi en visualisering av orden i dessa åtta låttexter.

[2] Webbsidan hade adressen <förnamn><efternamn>.se för en av bandmedlemmarna.



Figur 2: De två nedre textparen från figur 1, men där orden har behållt originalplaceringen genererad av t-SNE-algoritmen. (Titeln på låtarna står med liten stil vid sidan av graferna. Så för den som vill gissa vilken låttext som visualiseras, rekommenderar vi att inte kika på titeln i förväg.)

3.1 Dokumentvektorer

Skapandet av TF-IDF-vektorer för dokumenten i vår textsamling bestod av följande steg: (i) generera statistik över svenska dokumentfrekvenser utifrån en extern textmängd, (ii) identifiera vanliga begreppskluster i vår textmängd, och byta ut orden i dessa kluster mot en sträng som representerar klustret, (iii) skapa en TF-IDF-vektor för varje dokument, (iv) skapa den slutgiltiga dokumentvektorn genom att utöka TF-IDF-vektorn med word2vec-vektorer som representerar dokumentet.

(i) För att ha lite mer bakgrundsdata för svenska dokumentfrekvenser använde vi en stor extern korpus för IDF-uträkningen, närmare bestämt stycken från 1000 SOU:er[3], där varje stycke behandlades som ett dokument. (Det var en väldigt slumpmässigt vald textmängd, men en mängd som borde ge tillräckligt bra information om typiska svenska orddokumentfrekvenser för vårt syfte.)

(ii) Innan vi konstruerade TF-IDF-vektorerna, gjorde vi en pre-processning av texterna där vi bytte ut vissa ord i texten mot en sträng som representerar ett synonym- eller begreppskluster där ordet ingår. Detta gjordes både för bakgrundstexten och för de dokument vi ville visualisera. Exempelvis bestod ett av våra synonymkluster av orden ”arbeta/jobba/verka”. Det innebar att varje gång någon av dessa tre ord förekom i en text bytte vi ut ordet mot strängen ”arbeta/jobba/verka”. Dessa tre ord blev då behandlade som ett och samma ord när TF-IDF-vektorer skapades. Vi skapade synonymklustren genom att köra en dbscan-klustring (Ester et al., 1996) på word2vec-vektorer för orden i vår textsamling, och därefter manuellt gick igenom och rättade alla automatiskt skapade kluster. Totalt skapade vi 160 kluster. Ordvektorerna hittade vi genom att använda ett förtränat word2vec-ordrum[4] med 100 element långa vektorer.

(iii) Efter pre-processningen skapade vi TF-IDF-vektorer för alla dokument i vår textmängd. Vi gjorde inte någon kontroll och/eller borttagande av dubblerad text, trots att detta är ett vanligt fenomen i låttexter, exempelvis eftersom det ofta finns refränger. Anledningen var att om ord ofta återkom, som till exempel i refränger, så skulle detta också faktiskt avspeglas i en högre ordfrekvens för dessa ord. För att inte göra sådana ord helt dominerande använde vi emellertid inte standard TF utan dess logaritmiska värde.

(iv) Till den vanliga TF-IDF-vektorn konkatenerade vi sedan kombinerade ordvektorer. Detta för att fånga likhet mellan dokument utan överlappande ord eller begreppskluster. Samma word2vec-ordrum som ovan användes. Vi gav på flera sätt större vikt till ord med ett högt TF-IDF även för de kombinerade ordvektorerna. För det första konstruerade vi tre olika summerade ordvektorer, som vi konkatenerade

[3] https://github.com/UppsalaNLP/SOU-corpus

[4] http://vectors.nlpl.eu/repository/ *Word2Vec Continuous Skipgram* tränad på *Swedish CoNLL17 corpus*

till den vanliga TF-IDF-vektorn. Två vektorer bestående av summan av vektorerna för orden med de tre respektive tio högsta TF-IDF-värdena, och en vektor bestående av summan för alla ord i dokumentet. För det andra multiplicerade vi alla summerade vektorer med ordets TF-IDF-värde innan vi normaliserade vektorerna.

Alla beskrivna experiment utfördes med hjälp av maskininlärningsbiblioteket scikit-learn (Pedregosa et al., 2011).

3.2 Ordvisualisering

För varje par av dokument beräknades det euklidiska avståndet mellan dokumentens vektorer, och de tre par som var mest lika valdes ut. Paret bestående av gruppens mest kända låt och dess mest lika låttext i dokumentmängden valdes också ut. För dessa fyra par skapade vi sedan en enkel visualisering i Pyplot.

Vi skapade en bild vardera för de två texterna i paret, och placerade den ena till vänster och den andra till höger. I bilden skapade vi ett ordmoln med de 25 ord som hade högst TF-IDF-värde. Ju högre TF-IDF-värde, med desto starkare färg skrevs ordet, för ordmolnet till vänster i blå nyanser och för ordmolnet till höger i gröna nyanser. Kritiken mot att använda fontstorlek som indikation på ett ords signifikans till trots, visade vi ett ord med en större fontstorlek ju högre dess TF-IDF-värde var. Vi motiverar det med att ordmoln med varierande fontstorlek har blivit någon form av standard, på grund av dess popularitet. Därmed finns det en poäng i att på något sätt hålla sig till den standarden. Vi gjorde dock endast en liten ökning av fontstorleken med ökande TF-IDF-värde. Dessutom gjorde vi en liten, generell fontminskning av ord beroende på deras längd, för att minska risken att långa ord skulle kunna uppfattas som viktigare än korta ord. För att ge betraktaren en mer objektiv indikation på ett ords TF-IDF-värde än ordets färg och fontstorlek lade vi även dit en understrykning av ord, där längden på det streck med vilket ordet är understruket är direkt proportionellt mot ordets (logaritmiska TF)-IDF-värde.

För att lägga en verklig betydelse i ordens placering i molnet, skapade vi en utplacering som placerade ord med liknande betydelse nära varandra. Vi bestämde ordens placering genom att köra t-SNE-algoritmen (van der Maaten & Hinton, 2008) på word2vec-vektorerna som hörde till orden i texten. Att placera ut orden exakt på den plats som bestämdes av t-SNE-algoritmen skulle emellertid göra att de överlappade varandra och bli svårlästa. Vi provade att använda ett standardbibliotek, adjustText, som placerar om text för att undvika detta. Dock ledde det till att orden då istället förlorade den struktur de givits av t-SNE-algoritmen, trots att vi experimenterade med olika parametrar till adjustText. Vi implementerade därför istället en egen enkel algoritm för att flytta om orden. Algoritmen placerar först ut det ord som har högst TF-IDF-värde, och fortsätter sedan nedåt med fallande TF-IDF-värde. Om ett nytt ord som ska placeras ut överlappar ett ord som redan är utplacerat flyttas det nya ordet uppåt i grafen tills det nya ordet inte längre överlappar med ett tidigare utplacerat ord. Detta gör att de ord som är viktigast för texten oftare placeras i enlighet med dess innebörd, medan mindre viktiga ord kan flyttas runt.

Ord som ingår i ett synonymkluster indikeras med ett plus efter ordet. Endast det första ordet i ett kluster visas.

4 Resultat

De fyra par av texter som vi visualiserade visas i figur 1, och i figur 2 visas hur visualiseringen skulle ha sett ut om vi inte flyttat orden för att undvika överlapp.

Bilderna är ett försök att både skapa en visualisering av texterna där ordens placering har en innebörd, och, genom att placera texterna i par, även skapa en visualisering där två texter lätt kan jämföras. Vi vill att visualiseringen snabbt ska kunna svara på frågan: Vad innehåller den ena texten, och vad innehåller den andra?

Vi lämnar utvärderingen av huruvida visualiseringen ger en bra översikt över textparen till läsarna, särskilt till läsare som är väl förtrogna med musikgruppen i fråga. Går det exempelvis att gissa vilken gruppen är? Går det att gissa vilka låttexter som visualiseras i bilderna?

Tack

Vi vill tacka Nationella språkbanken (Vetenskapsrådet, 2017-00626), HUMINFRA (Vetenskapsrådet, 2021-00176) samt Dagstuhl-seminariet ”Visual Text Analytics”.

Referenser

Lukas Barth, Stephen G. Kobourov, & Sergey Pupyrev. 2014. Experimental comparison of semantic word clouds. In Joachim Gudmundsson & Jyrki Katajainen, editors, *Experimental Algorithms*, pages 247–258, Cham. Springer International Publishing.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, & Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Palo Alto, California. AAAI Press.

Adam Jatowt, Nina Tahmasebi, & Lars Borin. 2021. Computational approaches to lexical semantic change: Visualization systems and novel applications. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, & Simon Hengchen, editors, *Computational approaches to semantic change*. Language Science Press.

Kostiantyn Kucher & Andreas Kerren. 2015. Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*, pages 117–121.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, & Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Erich Schubert, Andreas Spitz, Michael Weiler, Johanna Geiß, & Michael Gertz. 2017. Semantic word clouds with background corpus normalization and t-distributed stochastic neighbor embedding. *ArXiv*, abs/1708.03569.

Laurens van der Maaten & Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Fernanda B. Viégas & Martin Wattenberg. 2008. Timelines tag clouds and the case for vernacular visualization. *Interactions*, 15(4):4952.

Datorlingvistikens vagga i Uppsala

Anna Sågvall Hein
Uppsala universitet
anna@lingfil.uu.se

Abstract

Computational Linguistics at Uppsala University has a long history. It goes back to the mid 1960-ies, when the University Computer Center, UDAC, was established. It should provide a powerful computer resource for research at the university and encourage the use of computers in new fields such as the Humanities. With this aim in mind, the director of UDAC approached the language departments. This initiative led to a contact with the Slavic department and a joint decision to support an ongoing thesis work in Slavic languages, directed towards computational linguistics. On completion of the thesis, the researcher was employed at UDAC with the mission to do research in Computational Linguistics and support Natural Language Processing. This was the embryo of the Natural Language Processing group at UDAC and its follower Center for Computational Linguistics at the Faculty of Humanities. We describe activities in this early period with a focus on corpus linguistics, process morphology and program development, where pioneering contributions were made by Lars Borin. We also emphasise the importance of the consistent and close cooperation between the Faculty of Humanities and UDAC, a prerequisite for the successful development of Computational Linguistics in Uppsala and the establishment of a chair in 1987.

1 Inledning

Datorlingvistiken har en lång historia i Uppsala. Den går tillbaka till 1960-talet, närmare bestämt till 1965, då Uppsala universitetsdatacentral, UDAC, inrättades. Universitetsdatacentralerna skulle tillgodose kända behov av datorkraft för forskning i teknik och naturvetenskap men också verka för att initiera användning av datorer inom andra områden.

Den nytillträdde chefen för UDAC, docent Werner Schneider, tog sig omedelbart an uppgiften. Han inriktade sig mot den Humanistiska fakulteten och bjöd in institutionerna till seminarier, där han informerade om vilka möjligheter den nya tekniken kunde erbjuda. Pågående och planerad forskning presenterades och institutionerna fick möjlighet att föreslå pilotprojekt. Av de språkvetenskapliga institutionerna var det Slaviska institutionen, företrädd av docent Carin Davidsson, som visade störst intresse.

Två pilotprojekt specificerades. Det ena gällde finalalfabetisk sortering av ett tjeckiskt ordboksmaterial. Det andra handlade om att utveckla datorbaserade verktyg för lingvistiska ändamål. Ett licentiandarbete i slaviska språk med datorlingvistisk inriktning pågick redan. Det bedrevs av en studerande på stipendium i Leningrad. Gemensamt beslutade man stödja detta projekt med programmeringshjälp och annat som behövdes. Satsningen föll väl ut och resulterade i en avhandling om automatisk analys av det ryska verbet (Sågvall, 1968) samt att licentiaten anställdes på UDAC för att bedriva forskning i datorlingvistik samt understödja språklig databehandling. Det blev embryot till Språkgruppen på UDAC (Natural Language Processing Group) och dess uppföljare Centrum för datorlingvistik vid Humanistiska fakulteten. 1987 beslutade regeringen om en professor i datorlingvistik.

En av pionjärerna var Lars Borin. Med sitt breda språkvetenskapliga och datavetenskapliga kunnande bidrog han på ett ovärderligt vis till utvecklingen av forskningsmiljön. Hans insatser gällde främst rysk korpuslingvistik, processmorfologi och programutveckling, intresseområden som får särskild uppmärksamhet i framställningen nedan.

Anna Sågvall Hein. 2022. Datorlingvistikens vagga i Uppsala. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 127–132. Available under CC BY 4.0

2 Språkgruppen (Natural Language Processing Group, NALP)

Intresset för språklig databehandling visade sig vara stort på fakulteten. Bland de språk som databehandlades i initialskedet fanns svenska, engelska, finska, tjeckiska, estniska, tyska, franska, persiska och ryska. Senare tillkom sanskrit och ungerska. I samverkan med språkvetare från de olika institutionerna byggde medlemmar i språkgruppen[1] upp språkliga resurser för såväl forskning som undervisning (NALP, 1978). De svarade också för databehandlingen vid genomförandet av undersökningarna (Sågvall et al., 1975). För engelskans del införskaffades the Brown Corpus.[2] Den förelåg i ett primitivt hålkortsformat men moderniserade så att den blev användbar för de svenska forskarna. Moderniseringen handlade främst om omkodning av teckenrepresentationen samt uppdelning av korpusen i meningar (sentences), så att användaren kunde arbeta med dem i stället för hålkortsbilder.

När verksamheten inleddes 1970 fanns inga färdiga program för språklig databehandling att tillgå. De fick skrivas av medlemmar i språkgruppen allt efter behov (Sågvall et al., 1974a).

2.1 Infrastruktur

1970 såg infrastukturen helt annorlunda ut jämfört med idag. Alla körningar gjordes på en stordator (IBM 370/155). Den stod i en maskinhall till vilken bara särskilda operatörer hade tillträde. Inmatning av såväl program som data skedde via hålkort. Den utfördes av stansoperatriser. Hålkorten representerade alfanumeriska tecken enligt en IBM-standard som var utformad för de västeuropeiska språken. Den uteslöts språk som ryska, polska, ungerska och tjeckiska. För dem krävdes transliteration. För ryskans del fastställdes en konvention i samarbete med den stansoperatris som skulle mata in materialet. Man bemödade sig om att finna sådana motsvarigheter som på grund av grafiska likheter kunde memoreras av den som inte kunde det kyrilliska alfabetet.

Även resultaten av körningarna kom ut på hålkort och hålkortsbilderna kunde skrivas ut på en radskrivare. Det var inget tilltalande format för språkliga data, särskilt inte för data i translittererad form. Utmatning med specialalfabeten var ett särskilt problem. Det löstes i vissa fall med hjälp av en printer-plotter. Den kunde programmeras till att rita godtyckliga bokstäver och så skedde för de grekiska, kyrilliska och franska alfabetena (med accenter). Plottern kunde också växla mellan olika fonter på en och samma rad, något som behövdes för utskrift av rysk-svenska ordlistor i ett pedagogiskt projekt. Tryckaccent för de ryska orden skrevs också ut. En annan lösning var att låta datorn producera binärkort som kunde läsas av en skrivmaskin med utbytbart typhuvud som var kopplad till en minidator. Metoden testades ut och användes för produktion av det tjeckiska lexikonet inför tryckningen.

2.2 Rysk morfologi

Till en början var det ryska som stod i centrum. Den automatiska analysen av det ryska verbet följdes av en modell för analys av hela den ryska böjningsmorfologin (Sågvall, 1973). Systemet, AUTLEX, tilldelade ryska ord lexikala beskrivningar som bland annat upptog ordklass, lemma och böjningsform. Ordklassindelningen byggde strikt på böjningskategorier och skiljde sig därigenom något från den traditionella ordklassindelningen. AULEX använde sig av ett stamlexikon och ord som saknades i lexikonet fick ingen analys medan homografa ord fick flera analyser, en för varje homografkomponent, AUTLEX användes flitigt för annotering av ryska korpusdata. Systemet opererade på ord i isolerad ställning och kunde således inte skilja ut kontextberoende homografer. För detta krävdes separat homografseparering. Den utfördes till en början helt manuellt, men gradvis introducerades metoder för att underlätta arbetet.

Den första större tillämpningen gällde ett pedagogiskt projektet *Målanalys i ryska* (Sågvall et al., 1976). Det genomfördes med bidrag från Universitetskanslersämbetet. Med hjälp av AUTLEX annoterades ett urval ryska litterära texter avsedda för textläsning på universitetets A-nivå. Genom statistiska beräkningar kunde de annoterade texterna ordnas i stigande svårighetsgrad efter ordrikedom. Vidare kunde man definiera den lexikala progressionen vid läsning av texterna i den fastställda ordningen. Det

[1] Medlemmarna i Språkgruppen växlade under årens lopp. 1978 bestod den av Annika Mattsson, Uwe Hein, Tone Tingsgård och Erling Wande. Gruppen leddes av Anna Sågvall Hein.

[2] A Standard Sample of Present-day Edited American English. 1964. The Brown university, Providence R.I., USA.

var grunden för utveckling av ett läromedel för rysk textläsning *Text och Ord* (Sågvall et al., 1974b).[3]

En annan större tillämpning gällde *textattribution*.[4] Den behandlade frågan om huruvida nobelpristagaren *Sjolochov* skrivit hela *Stilla flyter Don*. Det hade hävdats att en del av verket var skriven av en vitrysk officer *Krjukov*. Studien byggde på tre korpusar om vardera 50 000 ord, där en innehöll säker Sjolochovtext, en säker Krjukovtext och en tredje den omdiskuterade delen av verket.

Korpusarna matades in i datorn via hålkort och genomgick olika kvantitativa beräkningar. Det rörde sig om antal stavelser, medellängd per 1000 löpande ord, antal olika ordformer per 1000 löpande ord samt frekvens och frekvensdistribution. Motsvarande beräkningar gjordes på lexemnivå. På den syntaktiska sidan jämförde man meningsinledning i de olika korpusarna genom att undersöka sekvenser av lexem och ordklasser. En sammanställning av de olika statistiska beräkningarna visade att *Krjukov* kunde uteslutas som möjlig författare, men inte Sjolochov (Kjetsaa et al., 1984).

Den språkliga databehandlingen utfördes av Språkgruppen och senare Centrum för Datorlingvistik. Den innefattade igenkänning av lingvistiska begrepp som stavelser, ordformer, lexem, ordklasser, meningar och skiljetecken samt beräkningar på dessa enheter. Igenkänning av lexem och ordklasser gjordes genom analys med AUTLEX och homografseparering. Den underlättades genom ett nyutvecklat interaktivt program. Analysen föregicks av korpusanpassning av lexikonet.

2.3 Parsning

Forskning om generella metoder för parsning (lingvistisk analys) inleddes. Efter utprovning av *Augmented Transition Network Grammars* (Woods, 1973) inriktades forskningen mot nätverksbaserade analysmodeller och procedurella formalismer. Ett första konkreta resultat var implementering av en experimentversion av en chartparser. Den byggde på föreläsningsanteckningar om en lingvistisk processor som presenterades vid en internationell sommarskola i Pisa (Kay, 1974). Processorn simulerade en icke-deterministisk maskin och alternativa anlyser lagrades i en central datastruktur, benämnd chart. Charten var en riktad graf vars bågar bar lingvistisk information. Den gjorde det möjligt att hantera flera olika processer i ett och samma ramverk. De lingvistiska reglerna lagrades i s.k. väntelistor i bågarna.

