

Lexicography between NLP and Linguistics: Aspects of Theory and Practice

Lars Trap-Jensen
Euralex, Ljubljana 2018



DET DANSKE SPROG- OG
LITTERATURSELSKAB

DISCIPLINES

- ▶ General linguistics
 - ▶ Formal linguistics
 - ▶ Functional linguistics
- ▶ NLP
- ▶ Artificial intelligence
- ▶ Computer linguistics
- ▶ Corpus linguistics

OUTLINE OF PRESENTATION

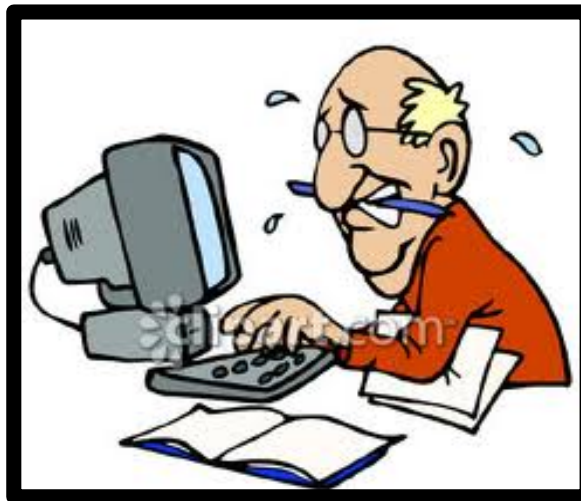
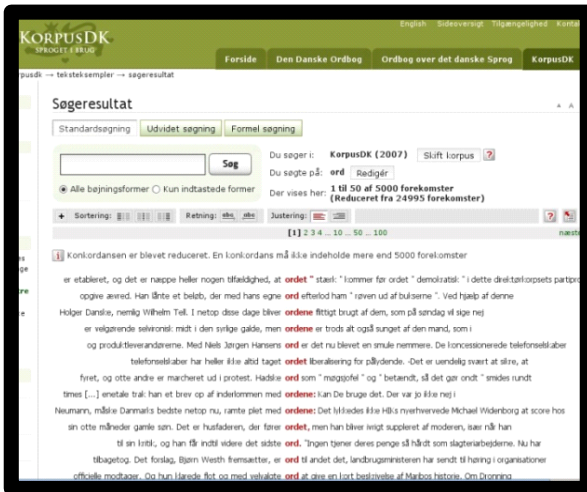
- ▶ The Descriptive Revolution
- ▶ The Corpus Revolution
- ▶ The Digital Revolution
- ▶ What now?

The Descriptive Revolution

1900



2000



THE ACADEMIC PRINCIPLE: Normative, educational

“[I see no room for] ... All coarse, plump and horny words and words that strive against decency .. for they do not need to be known to those who do not pay heed to them, and those who want to learn them will get to know them anyway”

Dictionary of the Royal Danish Society of Sciences and Letters,
J. Langebek 1740

THE ACADEMIC PRINCIPLE: Normative, educational

“... Even the most frequent use of a newly formed word, especially in spoken language, does not yield it any authority, and proves nothing for its usefulness in the pure language and good style, or for its acceptance in a dictionary, if it offends an ear cultivated towards fine language”

Chr. Molbech, 2. edition 1859

THE ACADEMIC PRINCIPLE: Normative, educational

“this vulgar tongue ... is threatening to force its way into the families .. having collected some of what belongs to it, I have, apart from a purely linguistic aim, in addition wanted to draw attention to the danger and tried to provoke resistance against the same and I assume that once people have opened their eyes to the indecent crossing of the line, all educated people will agree to ban the vulgar tongue from good society and leave it to the guttersnipes and the adherents of Grundtvig in whose taste it may fall”

V. Kristiansen: *Contributions to a Dictionary of the Common Language and So-Called Vulgar Tongue*, 1866

τοις παύσις αὐτῶν ἐπιπέσει. Ἐπει δὲ καὶ οὐκ ἔστιν ἡμεῖς,
og for hver Kant beregnes en Skilling.“ Flyveposten
1861, 244, 5.

φίς = ἢεν φιερετ σομ ικκε όερες“, M 2. βεδρε: εν ειנד σομ
σλιππες υδ αφ ενδεταρμεν μεδ εν σαγτε υίσλεν.

