FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Introducing MMDA

An interactive toolkit for corpus-based discourse analysis

Stefan Evert & Philipp Heinrich

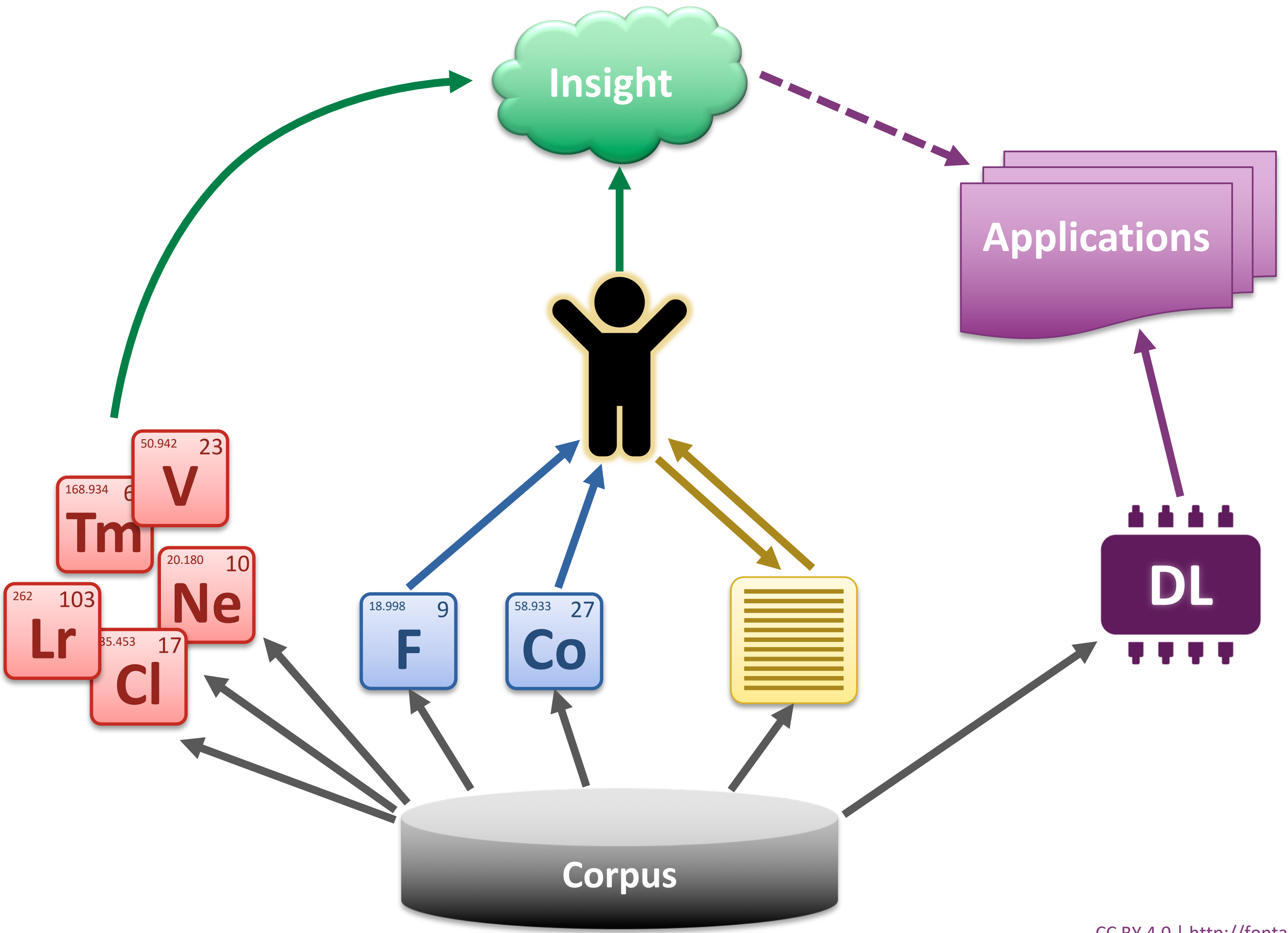
Computational Corpus Linguistics, FAU Erlangen-Nürnberg