Uppsalaparsern (Sågvall et al., 1975) medgav experiment med morfologisk och syntaktisk analys inklusive morfografematisk omskrivning och lexikonsökning. Med hjälp av en överordnad funktion kunde man välja vilken process man ville utföra. Parsern testades på en liten svensk grammatik. Implementeringen utgick från ett 80-tal LISP-funktioner, som Kay tillhandahöll. Den utgjorde en värdefull miljö för kompentensuppbyggnad.

Senare presenterade Kay (1977) en generalisering av principerna för processning i en chart parser. Det skedde genom en utveckling av charten. Inte bara lingvistisk information utan också regler representerades av bågar, passiva och aktiva. Genom ett samspel mellan aktiva och passiva bågar drevs processningen framåt. Kay ställde originalprogramvaran av den nya parsern till förfogande för språkgruppen och vid en forskningsvistelse i Uppsala medverkade han i implementeringen av den. Det gjordes i en lokal version av INTERLISP på UDAC:s stordator. Det blev utgångspunkt för utvecklingen av Uppsala Chart Processor, UCP.

UCP hanterade fonologisk, morfologisk och syntaktisk analys inklusive morfografematisk omskrivning och lexikonsökning. Den grafematiska omskrivningen och den morfologiska analysen provades ut på flera språk, i första hand finska (Sågvall Hein, 1977; Sågvall Hein, 1979) men också på ryska och serbokroatiska. Sågvall Hein (1980) föreslår en modell för finska där ordigenkänningen sker i två parallella processer, fonologisk analys i stavelser och morfologisk analys i morfer.

3 Centrum för datorlingvistik

På förslag från fakulteten inkorporerades språkgruppen i Humanistiska fakulteten som Uppsala Centrum för Datorlingvistik, UCDL (Sågvall Hein, 1981). Det skedde 1980. Enligt en fastställd instruktion

[3] *Text och ord* omfattade texterna ordnade efter svårighetsgrad åtföljda av kommenterade ordlistor med översättning, ett basordförråd samt ett ackumulerat ordförråd. Hela publikationen gjordes digitalt och skrevs ut med Benson Printer Plotter. Trots det primitiva formatet kom läromedlet att användas på universitetet och vid Arméns tolkskola under ett 10-tal år.

[4] Projektet var ett samarbete mellan Institutionen för slaviska språk vid Uppsala universitet och Slavisk-Baltisk Institutt vid Oslo universitet.

skulle verksamheten koncentreras till forskning i datorlingvistik samt verka för att genomföra språkvetenskapliga sektionens handlingsprogram för språklig databehandling samt främja historisk-filosofiska sektionens handlingsprogram för databehandling i textbaserade undersökningar. Humanistiska fakultetsnämnden och UDAC fick gemensamt ansvar för planering och uppföljning av verksamheten.

I mars 1981 togs ett viktigt steg i utvecklingen. Då fastställde Språkvetenskapliga sektionsnämnden en ämnesbeskrivning av datorlingvistik. Datorlingvistik är inriktad på simulering av språkligt beteende med dator. Till ämnet hör även allmän metodik för undersökningar av språk med hjälp av dator (språklig databehandling). En docenttjänst i datorlingvistik skapades. Dess innehavare skulle leda verksamheten på centret. Vidare fick centret tre deltidstjänster för programmering och systemarbete samt två språkkonsulttjänster, en i Nordiska språk (Olle Hammermo) och en i Slaviska språk (Lars Borin). Tidigare verksamhet inom språkgruppen följdes upp i Centrum för datorlingvistik med visst fokus på chartparsning och processmorfologi. Programutvecklingen ägnades också fortsatt uppmärksamhet.

3.1 Chartparsning

På parsningssidan inriktades forskningen primärt mot vidareutveckling av Uppsala Chart Parser (Carlsson, 1982; Sågvall Hein, 1980; Sågvall Hein, 1987)[5] samt uppbyggnad av en parser för svenska med tillhörande lexikon (Sågvall Hein, 1983; Sågvall Hein & Ahrenberg, 1985).

3.2 Processmorfologi

Forskning om generella metoder för processmorfologi inleddes. Utgångspunkt var *Tvånivåmorfologi* (Koskenniemi, 1983). Medan tidigare modeller för morfologisk processning varit inriktade på analys eller syntes, så möjliggjorde tvånivåmorfologin såväl analys som syntes. Med morfologisk analys förstås en process där en ytrepresentation, en bokstavssträng, tilldelas en morfologisk beskrivning (t.ex. AUTLEX), medan morfologisk syntes går den omvända vägen, dvs. från en morfologisk beskrivning till en ytrepresentation.

Tvånivåmorfologin erbjuder en generell, riktningsoberoende formalism samt ett processerande maskineri. Maskineriet utgörs av *finite-state-automater*, FSA. En FSA är en abstrakt maskin som vid en given tidpunkt befinner sig i ett av ett ändligt antal tillstånd. Automaten definieras av en lista över möjliga tillstånd, ett initialt tillstånd och ett finalt tillstånd, och regler som styr övergången från ett tillstånd till ett annat. Tillstånden kan ses som noder i ett övergångsnätverk. I tvånivåmorfologin utgörs noderna av tecken på två nivåer, lexikal nivå och ytnivå. Övergången från en nod till en annan styrs av parallella regler. De jämför strängar på de båda nivåerna och anger om de är tillåtna motsvarigheter till varandra eller inte. För utprovning av systemet ingick en beskrivning av finsk morfologi.

Tvånivåmorfologin fick stor spridning internationellt och implementerades tidigt i Uppsala. Lars Borin inledde utforskandet av tvånivåmorfologin på polsk böjningsmorfologi. I en kurs i språklig databehandling utgick han från en tvånivåbeskrivning för svenska (Blåberg, 1984) som han vidareutvecklade (Borin, 1985). Processmorfologi kom också att bli hans primära forskningsområde inom vilket han skrev sin doktorsavhandling (Borin, 1991).

3.3 Programvaruutveckling

Utvecklingen av programvara för språklig databehandling fortsatte med TEXTPACK (Rosén & Sjöberg, 1985) som bas. De dataformat som programmen arbetade med var i hög grad standardiserade, men åtkomstmetoderna var avhängiga av operativsystemet (IBM 370/155) och de accessoperatorer som programmeringsspråket (PL/I) tillhandahöll.

För att komma ifrån detta maskinberoende inledde Borin (1984) arbetet med att definiera och implementera ett generellt textbearbetningssystem, ett textdatabassystem för lingvister. Det skulle vara användarvänligt, generellt och portabelt. Han gjorde en djupdykning i det ungerska alfabetet för att demonstrera de problem man kan stöta på vid alfabetisering av språkliga data. I den föreslagna lösningen skulle man använda sig av ett användardefinierat alfabet med en 16-bitsrepresentation av varje tecken. Digram, trigram osv. skulle räknas som egna bokstäver och behandlas som odelbara enheter. De kommersiellt

[5] Utvecklingen av UCP bedrevs i projektet *Computer Simulation of the Text Comprehension Process* med stöd från Styrelsen för Teknisk Utveckling samt UDAC.

tillgängliga sorteringsprogrammen kunde inte komma åt problemen med bokstavsdiagram och längre enheter.

Med definitionen och den inledande implementeringen av detta första textdatabassystem såddes ett frö till kommande användning av språkdatabaser för lagring och åtkomst av språkliga data, något som kommit att bli standard i Uppsala.

4 Konklusion

De tidiga satsningarna på datorlingvistik i Uppsala bar frukt 1987, då regeringen beslutade inrätta en professur i ämnet. I propositionen konstateras att *Universitetet i Uppsala har utifrån tidigare gjorda satsningar goda förutsättningar att erbjuda en gynnsam forskningsmiljö i detta ämne, [...] Jag förordar att en professur i datorlingvistik inrättas*[6]. Ämnesbeskrivningen för docenturen behölls och tidigare forskning följdes upp. Nya forskningsområden var bland annat maskinöversättning och datoriserad språkgranskning.

Genom professuren legitimerades ämnet och viktiga steg kunde tas. Både forskarutbildning och grundutbildning anordnades. En plan för forskarutbildningen fastställdes 1990 och grundutbildning inleddes 1995 med *Språkteknologiprogrammet*. Det utgjorde den främsta rekryteringsbasen för forskarutbildningen. Programstudenterna var också efterfrågade på arbetsmarknaden.

Den första doktorn i datorlingvistik var Lars Borin. Med sin doktorsexamen representerar han ett av Uppsalas viktigaste bidrag till ämnets utveckling, nationellt och internationellt. Lars och hans meddoktorer gör, och har gjort, viktiga insatser inom universitetsvärlden och i samhället i övrigt. Noteras kan att utvecklingen av den artificiella intelligensen kan spåras tillbaka till maskininlärning och datadrivna metoder som bland annat introducerats inom datorlingvistiken. I sin doktorsavhandling presenterar Borin (1991) en sådan ansats.

Det har varit ett nöje att få följa Lars på vägen från doktorand i slaviska språk till doktor i datorlingvistik. Med tacksamhet ser vi på de bidrag till den tidiga utvecklingen av verksamheten som han givit.

Det långa och nära samarbetet mellan Språkvetenskapliga sektionen och Uppsala datacentral framstår som avgörande för den positiva utvecklingen av datorlingvistiken i Uppsala.

Referenser

Olli Blåberg. 1984. Svensk Böjningsmorfologi. En tvånivåbeskrivning. *Unpublished Master's Thesis, Department of General Linguistics, University of Helsinki.*

Lars Borin. 1984. Ett textdatabassystem för lingvister (A text database system for linguists)[In Swedish]. I: Anna Sågvall Hein (utg.). Föredrag vid De Nordiska datalingvistikdagarna 1983. Uppsala den 3-4 oktober. In *Proceedings of the 4th Nordic Conference of Computational Linguistics (NODALIDA 1983) Rapport UCDL-R-841. Uppsala universitet. Centrum för datorlingvistik.*, pages 37–47.

Lars Borin. 1985. Tvånivåmorfologi. Introduction och användarhandledning. Technical report, Rapport UCDL-L-3. Uppsala universitet. Centrum för datorlingvistik.

Lars Borin. 1991. *The automatic induction of morphological regularities*. Department of Linguistics, Uppsala University. Reports from Uppsala University, Linguistics (RUUL), 22.

Mats Carlsson. 1982. *Uppsala Chart Parser, 2. System Documentation.* UCDL R-81-1. Uppsala University. Center for Computational Linguisics.

Martin Kay. 1974. *Morphological and Syntactic Analysis.* Lecture notes from the 3rd International Summer School of Computational and Mathematical Linguistics. Pisa 1974.

Martin Kay. 1977. Reversible grammar. *Handbook from the 1977 Nordic Summer School in Computational Linguistics. Palo Alto.*

[6] Regeringens proposition 1986/87:80

Geir Kjetsaa, Sven Gustavsson, Bengt Beckman, & Steinar Gil. 1984. The Authorship of the Quiet Don. *Slavica Norvegica, vol.1. Oslo and Atlantic Highlands, N.J.*

Kimmo Koskenniemi. 1983. *Two-level Morphology: A General Computational Model för Word-form Recognition and Production*. Publications of the Department of General Linguistic, University of Helsinki, Finland.

NALP. 1978. Machine-readable text and dictionary material at UDAC. Technical report, Report No 4 1978. Uppsala University Data Center. Uppsala. Natural Language Processing Group.

Valentina Rosén & Margareta Sjöberg. 1985. TEXTPACK programpaket för språkvetenskaplig textbearbetning. Technical report, Centrum för datorlingvistik. Uppsala universitet. UCDL-L-85-2.

Anna-Lena Sågvall. 1968. *Ett system för automatisk morfologisk analys av det ryska verbet, applicerat på en c:a 80 sidor lång rysk text. Licentiatavhandling*. Uppsala universitet. Slaviska institutionen.

Anna-Lena Sågvall. 1973. *A System for Automatic Inflectional Analysis. Implemented for Russian*. Data linguistica 8. Stockholm. Almkvist & Wiksell.

Anna-Lena Sågvall, Berith Brännström, & Agneta Berghem. 1974a. Presentation av vid UDAC utvecklade verktyg för behandling av naturligt språk. Technical report, UDAC. Uppsala universitetsdatacentral. Uppsala, Sverige.

Anna-Lena Sågvall, Beritha Brännström, & Agneta Berghem. 1974b. *Text och Ord 1*. Slaviska institutionen. Uppsala universitet, Sverige.

Anna-Lena Sågvall, Berith Brännström, & Agneta Berghem. 1975. Processing Natural Language at UDAC. Technical report, Report No 2. September 1975. Uppsala universitetsdatacentral. Uppsala, Sverige.

Anna-Lena Sågvall, Berith Brännström, & Agneta Berghem. 1976. MIR: A Computer Based Approach to the Acquisition of Russian Vocabulary in Context. *System*, 4(2):116–127.

Anna Sågvall Hein & Lars Ahrenberg. 1985. A Parser for Swedish. Status Report for SVE. UCP. Technical report, Rapport UCDL-R-85-2. Uppsala universitet. Centrum för datorlingvistik.

Anna-Lena Sågvall Hein. 1977. Chartanalys och morfologi. In Martin Gellerstam, editor, *Nordiska Datalingvistikdagar 1977. Föredrag från en konferens i Göteborg.*

Anna-Lena Sågvall Hein. 1979. Natural Language Processing Group (NALP) at Uppsala Univer-sity Data Center (UDAC). In *Nordic Linguistic Bulletin. Vol 3. No 1.*

Anna Sågvall Hein. 1980. An overview of the Uppsala Chart Parser version 1 (UCP-1). *Report no. UCDL-R-80-1 Center for Computational Linguistics. Uppsala University, Department of Linguistics.*

Anna Sågvall Hein. 1981. UCDL. Centrum för Datorlingvistik vid Uppsala universitet. En presentation. Technical report, Rapport UCDL-R-81-3. Centrum för datorlingvistik. Uppsala universitet, Sverige.

Anna Sågvall Hein. 1983. A Parser for Swedish. Status Report for SVE. UCP. Technical report, Rapport UCDL-R-83-2. Uppsala Universitet. Centrum för datorlingvistik.

Anna Sågvall Hein. 1987. Parsing by means of Uppsala Chart Processor (UCP). In *Natural Language Parsing Systems*, pages 203–266. Springer Verlag Berlin Heidelberg.

William A Woods. 1973. An experimental parsing system for Transition Network Grammars. pages 111–154. Algorithmics Press. New York.

From open parallel corpora to public translation tools: The success story of OPUS

Jörg Tiedemann
Department of Digital Humanities Language Technology
University of Helsinki, Finland
jorg.tiedemann@helsinki.fi

Abstract

This paper describes the success of OPUS, starting from a small side-project but leading to a full-fledged ecosystem for training and deploying open machine translation systems. We briefly present the current state of the framework focusing on the mission of increasing language coverage and translation quality in public translation models and tools that can easily be integrated in end-user applications and professional workflows. OPUS now provides the biggest hub of freely available parallel data and thousands of open translation models have been released supporting hundreds of languages in various combinations.

1 Introduction

The starting point of OPUS is clearly connected to Uppsala and the language technology research group at the former department of linguistics. Work with parallel corpora has been pushed by projects on machine translation (Tjong Kim Sang, 1999) and multilingual corpus-driven linguistics and lexicography (Borin, 1998; Borin, 2002). The significant value of aligned multilingual data sets had been recognized by the leading researchers in the group and various resources came out of their efforts together with applications in translation studies, bilingual lexicon induction and machine translation development (Borin, 2000a; Borin, 2000b; Sågvall Hein et al., 2002). Inspired by those projects, OPUS filled the gap of public data sets that can be freely shared and used in research and development. Initially starting with software localization data, OPUS slowly grew into a massive collection of parallel translation data covering hundreds of languages and thousands of language pairs coming from a wide variety of domains.

The mission of OPUS was clear from the beginning: Data sets in the collection shall be open and free and support reproducible science to push cross-lingual NLP research and machine translation in particular. The essential principle is to provide a consistent interface to data sets that are readily prepared for further work without losing information from the original source. Wide language coverage has been a goal from the start with a complete alignment across all languages included.

The collection now represents a crucial foundation for wide-coverage machine translation. Taking advantage of the huge resource, we launched OPUS-MT (Tiedemann & Thottingal, 2020), an initiative to systematically exploit the data set to train open neural machine translation (MT) models that can be shared and re-used as well. The project tackles the growing responsibility of language technology providing essential tools for fair information access without language barriers and avoiding commercial exploitation. Our focus is on transparency and the paper describes our efforts in building the infrastructure that enables the use of free and independent machine translation in end-user applications and professional workflows.

Below, we briefly provide the background on OPUS and present tools for finding and processing the data. We then introduce OPUS-MT and its components before discussing the integration of pre-trained translation models in development platforms, end-user applications and translation workflows. Finally,

Jörg Tiedemann. 2022. From open parallel corpora to public translation tools: The success story of OPUS. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 133–138. Available under CC BY 4.0

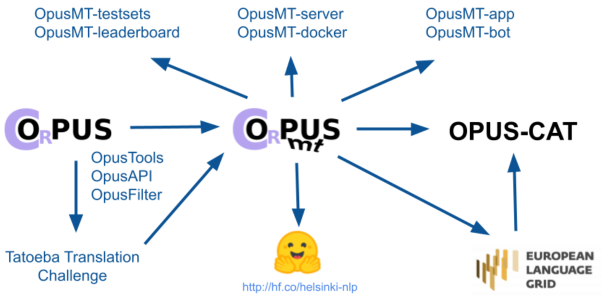


Figure 1: OPUS and OPUS-MT and its connections to other components, platforms and applications.

we also present the importance of benchmarking and monitoring the progress and briefly mention ongoing work on scaling up language coverage and optimizing translation models in terms of speed and applicability. Figure 1 illustrates the connections between various components.

2 OPUS – The Open Parallel Corpus

OPUS[1] has been a major hub for parallel corpora since 2004 (Tiedemann & Nygaard, 2004; Tiedemann, 2009; Tiedemann, 2012). The current release covers over 600 languages compiled into sentence-aligned bitexts for more than 40,000 language pairs. Over 20 billion sentences and sentence fragments correspond to 290 billion tokens and the data set contains about 12 TB of compressed files. Despite the typical Zipfian distribution, there are over 300 language pairs with more than one million sentence pairs, a good base for high quality machine translation.

OPUS tries to follow a consistent format with a simple standalone XML format for language content and standoff annotation in XCES Align to annotate links between translated sentences. The latter enables a space-efficient way of storing bilingually-aligned multilingual data sets without duplicating essential content. For convenience, other common data formats are generated from the native OPUS format including plain text versions with aligned sentences on corresponding lines and translation memory exchange files (TMX) that are common in professional translation platforms. Additionally, OPUS also releases token frequency counts, word alignment files and rough bilingual dictionaries extracted from automatically aligned bitexts.