λιγε ομ λιγε — φίς φικ φιερενς δατερ.
P. Laale ved Ley 23.

Dat veniam corvis, vexat censura columbas, άν ταγερ
φιδενε παρε, ογ λαδερ φιερενε παρε.
O. Lades Phraser b 3.

δεν, σομ φοερστ βλι'ερ φιδεν παρ, ερ φιδενς φα'ρ.
Alm. Mundheld.

σενσκ: φίς. ενγελσκ: φιδε ογ φονσε ογ φιζζ'λε, Gtose.

φισε = ατ σλιππε εν ειנד υδεν στοει. σενσκ: φισα.

φισερ = εν, δερ φισερ.

Similis simili gaudet, φιδερεν φικ φιερερενς δατερ.
λιγε ομ λιγε.

O. Lades Phraser F.

ενγελσκ: το φιζζ' ογ φιζζ'λε = το μακε α ιδδινγ σοενδ. Webster.

fiske,** i overført Betydn., = fange, faa fat paa, erhverve,
sætte sig i Besiddelse af.

Men hør engang. Du lade Drog,
Som Herre-Trappen visker,

THE DESCRIPTIVE REVOLUTION

“First of all, I cannot ask, “should this or that word be used?”. I ask instead: “is it used or has it been in use?” If so, I will include the word in so far as it falls within the scope of the dictionary”

V. Dahlerup, 1907



THE CORPUS REVOLUTION: Data

- ▶ Brown (1960'erne): 1 million
- ▶ Press65 (1960'erne): 1 million
- ▶ Lancaster/Bergen/Oslo (1970s): 1 million
- ▶ COBUILD (1985): 18 million
- ▶ DDO (1990s): 40 million
- ▶ BNC (1990s): 100 million
- ▶ DeReKo (2017): 42 billion
- ▶ Google Books Corpus (2011): 200+ billion (English)

THE CORPUS REVOLUTION: methods

- ▶ 1960/1970s: concordances
- ▶ MI, Church & Hanks, 1989
- ▶ c. 1990: annotated corpora (lemmatized, POS-tagged)
- ▶ c. 2000: syntactic markup (lexical profiles)
 - ▶ pre-processing: overview

THE CORPUS REVOLUTION: methods

- ▶ Sense division
- ▶ Lemma selection
- ▶ Multi-word expressions: idioms, collocations
- ▶ Valency, morfological/syntactic restriction
- ▶ Neologisms (words, senses, expressions)
- ▶ Domains, language usage
- ▶ Examples, quotations