www.linguistik.fau.de

EMERGING
FIELDS
INITIATIVE

THE VISION

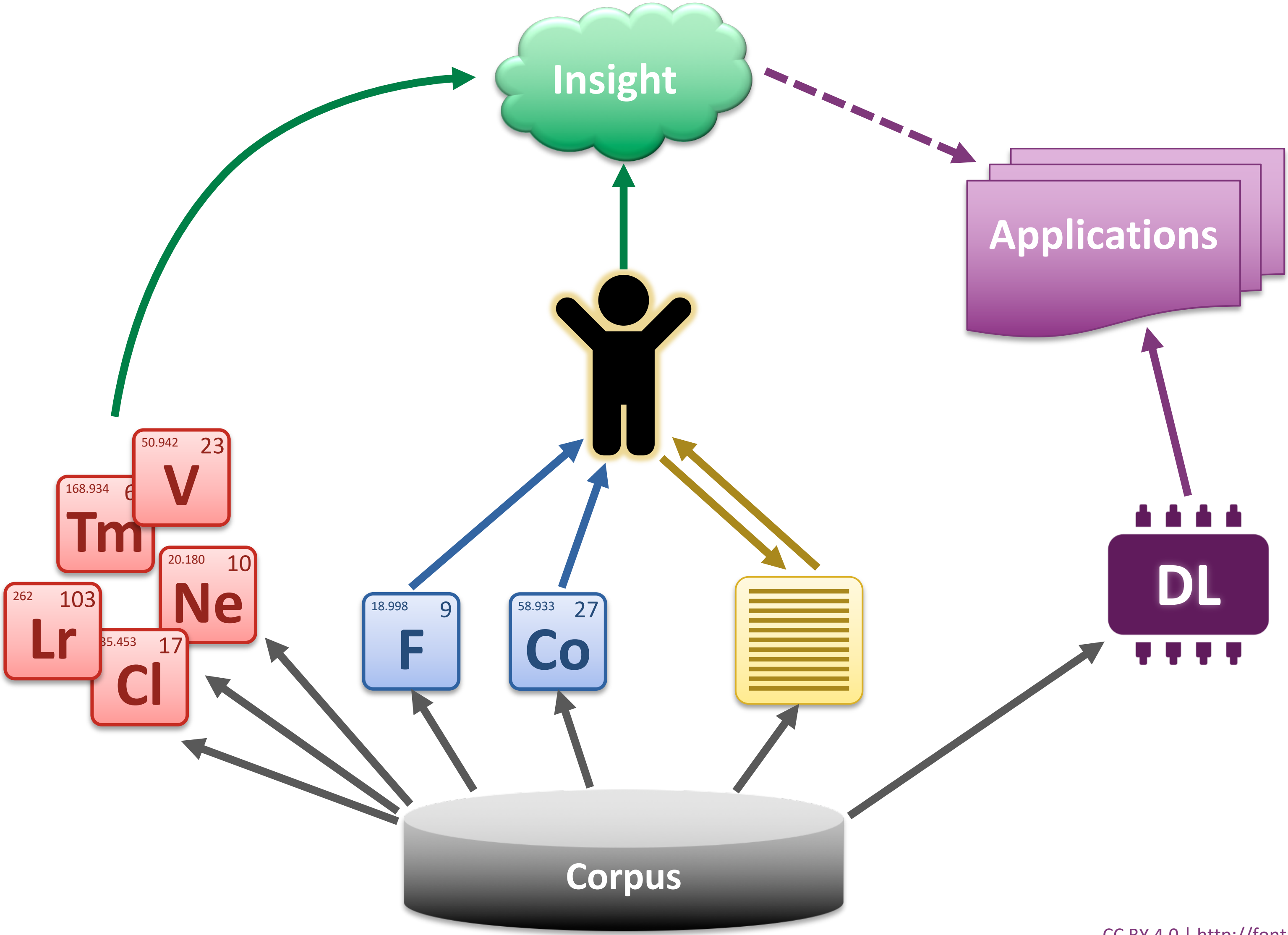
The future of applied CL



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

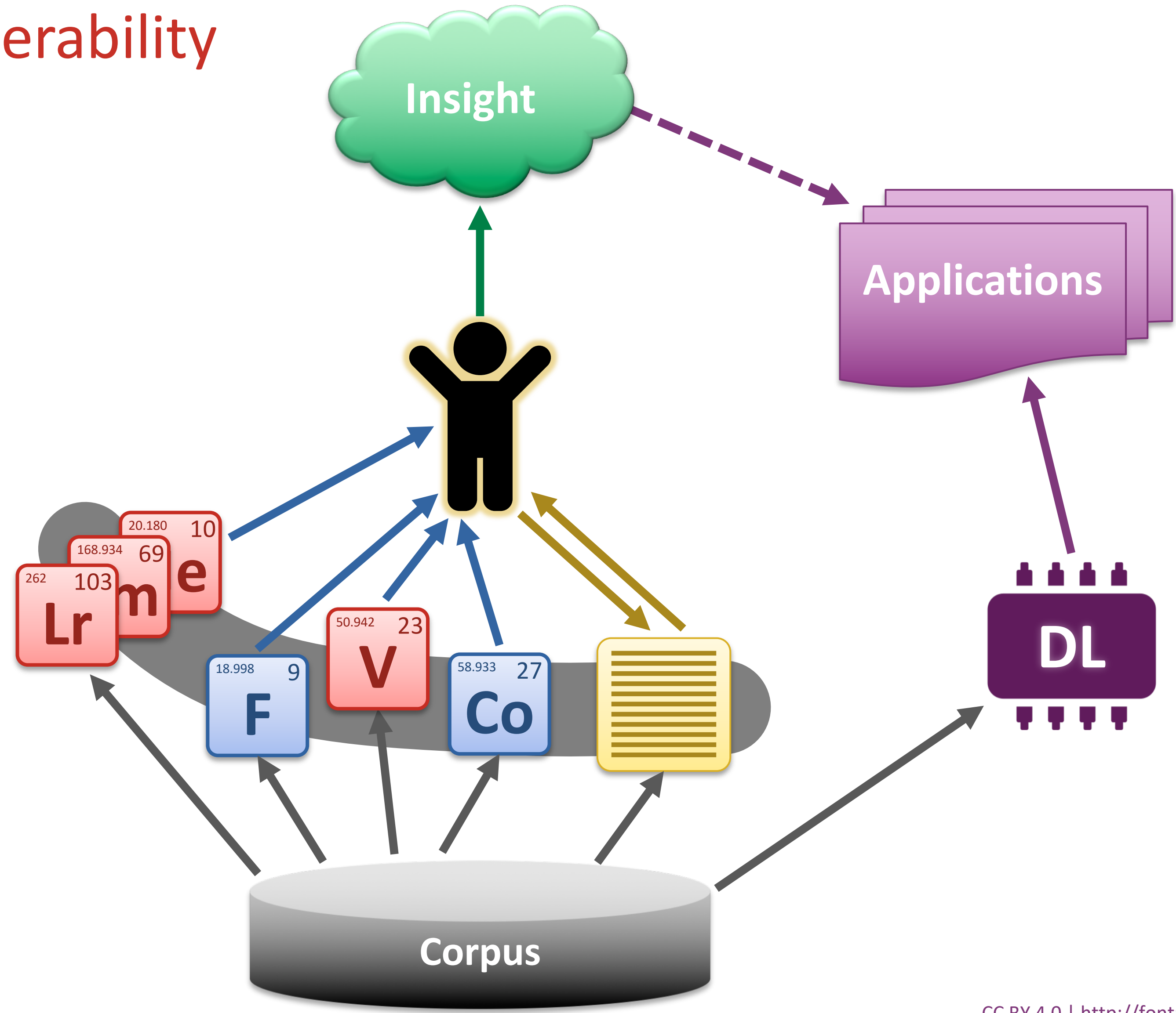
The future of applied CL



FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

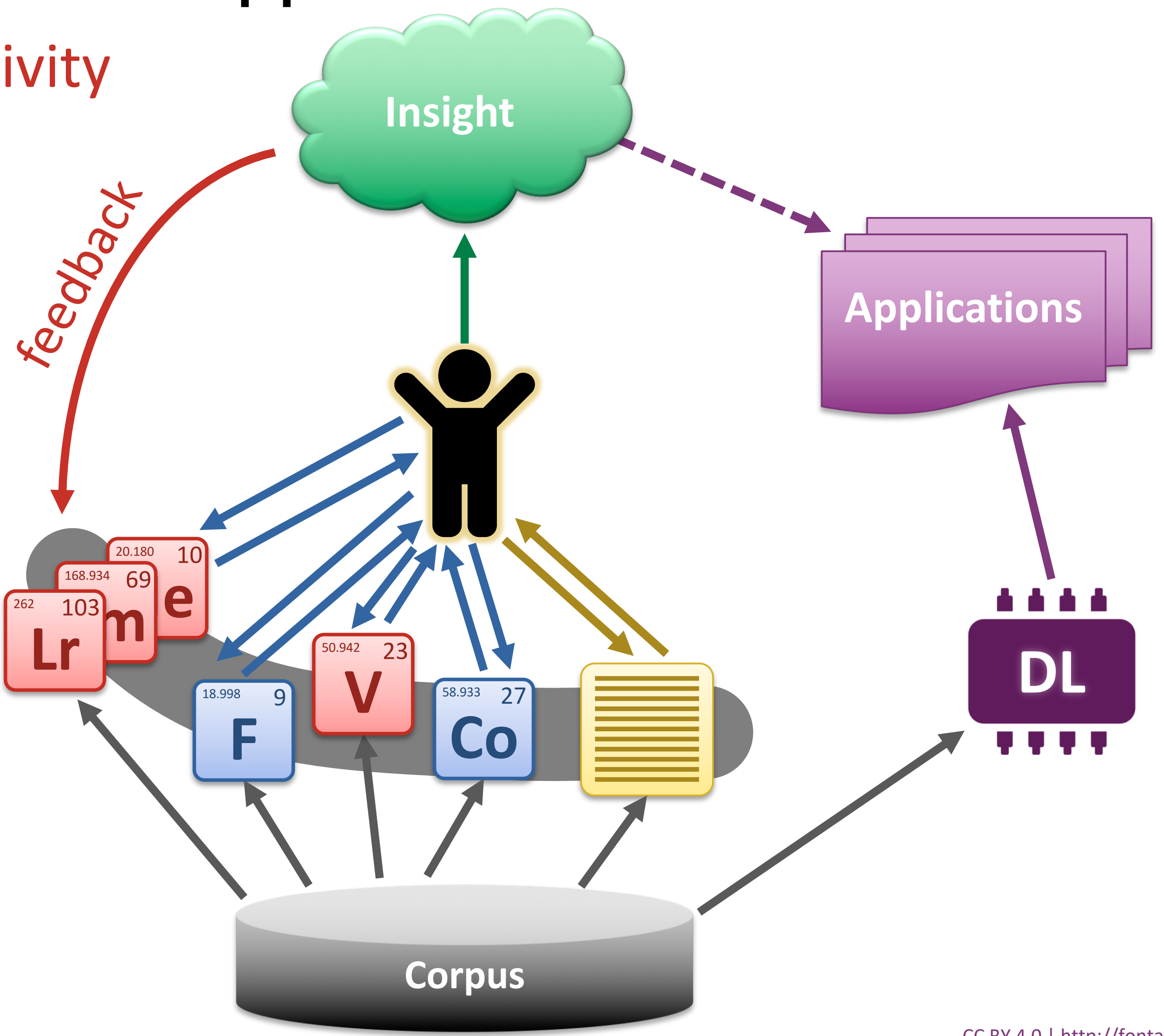
The future of applied CL

1) Interoperability



The future of applied CL

2) Interactivity

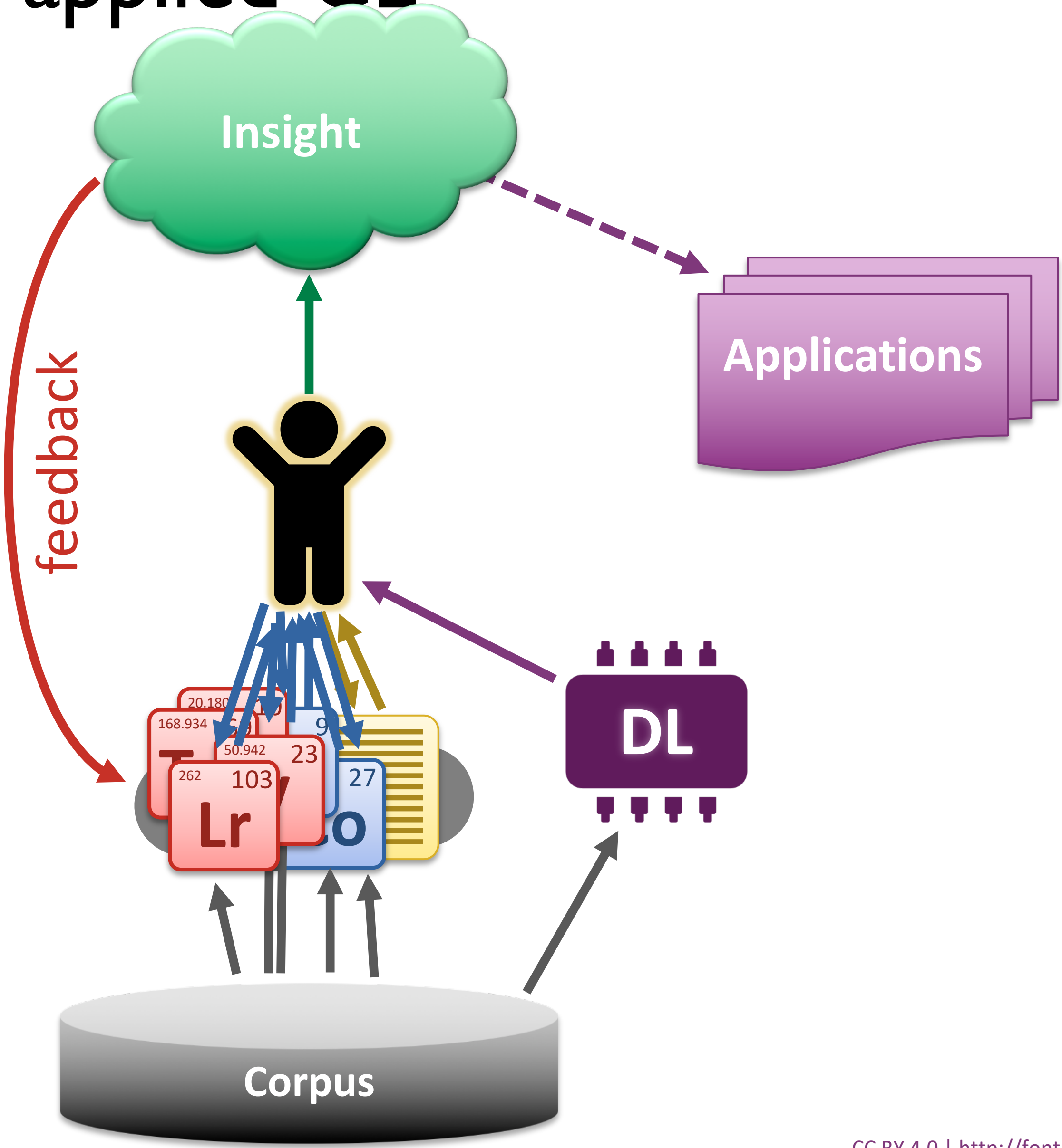


FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