Recently, we also released a compilation of the data under the label of the *Tatoeba Translation Challenge*[2], TTC for short (Tiedemann, 2020). The purpose of this release is to provide a streamlined collection for MT training pipelines. The latest release of the TTC includes 29 billion translation units in 3,708 bitexts covering 557 languages altogether. We made an effort to unify different sources, to improve the consistency in language labeling and to remove noise and duplicates. Dedicated development and test sets are also provided to make the application of TTC as straightforward as possible in standard machine learning setups.

2.1 Finding and processing OPUS data with the OPUS-API and OpusTools

An important ingredient for OPUS is automation. Making resources available requires efficient ways of finding and accessing them. The OPUS-API[3] provides an online API for searching resources and enables

[1] `https://opus.nlpl.eu`
[2] `https://github.com/Helsinki-NLP/Tatoeba-Challenge`
[3] `https://opus.nlpl.eu/opusapi/`

queries for specific languages and corpora. It provides the essential information about released data sets and returns download links to fetch data from the external data storage. The API responds in simple JSON format, which can easily be used programmatically when looking for resources.

We make use of the OPUS-API ourselves with the implementation of the OpusTools package[4] (Aulamo et al., 2020a). This software library provides a Python interface with methods for locating, downloading and converting OPUS data sets. Command-line scripts such as *opus_read* provide convenient functions to query the database and to fetch data from the original storage. Furthermore, the tools read from compressed release-packages and can be used to convert data sets into various formats such as TMX and plain text on the fly. Sentence alignments can also be filtered based on alignment confidence score, alignment type or language flag. For the latter, the package includes tools for automatic language identification.

2.2 Cleaning parallel data with OpusFilter

OpusFilter[5] (Aulamo et al., 2020b) integrates the functionality provided by OpusTools and the OPUS-API but adds a modular system for filtering and preparing parallel data sets. It provides a wide variety of modules for data preparation and noise reduction. A YAML configuration file defines the pipeline to transform raw corpus files to clean training and test set files. The same pipeline can be generalized over multiple language pairs. The toolbox can easily be extended and currently supports different kinds of segment-level processing steps such as tokenization and subword splitting as well as filters based on automatic language identification, word alignment scores, language models and sentence embeddings. Furthermore, scores can be analyzed and visualized, and custom classifiers can be trained to make domain-specific filter decisions (Vázquez et al., 2019).

3 Machine translation with OPUS-MT

The natural next step after collecting and compiling parallel data is to systematically exploit them in learning machine translation models. OPUS-MT[6] aims to provide training pipelines and solutions for deploying MT models derived from OPUS data. The goal is to develop a major hub for open state-of-the-art models with a large language coverage and straightforward use in end-user applications and further research and development. The framework is based on Marian, an efficient implementation of neural machine translation (NMT) in pure C++ and with minimal dependencies (Junczys-Dowmunt et al., 2018).

OPUS-MT training pipelines come in the form of makefile recipes that enable massive and systematic experiments on high-performance computing facilities. Automation provided by the recipes cover all necessary sub-tasks for preparing data sets, training models, testing their performance and finally releasing pre-trained NMT models. Special care has been taken to allow the creation of multilingual models that support more than one language as input or output. The recipes transparently handle different language combinations and combine data sets as necessary adding language flags if required (Johnson et al., 2017). Subword segmentation using SentencePiece (Kudo & Richardson, 2018) is fully integrated, and automatic word alignment (Östling & Tiedemann, 2016) can be used to train transformer models with guided alignment features. Batch jobs can easily be created to run on SLURM-based task management systems.

OPUS-MT further provides pipelines for data augmentation using back-translation (Sennrich et al., 2016) or pivot-based triangulation. Fine tuning is also supported in order to adapt to specific domains, user-specific data sets or selected language pairs in mulitlingual models.

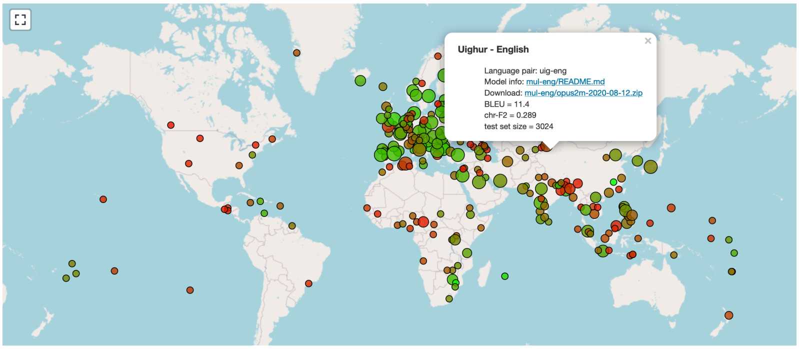
3.1 Integrating OPUS-MT

Important for the success of pre-trained models is the ease of use and deployment. OPUS-MT strives to make the models accessible and useful for a wide range of users. Substantial efforts have been made to provide simple deployment procedures and integration routines for all our models.

[4] https://github.com/Helsinki-NLP/OpusTools

[5] https://github.com/Helsinki-NLP/OpusFilter

[6] https://github.com/Helsinki-NLP/OPUS-MT



Uighur - English

Language pair: uig-eng
Model info: mul-eng/README.md
Download: mul-eng/opus2m-2020-08-12.zip
BLEU = 11.4
chr-F2 = 0.289
test set size = 3024

Figure 2: Language coverage of translation models visualized on an interactive map. Geolocations of languages are taken from Glottolog and dot colors indicate the translation quality in terms of an automatic evaluation metric measured on the Tatoeba test set in this case on a scale from green (best) to red (worst). Smaller circles refer to smaller, less reliable test sets.

First of all, we provide methods to create translation servers using web applications that provide service APIs through web sockets and requests. Servers can easily be configured using JSON files and the API also uses JSON for communication. Multiple translation servers can be combined and accessed via the same interface and caching is implemented to decrease the workload of the server. All pre-trained models we release can be integrated in the server solutions we offer.

Another important integration is the conversion of the native Marian NMT models to PyTorch, which opens up their use in a wide range of applications through the popular transformers library provided by Huggingface.[7] Conversion scripts are available to prepare OPUS-MT models for the public model hub making them available to the NLP community and also accessible through the online inference API.

Similarly, we also integrate OPUS-MT models in the European Language Grid (ELG). Dockerized OPUS-MT servers run on the ELG platform making it possible to directly access translation models from the ELG cloud services and the APIs provided by the infrastructure. The same docker images can also be downloaded from DockerHub and may run locally or on other cloud infrastructures.

Addressing the needs of professional translators is done by OPUS-CAT,[8] a collection of tools and plugins that add the power of OPUS-MT to computer-assisted translation (CAT) workflows in Trados Studio, memoQ, and OmegaT. In some CAT tools, such as Wordfast, OPUS-CAT can be used by connecting directly to its API through a custom MT provider functionality. OPUS-CAT also includes a Chrome browser extension, which makes it possible to use OPUS-MT in browser-based CAT tools. The Chrome extension currently supports Memsource[9] and XTM.[10] Different to other solutions, OPUS-CAT runs MT locally and does not require to send data to any external service. This has huge advantages in terms of data privacy and security and also enables fine-tuning of local translation engines on custom data without compromizing data safety.

3.2 Benchmarks and evaluation

Monitoring language coverage and translation quality is important to keep track of the progress in our mission to improve language support and cross-lingual information accessibility using open MT solutions. Therefore, we systematically run benchmarks on all our models using a wide range of test sets coming from established evaluation campaigns. Automatic evaluation is certainly not sufficient but still

[7] `https://huggingface.co/transformers/`
[8] `https://helsinki-nlp.github.io/OPUS-CAT/`
[9] `https://www.memsource.com/`
[10] `https://xtm.cloud/`

provides a good indication of quality especially if several benchmarks are used in parallel instead of relying on single test sets.

We aim at a comprehensive collection of benchmarks[11] and results are stored in a public repository,[12] which can be explored in a public leaderboard.[13] Translation results are also kept in the same repository to make it possible to run further qualitative studies on actual output of each model. Finally, we also create dynamic maps that visualize language coverage according to geographic locations of languages supported by OPUS-MT (see Figure 2).

4 Conclusions

Above, we have shown how OPUS developed from a small data collection initiative to a mature ecosystem for research on large coverage machine translation. All the components connected to OPUS provide a complete framework for systematic experiments and state-of-the-art neural MT development. The main building blocks refer to data collection, data curation, model training, system evaluation as well as deployment and MT integration tasks. The data collection itself is extensive but the coverage of released MT models is also impressive already. A lot of further work is on-going including the implementation of modular multilingual machine translation and the development of speed-optimized compact translation models using various kinds of knowledge distillation and quantization. Further integration into end-user applications on various devices are planned as well and translation quality and language coverage are constantly improved.

References

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, & Jörg Tiedemann. 2020a. OpusTools and Parallel Corpus Diagnostics. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3782–3789, Marseille. European Language Resources Association.

Mikko Aulamo, Sami Virpioja, & Jörg Tiedemann. 2020b. OpusFilter: A Configurable Parallel Corpus Filtering Toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156. Association for Computational Linguistics.

Lars Borin. 1998. ETAP: Etablering och annotering av parallellkorpus för igenkänning av översättningsekvivalenter (ETAP: Creating and annotating a parallel corpus for the recognition of translation equivalents). *ASLA Information*, 24(1):33–40.

Lars Borin. 2000a. ETAP project status report December 2000. Technical report, Uppsala University, Department of Linguistics.

Lars Borin. 2000b. You'll Take the High Road and I'll Take the Low Road: Using a Third Language to Improve Bilingual Word Alignment. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Lars Borin. 2002. Alignment and tagging. In *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999*, Language and computers: studies in practical linguistics, pages 207–218. Amsterdam: Rodopi.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, & Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, & Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne. Association for Computational Linguistics.

[11] `https://github.com/Helsinki-NLP/OPUS-MT-testsets`

[12] `https://github.com/Helsinki-NLP/OPUS-MT-leaderboard/`

[13] `https://opus.nlpl.eu/leaderboard/`

Taku Kudo & John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels. Association for Computational Linguistics.

Robert Östling & Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Anna Sågvall Hein, Eva Forsbom, Jörg Tiedemann, Per Weijnitz, Ingrid Almqvist, Leif-Jöran Olsson, & Sten Thaning. 2002. Scaling Up an MT Prototype for Industrial Use - Databases and Data Flow. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation, (LREC'2002)*, volume V, pages 1759–1766, Las Palmas de Gran Canaria.

Rico Sennrich, Barry Haddow, & Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin. Association for Computational Linguistics.

Jörg Tiedemann & Lars Nygaard. 2004. The OPUS Corpus - Parallel and Free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon. European Language Resources Association (ELRA).

Jörg Tiedemann & Santhosh Thottingal. 2020. OPUS-MT - Building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*. European Association for Machine Translation.

Jörg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Erik Tjong Kim Sang. 1999. Aligning the Scania Corpus. *Working Papers in Computational Linguistics & Language Engineering*, 18.

Raúl Vázquez, Umut Sulubacak, & Jörg Tiedemann. 2019. The University of Helsinki Submission to the WMT19 Parallel Corpus Filtering Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 294–300, Florence. Association for Computational Linguistics.

Binomials in Swedish corpora – ‘Ordpar 1965’ revisited

Martin Volk and Johannes Graën
Department of Computational Linguistics
University of Zurich
`volk|graen@cl.uzh.ch`

Abstract

This paper describes a corpus study on Swedish binomials, a special type of multi-word expressions. Binomials are of the type *X conjunction Y* where *X* and *Y* are words, typically of the same part-of-speech. Bendz (1965) investigated the various use cases and functions of such binomials and included a list of more than 1000 candidates in his appendix. We were curious to what extent these binomials can still be found in modern corpora. We therefore checked this list against the Swedish Europarl and OpenSubtitles corpora. We found that many of the binomials are still in use today even in these diverse text genres. The relative frequency of binomials in Europarl is much higher than in OpenSubtitles.

Foreword

It is a great honor to contribute to this Festschrift for Lars Borin. Lars has been our source of inspiration for many years. He worked on parallel corpora (Borin, 2002) and word alignment (Borin, 2000) before we thought about it. He published on named entity recognition for the digital humanities (Borin et al., 2007) and (Borin et al., 2014), on the architecture and processing pipeline of Swedish Språkbanken (Borin et al., 2012) and (Borin et al., 2016), and many other topics.

Over the years we profited enormously from meetings and discussions with Lars. In addition to being a knowledgeable discussion partner, we would like to thank him for being a great colleague and friend. We dedicate our little corpus study to him.

1 Introduction

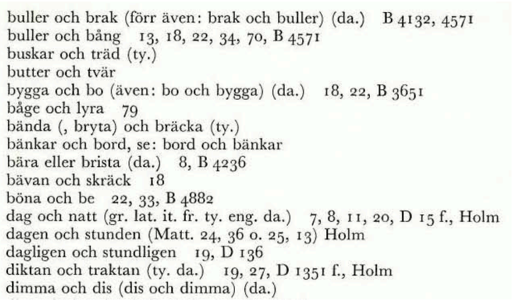
Binomials are an interesting case of multi-word expressions. Binomials are patterns of the type *X conjunction Y* where X and Y are words, typically of the same part-of-speech, connected by a conjunction (most often by *and* or its equivalent in the respective languages).

Some time ago, we investigated adverbial binomials (Volk et al., 2016; Volk & Graën, 2017; Graën & Volk, 2021) in a number of languages. Our study was initially motivated by the observation that some such binomials are homonyms with coordinated prepositions (e.g. DE: *ab und zu*, EN: *on and on*, SV: *till och med*) and require special treatment in automated language processing. We searched our corpora for special PoS patterns (e.g. with adjectives, adverbs, particles and prepositions) and measured the idiomaticity of candidates with the reversibility score, cf. Mollin (2014)[1], with a mutual information score and with an entropy measure in order to identify larger fixed expressions containing the binomial (e.g. FR: *d’ores et déjà*, DE: *mehr als je zuvor*, EN: *both internally and externally*, SV: *om och om igen*).

For Swedish we found that *till och med* is by far the most frequent adverbial binomial, followed by *först och främst* and *helt och hållet*. Modern Machine Translation systems are surprisingly good at translating these fixed expressions, but parsers often have difficulties integrating them correctly into the syntactic structure.

[1] If *X conj Y* is much more frequent than *Y conj X*, then this is evidence for a fixed (i.e. idiomatic) expression.

Martin Volk and Johannes Graën. 2022. Binomials in Swedish corpora – ‘Ordpar 1965’ revisited. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 139–144. Available under CC BY 4.0



buller och brak (förr även: brak och buller) (da.) B 4132, 4571
buller och bång 13, 18, 22, 34, 70, B 4571
buskar och träd (ty.)
butter och tvär
bygga och bo (även: bo och bygga) (da.) 18, 22, B 3651
båge och lyra 79
bända (, bryta) och bräcka (ty.)
bänkar och bord, se: bord och bänkar
bära eller brista (da.) 8, B 4236
bävan och skräck 18
böna och be 22, 33, B 4882
dag och natt (gr. lat. it. fr. ty. eng. da.) 7, 8, 11, 20, D 15 f., Holm
dagen och stunden (Matt. 24, 36 o. 25, 13) Holm
dagligen och stundligen 19, D 136
diktan och traktan (ty. da.) 19, 27, D 1351 f., Holm
dimma och dis (dis och dimma) (da.)

Figure 1: Excerpt from the Ordpar appendix in (Bendz, 1965), page 93. The examples display ordering variants, language information (da., ty., gr., lat., etc.), and optional parts (e.g. *bryta*) as well as references to the Bible and other publications.

A recent publication in Språktidningen (Landberg, 2022) reminded us that the topic of word pairings is still relevant for linguistics and also of interest for the social sciences. The article highlighted the use of coordinated words in political discourse. We therefore returned to an open edge of our previous work.

During our past investigation we had encountered (Bendz, 1965), an early work on Swedish binomials which includes an appendix with more than 1000 candidates. For the current paper, we turned this list into a computer-digestible format (and termed it "the Ordpar list" after the book's name). And we checked the binomials of this list in two corpora of Swedish. We are interested to find out which of these binomials are used in modern Swedish.

1.1 The Ordpar list

Bendz (1965) introduces his list as: "Nedan följer en förteckning över svenska ordpar (samt några längre sammanställningar) av högst olika ålder, typ, stilvalör och frekvens. … Listan omfattar endast ordpar vilkas komponenter är samordnade med 'och', 'eller', 'men' eller asyndetiskt … Den är givetvis inte på något sätt fullständig, men, särskilt vad beträffar synonymier, tillräckligt rikhaltig för att vara representativ." (Bendz, 1965, page 90).

After the first step of scanning, digitizing and structuring the list, we counted 1183 raw entries for Swedish[2]. The vast majority (987) are *X och Y* cases, where X and Y consist of only one token. See figure 1 for examples. Only 13 entries are of the type *X eller Y* (here: *bära eller brista*), and 5 entries are of *X men Y*.

In addition, there are entries with alternatives of the three conjunctions (12 entries with *och+eller* (e.g. *tid och (eller) tillfälle*), 4 entries with *eller+och* (e.g. *mer eller (och) mindre*), and 2 entries for *men+och* (e.g. *sakta men (och) säkert*)). We assume that the order of the conjunctions indicates a preference.

This leaves 160 entries in the Ordpar list with other characteristics. Many of them contain another token either as a modifier (e.g. *fläsk och bruna bönor*), in a larger expression (e.g. *mellan barken och trädet*), or in a listing (e.g. *guld, rökelse och myrra*). In a number of cases, the additional token is in parentheses and constitutes a variant (e.g. *blommor (blomst) och blad(er)*). We decided to unfold these cases into new entries. This means, we interpret this entry as standing for *blommor och blad, blomst och blad, blommor och blader, blomst och blader*. In this way we processed 139 variant entries and unfolded them to 300 generated candidates.