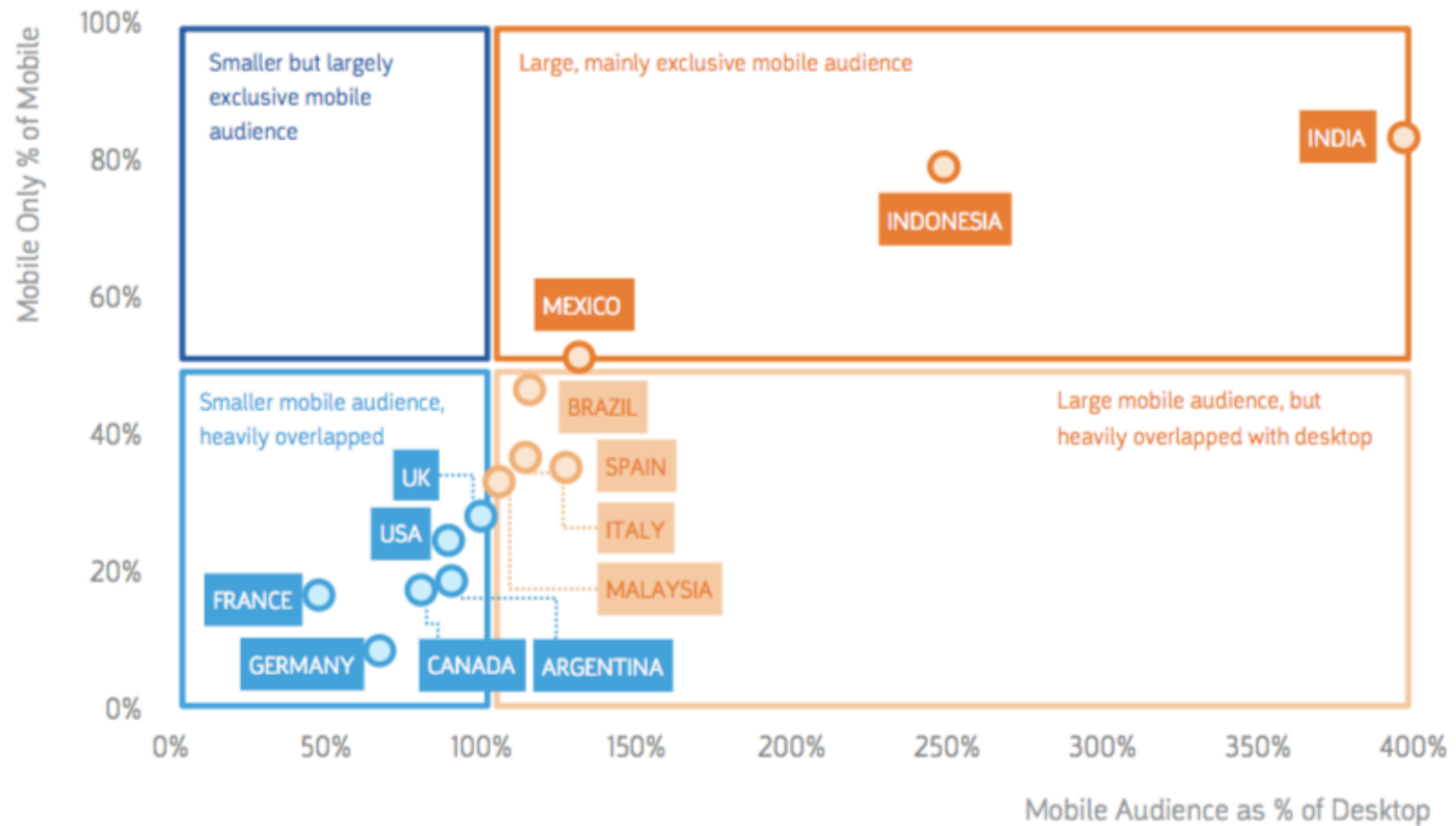
THE CORPUS REVOLUTION: Summary

- ▶ Improves the descriptive paradigm
 - ▶ gradual development, 1900-
 - ▶ better descriptions → better dictionaries
- ▶ Change of technology
 - ▶ from support function to pre-processed patterns
- ▶ Changed role of lexicographer
 - ▶ from all-powerful to gate keeper