The future of applied CL

3) Integration



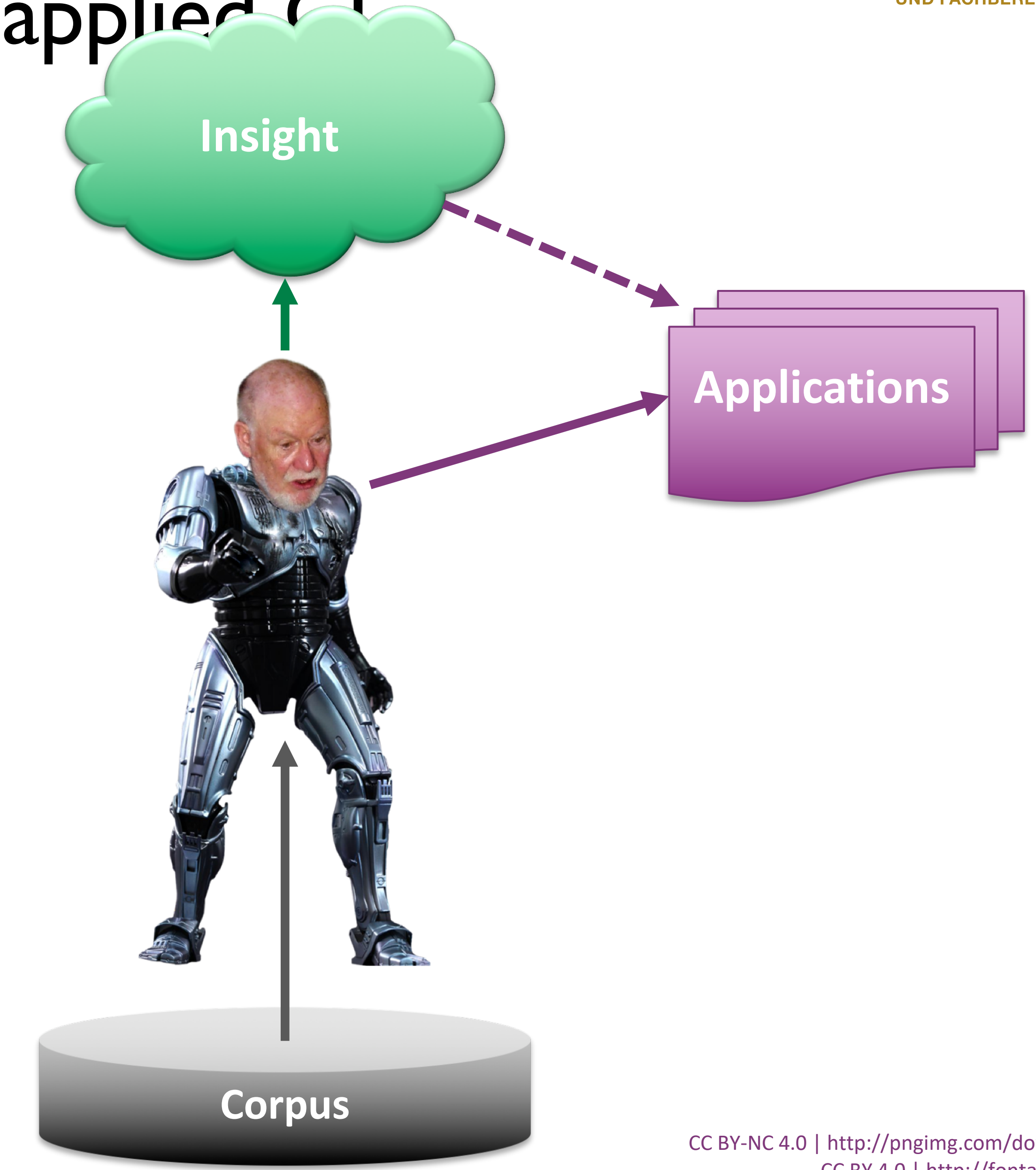
FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

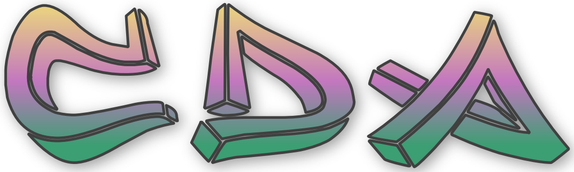
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

The future of applied CL

Hermeneutic

Cyborg

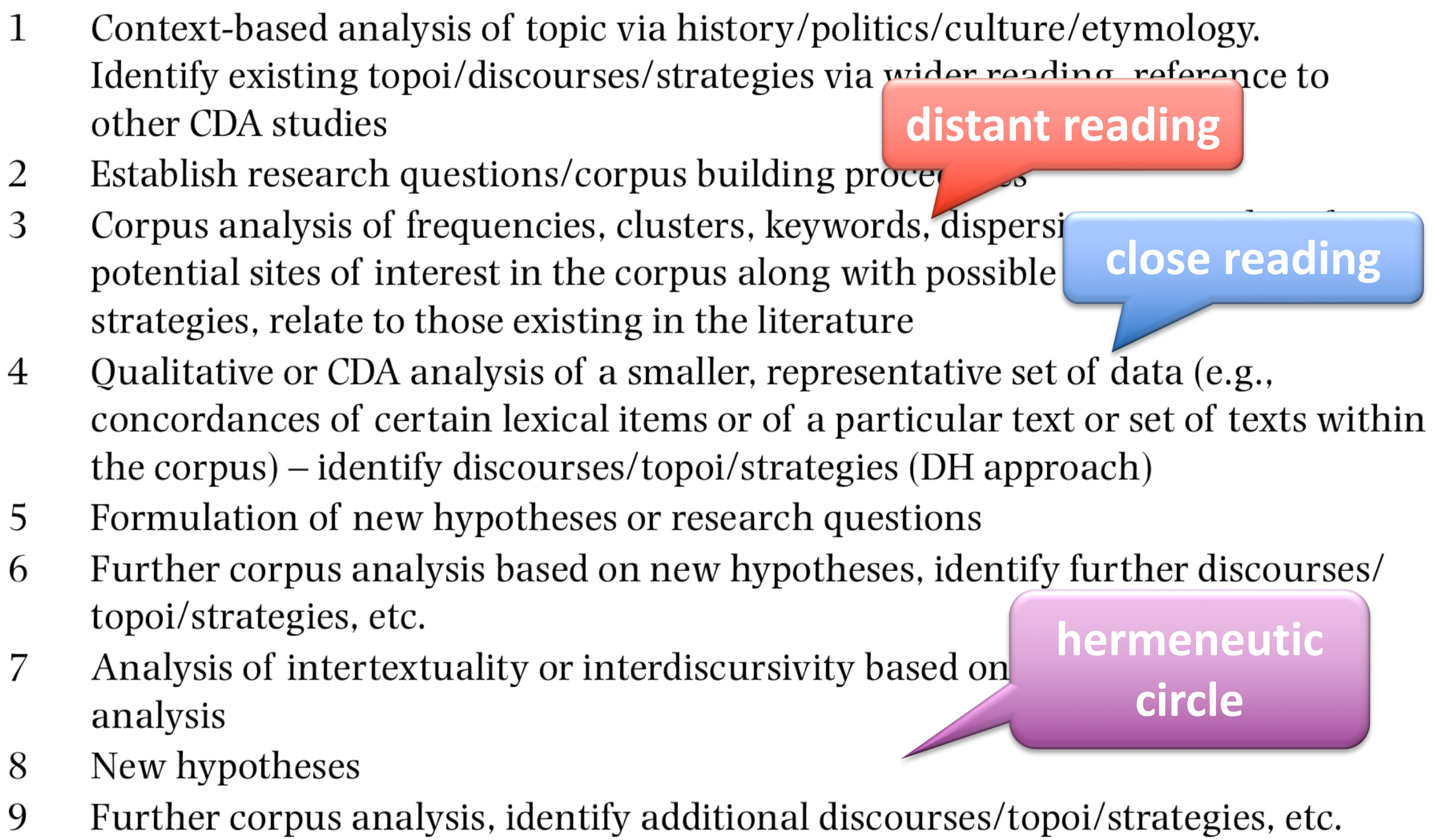



CDA

Corpus-based discourse analysis

- Rooted in critical discourse analysis (Foucault 1969)
  - socio-political discourses = statements in conversation
  - approach: categorization of textual units, but categories not known *a priori* → emerge in hermeneutic process
- CDA (Baker 2006, Mautner 2009) applies process on basis of concordances, collocations and keywords
  - systematic analysis of large corpora possible
  - aims to combine “distant” and “close” reading aspects
  - successful application to refugees (Baker et al. 2008), gender (Baker 2014), climate change (Grundmann & Krishnamurthy 2010), LGBT (Love & Baker 2015), multi-resistant pathogens (Evert/Dykes/Peters 2018)