It should be noted that the original Ordpar list contains 88 entries in both orders (e.g. it contains both *kall och klar* and *klar och kall*) which we counted as two entries so far. For some of these entries Bendz

[2] Bendz (1965) also features shorter lists of binomials for German, English, French, Italian, Spanish, Latin, and Greek, which we ignore here. We make the digitized Swedish list available at `https://pub.cl.uzh.ch/purl/ordpar_list`

X	C	Y	f(X,C)	f(C,Y)	f(X,C,Y)	local-MI	f(Y,C,X)	OS rank
till	och	med	8 037	16 756	6 621	0,002 031	7	1
först	och	främst	3 916	3 938	3 857	0,001 442		13
helt	och	hållet	3 188	2 532	2 532	0,000 970		5
i	och	med	4 089	16 756	3 008	0,000 909	13	31
var	och	en	2 078	13 899	1 768	0,000 558		7
saker	och	ting	1 334	1 219	1 217	0,000 509		2
klart	och	tydligt	1 357	1 476	1 121	0,000 456	116	17
mer	eller	mindre	812	910	799	0,000 346	1	8
grund	och	botten	843	625	624	0,000 273		27
helt	och	fullt	3 188	694	590	0,000 222	2	53
kvinnor	och	barn	2 818	813	589	0,000 220	49	11
sätt	och	vis	1 764	396	386	0,000 156		6
förr	eller	senare	230	257	228	0,000 110		3
är	och	förblir	955	280	236	0,000 100		90
en	och	samma	554	602	229	0,000 095		25
rättigheter	och	skyldigheter	5 293	335	258	0,000 090	20	426
tack	och	lov	247	167	167	0,000 081		4
fullt	och	fast	177	194	140	0,000 068	3	107
rätt	och	riktigt	801	216	155	0,000 066	2	225
lugn	och	ro	207	134	132	0,000 065		9

Table 1: List of Ordpar binomials in **Europarl** ranked by the local mutual information score *local-MI*(X, C, Y). Please note that $f(X, C, Y)$ is the frequency of the triple *X conjunction Y* while $f(Y, C, X)$ displays the frequency of the reverse order, if at all present in the corpus. The rightmost column displays the rank in the OpenSubtitles corpus.

identifies the preferred order (e.g. *"reda och ordning – oftast: ordning och reda"*). In a few cases, the preference information comes with a temporal judgment (e.g. *"jämt och samt – förr även: samt och jämt"*). We mark all these double-order entries since we will check both orders for all binomials always in order to compute the reversibility score.

To many entries, Bendz (1965) added information about corresponding occurrences of the binomial in other languages: Danish (200 entries), German (132), Latin (27), English (19), and single digit numbers for Dutch, French, Italian, Spanish, Hebrew and Greek. And many binomials in the list also have references to the Bible (e.g. *anda och sanning; vägen, sanningen och livet*), to Svenska Akademiens Ordbok and to other previous work. Interestingly Bendz' Ordpar list does not contain any reduplication binomials (as e.g. *mer och mer, igen och igen*).

Our goal is to check the occurrence frequencies of all Ordpar entries in the Swedish parts of the Europarl (Koehn, 2005) and the OpenSubtitles corpus (Lison & Tiedemann, 2016), and see by comparison whether binomials show a stable distribution pattern. The input to our investigations in the following sections are 1101 binomial candidates of the type *X conjunction Y* explicitly or implicitly in the Ordpar list.

2 Binomials in the Swedish Europarl corpus

We cleaned the Europarl Corpus for corpus linguistics studies (Graën et al., 2014). And we processed both the Europarl and the OpenSubtitles corpus with Stanza (Qi et al., 2020), which leaves us with 35 and 240 million token, respectively. After cleaning and unfolding, the Ordpar list contains 1055 binomials of type *X och Y* (some of which are form variants as described above). We searched all 1055 binomials in both orders (i.e. as *X och Y* and as *Y och X*) in both corpora.

In order to measure the idiomaticity we computed the local mutual information scores (local-MI) for

X	C	Y	f(X,C)	f(C,Y)	f(X,C,Y)	local-MI	f(Y,C,X)	EP rank
till	och	med	19 649	27 858	16 017	0,000 851		1
saker	och	ting	5 816	4 214	4 187	0,000 266		6
förr	eller	senare	2 916	3 056	2 816	0,000 189		13
tack	och	lov	3 498	3 012	2 789	0,000 185		17
helt	och	hållet	2 480	2 135	2 119	0,000 145		3
sätt	och	vis	2 671	2 219	2 130	0,000 145	2	12
var	och	en	4 493	49 610	2 650	0,000 126	13	5
mer	eller	mindre	1 512	1 602	1 379	0,000 097	2	8
lugn	och	ro	2 357	1 423	1 328	0,000 091	1	20
in	och	ut	11 382	2 645	1 590	0,000 090	573	111
kvinnor	och	barn	2 304	3 413	1 291	0,000 081	23	11
fram	och	tillbaka	3 535	2 008	1 178	0,000 074	4	31
först	och	främst	2 150	1 081	1 056	0,000 073		2
liv	och	död	4 664	1 697	987	0,000 061	8	51
vänta	och	se	1 211	14 116	897	0,000 050		39
kött	och	blod	1 283	1 165	701	0,000 048	10	122
klart	och	tydligt	907	834	632	0,000 046	27	7
in	eller	ut	899	662	593	0,000 044	181	199
dag	och	natt	3 009	652	621	0,000 041	212	68
gott	och	ont	1 112	630	542	0,000 039	26	57

Table 2: List of Ordpar binomials in **OpenSubtitles** ranked by the local mutual information score *local-MI*(X, C, Y). Here also $f(X, C, Y)$ is the frequency of the triple *X conjunction Y*, and $f(Y, C, X)$ is the frequency of the reverse order, if at all present in the corpus. The rightmost column displays the rank in the Europarl corpus as an indication of the varying prominence of the binomial in different text genres.

all candidates “X C Y” where C is the conjunction.[3] For this we used the bigram frequencies of “X C” and “C Y” in comparison with the frequency of the triple “X C Y”. Our formula is

$$\textit{local-MI}(X, C, Y) = \frac{f(X, C, Y)}{N} \times \log_2 \frac{N \times f(X, C, Y)}{f(X, C) \times f(C, Y)}$$

with N being the number of tokens in the corpus. In this way, the local-MI score predicts the probability of “X C” being followed by Y, and the likelihood of “C Y” being preceded by X. A shorter way of writing this with “Observed” versus “Expected” variables (the respective frequencies divided by N) is

$$\textit{local-MI}(X, C, Y) = O \times \log_2 \frac{O}{E}$$

We first search with lower-cased tokens from the corpus which results in 29 345 binomial occurrences of 431 different types in Europarl and 69 849 occurrences of 834 types in OpenSubtitles.

Table 1 shows the Ordpar binomials with the highest local mutual information scores in the Europarl corpus. Out of the 431 different binomials 86 occur in both orders, 49 with a combined frequency of 10 or more. We observe that the top candidates have clear ordering preferences with very few occurrences in the reverse order. On the opposite end with more balanced variants (and thus allegedly more composite meaning) are *nu och då* vs. *då och nu* (9 vs. 7), *sedlar och mynt* (78 vs. 48), and *höger och vänster* (38 vs. 24). Since our Europarl corpus is lemmatized, we may alternatively search the Ordpar binomials via the lemmas in the corpus. In theory this will conflate e.g. *hel och full* (173 hits) and *helt och fullt* (425

[3] See Evert (2008, page 1226) for a motivation for this measure.

PoS	Europarl		OpenSubtitles	
	frequency	ratio	frequency	ratio
ADJ	92 979	21,2 %	96 238	11,6 %
ADP	14 180	3,2 %	30 465	3,7 %
ADV	17 414	4,0 %	74 891	9,0 %
NOUN	277 744	63,3 %	334 606	40,2 %
PRON	7 479	1,7 %	142 737	17,2 %
VERB	29 242	6,7 %	153 321	18,4 %
Total	439 038		832 258	

Table 3: List of absolute and relative frequencies of different PoS in binomial configurations.

hits) into the same binomial and combine their frequency counts. But since the Ordpar binomials are not lemmatized systematically, this does not work. Therefore we search only over the word forms in the corpus.

The PoS tags for the found binomials are not perfectly reliable since often the binomials constitute a special syntactic environment. But the top frequencies of the PoS pairs for the Ordpar binomials might still give us an indication of the most typical ones. The most frequent PoS pairs are adverbs (14 456 hits), nouns (5156), adjectives (2698), pronouns (1860), and verbs (449).

For the other conjunctions (*eller, men*) we run the same counting experiments. The unfolded Ordpar list has 32 binomials with *eller* which we investigate in either order (via 59 trigger words involved). We find a total of 18 binomials in Europarl (when searching via tokens, i.e. word forms), but only four of them add up to more than 10 hits: *mer eller mindre* (799 hits), *förr eller senare* (228 hits), *liv eller död* (21 hits), and *nu eller aldrig* (12 hits). The unfolded Ordpar list contains 8 different binomials with the Swedish conjunction *men*, 3 of which occur in Europarl: *sakta men säkert* (101 hits), *långsamt men säkert* (13 hits) and *hårt men rättvist* (1 hit).

Overall this means that Ordpar binomials occur a total of 29 345 times (28 111 for *och*, 1095 for *eller*, 115 for *men* and 24 for other conjunctions, namely *om*, *som* and *över*) in Europarl (when counted via word forms) which corresponds to a relative frequency of 838.4 per 1 million tokens in the Europarl corpus.

3 Binomials in Swedish OpenSubtitles

For comparison we investigate the Swedish OpenSubtitles corpus (240 million tokens) with the same methods as above.

Table 2 shows the top ranked binomials from the Ordpar list. Many binomials differ in rank because of the different text genre. For example *liv och död* which is on rank 14 in OpenSubtitles, is only on rank 51 in Europarl. It is also noteworthy that *in och ut* has a high local-MI score even though it also has a frequent reversibility option (1590 vs. 573).

Based on the Ordpar list we find 69 849 binomials in OpenSubtitles. This results in a relative frequency of 290.6 Ordpar binomials per 1 million tokens in OpenSubtitles.

We also calculated the frequencies of binomial candidates in both corpora, searching for the pattern *X conjunction Y* with X and Y being of the same PoS (shown in Table 3). This gives us an indication of the typical types of binomials in each genre. In comparison, we clearly see that binomials are much more popular in Europarl and that adjective and noun binomials are used considerably more often in the Europarl debates (with OpenSubtitles being 8 times in size).

4 Conclusion

The Ordpar list as compiled by (Bendz, 1965) more than 50 years ago still serves as a source of inspiration for the investigation of binomials today. It is fascinating to observe that many of the idiomatic binomials in the Ordpar list are still in use in modern texts as diverse as parliamentary proceedings and movie subtitles.

As suspected we find more binomials in parliamentary proceedings (Europarl) than in subtitles. We suspect that this is due to the constraint that subtitles must be short and concise, whereas binomials are often repetitive (e.g. *först och främst* instead of *först*) and thus add to the length.

References

Gerhard Bendz. 1965. *Ordpar*. P. A. Nordstedt & Söners Förlag, Stockholm.

Lars Borin, Dimitrios Kokkinakis, & Leif-Jöran Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In *Proceedings of The ACL Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, Prague.

Lars Borin, Markus Forsberg, & Johan Roxendal. 2012. Korp — the corpus infrastructure of språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 474–478, Istanbul. European Language Resources Association (ELRA).

Lars Borin, Dana Dannélls, & Leif-Jöran Olsson. 2014. Geographic visualization of place names in Swedish literary texts. *Literary and Linguistic Computing*, 29(3).

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, & Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC)*, pages 17–18, Umeå.

Lars Borin. 2000. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 97–103, Saarbrücken.

Lars Borin, editor. 2002. *Parallel Corpora, Parallel Worlds. Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22-23 April, 1999*, volume 43 of *Language and Computers*. Rodopi, Amsterdam.

Stefan Evert. 2008. Corpora and collocations. In A. Lüdeling & M. Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 2, pages 1212–1248. Walter de Gruyter.

Johannes Graën & Martin Volk. 2021. Binomial adverbs in Germanic and Romance languages. a corpus-based study. In Julia Lavid-López, Carmen Maíz-Arévalo, & Juan Rafael Zamorano-Mansilla, editors, *Corpora in Translation and Contrastive Research in the Digital Age. Recent advances and exploration*, chapter 13, pages 326–342. John Benjamins.

Johannes Graën, Dolores Batinic, & Martin Volk. 2014. Cleaning the Europarl corpus for linguistic applications. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 222–227. Stiftung Universität Hildesheim.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, volume 5, pages 79–86. Asia-Pacific Association for Machine Translation.

Joel Landberg. 2022. Ett "och" betyder så mycket. *Språktidningen*, pages 24–30.

Pierre Lison & Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Sandra Mollin. 2014. *The (Ir)reversibility of English Binomials. Corpus, Constraints, Developments*, volume 64 of *Studies in Corpus Linguistics*. John Benjamins.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, & Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Martin Volk & Johannes Graën. 2017. Multi-word adverbs – how well are they handled in parsing and machine translation? In *Proceedings of The 3rd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT)*, London.

Martin Volk, Simon Clematide, Johannes Graën, & Phillip Ströbel. 2016. Bi-particle adverbs, PoS-tagging and the recognition of German separable prefix verbs. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, pages 297–305.

ICALL: Research versus reality check

Elena Volodina
Språkbanken Text
University of Gothenburg, Sweden
elena.volodina@svenska.gu.se

David Alfter
CENTAL
Université catholique de Louvain
david.alfter@uclouvain.be

Abstract

Intelligent Computer-Assisted Language Learning has been one of Lars Borin's research interests. The work on the *Lärka* language learning platform has started under his coordination. We see it our mission to make the platform live and prosperous, and through it to stimulate research into Swedish as a second language. Below, we name some weaknesses we have identified in *Lärka* while working with a course of beginner Swedish and outline our plans for tackling those.

1 Introduction

Research agenda is – by definition – driven by the research needs. However, the aim of the research is to explain the world around us as well as to explore and outline new ways of approaching certain phenomena (Hempel, 1967). It should, thus, aim to align with the real world needs, at least to a certain degree. This is especially true of research aimed at language learning.

Språkbanken Text[1] has for at least a decade been involved in research around Intelligent Computer-Assisted Language Learning (ICALL), with the major focus on Swedish as a second language (L2 Swedish), for example, Volodina et al. (2016), Pilán (2018), Alfter (2021). As part of these research activities, a platform for language learning, *Lärka*,[2] has been implemented (Volodina et al., 2014; Volodina & Pijetlovic, 2015; Alfter et al., 2019a). *Lärka* is used for staging experiments, testing prototypes, for demo purposes as well as for collection of research data and learner logs for further analysis.

Probably, the most significant impact of *Lärka* on research could be expected from research data collection, which could feed into new research insights relevant for the field of L2 Swedish and Second Language Acquisition (SLA). However, the major problem has up to now been to attract language learners to use *Lärka*. While *Lärka* exercise types demonstrate the capacity to use language technology for language learning purposes, the basic pedagogical aspects, unfortunately, are not considered, and that is a hypothetic reason why neither teachers nor learners see benefits of using the platform to the extent that can help generate new research data. The basic question is therefore – how can this be changed?

The recent effort by the authors to set up a self-study course in basic Swedish for Ukrainian refugees, *SwedishFromScratch*[3] (SFS), has become a well-needed reality check. *SFS*, the course which we make available through a dedicated channel on the social media platform Telegram[4] – while having the initial aim to help one of the authors' relatives fleeing from the war in Ukraine – has collected over 1800 followers in a couple of months. We receive messages from the users with thanks, reported problems and requests for desired features and materials (all of which witness active usage). That extent of acceptance has taken us by surprise and put the previous work on *Lärka* into a new perspective, which we are trying to analyze and present in this paper.

[1] https://spraakbanken.gu.se/

[2] https://spraakbanken.gu.se/larkalabb/

[3] https://spraakbanken.gu.se/en/research/themes/icall/swedish-from-scratch

[4] https://t.me/quizlet4swedish

Elena Volodina and David Alfter. 2022. ICALL: Research versus reality check. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 145–151. Available under CC BY 4.0

Below, we describe *Lärka* and the course *SwedishFromScratch* (Sections 2 and 3), present our insights from comparison of the two (Section 4), and conclude by our plans to move forward (Section 5).

2 Lärka in a nutshell

Lärka is a research platform that offers, among other things, automatically generated exercises based on authentic corpus material (Volodina et al., 2012; Pilán et al., 2014; Lindström Tiedemann et al., 2016), a text evaluation tool (Pilán et al., 2016), a lexicographic annotation tool (Alfter et al., 2019b), and the Swedish L2 profile (Volodina et al., 2021).

For exercise generation, there are highly customizable linguistic knowledge exercises for parts-of-speech, syntactic relations, and semantic roles (Volodina et al., 2014; Pilán & Volodina, 2014). For language learners, there are multiple-choice exercises (vocabulary and inflection), a listening exercise (Volodina & Pijetlovic, 2015), and two gamified exercises: Wordguess, a hangman-type game based on dictionary translations, and a particle verb exercise based on parallel multilingual corpora (Alfter & Graën, 2019).

3 SwedishFromScratch: an outline

SwedishFromScratch is a free course that can be followed by anyone using Telegram channel where notifications about new lessons are sent to subscribers. Up to now the course has been a combination of several types of lessons:

1. *Quizlet lessons*[5] primarily introduce Swedish vocabulary and phrases with their pronunciation and translation into Russian. We select vocabulary for training ourselves and create *Quizlet* flashcards manually. A set of exercises are automatically generated by *Quizlet* based on the flashcards, e.g. listening, spelling, translation, matching, test. Besides, we also introduce and train some grammar phenomena using *Quizlet* functionalities. The collectiion of lessons is cotinuously growing.

2. *Clilstore lessons*[6] present texts for practicing reading which we write ourselves (with a few exceptions) to avoid copyright issues. On the starting page, one can select language *sv* and order by *Title*. The numbered texts (currently at A1-B1 levels) belong to this course. Once inside a text, the *Clilstore* platform (Gimeno & Dónaill, 2008) allows to click on words, which opens a window for dictionaries to the right. The user can set a language for translation (Russian is this case) and a dictionary, e.g. Lexin (Hult et al., 2010) at the beginner levels. Currently, there are 24 *Clilstore* texts connected to the *SFS* course, and the collection is constantly growing.

3. *Grammar explanations* are reused from several sources, among others *SFI grammar*[7] that we translate into Russian,[8] and *online encyclopedia of Swedish in Russian* (Maslova-Lashanskaya, 1953).[9]

4. *Grammar exercises*[10] are based on the course texts we publish in *Clilstore*. The texts are uploaded to the generator, automatically analyzed for parts-of-speech, and gapped items are generated for a selected word class. To avoid errors, automatic annotation is manually checked, which is currently a bottleneck in the process. Only the first six texts out of the 24 available are prepared for the grammar exercise training module.

5. *Third-party open materials* are used to complement the course, e.g. *UR språkplay*[11] a set of films and series with subtitles for each phrase you hear, and a translation into the language you choose.