THE DIGITAL REVOLUTION

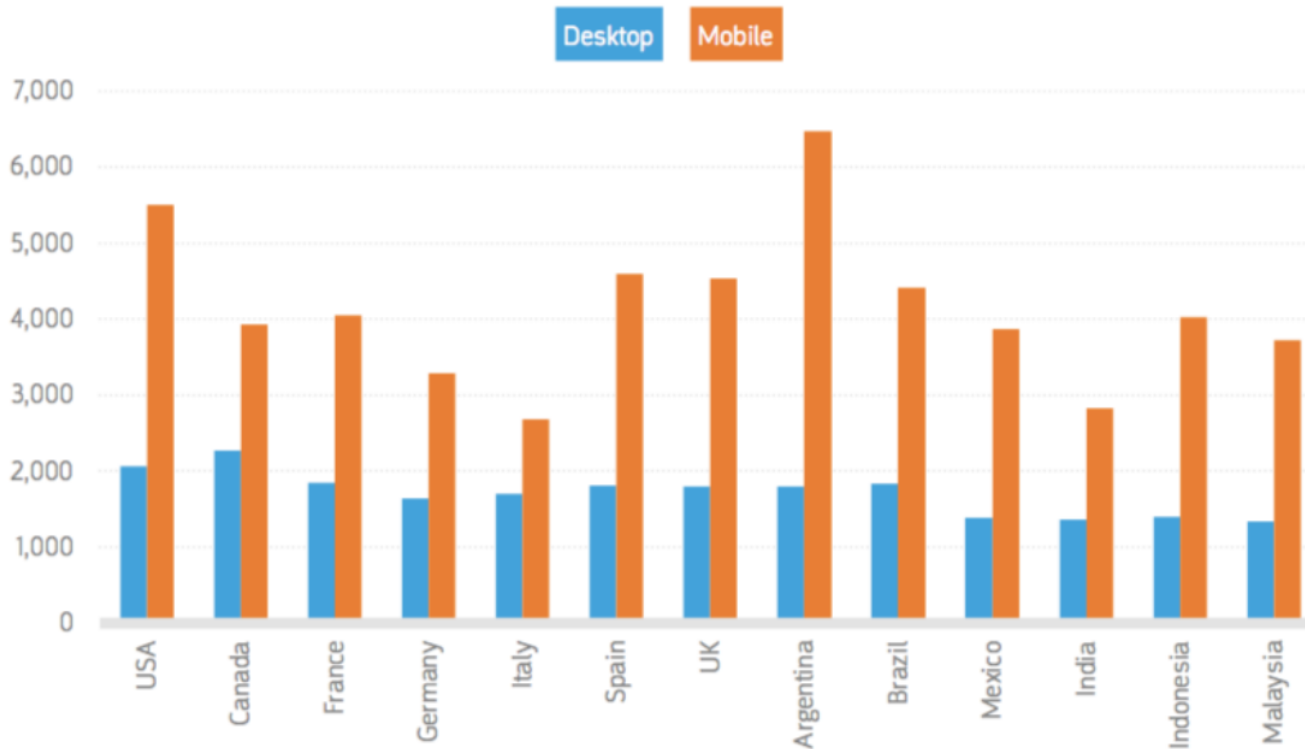
- ▶ 1980s: digitization of SAOB, OED, ..
- ▶ 1990s: CD-ROM, PDA
- ▶ 2000s
 - ▶ Online dictionaries
 - ▶ Smartphones (iPhone: 2007), tablet computers (iPad: 2010)

'Mobile-Firstness' of Markets' Total Digital Populations



Mobile users consume more than 2x minutes vs. desktop users

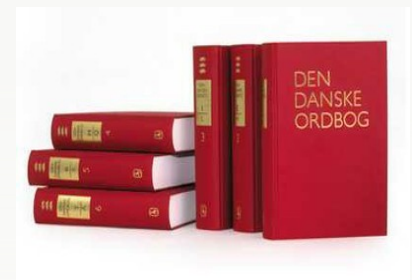
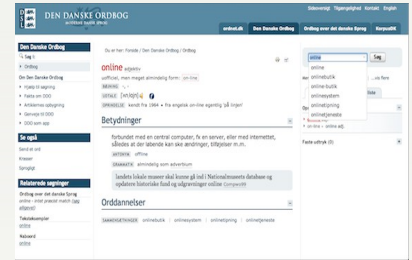
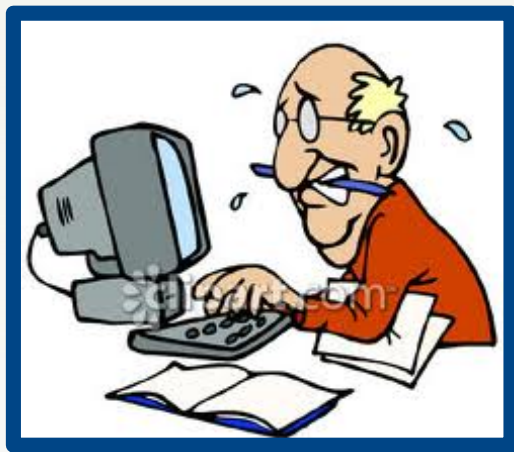
Average Minutes per User by Platform



When looking at each region's desktop users and mobile users separately, mobile users universally consume more digital minutes per person – more than double in the majority of countries.

Argentina continues to deliver the largest number of mobile minutes per user, while Canada has the highest level of per-user desktop consumption.