Corpus-based discourse analysis



1 Context-based analysis of topic via history/politics/culture/etymology. Identify existing topoi/discourses/strategies via wider reading, reference to other CDA studies

2 Establish research questions/corpus building procedures

3 Corpus analysis of frequencies, clusters, keywords, dispersi potential sites of interest in the corpus along with possible strategies, relate to those existing in the literature

4 Qualitative or CDA analysis of a smaller, representative set of data (e.g., concordances of certain lexical items or of a particular text or set of texts within the corpus) – identify discourses/topoi/strategies (DH approach)

5 Formulation of new hypotheses or research questions

6 Further corpus analysis based on new hypotheses, identify further discourses/topoi/strategies, etc.

7 Analysis of intertextuality or interdiscursivity based on analysis

8 New hypotheses

9 Further corpus analysis, identify additional discourses/topoi/strategies, etc.

distant reading

close reading

hermeneutic circle

(Baker et al. 2008, 295, Tab. 7)

Corpus-based discourse analysis

POLITICS SCIENCE ACTION

US-cc:	UK-cc:	US-gw:	UK-gw:	US-ge:	UK-ge:
change	change	warming	warming	greenhouse	greenhouse
CHANGE	CHANGE	WARMING	WARMING	GREENHOUSE	GREENHOUSE
CLIMATE	CLIMATE	GLOBAL	GLOBAL	EFFECT	EFFECT
GLOBAL	GLOBAL	CLIMATE	CLIMATE	GASES	GASES
PANEL	TACKLE	SCIENTISTS	EFFECTS	GAS	GAS
INTERGOVERNMENTAL	LEVY	GREENHOUSE	CHANGE	EMISSIONS	EMISSIONS
WARMING	TACKLING	TREATY	CAUSED	WARMING	CARBON
ENERGY	EFFECTS	CAUSE	THREAT	CARBON	GLOBAL
EFFECTS	IMPACT	EMISSIONS	WORLD	GLOBAL	WARMING
RESEARCH	ENERGY	GASES	EMISSIONS	DIOXIDE	DIOXIDE
INTERNATIONAL	ACTION	POLLUTION	TACKLE	REDUCE	ATMOSPHERE
KYOTO	PANEL	EFFECTS	EFFECT	ATMOSPHERE	CAUSED
ENVIRONMENTAL	WORLD	KYOTO	POLLUTION	SCIENTISTS	REDUCE
ISSUE	THREAT	REDUCE	COMBAT	HEAT	RUNAWAY
REPORT	ISSUES	THREAT	CARBON	CAUSED	CLIMATE
NATIONS	MR	CONTRIBUTE	SCIENTISTS	OZONE	CONTRIBUTE
SCIENTISTS	COMBAT	FIGHT	IMPACT	CLIMATE	OZONE
POLICY	INTERGOVERNMENTAL	TREND	GREENHOUSE	PERCENT	CAUSE
ADDRESS	BILL	ENERGY	US	REDUCING	SCIENTISTS
ISSUES	HELP	ISSUE	FIGHT	RAIN	CHANGE
HUMAN	NEED	BUSH	DUE	CONTRIBUTE	CAUSING
CONFERENCE	PEOPLE	REAL	GASES	KNOWN	WORLD

(Grundmann & Krishnamurthy 2010, 139, Tab. 9)

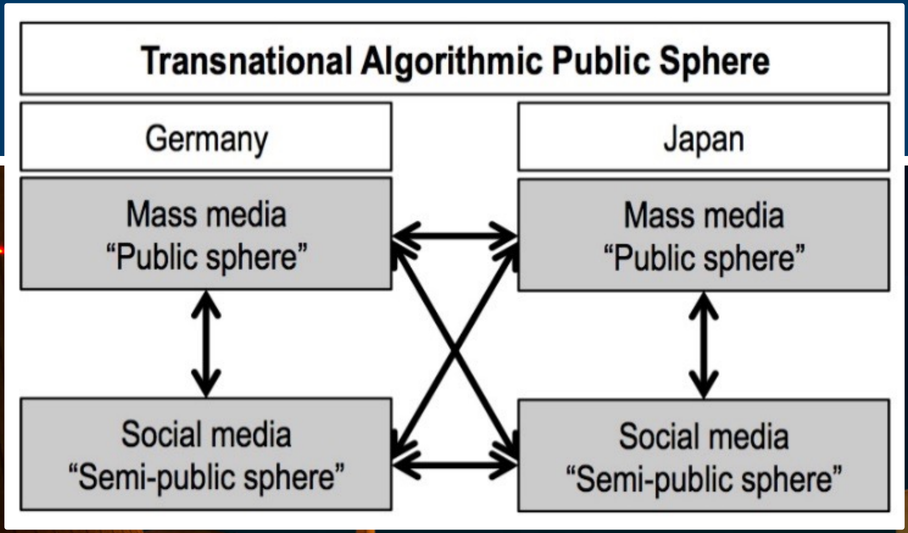
EMERGING FIELDS INITIATIVE

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

EFE – Exploring the Fukushima Effect

www.linguistik.fau.de/projects/efe

Stefan Evert, Fabian Schäfer, Christina Holtz-Bacha, Marc Stamminger



Transnational Algorithmic Public Sphere

Germany

Japan

Mass media
“Public sphere”

Mass media
“Public sphere”

Social media
“Semi-public sphere”

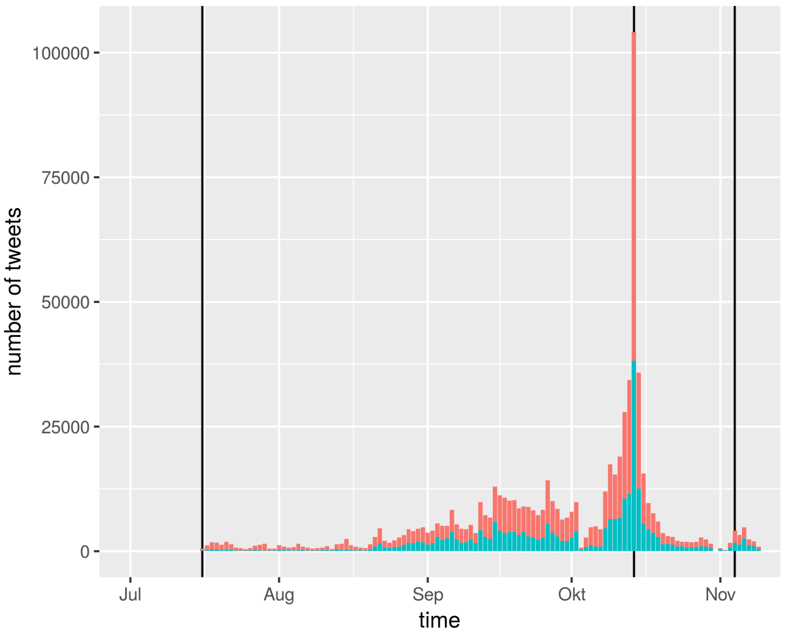
Social media
“Semi-public sphere”