[5] https://quizlet.com/class/22036236/

[6] https://clilstore.eu/clilstore/

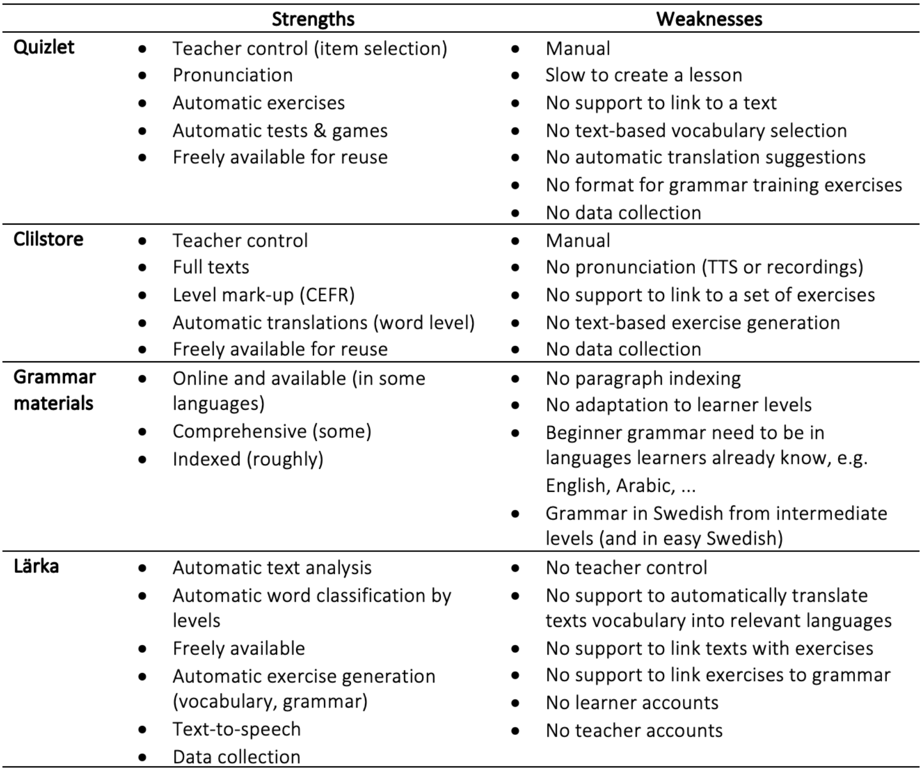
[7] https://sfipatxi.wordpress.com/grammatik/

[8] https://elenavolodina.github.io/SwedishFromScratch/Grammar

[9] https://svspb.net/bok/

[10] A prototype of an exercise: https://spraakbanken.gu.se/larkalabb/sfs/?lesson_number=3&pos=NOUN&tl=english

[11] https://www.ur.se/sprakplay/#/programs



	Strengths	Weaknesses
Quizlet	• Teacher control (item selection) • Pronunciation • Automatic exercises • Automatic tests & games • Freely available for reuse	• Manual • Slow to create a lesson • No support to link to a text • No text-based vocabulary selection • No automatic translation suggestions • No format for grammar training exercises • No data collection
Clilstore	• Teacher control • Full texts • Level mark-up (CEFR) • Automatic translations (word level) • Freely available for reuse	• Manual • No pronunciation (TTS or recordings) • No support to link to a set of exercises • No text-based exercise generation • No data collection
Grammar materials	• Online and available (in some languages) • Comprehensive (some) • Indexed (roughly)	• No paragraph indexing • No adaptation to learner levels • Beginner grammar need to be in languages learners already know, e.g. English, Arabic, ... • Grammar in Swedish from intermediate levels (and in easy Swedish)
Lärka	• Automatic text analysis • Automatic word classification by levels • Freely available • Automatic exercise generation (vocabulary, grammar) • Text-to-speech • Data collection	• No teacher control • No support to automatically translate texts vocabulary into relevant languages • No support to link texts with exercises • No support to link exercises to grammar • No learner accounts • No teacher accounts

Figure 1: Pros and cons of the used tools.

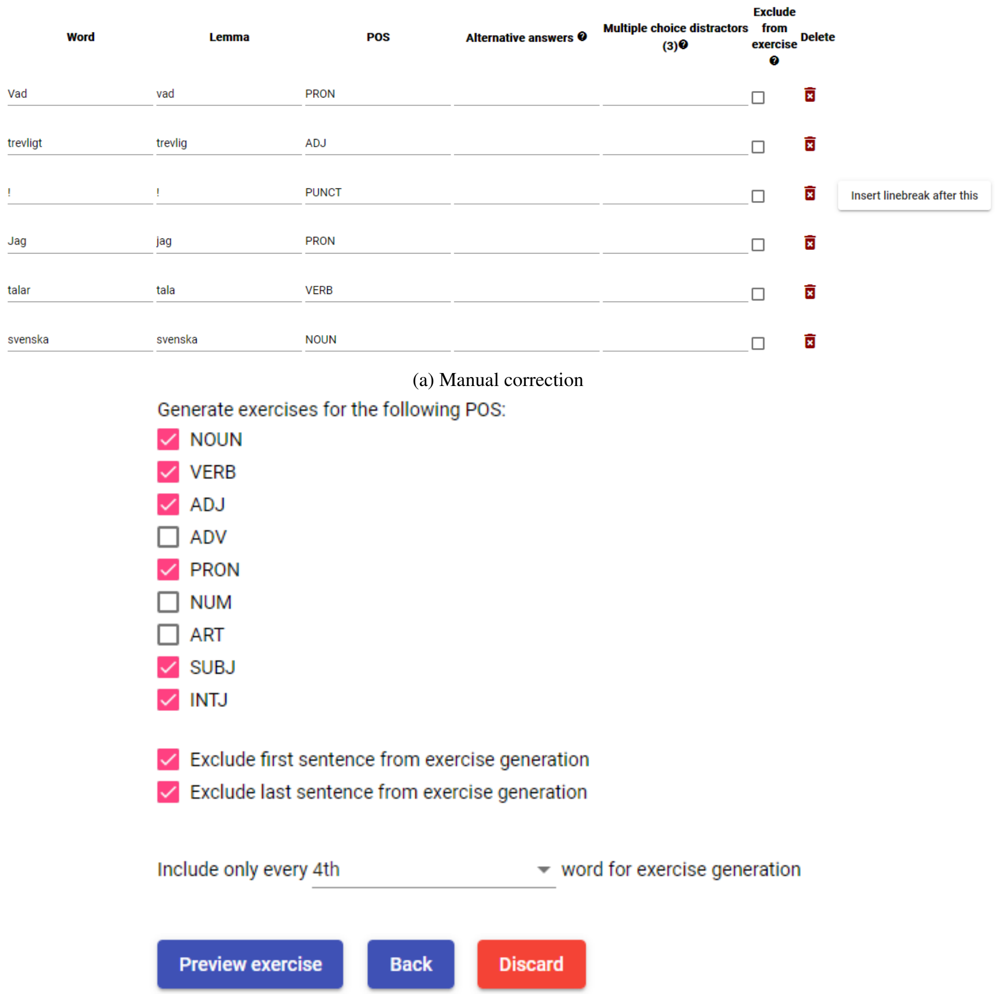
All lessons are sequenced in the order they are suggested to be learnt. Learners may follow a different path, if they wish, repeat lessons, skip lessons or take a break in the course. Since this is a self-study course, learners can also adapt the course to their time restrictions.

In the process of preparing lessons, we function as teachers, search for tools that can address the course needs, and continuously evaluate *Lärka* from that point of view.

4 Lessons learned

Previous research, e.g. Arhar Holdt et al. (2020), Burstein et al. (2009), suggests that teachers are likely to adopt (technical) innovations only if those can fit logically into teachers' daily routines. In this connection, Burstein et al. (2012) identify five critical components for technical innovations aimed at language classrooms, three of which we mention here:

- Relevance to the curriculum standards and lesson objectives, i.e. suggested technological solutions should be able to provide support with immediately relevant tasks.
- Ability to keep learners motivated and focused on the learning goals, i.e. the innovations should not distract from the learning goals, but help concentrate on them.
- Potential to independent practices, i.e. technical solutions should facilitate independent learner work, e.g. creation of activities that can be given to learners for self-study or homework.



Word	Lemma	POS	Alternative answers	Multiple choice distractors (3)	Exclude from exercise	Delete
Vad	vad	PRON			□	
trevligt	trevlig	ADJ			□	
!	!	PUNCT			□	
Jag	jag	PRON			□	
talar	tala	VERB			□	
svenska	svenska	NOUN			□	

(a) Manual correction

Generate exercises for the following POS:

- ☑ NOUN
- ☑ VERB
- ☑ ADJ
- □ ADV
- ☑ PRON
- □ NUM
- □ ART
- ☑ SUBJ
- ☑ INTJ

☑ Exclude first sentence from exercise generation

☑ Exclude last sentence from exercise generation

Include only every 4th word for exercise generation

(b) Exercise options

Figure 2: Authoring tool prototype.

All of the above criteria are met by the *SwedishFromScratch*, and only the last one can be claimed to be met by *Lärka*. Most critically, the exercises in *Lärka* use random words of a certain level of proficiency for exercise generation. This does not encourage learning, since the vocabulary scope is too wide. In the best case, random selection of vocabulary items facilitates testing or brings gaming experience. For learning, there is a need to limit the number of words for each session, so that learning can actually occur. Therefore, one important lesson from our *SFS* initiative is that we should offer teachers or learners control over what they use for a lesson, e.g. a possibility to enter the words for training themselves, limiting the scope to the relevant items.

Another issue concerns the context of *Lärka* exercises. Sentences, that are currently a unit used for exercise generation, are exempt of copyright issues. However, they cannot frame a focus for a whole lesson. Texts are a much better alternative in this respect. Even available texts on the Internet may still have copyright problems if used for other purposes than the ones intended by the authors. Thus, the option to upload their own texts or texts that they are sure can be used should at least be offered to the users (cf *Clilstore* concept). This way we could collect through *Lärka* free reading materials for future uses and to generate lesson materials based on those texts, for example covering vocabulary training, morphology and grammar exercises, as well as syntax-focused exercises.

Figure 1 provides a summary of further pros and cons over the various tools discussed here. A look at those suggests that each of these tools offers a unique type of language learning material generation and sharing, but is insufficient on itself for setting up any comprehensive language learning course.

5 Research and practice: agenda

Returning to the question we asked at the beginning of this article – how can we attract more users to *Lärka* and stimulate growth of learner-produced data for research – we can summarize the following:

In the best of worlds, *Lärka*'s NLP-based algorithms should be coupled with the functionalities offered by *Clilstore* (texts with translation) and *Quizlet* (flashcards and exercise generation) or similar to achieve maximal flexibility and usefulness for teachers and learners, and as a consequence to stimulate users to produce research data. Importantly, this would form a sound pedagogical basis for creating self-study courses, collecting all the currently distributed modules (in *SFS* or any other potential course) into one space, offering

– for teachers – an easier way to author a language lesson or a course;
– for learners – a convenient way to follow the course;
– for researchers – a safe and stable way to collect user-generated data (with access to learners).

Language Muse (Burstein & Sabatini, 2016; Burstein et al., 2017) has used this type of approach and it has shown to be working for both teachers, learners and researchers. Some planned practical step to achieve the above in *Lärka* is to provide an authoring tool that:

1. Incorporates manual correction of automatic processes such as POS annotation (Figure 2).
2. Generates gap cloze items and potentially other items out of texts.[12]
3. Follows standard conventions such as leaving the first and last sentence in a text intact, replacing only every x-th word, . . .
4. Gives more control to the course preparer by generating exercises semi-automatically based on choices made by the teacher.
5. Can be extended to allow for more diverse types of exercises.

The current prototype (Figure 2) offers some of the above-mentioned control to the course author.

To summarize, we have two major conclusions to draw from this experience. *First* is a general one: researchers need to have periodical **reality checks** to understand how to adjust their research agenda to reality in order to prevent their research from turning into a selfish playground. *Second* conclusion has relevance to the ICALL field: by giving (certain) control to teachers and learners, ICALL platforms (e.g. *Lärka*) may become a win-win platform for both pedagogical and research scenarios - an insight that we would not have gained without the **reality check** with the *SFS* course.

Acknowledgements

This work has been supported by Nationella språkbanken and HUMINFRA, both funded by the Swedish Research Council (2018-2024, contract 2017-00626; 2022-2024, contract 2021-00176) and their participating partner institutions.

References

David Alfter & Johannes Graën. 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 321–326.

[12] https://spraakbanken.gu.se/larkalabb/sfs/?lesson_number=4&pos=PRON&tl=english

David Alfter, Lars Borin, Ildikó Pilán, Therese Lindström Tiedemann, & Elena Volodina. 2019a. Lärka: from language learning platform to infrastructure for research on language learning. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, number 159, pages 1–14. Linköping University Electronic Press.

David Alfter, Therese Lindström Tiedemann, & Elena Volodina. 2019b. LEGATO: A flexible lexicographic annotation tool. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 382–388.

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective. Doctoral Thesis.* Univerisity of Gothenburg, Data Linguistica 31.

Špela Arhar Holdt, Rina Zviel-Girshin, Elżbieta Gajek, Isabel Durán-Muñoz, Petra Bago, Karën Fort, Ciler Hatipoglu, Ramunė Kasperavičienė, Svetla Koeva, Ivana Lazić Konjik, et al. 2020. Language teachers and crowdsourcing: Insights from a cross-European survey. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(1):1–28.

Jill Burstein & John Sabatini. 2016. The Language Muse Activity Palette. *Adaptive educational technologies for literacy instruction*, pages 275–280.

Jill Burstein, Jane Shore, John Sabatini, Brad Moulder, Steven Holtzman, & Ted Pedersen. 2012. The language musesm system: Linguistically focused instructional authoring. *ETS Research Report Series*, 2012(2):i–36.

Jill Burstein, Nitin Madnani, John Sabatini, Dan McCaffrey, Kietha Biggers, & Kelsey Dreier. 2017. Generating Language Activities in Real-Time for English Learners using Language Muse. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 213–215. ACM.

Jill Burstein. 2009. Opportunities for natural language processing research in education. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 6–27. Springer.

A Gimeno & Ó Dónaill. 2008. C. & Zygmantaite, R.(2013). Clilstore Guidebook for Teachers. *Tools for CLIL Teachers*.

Carl G Hempel. 1967. Philosophy of natural science. *British Journal for the Philosophy of Science*, 18(1).

Ann-Kristin Hult, Sven-Göran Malmgren, & Emma Sköldberg. 2010. Lexin-a report from a recycling lexicographic project in the North. In *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6-10 July 2010)*.

Therese Lindström Tiedemann, Elena Volodina, & Håkan Jansson. 2016. Lärka: ett verktyg för träning av språkterminologi och grammatik. *LexicoNordica*, 23:161–181.

S.S. Maslova-Lashanskaya. 1953. *Shvedsky yazyk: Chast pervaya. Otvetstvennyj redaktor prof. M.E. Steblin-Kamensky*. L.: Izd-vo LGU Zhdanova.

Ildikó Pilán & Elena Volodina. 2014. Reusing Swedish FrameNet for training semantic roles. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1359–1363.

Ildikó Pilán, Elena Volodina, & Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, pages 174–184.

Ildikó Pilán, David Alfter, & Elena Volodina. 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners writings. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 120–126.

Ildikó Pilán. 2018. *Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning. Doctoral Thesis.* Univerisity of Gothenburg, Data Linguistica 29.

Elena Volodina & Dijana Pijetlovic. 2015. Lark Trills for Language Drills: Text-to-speech technology for language learners. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 107–117.

Elena Volodina, Lars Borin, Hrafn Loftsson, Birna Arnbjörnsdóttir, & Guðmundur Örn Leifsson. 2012. Waste not, want not: Towards a system architecture for ICALL based on NLP component re-use. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*, pages 47–58.

Elena Volodina, Ildikó Pilán, Lars Borin, & Therese Tiedemann Lindström. 2014. A flexible language learning platform based on language resources and web services. In *Proceedings of LREC 2014, Reykjavik, Iceland*, pages 3973–3978.

Elena Volodina, Ildikó Pilán, & David Alfter. 2016. Classification of Swedish learner essays by CEFR levels. *CALL communities and culture–short papers from EUROCALL*, 2016:456–461.

Elena Volodina, Yousuf Ali Mohammed, & Therese Lindström Tiedemann. 2021. CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 178–189.

*Lyx*ig *språk*lig födelse*dag*spresent from the Swedish Word Family

Elena Volodina **Yousuf Ali Mohammed**
Språkbanken Text, University of Gothenburg, Sweden
name1.name2.surname@svenska.gu.se

Therese Lindström Tiedemann
University of Helsinki, Finland
therese.lindstromtiedemann@helsinki.fi

Abstract

Morphology and lexical resources are known to be two of Lars Borin's biggest research passions. We have, therefore, prepared a short description of a new kind of a lexical resource for Swedish, the *Swedish Word Family*. The resource is compiled based on learner corpora, and contains lexical items manually analyzed for derivational morphology.

1 Introduction

In the past couple of years, we have compiled a Swedish Word Family (SweWF) resource to study word formation mechanisms in the Swedish learner language (publication is in preparation). SweWF can boast unique features that differentiate it from the majority of other word (and morpheme) family resources, e.g. Körtvélyessy et al. (2020), Nikolaev et al. (2019), Hiebert et al. (2018), Zeller et al. (2013), Baayen et al. (1996), Bauer & Nation (1993):

1. we include compounds among the family members
2. we include multi-word expressions in the families
3. all lexical items in the SweWF resource contain other types of associated information, so that it is possible to use lemmas, lemgrams and sense-based units for comparisons with, for example, reference L1 corpora, depending on the type of units used in the reference corpora
4. the resource is descriptive in nature and as such it is possible to use it for both teaching and research.

The Swedish Word Family, based on the first version of CoDeRooMor (Volodina et al., 2021), contains 16 230 sense-based lemgrams organized into 4 400 word families (i.e. through shared roots). The size of each particular family varies between 1 and 281 members. Most numerous are the families where the root, if taken individually, is a preposition/particle, e.g *ut* ('out, towards'), including such family members as *ut*bildning ('education'), *ut*släpp ('exhaust, release'), söder*ut* ('southward').

The distribution of word families by their size in L2 Swedish corpora is shown in Figure 1 and Table 1. Families with few members (1-9) constitute the majority of all families (87%), with only 13% of families containing 10 or more members. Around 42% of word families (1868) in the Swedish word family resource are singletons containing only one family member (e.g. one word like *asfalt, astma, alzeimers*). Less than one percent of the families contain 61-281 members and these groups consist of roots that are nearly exclusively either Scandinavian or common Germanic words.

The Swedish word family resource shows that word families with fewer members more often tend to contain loanwords in contrast to the word families with larger number of members.

Elena Volodina, Yousuf Ali Mohammad and Therese Lindström Tiedemann. 2022. *Lyx*ig *språk*lig födelse*dag*spresent from the Swedish Word Family. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 153–160. Available under CC BY 4.0

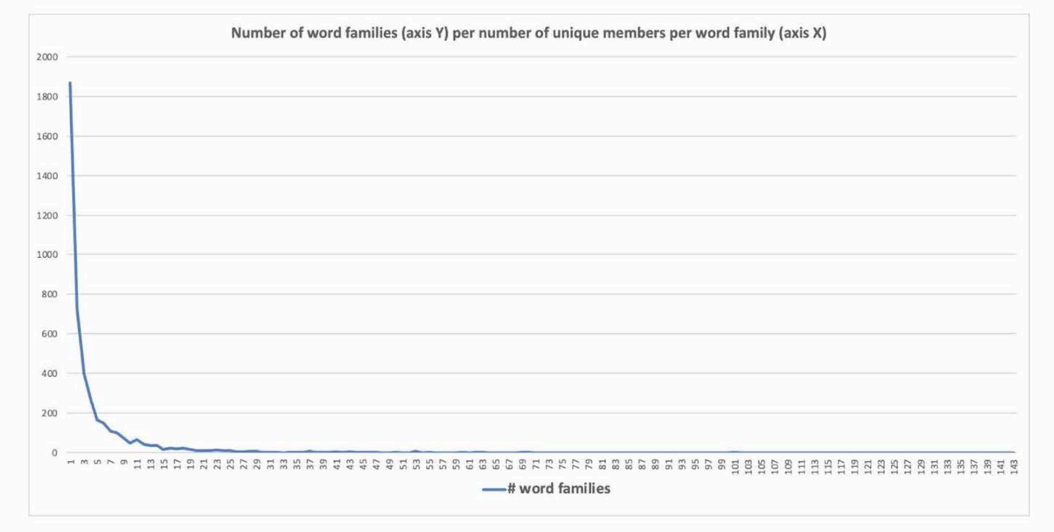


Figure 1: Distribution of word families by number of members.