DIGITAL PUBLICATION



THE DIGITAL REVOLUTION

- ▶ Contents and presentation are separated
- ▶ New information types: audio pronunciation, real sound, video and images
- ▶ User involvement: feedback, crowdsourcing, logfiles
- ▶ Assistance to learners and insecure spellers: "Did you mean ..", auto completion
- ▶ Hyperlinks:
 - ▶ Internal: referrals, show more/less, content
 - ▶ External: access, link and share with others
- ▶ Space economy: abbreviations, tilde, condensed style

THE DIGITAL ERA: What now?

- ▶ Economic crisis:
 - ▶ Users are unwilling to pay for online service
 - ▶ No obvious new business model
- ▶ User crisis?
- ▶ "The biggest problem in lexicography is that lexicographic products are no longer perceived as relevant for the vast majority of people. Most people, in fact, do not use dictionaries, and if they need to find help when communicating or when looking for data, they simply use the Internet instead" (Simonsen 2017: 419)

THE DIGITAL ERA: What now?

Competition from NLP and language technology

- ▶ WordNet, FrameNet, VerbNet, ...
- ▶ BabelNet, TheFreeDictionary, ...

Crowdsourcing, collaborative works

- ▶ Wikipedia, Wiktionary, ...

THE DIGITAL ERA: What now?

Dilemma

- ▶ self-containment
- ▶ co-operation

DEFINITIONS OF 'kind'

(1) behaving in a way that shows you care about other people and want to help them (**Macmillan**)

(2) generous, helpful, and thinking about other people's feelings (**Cambridge English Dictionary**)

(3) caring about others; gentle, friendly and generous (**OALD**)

(4) saying or doing things that show that you care about other people and want to help them or make them happy (**LDOCE**)

DEFINITIONS

(1) saying kind things to someone who has problems and behaving in a way that shows you care about them
(**LDOCE**, *sympathetic*)

(2) kind, helpful, and sympathetic towards other people (**Macmillan**, *caring*)

(3) behaving in a pleasant, kind way towards someone
(**Cambridge**, *friendly*)

(4) (of a person) kind, friendly and sympathetic
(**OALD**, *warm-hearted*)

DEFINITIONS: REGULAR POLYSEMY

hospital, school, office, supermarket

(1) a building or room: *'she went into the office'*

(2) the people working there: *'the hospital decided to close the clinic'*

(3) an institution or business: *'the highest-ranking schools in the country'*

DEFINITIONS: REGULAR POLYSEMY

glass, cup, bottle, bowl, plate

(1) physical object ('a glass', 'a bottle')

(2) its content ('they had two glasses and left')

secondary school - definition and synonyms

Show less



NOUN [COUNTABLE] EDUCATION




Word Forms

 [Using the thesau...](#)



Contribute to our Open Dictionary

a school for children between the ages of 11 and 16 or 18

 Synonyms and related words

Schools: *academy, approved school, boarding school...*

Explore Thesaurus

This is the British English definition of **secondary school**. View American English definition of **secondary school**.

Change your default dictionary to American English.

View the pronunciation for **secondary school**.

From the Blog

What is binge-watching?

Is it 'suffragette' or 'suffragist'?

'Black and white' is a surprisingly colourful expres...

< Semdel SemID="70074696" >

< Semem >

< Restspec >

< Sysfag > geg

< Denbet DanNetSemID="21075434" DanNetSemType="Semem" > person fra

Slovenien

< Genprox > person

< Dok DokStatus="a" >

< Citat >

< txt > Masser af slovenere havde taget opstilling uden for parlamentsbygningen i den slovenske hovedstad

< Kilde >

< DDOkilde > DR-nyh89

< Kildeid ddo-korpus="Mlaj" > 1210044511

CONCLUSIONS: WHAT TO DO?

- ▶ Accessibility
- ▶ Unique identification
- ▶ Data structure
- ▶ Consistency

CONCLUSIONS: WHAT TO DO?

- (a) Build a lexical database that is independent of any particular end product
- (b) Use a standard format to make your data easy to export and modify when exchanged with others.
- (c) Use ID numbers to uniquely identify the central elements of the database, usually the lexical units
- (d) Use elements and attributes in the database that could be useful for NLP purposes: genus proximum, systematic domain assignment, ontological type, super senses, etc.
- (e) Use attributes or elements to indicate position or relation to external NLP resources: WordNet, FrameNet, VerbNet etc.

Thank you for your attention!