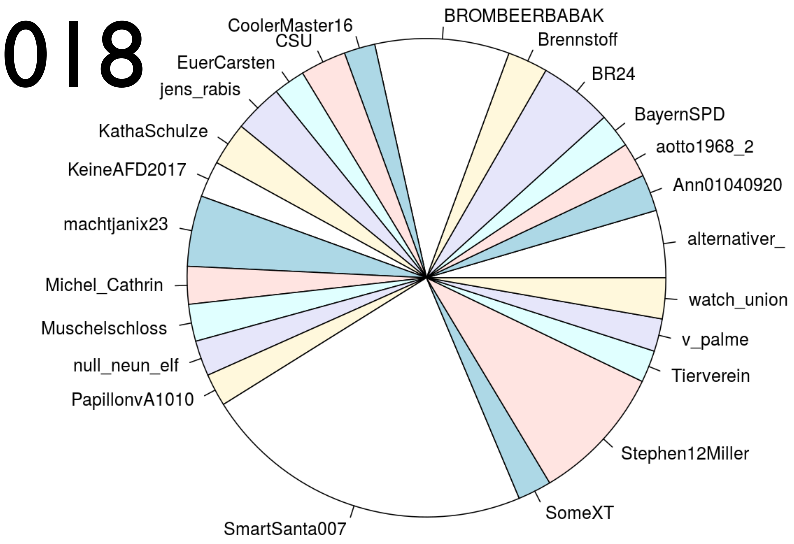
https://www.flickr.com/photos/25058419@N05/5172953200
© 2010 Bjoern Schwarz (CC-by 2.0)

Case study: Bavarian state elections 2018

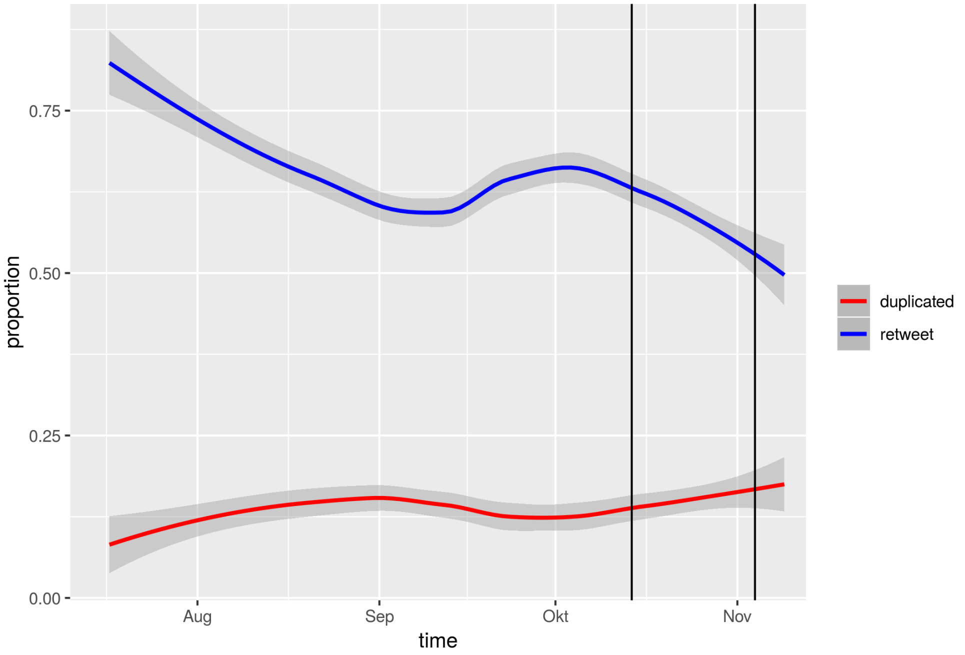
- Interdisciplinary research + teaching project with communication science & political science
  - *Wahlkampfreader 2017* (Holtz-Bacha, ed.)
  - Seminar *Heimatstolz & Vorurteil* (WS 2018/19)
- Twitter corpus: 5.5M tokens
  - collection: 16.07.–16.11.2018, keywords + accounts
  - 213,997 unique tweets (cf. Schäfer et al. 2017)
- Newspaper corpus: 2.6M tokens
  - 4,602 articles from FAZ, SZ, BILD, Münchner, Nürnberger, …
  - Web scraping with similar keywords

Case study:
Bavarian state elections 2018





ratio of retweets and duplicated content



Twitter data set

Case study:
The age of austerity

- Newspaper corpus: 18.1M tokens
  - from LexisNexis: keyword *austerit*, 2010–2016
  - 18,353 unique articles (→ complicated deduplication)
  - sources: Guardian (12,137), Daily Telegraph (6,216)

- Twitter corpus: 4.4M tokens
  - extracted from 10% sample (2008–2015)
  - 243,058 unique geo-localized tweets (UK, US, Greece, India, Ireland, …)

Workshop
Texts and Images of Austerity in Britain.
A Multimodal Multimedia Analysis