Nr of word familes	Family size	Percent of all word families	Examples
1956	1	44.45	*asfalt, astma, alzheimers*
1886	2-8	42.86	*lyx: lyxa, lyxig, lyxliv, ...*
370	9-20	8.41	*fam: familj, familjfoto, ...*
159	21-60	3.61	*nord: nordisk, Nordamerika, ...*
15	61-100	0.34	*språk: språklig, språkljud, ...*
14	101-281	0.32	*dag: måndag, daglig, ...*

Table 1: Statistics over family sizes.

2 Hypotheses and case studies

The Swedish Word Family resource presents an opportunity to trace various trends in the language, including linguistic, cultural and cognitive ones. In this paper we are looking into the following hypotheses:

> (a) Simpler words (consisting of a minimal number of morphemes) within each family appear at earlier levels and are more frequent.
> (b) Related to above: relations between word family members are complexity ordered through word formation mechanisms (inspired by Lango et al. (2021)) which is reflected in the order of appearance of the new word family items in receptive and productive data.

2.1 Case Study 1: distribution of simplex root lexemes

To look into this hypothesis, we start from an analysis of the distribution of distinct simplex root lexemes over the two corpora. By **simplex root lexeme** we understand lexical items that consist of one morpheme only, namely strictly of one root, e.g. *dag* ('day'). By being **distinct** we mean that we account for each simplex root lexeme only once, at the level where they occur for the first time. For example, *dag* may have been used for the first time at A1 level, but is repeatedly used at all other levels. We count *dag* only once, at the level of its first occurrence (i.e. A1). We, thus, do not count *dag* among root lexemes at levels above A1.

	Total	A1	A2	B1	B2	C1	C2
SweLL pilot (productive)	1108 (23%)	254 (52%) *snö se*	347 (39%) *mjölk lär*	207 (22%) *värd yr*	189 (15%) *blind falsk*	108 (10%) *botten armé*	3 (5%) *mord ström*
Coctaill (receptive)	2195 (16%)	732 (36%) *fisk gift*	499 (20%) *paj skuld*	447 (12%) *café dräng*	361 (10%) *feg hinder*	157 (7%) *slarv tvist*	0

Table 2: Distribution of root lexemes across the two corpora per level: absolute number of simplex root items (and their percentage of all new items at each level).

There are a total of 2298 distinct simplex root lexemes in the Swedish Word Family resource. These are split between receptive (2195 items) and productive (1108 items), with an overlap of 1063 items as shown in Table 2. These simplex root lexemes are more represented at earlier levels and gradually decline as the level of proficiency grows. Figure 2 shows that more than half of the new vocabulary at A1 level is constituted of lexemes consisting of only a root. This percentage drops gradually to 10% at C1 level and to 5% at C2 level. The same tendency can be seen in the receptive corpus where most of the simplex root lexemes are at A1 level with 36% and the number gradually drops to 7% at C1 level.

An interesting question is whether simplex root lexemes within respective word families tend to precede items that are more complex in terms of word formation (derivations and compounds). That is, whether learners first get acquainted with, e.g. the simplex root lexeme *dag*, and then learn its derivations (e.g. dag*lig*, dag*is*) and compounds (e.g. *mån*dag, *vår*dag). We have looked into several word families to examine this assumption.

2.2 Case study 2: *dag*-family

The *dag*-family (‘day’) is one of the most numerous word families in the Swedish Word Family resource with a total of 101 members in the Coctaill corpus, 32 of which are also represented in the SweLL-pilot learner essays. As hypothesized, the simplex item consisting of only the root, *dag*, is introduced before other items at A1 level together with some derivatives and compounds, namely days of the week (*lör*dag, *fre*dag, etc.) and some of the words describing everyday routines, such as *mid*dag (‘dinner’), dag*is* (‘kindergarden’), as well as parts of the day, *eftermid*dag (‘afternoon’), see Figure 3. It is obvious from the word cloud in Figure 3 that the root lexeme *dag* is by far the most frequent in the *dag*-family at the A1 level, judging by its relative size.

The *dag*-family is growing through numerous complex patterns of compounding and derivation, with up to five roots within one item, e.g. *här-om-dag-en* (3 roots; ‘the other day’), *föd-else-dag-s-pre-sent* (3 roots; ‘birthday present’), *sön-dag-s-efter-mid-dag* (5 roots; ‘Sunday afternoon’). An interesting fact is that most family members are nouns, with only a few adjectives (dag*lig*, *gammal*dags), adverbs (dag*ligen*, *härom*dag*en*) and proper names, of which both designate names of newspapers (Dag*ens nyheter*, *Svenska* Dag*ladet*); and there are no verbs apart from the multi-word expression *sova mid*dag (‘have an afternoon nap’).

One more interesting observation is that the most radical expansion of the family happens at A2 and B1 levels. Gradually, the new items decrease after these two levels, with as little as only seven new items at C1 level. This is most probably due to the “topical” nature of the word *dag* and the fact that daily routines have already been well covered at earlier levels.

2.3 Case study 3: *språk*-family

A predictable pattern of “easy first” can be traced also in the *språk*-family, exhibiting 62 family members, with 57 of them appearing in the Coctaill corpus (i.e. in course book texts). The root lexeme *språk* appears first at A1 level in both corpora and is the only representative of the family at that level (see center of Figure 4). We can assume that its presence in the texts has a priming effect on learners, making it possible to combine that root with a number of other roots and affixes at the next levels.

There are a few distinct patterns that we observe in the *språk*-family development across levels:

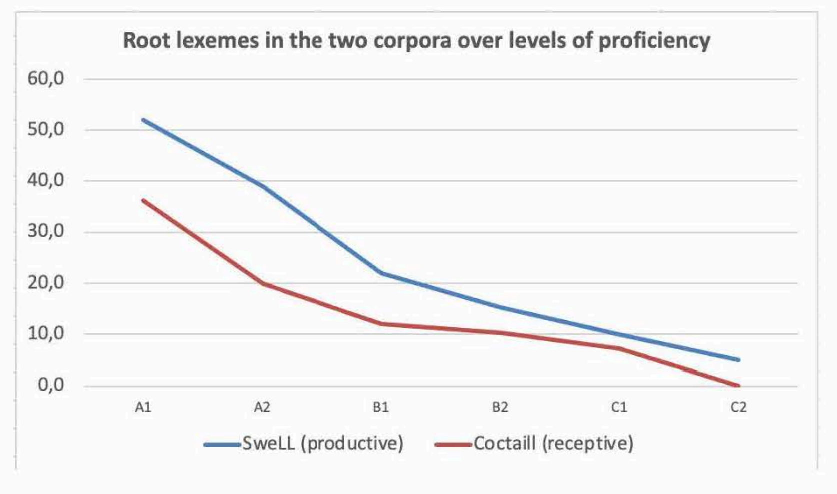


Figure 2: Simplex root lexemes in receptive and productive corpora.

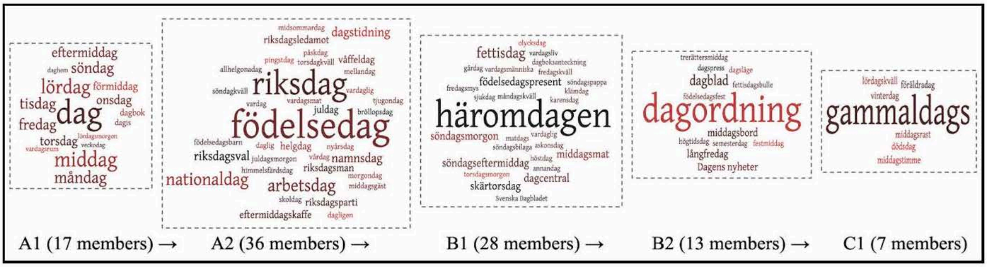


Figure 3: Simplex root lexemes in receptive and productive corpora.

- (numeral/adjective root) + *språk* + adjectival suffix *-ig*, e.g. *en*språk*ig* ('monolingual'), *fri*språk*ig* ('free-spoken'), *fler*språk*ig* ('multilingual') → in turn, leading to a derivation pattern with suffix *-het* thereof at the advanced levels, e.g. *fler*språk*ighet* ('multilinguality').

- compound nouns, ending in *språk*, that describe various types of languages, e.g. *tal*språk ('spoken language'), *riks*språk ('state language'), *yrkes*språk ('professional language'). The left hand element is usually a noun used in attributive function. This seems to be one of the most productive patterns of word-formation within this word family.

- *språk* used attributively to characterize nouns used as a right-hand element in compounds, e.g. språk*kurs* ('language course'), språk*familj* ('language family'), språk*politik* ('language policy'). This pattern of word formation is highly productive, making the word family expand vastly by C1 level, as shown in Figure 4.

The outlined tendencies observed in course books echo analysis of the word formation networks in Czech and other languages, a representation of one of them shown in Figure 5. While Lango et al. (2021) aimed at a general language description and analysis, we see similar patterns in the learner language. A hypothesis that more complex patterns follow easier patterns requires, however, more thorough examination of far larger number of word families than we provide here, correlating those with each other

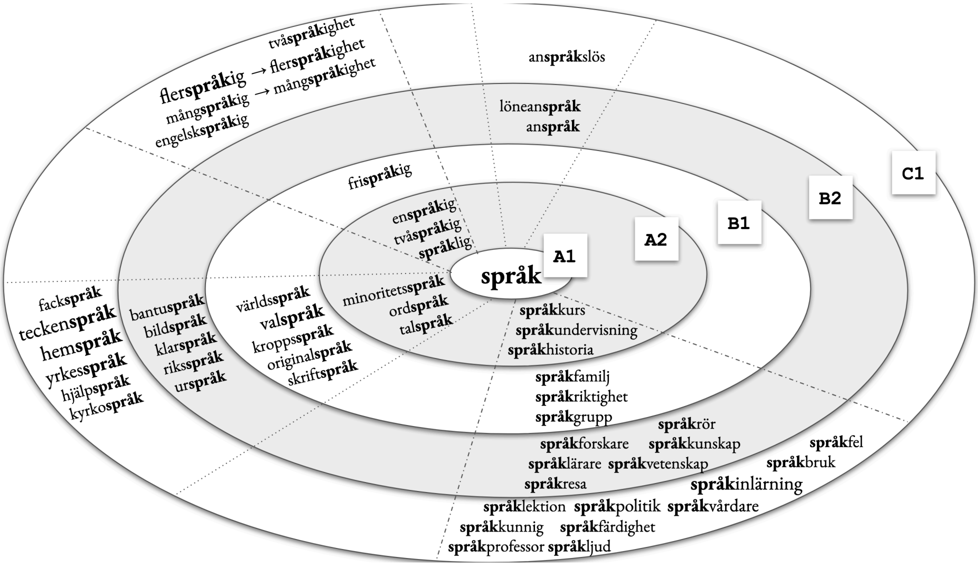


Figure 4: Distribution of *språk*-family in the Coctaill (receptive) data.

and with derivational families (i.e. families in a boarder sense, where lexical items are centered around a shared affix, infix or other morpheme types).

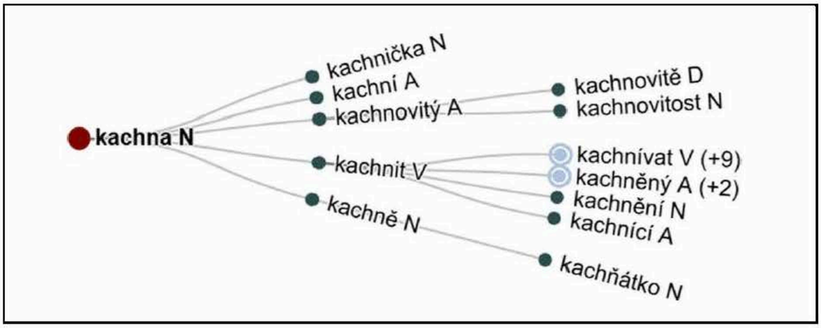


Figure 5: A word formation network for Czech. A reprint from Lango et al. (2021).

All in all, the above analysis of the *språk*-family demonstrates a clear case where quite a few word-formation patterns give rise to numerous new lexical items. An interesting fact is that most of the *språk*-family members have been used very infrequently with only a few exceptions, such as *fler*språk*ig*, *hem*språk and *tecken*språk, all of which appear at C1 level. And, of course, the core item itself, *språk*, dominates at all levels, not only at the level of first occurrence; which is probably easily explainable by the course book orientation. The predominant word formation mechanism is compounding.

2.4 Case study 4: *lyx*-family

Most families we have looked into seem to follow the above outlined path, i.e. introducing simpler words before more complex ones. However, an important step when examining a hypothesis is to find

	B1	B2	C1
Coctaill	lyx*vara* ('luxury product')	*lyx* ('luxury'); lyx*a* ('to afford, indulge'); lyx*ig* ('luxurious')	—
Swell	—	—	lyx*liv* ('luxury life')

Table 3: *lyx*-family

counterexamples that could trigger new insights. One of such counterexamples is represented by the *lyx*-family. It contains only 5 members, distributed in the data as shown in Table 3.

A more intuitive order of introduction, following the principle of 'easy first', would be:
lyx → *lyxa*, *lyxig*, *lyxliv, lyxvara*

However, we can see that the order of appearance of the words in the *lyx*-family is counter-intuitive: a compound lyx*vara* ('luxury item') appears before the simplex root item *lyx* ('luxury') and its derivatives: lyx*ig* ('luxurious') and lyx*a* ('to indulge, to treat yourself to luxury'). Several explanations come to mind (apart from the obvious one that natural languages are idiosyncratic, and do not tend to adhere to rules):

The first one is connected to the reasoning about 'what makes an item easy'. Up to now we operated under the assumption that the simplest morphological structure is the main characteristics of an easy item. This is, of course, a simplified view on reality. To complicate this, semantics could be another constituent of the simplicity equation that needs to be considered, in this case, the dichotomy between the concreteness and the abstractness of the words in the *lyx*-family. All of the *lyx*-items at B2 level have an abstract meaning whereas the B1 item lyx*vara* is concrete. From the cognitive point of view it may be easier to acquire a concrete item (lyx*vara*) than the other ones.

The second likely explanation could be the priming effect of the second family that lyx*vara* belongs to, namely, *var*-family. If we examine how var*a* ('product, item') is used in the Coctaill texts, we will see that up to B1 ten (10) family members are introduced, as shown in Table 4, and all of them are compounds except the root lexeme itself (var*a*, 'product, item'). Shopping seems to be one of the central topics in texts at B1 level, since various types of products are introduced. The word formation pattern is very similar between five of the six items containing the root *var* at B1 level: 'a modifier describing the type of product + var*a*'; lyx*vara* falls into this pattern, and becomes the first member of the *lyx*-family to get introduced to language learners. It is possible to speak about a priming effect of statistically recursive orthographic chunks (in this case 'modifier+var*a*'), which, after repeated appearances, start to distinguish themselves as separate morphemes (i.e. var*a* as a separate item, distinct from a range of modifiers, and gradually, *lyx* becomes recognized as an independent lexical item).

Interestingly, even in the case of the *var*-family, we see that the first item introduced at A1 level is not the root item var*a*, but the compound var*uhus* ('store', literally 'house with goods'), see Table 4. Here it would make sense to check the pattern of introduction of the *hus*-family members, and there is a good reason to believe that this one would lead to another quest and become a never-ending story. But we can also see that the issue of what is easier for a learner would be interesting to test by testing learners on high-frequency roots, compounds and two-morpheme derivations.

Besides, the topical focus of texts influences which items that are introduced at which levels, which is a predictable consequence of sequencing language education: learners first need to learn how to introduce themselves and attend to their immediate needs, and gradually to lift their attention to the world around them and topics that are no longer centered on learners (the *hus*-family and the *var*-family are two clear examples of how central topics change over the proficiency levels in relation to learner needs) (as we also see in the CEFR level can-do statements, of Europe (2018)).

Finally, there is hypothetically a good reason why most word families do not count compounds: compounds add factors that are difficult to account for, such as exposure to the other families and influence thereof, and since word families are often used in relation to frequency bands, compounds would be complicated to include in the total frequency counts since they can combine a high-frequency and a low-

	A1	A2	B1
Coctaill	var*uhus* ('store')	var*a* ('product, item'); *möbel*var*uhus* ('furniture shop')	mat*vara* ('food item') *märkes*var*a* ('branded item') *rå*var*a* ('raw product') *bytes*var*a* ('return item') *lyx*var*a* ('luxury item') *mat*var*uhus* ('food store')
Swell	—	—	—

Table 4: *var(a)*-family

frequency word family. If we disregard the compounds in the *lyx*-family, the pattern will become the simplex first, all words in the family appearing at the same level with the morphologically simplest one among them (see Table 3, B2 level). Compounds are debated in cognitive research on morphology, some studies positing that compounds are processed as whole-word units, whereas others show evidence that access to constituent morphemes prior to the whole compound word ensures easiness of mental access to the item (see review of the studies in Leminen et al. (2019)). Regardless, the suggestion to remove compounds from the analysis of word families is not viable for Swedish, since compounding is the most widely spread word formation mechanism (cf Svensson (2022)) and many new roots/words are learnt initially from compounds, like var*a* from var*uhus* and *lyx* from *lyx*var*a*. In fact, even placenames, which are made up of compounds, have been recognised to help L2 Swedish learners learn new roots, such as *torg* ('square') from placenames like *Rådmans*torg*et* (Löfdahl et al., 2015).

To conclude, we have traced the order of learning the word *lyx* as follows:
$hus\ (A1) \rightarrow varuhus\ (A1) \rightarrow vara\ (A2) \rightarrow lyxvara\ (B1) \rightarrow lyx\ (B2)$

3 Conclusion and future work

We have shown that, using a descriptive resource like Swedish Word Family, it is possible to research linguistic questions and to study language learning paths. However, there are many more application scenarios, for example, to utilize SweWF as a graded resource for Intelligent Computer-Assisted Language Learning, and in particular for the area of Computer-Adaptive Language Testing, e.g. based on derivation patterns that are typical at later levels of development, for coining non-existent words for word knowledge testing items; and many, many others.

Acknowledgements

This work has been supported by a research grant from the Swedish Riksbankens Jubileumsfond Development of lexical and grammatical competences in immigrant Swedish, P17-0716:1, and by Nationella språkbanken and HUMINFRA, both funded by the Swedish Research Council (2018-2024, contract 2017-00626; 2022-2024, contract 2021-00176) and their participating partner institutions.

References

R Harald Baayen, Richard Piepenbrock, & Leon Gulikers. 1996. The CELEX lexical database (CD-ROM). In *Linguistic Data Consortium*. University of Pennsylvania.

Laurie Bauer & Paul Nation. 1993. Word families. *International journal of Lexicography*, 6(4):253–279.

Elfrieda H Hiebert, Amanda P Goodwin, & Gina N Cervetti. 2018. Core vocabulary: Its morphological content and presence in exemplar texts. *Reading Research Quarterly*, 53(1):29–49.

Lívia Körtvélyessy, Alexandra Bagasheva, & Pavol Štekauer. 2020. *Derivational networks across languages*. De Gruyter Mouton.

Mateusz Lango, Zdenek Zabokrtsky, & Magda Sevcikova. 2021. Semi-automatic construction of word-formation networks. *Language Resources and Evaluation*, 55, 03.

Alina Leminen, Eva Smolka, Jon A Dunabeitia, & Christos Pliatsikas. 2019. Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *Cortex. Elsevier*, 116:4–44.

M. Löfdahl, S. Tingsell, & L. Wenner. 2015. Lexikon, onomastikon och flerspråkighet [= Lexicon, onomasticon and multilingualism]. In E. Aldrin, M. L. Gustafsson, M. Löfdahl, & L. Wenner, editors, *Innovationer i namn och namnmönster*. NORNA-förlaget.

Alexandre Nikolaev, Sameer Ashaie, Merja Hallikainen, Tuomo Hänninen, Eve Higby, JungMoon Hyun, Minna Lehtonen, & Hilkka Soininen. 2019. Effects of morphological family on word recognition in normal aging, mild cognitive impairment, and Alzheimer's disease. *Cortex. Elsevier*, 116:91–103.

Council of Europe. 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. *Companion volume with new descriptors.*

Anders Svensson. 2022. Tre av fyra nyord är substantiv [=Three out of four neologisms are nouns]. *Språktidningen 2 Jan. 2022*.

Elena Volodina, Yousuf Ali Mohammed, & Therese Lindström Tiedemann. 2021. CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 178–189, Reykjavik (Online). Linköping University Electronic Press.

Britta Zeller, Jan Šnajder, & Sebastian Padó. 2013. DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1201–1211.

Annotating the Narrative: A Plot of Scenes, Events, Characters and Other Intriguing Elements

Mats Wirén[1], **Adam Ek**[2] **and Murathan Kurfalı**[1]

[1]Department of Linguistics
Stockholm University
mats.wiren@ling.su.se
murathan.kurfali@ling.su.se

[2]Department of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
adam.ek@gu.se

Abstract

Analysis of narrative structure in prose fiction is a field which is gaining increased attention in NLP, and which potentially has many interesting and more far-reaching applications. This paper provides a summary and motivation of two different but interrelated strands of work that we have carried out in this field during the last years: on the one hand, principles and guidelines for annotation, and on the other, methods for automatic annotation.

1 Introduction

This paper summarizes and motivates recent work of ours on analysing aspects of narrative structure in prose fiction. If you (the reader) are Lars Borin, we hope that you will find this exposition interesting. We think that it fits best with your interests in digital humanities and digital language infrastructure (Borin et al., 2017), to which it is meant as a contribution. If you are not Lars Borin, we still hope that you will find this exposition interesting. In particular, if we manage to convince you that analysis of narrative structure is an excellent testbed not just for people interested in literature but for NLP in general, we will be happy.

In terms of methodology, our work is rooted in linguistics and NLP, but we have also adopted relevant concepts from narratology, the field in which narratives in all their forms are studied. A central notion here is the distinction between three layers of abstraction (for which the terminology differs): a *narrative text* as told by a narrator (and which is what the reader sees); the *story*, which corresponds to the chronological and causal sequence of events in the fictional world; and the *discourse*, which is the sequential organization of the story by the narrator (with events typically presented in non-chronological order). These three layers can be associated with a "who?", a "what?" and a "how?", respectively, and the question "Who tells what, and how?" can thus be taken as an outline of the goal of narrative analysis (Jahn, 2021, Figure 2, page 17). Incidentally, this is a much harder question than the one typically associated with traditional NLP, "Who did what to whom?", something that we will return to below.

We have approached the analysis of narrative structure in two ways: First, by developing a series of annotation guidelines and using them for successive manual annotations; and secondly, by developing methods for automatic analysis based on our own annotations as well as those of others. In the following, we will give an account of these two strands of work, and will end by discussing why we are doing this.

2 Annotation guidelines

Annotation of narrative phenomena in prose fiction (whether manual or automatic) is relatively recent in computational linguistics; among the first approaches is Elson et al. (2010). In contemporary narratology, annotation barely plays any role at all (Reiter et al., 2019a, page 17). However, a recent event which spurred interest in this area was the Shared Task on Systematic Analysis of Narrative Texts through Annotation (SANTA), first announced in 2017 and organized in Germany (Reiter et al., 2019b; Reiter et al., 2019a). The format of SANTA emulated a shared task in NLP, but the difference was that what was compared was not systems but annotation guidelines. More specifically, the task in SANTA was annotation guidelines for *narrative levels*, a term introduced by Genette (1983), meaning roughly subordination

Mats Wirén, Adam Ek and Murathan Kurfalı. 2022. Annotating the narrative: A plot of scenes, events, characters and other intriguing elements. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 161–166. Available under CC BY 4.0

in the sense of stories in stories, or phenomena such as direct speech between characters which can be thought of as a form of embedding (of quotations) relative to the narrator. The organizers provided a selection of theoretical material as a background, but did not recommend any particular narratological approach; rather, the participants were given free hands in terms of theoretical assumptions.

Annotation of narrative structure is much like annotation in NLP and should fulfil the same basic objectives (or so we will assume). In particular, it should capture the crucial aspects of the phenomena targeted, and being simple enough both for annotators to perform and for machines to learn. However, it has two distinctive characteristics. First, narrative categories may cover very large portions of the text, in contrast to syntactic annotation and others which are typically constrained to the sentence level. Furthermore, the context relevant for determining a category can also be very large. In these respects, it is similar to co-reference annotation, which is generally seen as a document-level task. Secondly, annotators typically need to decide not just on the categories themselves but also on the spans of these categories. This is reminiscent of named entities, for which annotators first need to identify the segment that it spans and then decide on its category.[1]

Guidelines for the first round of SANTA were submitted in June 2018. Eight groups from Germany, Ireland, Canada, the U.S. and Sweden (us) made submissions. The participation was broad, with computer scientists, computational linguists and literary scholars. In September 2018, the groups convened for a workshop in Hamburg where each of the guidelines were presented, discussed and ultimately ranked. To this end, the guidelines were evaluated both qualitatively and quantitatively. The former included conceptual coverage (the extent of the theoretical underpinnings), applicability (usability for annotators), and usefulness (roughly, helpfulness for the purpose of understanding the narrative structure). The quantitative measure was inter-annotator agreement. This was based on parallel annotations made prior to the workshop by letting participants apply their own guideline, someone else's guideline, and having a group of students (supervised by the organizers) annotate for everyone. The guidelines, expert reviews of these administered by the organizers, and the evaluation results were published in a special issue of *Journal of Cultural Analytics* (Gius et al., 2019; Willand et al., 2019); our guideline is included as Wirén et al. (2019).

Development of annotation guidelines is an inherently iterative process. To take advantage of the insights gathered in the first round, the organizers decided to do a second round in which the groups were offered to submit revised guidelines, in May 2019. Although it was not possible to hold a second workshop, the organizers administered new parallel annotations according to the revised guidelines, and made a quantitative evaluation of these. The results and final conclusions were published in another special issue of *Journal of Cultural Analytics* (Gius et al., 2021).

Overall inter-annotator agreement improved between round 1 and 2. Also, whereas the best score in round 1 was 0.30 (ours: 0.23), the best score in round 2 was 0.46 (which was our guideline). These scores were measured using the γ (gamma) metric (Mathet et al., 2015), where 1 means no disagreements and 0 corresponds to random agreement. The γ metric was chosen since it takes care of agreement both with respect to categories and spans, as mentioned in Section 1. In summary, the 0.46 score is not a great result, but it was considered acceptable since this was a new task, and it was at least substantially higher than the best agreement in the first round.

A breakdown of our annotation scheme is shown in Table 1; it is further described in Wirén & Ek (2021). The tagset is hierarchically structured in four layers, ordered by an inclusion relation. The scheme covers *voice*, that is, whether the narrator is ever present in the story or not (Genette, 1983, Chapter 5); *focalization*, how much information the narrator has access to (Genette, 1983, page 189 ff.); and identification of passages told by the narrator and passages containing the characters' direct speech, respectively. In addition, the scheme allows for embeddings among these latter two kinds of passages (stories in stories, etc.). Our annotation of fictional dialogue is relatively fine-grained. We distinguish between turns and

[1] It might be argued that, in distinction to NLP annotation, the graphic formatting of printed prose fiction would tell us something about narrative structure; for example, turn changes among speakers in fictional dialogue seem to regularly be accompanied by new paragraphs. However, conventions for this vary wildly, and a basic assumption in SANTA is that narrative annotation should be based solely on the contents of the text, and not on formatting devices such as text structure in TEI XML or paragraphs or chapters in a printed book (Reiter et al., 2019a, page 15).

Table 1: Hierarchical structure of the annotation scheme.

Layer	Tag	Description
1	<VOICE_1>, <VOICE_3>	Narrator's presence in the story
2	<FOC_UNR>, <FOC_INT>, <FOC_EXT>	Perspective of the narrator
3	<NARRATOR>	Narrator's discourse
4	<CHARACTERS>	Characters' discourse
4.1	<TURN>	*Turn* = one or several lines with the same speaker
4.1.1	<*Speaker–Addressee*>	*Line* = one or several utterances with the same speaker and the same addressee, and tagged with these
4.1.1.1	<NC>	Speech-framing construction

lines, and annotate the identities of speakers and addressees. We also annotate *speech-framing constructions*, which provide the narrators cues about the circumstances of the speech as opposed to the speech itself (somewhat related to the notion of speech-framing expressions in Caballero & Paradis (2017)).

3 Automatic annotation

The SANTA shared task had a twofold aim: increasing the understanding of narrative levels, and generating annotated data for the purpose of machine learning. (The plan is to have a second, follow-up shared task which will be devoted to automatic annotation of narrative levels based on annotations from SANTA.) However, even before SANTA we began to explore automatic annotation of narrative structure. Our interest in this began with direct speech, an independent narrative mode which has an interesting "double orientation" (Koivisto & Nykänen, 2016): it can be understood both in relation to our experiences of real-life conversations (by mimicking aspects of this) and in relation to the fictional world.

In the first system that we constructed, by Ek et al. (2018), our goal was to keep track of the identities of speakers and addressees, as this is one key aspect of the structure of a story. To this end, we used an averaged perceptron with handcrafted features for both speakers and addressees. The system relied on three contexts: (i) the passage of direct speech in which we want to identify the speakers and addressees; (ii) the narrative passage immediately preceding this; and (iii) a global context consisting of all dialogue and narration preceding these. We used information about the frequency with which different characters occurred in all these, as well as mention with speech verbs, such as "... said Adam". We compared our system against three baselines and found that it outperformed all of them. Another finding was that the system was better at predicting speakers than addressees.

In a spin-off from this work, Ek & Wirén (2019) were interested in identifying speech-framing constructions as mentioned in Section 2, in effect elements of narration appearing inside passages of dialogue. For example, in "Machine learning is fun, said Adam, watching the audience gleefully", the part "said Adam, watching the audience gleefully" is the narrators cues about the circumstances of the speech (including a speech tag indicating the identity of the speaker). Our model used logistic regression with handcrafted features, mainly various syntactic cues. Since this was a new task, there were no previous systems to compare with, but the system outperformed three baselines with a large margin.

Kurfalı & Wirén (2020) described an attempt towards a generalized solution to this task, applicable to a multitude of languages. Operating under a low-resource assumption of complete absence of manually labelled data, we firstly devised a set of heuristics to identify the cases in a large book corpus where quotation marks are used sufficiently consistently to automatically elicit enough training data. Then, as for classifier, we replaced the linguistics features, which we cannot assume to be available across languages, with the multilingual contextual embeddings to arrive at a feature-independent multilingual model. The results, obtained on manually annotated datasets in four different languages (including Swedish, with data provided by Stymne & Östman (2020)), suggest that the proposed methodology achieves comparable results to the supervised monolingual baselines.

A different kind of problem that we have worked on is segmentation of scenes. This concept was introduced in narratology by Genette (1983), where the basic idea was that in a scene, the amounts of time in the story and the discourse are proportional to each other. Put differently, the narration in a scene should be roughly chronological with a uniform pace, without time leaps, flashbacks or other temporal discontinuities. Scenes are basic building blocks of a narrative discourse, and are in turn composed of events, considered to be the smallest spatiotemporal units.

Kurfalı & Wirén (2021) explored this problem through participation in the Shared Task on Scene Segmentation (STSS) (Zehe et al., 2021). To obtain an operational definition of a scene, the organizers had adopted a more general definition than above which also involved the set of characters (which should be stable), space (largely unchanged) and action (coherent and continuous). An extensive annotation guideline had been developed to make these notions more precise, and a corpus of 15 dime novels in German had been annotated. In addition to scenes, there are also non-scenes, typically in the form of summaries where the progress of time is compressed, though sequences of non-scenes were not targeted here. The task of scene segmentation thus consisted in dividing a text into scenes and non-scenes and labelling them accordingly. An operationalization of this which we adopted is to identify and label scene transitions of three types: `SCENE-SCENE`, `SCENE-NONSCENE` and `NONSCENE-SCENE`.

We modelled scene segmentation as a sequence classification task, similar to named-entity recognition or part-of-speech tagging, with the difference that sentences are the target linguistic units rather than tokens. Following Cohan et al. (2019), a sequence of N sentences was concatenated by BERTs special delimiter token $[SEP]$, which was used for representing the sentence it preceded. The best F1 score obtained in the shared task was 0.37 (which was our system; the second-best system had 0.16). Our system was comparatively successful in detecting scene-to-scene transitions; yet, it completely failed in discovering the boundaries between non-scenes and scenes. Overall, the result was not great, but it was expected that this new task would be difficult. Also, the F1 score is very strict in the sense that it counts a scene boundary as correct only if it is predicted at exactly the right position, whereas an offset of even a single sentence is counted as a complete miss.

4 Discussion

We have found it very fruitful to engage in work (especially the shared tasks) on both guidelines and methods for annotation of narrative discourse, as we have been able to transfer ideas from the one to the other. For example, our detailed annotation for direct speech has inspired our work on methods for automating this kind of analysis. In general, we are motivated both by an interest in trying to make narratological concepts more precise and in applications of this kind of analysis, whether in literary studies or in other areas where story-telling may play a role (journalism, history, courtroom interactions, to just take a few examples).

It was suggested in Section 1 that the question “Who tells what, and how?” is much harder to answer than “Who did what to whom?”. This is in fact also an important reason why we are pursuing this line of work: we believe that analysis of narrative structure provides an excellent testbed for state-of-the-art NLP technology. The Shared Task on Scene Segmentation was a demonstration of this: although our system outperformed the other systems, the overall results were not impressive, thus giving a sense of the big challenges posed by problems like this.

Scene segmentation is a crucial task since it concerns the basic building blocks of a narrative discourse. However, out of the tasks that we have explored so far, we believe that it is the one with the largest room for improvement, as suggested by the gap between the current state-of-the-art and human performance. Despite several different architectures based on language models, they seem to fall short of achieving acceptable performance unaided. One possible source of aid can come from the task of event detection on the assumption that scenes and non-scenes consist of different event types (for example non-scenes can be expected to contain more state-like events).

References

Lars Borin, Nina Tahmasebi, Elena Volodina, Stefan Ekman, Caspar Jordan, Jon Viklund, Beáta Megyesi, Jesper Näsman, Anne Palmér, Mats Wirén, Kristina Björkenstam, Gintarė Grigonytė, Sofia Gustafson Capková, & Tomasz Kosiski. 2017. Swe-Clarin: Language resources and technology for Digital Humanities. In *Digital Humanities 2016. Extended Papers of the International Symposium on Digital Humanities (DH 2016) Växjö. Edited by Koraljka Golub, Marcelo Milra. Vol-2021*, Aachen. M. Jeusfeld c/o Redaktion Sun SITE, Informatik V.

Rosario Caballero & Carita Paradis. 2017. Verbs in speech framing expressions: Comparing English and Spanish. *Journal of Linguistics*, 54(1):45–84.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, & Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699.

Adam Ek & Mats Wirén. 2019. Distinguishing narration and speech in prose fiction dialogues. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, pages 124–132, Copenhagen.

Adam Ek, Mats Wirén, Robert Östling, Kristina Nilsson Björkenstam, Gintarė Grigonytė, & Sofia Gustafson-Capková. 2018. Identifying speakers and addressees in dialogues extracted from literary fiction. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC)*, Miyazaki.

David Elson, Nicholas Dames, & Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala. Association for Computational Linguistics.

Gérard Genette. 1983. *Narrative Discourse: An Essay in Method*. Cornell University Press.

Evelyn Gius, Nils Reiter, & Marcus Willand. 2019. A Shared Task for the Digital Humanities. Chapter 2: Evaluating Annotation Guidelines. *Journal of Cultural Analytics*, 4(3):1–11.

Evelyn Gius, Marcus Willand, & Nils Reiter. 2021. On Organizing a Shared Task for the Digital Humanities — Conclusions and Future Paths. *Journal of Cultural Analytics*, 6(4):1–28.

Manfred Jahn. 2021. *Narratology 2.3: A Guide to the Theory of Narrative*. English Department, University of Cologne, Cologne.

Aino Koivisto & Elise Nykänen. 2016. Introduction: Approaches to fictional dialogue. *International Journal of Literary Linguistics*, 5.

Murathan Kurfalı & Mats Wirén. 2020. Zero-shot cross-lingual identification of direct speech using distant supervision. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 105–111, Online. International Committee on Computational Linguistics.

Murathan Kurfalı & Mats Wirén. 2021. Breaking the narrative: Scene segmentation through sequential sentence classification. In *Proceedings of the Shared Task on Scene Segmentation, co-located with the 17th Conference on Natural Language Processing (KONVENS)*, volume 3001, `http://ceur-ws.org/Vol-3001/`, pages 49–53, Düsseldorf.

Yann Mathet, Antoine Widlöcher, & Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Nils Reiter, Marcus Willand, & Evelyn Gius. 2019a. A Shared Task for the Digital Humanities. Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks. *Journal of Cultural Analytics*, 4(3):1–24.

Nils Reiter, Marcus Willand, & Evelyn Gius. 2019b. Foreword to the Special Issue "A Shared Task for the Digital Humanities: Annotating Narrative Levels". *Journal of Cultural Analytics*, 4(3):1–4.

Sara Stymne & Carin Östman. 2020. SLäNDa: An annotated corpus of narrative and dialogue in Swedish literary fiction. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 826–834, Marseille. European Language Resources Association.

Marcus Willand, Evelyn Gius, & Nils Reiter. 2019. A Shared Task for the Digital Humanities. Chapter 3: Description of Submitted Guidelines and Final Evaluation Results. *Journal of Cultural Analytics*, 4(3):1–15.

Mats Wirén & Adam Ek. 2021. Annotation Guideline No. 7 (revised): Guidelines for annotation of narrative structure. *Journal of Cultural Analytics*, 6(4):164–186.

Mats Wirén, Adam Ek, & Anna Kasaty. 2019. Annotation Guideline No. 7: Guidelines for annotation of narrative structure. *Journal of Cultural Analytics*, 4(3):1–22.

Albin Zehe, Leonard Konle, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, et al. 2021. Detecting scenes in fiction: A new segmentation task. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3167–3177.

The other SAT-Solver:
Applying lexicons to SweSAT word questions

Niklas Zechner
Språkbanken Text
Gothenburg University
niklas.zechner@gu.se

Abstract

SweSAT synonyms is a collection of questions from the Swedish SAT (*högskoleprovet*), and part of the SuperLim suite of machine learning datasets. Each question consists of one word or short phrase, and five possible explanations, each of which is also a word or short phrase. In this study, two different lexicons are applied in trying to answer to these questions. The first is Bring's Thesaurus, in a version partly updated from the 1930 original. The second is SALDO, a Swedish association lexicon. We find that although coverage is limited by the presence of multiword expressions, each is reasonably accurate for words where a match is found in the lexicon, and by combining them we get an overall accuracy of 47% when counting all words.

1 Introduction

Understanding synonyms is an important part of natural language understanding. As an example of this, SuperLim provides questions from the word section of the Swedish SATs, where the task is to identify which of five words or expressions are synonymous with one given word or expression. We attempt to do that by using two different lexicons, with different types of associations between words.

2 Data

2.1 SuperLim

The SuperLim suite (Adesam et al., 2020) consists of (so far) 13 datasets for evaluation of language understanding models in Swedish. One of those is SweSAT Synonyms, based on questions from the Swedish SAT (*högskoleprovet*). For each question, there is one focus word, and five alternatives to choose from. The focus word and the alternatives can be single words, or short expressions, and the task is to find the alternative which is a synonym or alternative for the focus word.

One purpose of SuperLim is to provide a unified test set for several language understanding tasks in Swedish. Since the resource is relatively new, there are few results to compare with yet, and the only attempt we find for the SweSAT task is that of Rekathati (2021).

2.2 Bring

Svenskt ordförråd ordnat i begreppsklasser is an adaptation to Swedish of Roget's Thesaurus, written by S. C. Bring in 1930. In a digitised and modernised version, it is provided by Språkbanken as Blingbring (here used in version 0.3) (Borin et al., 2014). It contains just over 1000 semantic categories, each with a number of subcategories. The subcategories are either noun, verb, or adjective categories. The categories, subcategories, and words are all ordered so that the most similar words should be close together.

Niklas Zechner. 2022. The other SAT-Solver: Applying lexicons to SweSAT word questions.
In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 167–169. Available under CC BY 4.0

2.3 Saldo

Saldo (Swedish Associative Thesaurus, here used in version 2.3) (Borin et al., 2013) is an electronic lexicon resource for modern Swedish. For each entry, it contains links to a primary descriptor, a more basic word considered an "ancestor sense" of the word in question. Many words also have one or more secondary descriptors, other words whose associations give a better idea of the semantics of the word in question. Since the descriptors are always words considered more basic, the full set of words form a tree (using the primary descriptor) or a directed acyclic graph (using all descriptors), leading back to a single pseudo-word root node.

3 Experiment

The SweSAT dataset contains 822 questions, each with five alternatives, only one correct. Random guessing would therefore have an expected accuracy of 20%. Taking Rekathati (2021) as the state of the art, it reports an accuracy of 42.82%.

3.1 Bring

For each question, we look up the focus word in Bring, and compare with each of the options. If there is an option word which appears in the same subcategory as the focus word, we choose that. Otherwise, we look at the larger categories. If there is also no match, we choose the option which is closest to the focus word in the entire word list. In cases where more than one option word is found in a category, we pick the first match (so effectively at random). In some cases, a word appears in more than one category in Bring; we count a match if both words appear together in any category.

Out of the 822 focus words, 258 (31%) did not appear in Bring. Many of these words are adjectives, which seem to be underrepresented in Bring, and many are phrases. For 88 words (11%), although the focus word appeared in Bring, none of the option words did.

For 243 words (30%), a match was found in a subcategory. Of those, 192 (79%) were correct.

For 61 words (7%), a match was found in a large category. Of those, 26 were correct (43%), so still much better than random, and in fact on par with the state of the art.

For 172 words (21%), matches were found in Bring, but none within the same large category. In principle, the categories in Bring are arranged so that more similar categories are close, which should mean that we can find the most likely option by looking at the closest word. This gives only 25 correct answers (15%), considerably worse than random. This might be a coincidence, but we can also speculate about a possible explanation: For option words which are found in Bring, words that are actually similar are likely to share a category, so if the word is found but in a different category, that word is less likely to be correct than a word which is not found at all. If we flip the strategy around, and instead choose an absent word over one which is present, we get 46 correct answers (27%). Better than random, but not a particularly reliable method.

This means that in total, if we were to give an answer to each question, using the absent-above-distant method, and assuming that a random guess has a 20% chance, we would get 333 correct (41%), just shy of the state of the art. If we only consider questions where there is a match in a subcategory, we get 79% correct while answering 30% of the questions, which is at least a useful accuracy. If we include those with a match in a large category, we get 72% correct out of 37% answered.

3.2 Saldo

In Saldo, entries can be considered nodes in a graph, where each entry is connected to its descriptors. To see the distance between two words, we traverse through its list of ancestors (i.e. descriptors, the descriptors' descriptors, etc.) until we find one shared by both words. For each question, we choose the option word which is closest to the focus word.

We find 291 correct (35%), 330 incorrect (40%), 96 (12%) where the focus word does not appear in Saldo, and 105 (13%) where the focus word but none of the options appear in Saldo. This means that if we only consider questions where a match is found, we get 47% correct while answering 76% of the questions. If we guess at random for the unknown questions, that would give an accuracy of 40%.

3.3 Simple tricks

A common piece of advice for multiple-choice questions is to "always guess C". How well would that work here? It turns out 164 of the answers were option C, which is 20% (as close as we can get, with 822 questions). The most common answer in this dataset was A, at 172; the difference from random is not statistically significant.

Another trick is to avoid the same answer twice in a row. The dataset contains information on which test the question comes from, so we rule out all first questions and look at how many of the remaining have the same correct option as the previous question. The result is 139 of 773 (18%), which may look meaningful, but is not statistically significant.

Judging from personal experience, it seems that the correct answer is often longer than the others. Is this a real effect, or just coincidence and confirmation bias? We try applying the "longest answer" method, and find that it gives 197 correct answers (24%). In this case, the difference from random guessing is statistically significant (p < 5%).

3.4 Combining methods

Since the two lexicons find different words, we can put them together to find a greater number of matches and hopefully a better accuracy. Because Bring has a higher accuracy but lower coverage, it makes sense to start with. Then we apply Saldo to the remaining words. For those words left unidentified by both Bring and Saldo, we use the longest-word method.

Bring, as before, gets 218 correct out of 822, leaving 518. Saldo finds an answer to 321 of those, of which 117 are correct. That leaves 197. We pick the longest option on those, and get 49 correct. This all adds up to 284 correct (47%).

4 Conclusion

When S. C. Bring wrote his thesaurus, he was 88 years old. His work has now surpassed that age, but it is still a very useful resource in tasks like this. Using it, we are not able to find all the words and expressions in the material, but for those where we do, the accuracy is quite encouraging. Saldo, on the other hand, a more modern style of lexicon, is able to find a considerably larger number of words, but with a much lesser accuracy. It seems likely that other, similar lexicons, perhaps with different types of associations, may be able to reach higher accuracies. By combining both, we are able to reach an accuracy of 47%, improving on the previously best known result, using a fast, lightweight, and transparent method.

References

Yvonne Adesam, Aleksandrs Berdicevskis, & Felix Morger. 2020. SwedishGLUE – towards a Swedish test set for evaluating natural language understanding models. Technical report, University of Gothenburg.

Lars Borin, Markus Forsberg, & Lennart Lönngren. 2013. Saldo: a touch of yin to wordnet's yang. *Language resources and evaluation*, 47(4):1191–1211.

Lars Borin, Jens Allwood, & Gerard de Melo. 2014. Bring vs. mtroget: Evaluating automatic thesaurus translation. In *Proceedings of LREC 2014, May 26-31, 2014 Reykjavik, Iceland*. European Language Resources Association.

Faton Rekathati. 2021. The KBLab Blog: Introducing a Swedish sentence transformer. Kungliga biblioteket, August, 23.

Mot en mänskligare maskinöversättning

Robert Östling
Institutionen för lingvistik
Stockholms universitet
robert@ling.su.se

Abstract

Over the lifetime of Lars Borin, machine translation has made a gigantic leap – from simple rule-based systems residing on vacuum tube computers, to the latest zero-shot translation systems. The amount of text data used by modern systems can reach hundreds of billions of words, but is this really necessary? What is the lower limit on training data for a translation system? Here I suggest a simple experiment, entirely without computers, that could go some way towards answering this question.

1 Introduktion

I forntiden, över tre år före Lars födelse, utfördes det första experimentet med automatisk maskinöversättning. På en dator stor som ett rum översattes några väl valda meningar från ryska till engelska, och datorlingvistiken var född.

Det var då. I detta nådens år 65 efter Lars födelse, kan vi mata in en massa text i en dator stor som ett rum, och några gigawattimmar senare få ut ett program som översätter – om vi ber den på rätt sätt. Vissa tolkar detta som datorlingvistikens död, men den frågan får vi återkomma till vid annat tillfälle.

Vad krävs egentligen för att översätta från ett språk till ett annat? Några hundra miljarder ord av språkröra från internet räcker tydligen gott, men nog borde det väl gå att klara sig med betydligt mindre? Den här frågan kan och har utforskats inom olika projekt för maskinöversättning av lågresursspråk, men tolkningen av negativa resultat ställer till problem. Även om ett visst experiment misslyckades med att producera en godtagbar översättning, hur vet vi att det inte finns någon annan algoritm som skulle ha klarat uppgiften?

2 Metod

I det här fallet kan det vara lättare att jobba med människor, allra helst någon lingvistiskt skolad. Metoden är enkel: sätt människan med en parallell text på hennes modersmål samt ett annat språk, och be henne sedan att översätta meningar från en annan källa mellan språken. Om hon trots tid och ansträngningar misslyckas, så beror det förmodligen på att parallelltexten inte är tillräckligt lång och/eller varierad för att uppgiften ska gå att lösa.

Genom en sådan studie skulle vi kunna etablera en mänsklig baslinje för lågresursöversättning, och min gissning är att det vore svårt för en dator att göra speciellt mycket bättre ifrån sig. Olyckligtvis har jag inte tillgång till den mängd uttråkade lingvister som skulle krävas, men vi kan approximera översättning från modersmålet till det okända språket med en enklare metod. Eftersom allt vi vet om målspråket kommer från parallelltexten, är vi begränsade till de ord och konstruktioner som förekommer i texten. Under (det något optimistiska) antagandet att vår försöksperson lyckas korrekt tolka varje språklig detalj i parallelltexten, reduceras uppgiften sedan till att pussla ihop meningar genom att enbart använda de ord och konstruktioner som förekommer i parallelltexten.

Robert Östling. 2022. Mot en mänskligare maskinöversättning. In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors), *Live and Learn – Festschrift in honor of Lars Borin*, pages 171–173. Available under CC BY 4.0

Text	Källa	Konstruktion
de förlorad**e** får**en**	Matt 10:6	nominalfras i pluralis, bestämd form med adjektiv
nyligen ville judarna stena dig	Joh 11:8	lexikon
under **förgångna** släktens tider	Ef 3:5	lexikon
mycket stor glädje	Matt 2:10	modifiering av adjektiv
de fyrtio **åren** i öknen	Apg 7:42	lexikon
ständigt, var dag, voro de	Apg 2:46	lexikon
hört konungens ord	Matt 2:9	lexikon
skall edert **tal** vara	Matt 5:37	lexikon
talen I eder emellan **om att**	Matt 16:8	bisatsinledning för diskussionsämne
det bliver klart **väder**	Matt 16:2	lexikon
förändra de stadgar	Apg 6:14	lexikon
kall**as**	Matt 1:16	passivsuffix
för . . . **har** han förkort**at**	Mark 10:20	verb i perfekt efter adverbial (V2)
kom i skarpt **ordskifte** med dem	Apg 15:2	lexikon
bland sig **utvälja** några män	Apg 15:22	lexikon
de äro blinda **ledare**	Matt 15:14	lexikon
öst**erns länder**	Matt 2:1	lexikon
sydväst och **nord**väst	Apg 27:12	lexikon
var konung **över** Judeen	Luk 1:5	ledare för ett land

Tabell 1: Källor till ord, fraser och konstruktioner i Nya Testamentet.

Vi illustrerar detta med hjälp av den flitigast översatta texten, det Nya Testamentet.[1] Ett verktyg som producerar acceptabla översättningar enbart baserat på denna korta text skulle innebära att maskinöversättning blir möjligt för tusentals nya språk.

Som exempel plockar vi en mening på måfå från dagens tidning:

> *De senaste åren har klimatfrågan dominerat i de nordiska grannländernas valkampanjer.*
> (Dagens Nyheter, 2022-09-01)

Som moderna läsare kan vi förstå och förklara vad den här meningen betyder, så nästa steg är att leta efter ord och konstruktioner som kan användas för att uttrycka detta. Vi väljer i det här fallet Nya Testamentet från 1917 års bibelöversättning som vår parallelltext, så uppgiften är nu att uttrycka betydelsen hos ovanstående mening enbart med hjälp av språkligt material från denna text.

Tabell 1 visar de källor i Nya Testamentet som används. Jag har, utan att leta överdrivet noga, försökt att använda den *första* lämpliga förekomsten av ett visst ord, fras eller grammatisk konstruktion.

Vi börjar med nominalfrasen ”de senaste åren”. Bestämd form, pluralis, modifierat med ett adjektiv. Texten innehåller ”de förlorade fåren” (Matt 10:6), som visar att vi kan använda samma ordföljd och artikel (”de”). Nästa steg är att leta lite lexikalt material. För ”senaste” är det närmaste jag hittar ”förgångna” (Ef 3:5) som modifierar ett substantiv (”förgångna släktens tider”). Adverbet ”nyligen” (Joh 11:8) kan modifiera ”förgångna” för att komma närmare betydelsen hos ”senaste”, och detta görs (som i Matt 2:10) genom att sätta adverbet direkt före adjektivet. Till sist har vi ”åren”, som förekommer i texten. Nu har vi en början på översättningen:

> *De nyligen förgångna åren . . .*

[1] Lars har för övrigt sett till att Gustav Vasas bibelöversättning nu finns i Språkbankens valv. För detta är vi som arbetar med Bibeln som parallellkorpus tacksamma!

Jag kommer inte att gå igenom resten av meningen i lika stor detalj, läsaren kan själv lägga pusslet med hjälp av tabellen. Några kommentarer kan däremot behövas om koncept som var okända under biblisk tid. ”Klimatfrågan” är ett bra exempel. Här kan vi uppenbarligen inte hitta ett passande lexem i Nya Testamentet, men en översättare är naturligtvis fri att parafrasera svåröversatta stycken. Jag har här valt att översätta ”har klimatfrågan dominerat” med ”har vi ständigt hört tal om att väder förändras”. I princip skulle ett maskinöversättningssystem kunna göra samma sak, även om det ställer höga krav på omvärldskunskap och språkgenereringsförmåga.

En annan svår nöt är hur ”valkampanjer” kan uttryckas helt utan den vokabulär som hör till en modern representativ demokrati. Det närmaste jag kunde hitta var det inte helt exakta och något ålderdomliga ”ordskiftet om att utvälja ledare”. Själva ordet ”ledare” förekommer enbart i en mer bokstavlig betydelse av en blind ledare som leder en annan blind ner i en grop (Matt 15:14). Även om den liknelsen ibland kan tyckas passande under valkampanjer så kanske en annan översättare hade föredragit ”konungar” (Matt 10:18) eller ”härskare” (Luk 22:25).

Slutresultatet blir som följer, och jag överlåter till läsaren att bedöma hur väl det förmedlar betydelsen i originalmeningen:

> *De nyligen förgångna åren har vi ständigt hört tal om att väder förändras, i ordskiftet om att utvälja ledare över nordens länder.*

För att återknyta till frågan om vilken storlek på träningsmaterial som krävs, kan vi studera kolumnen **Källa** i Tabell 1. Större delen av materialet förkommer redan i den första boken, Matteusevangeliet (Matt), och nästan allt i något av de fem första böckerna (Matt, Mark, Luk, Joh, Apg) som tillsammans utgör omkring 100 000 ord i den här översättningen. Som jämförelse är det bara 0,2% av textmängden i Europarl-korpusen, vilken i sig får räknas som en parallellkorpus av medelstorlek med nutida mått mätt.

3 Slutsats

Även en relativt kort text innehåller tillräckligt stor vokabulär och språkliga konstruktioner för att, med viss parafrasering, uttrycka meningar från en helt annan tid och domän. Det innebär alltså inga hinder i princip mot att konstruera ett verktyg för maskinöversättning till de omkring 2000 språk som Nya Testamentet finns översatt till. Vi har här antagit tillgång till en komplett och korrekt lingvistisk analys av parallelltexten på målspråket, och nästa steg vore att utforska i vilken utsträckning det går att uppnå i praktiken. Allt som behövs är en lingvist med mycket tid.

Acknowledgments

We are grateful to all the authors of the volume who took time to sign the best type of congratulation cards to Lars – in the form of the articles. Special thanks go to Gerlof Bouma for all his help with preparing this festschrift for publication, and to all people who helped us translate the phrase “Live and learn” into their languages. Thanks to Karin Wenzelberg who has approached the task of creating the cover with professional creativity. We would also like to express appreciation for Shalom Lappin and Bernard Comrie for their help with earlier versions of the manuscript template.

The work has been supported by Nationella språkbanken and HUMINFRA, both funded by the Swedish Research Council (contracts 2017-00626 and 2021-00176). This funding allowed us to perform part of the work on the volume during office hours. Finally, we thank our families who had to tolerate our absence in the evenings and weekends in connection to the “secret work” on the Festschrift.

We did our best to reach out to as many of Lars’ colleagues and friends as possible (and as secretly from Lars as possible), but his network is as endless as his research interests, and we deeply apologize for not being able to include everyone.

GU-ISS, Forskningsrapporter från Institutionen för svenska, flerspråkighet och språkteknologi, är en oregelbundet utkommande serie, som i enkel form möjliggör spridning av institutionens skriftliga produktion. Det främsta syftet med serien är att fungera som en kanal för preliminära texter som kan bearbetas vidare för en slutgiltig publicering. Varje enskild författare ansvarar för sitt bidrag.

GU-ISS, Research reports from the Department of Swedish, Multilingualism, Language Technology is an irregular report series intended as a rapid preliminary publication forum for research results which may later be published in fuller form elsewhere. The sole responsibility for the content and form of each text rests with its author.