25.09.-29.09.2017

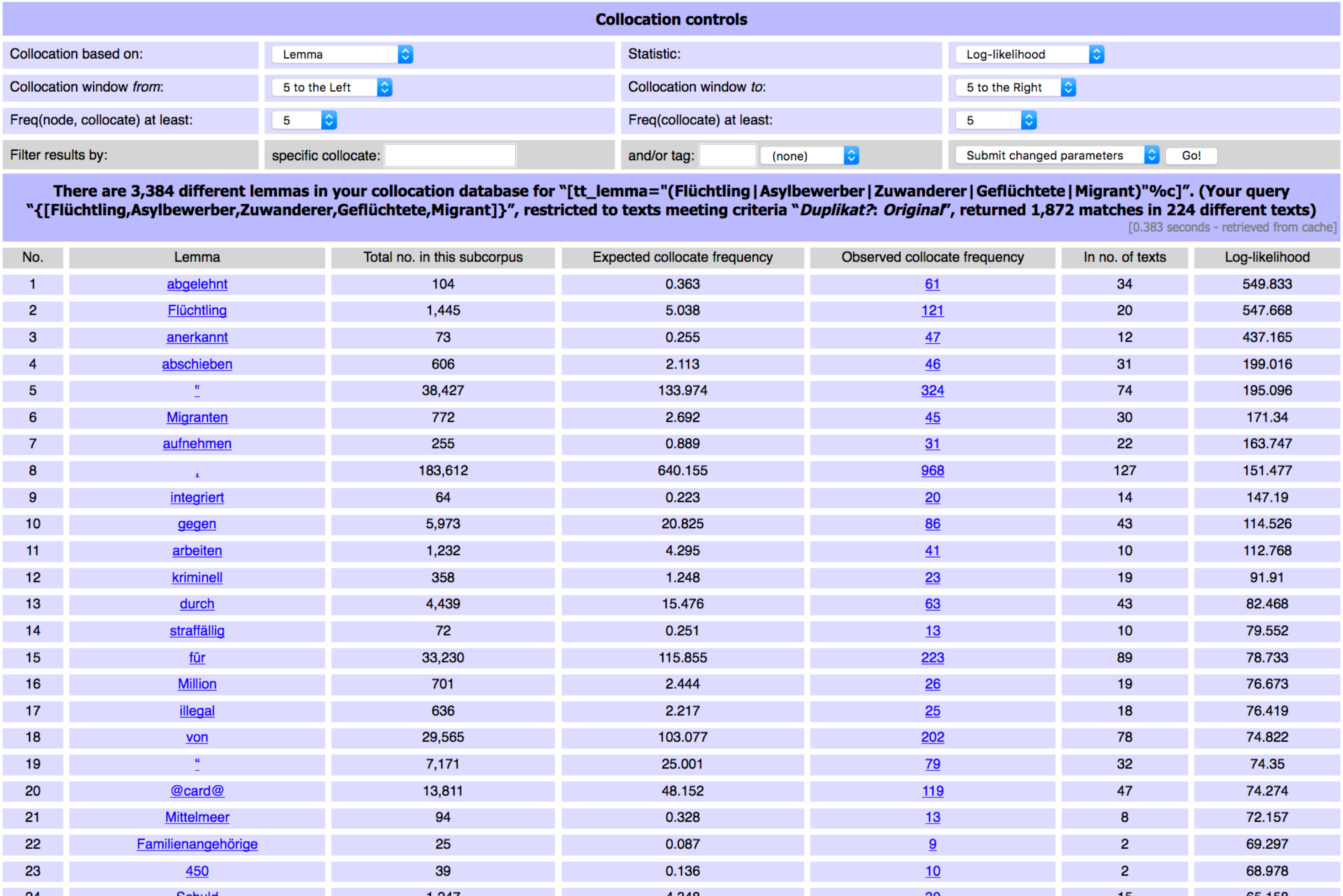
more details

Gefördert von
UNIVERSITÄTSBUND Erlangen-Nürnberg FAU
EMERGING TALENTS INITIATIVE
Visiting Professorship Programme

CDA with CQPweb

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE



Collocation controls

Collocation based on:	Lemma	Statistic:	Log-likelihood
Collocation window *from*:	5 to the Left	Collocation window *to*:	5 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters Go!

There are 3,384 different lemmas in your collocation database for “[tt_lemma="(Flüchtling|Asylbewerber|Zuwanderer|Geflüchtete|Migrant)"%c]”. (Your query “{[Flüchtling,Asylbewerber,Zuwanderer,Geflüchtete,Migrant]}”, restricted to texts meeting criteria “*Duplikat?*: *Original*”, returned 1,872 matches in 224 different texts)

[0.383 seconds - retrieved from cache]

No.	Lemma	Total no. in this subcorpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	abgelehnt	104	0.363	61	34	549.833
2	Flüchtling	1,445	5.038	121	20	547.668
3	anerkannt	73	0.255	47	12	437.165
4	abschieben	606	2.113	46	31	199.016
5	"	38,427	133.974	324	74	195.096
6	Migranten	772	2.692	45	30	171.34
7	aufnehmen	255	0.889	31	22	163.747
8	,	183,612	640.155	968	127	151.477
9	integriert	64	0.223	20	14	147.19
10	gegen	5,973	20.825	86	43	114.526
11	arbeiten	1,232	4.295	41	10	112.768
12	kriminell	358	1.248	23	19	91.91
13	durch	4,439	15.476	63	43	82.468
14	straffällig	72	0.251	13	10	79.552
15	für	33,230	115.855	223	89	78.733
16	Million	701	2.444	26	19	76.673
17	illegal	636	2.217	25	18	76.419
18	von	29,565	103.077	202	78	74.822
19	“	7,171	25.001	79	32	74.35
20	@card@	13,811	48.152	119	47	74.274
21	Mittelmeer	94	0.328	13	8	72.157
22	Familienangehörige	25	0.087	9	2	69.297
23	450	39	0.136	10	2	68.978
24	Schuld	1,247	4.348	30	15	65.158

MMDA

MMDA:
An interactive toolkit for CDA

- State-of-the art quantitative techniques:
  semantic word embeddings, sentiment analysis, …
- Visualization:
  collocations displayed as semantic map
- Interactivity:
  interactive manipulation of algorithmic parameters
- Integration:
  CDA categorization procedure carried out within MMDA toolkit → database of discursive positions

Operationalizing CDA …

- Traditional CDA groups collocates (or keywords) into sets that indicate discursive positions to analyst

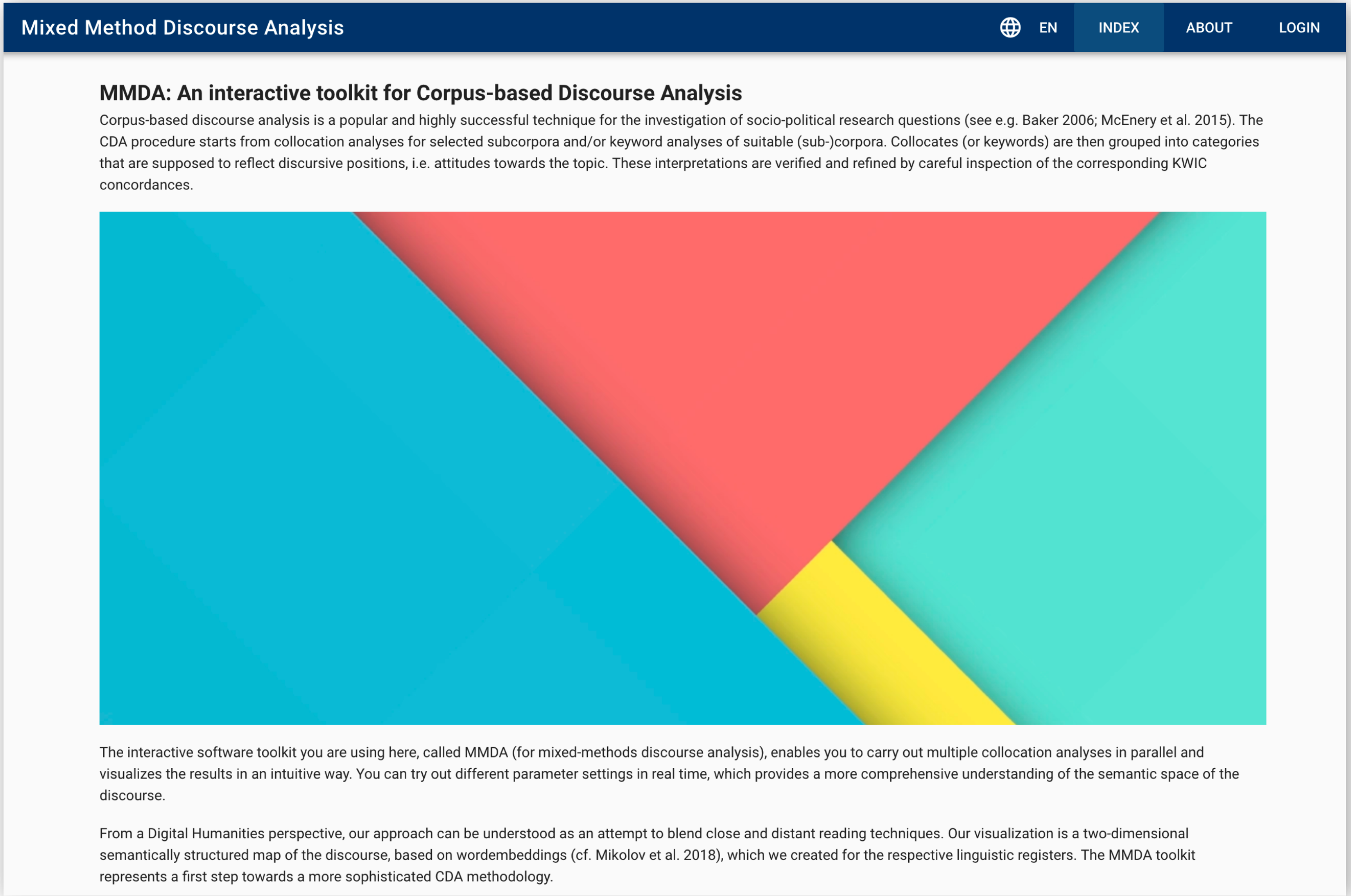
Our operationalization:

- **discourseme** = set of closely related lexical items
  - formed by grouping collocates of given discourse topic
  - topic is also operationalized as a discourseme!
- **discursive position** = constellation of discoursemes
  - minimally: topic discourseme + group of collocates

MMDA architecture

- NLP pipeline
  - German: SoMaJo, SoMeWeTa, TreeTagger / SMOR
  - English: Stanford CoreNLP, TweetNLP, GabLemmatizer
- Word embeddings
  - word2vec (Gensim) on Wikipedia + Twitter data
  - pre-trained FastText embeddings
- Interactive Web application
  - semantic map: PyMagnitude, t-SNE layout
  - corpus indexing & analysis: CWB, UCS toolkit, Pandas
  - frontend: Vue.js | backend: Flask
  - persistent database: sqlalchemy + SQLite

MMDA demo



MMDA: An interactive toolkit for Corpus-based Discourse Analysis

Corpus-based discourse analysis is a popular and highly successful technique for the investigation of socio-political research questions (see e.g. Baker 2006; McEnery et al. 2015). The CDA procedure starts from collocation analyses for selected subcorpora and/or keyword analyses of suitable (sub-)corpora. Collocates (or keywords) are then grouped into categories that are supposed to reflect discursive positions, i.e. attitudes towards the topic. These interpretations are verified and refined by careful inspection of the corresponding KWIC concordances.

The interactive software toolkit you are using here, called MMDA (for mixed-methods discourse analysis), enables you to carry out multiple collocation analyses in parallel and visualizes the results in an intuitive way. You can try out different parameter settings in real time, which provides a more comprehensive understanding of the semantic space of the discourse.

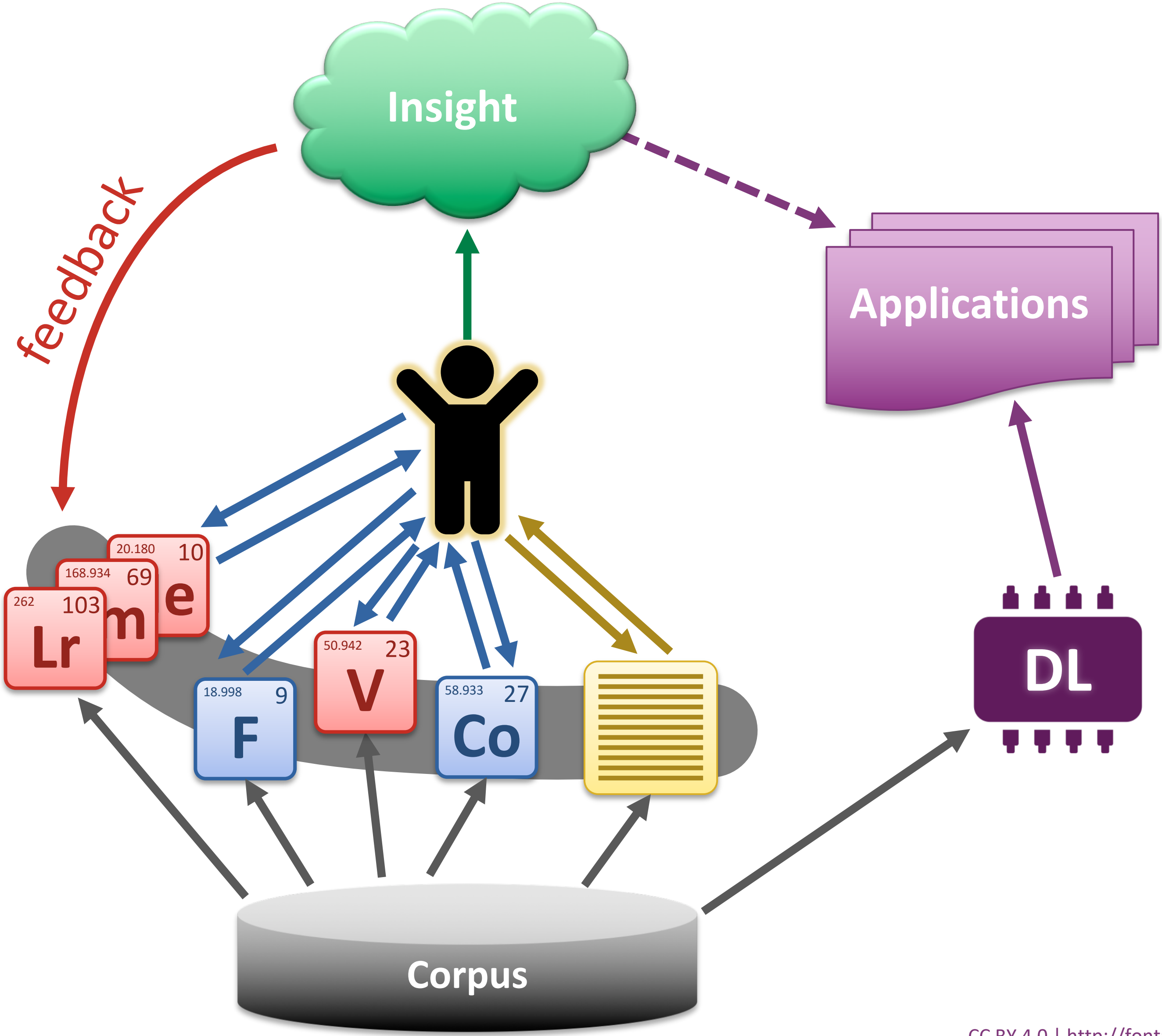
From a Digital Humanities perspective, our approach can be understood as an attempt to blend close and distant reading techniques. Our visualization is a two-dimensional semantically structured map of the discourse, based on wordembeddings (cf. Mikolov et al. 2018), which we created for the respective linguistic registers. The MMDA toolkit represents a first step towards a more sophisticated CDA methodology.

ROADMAP

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

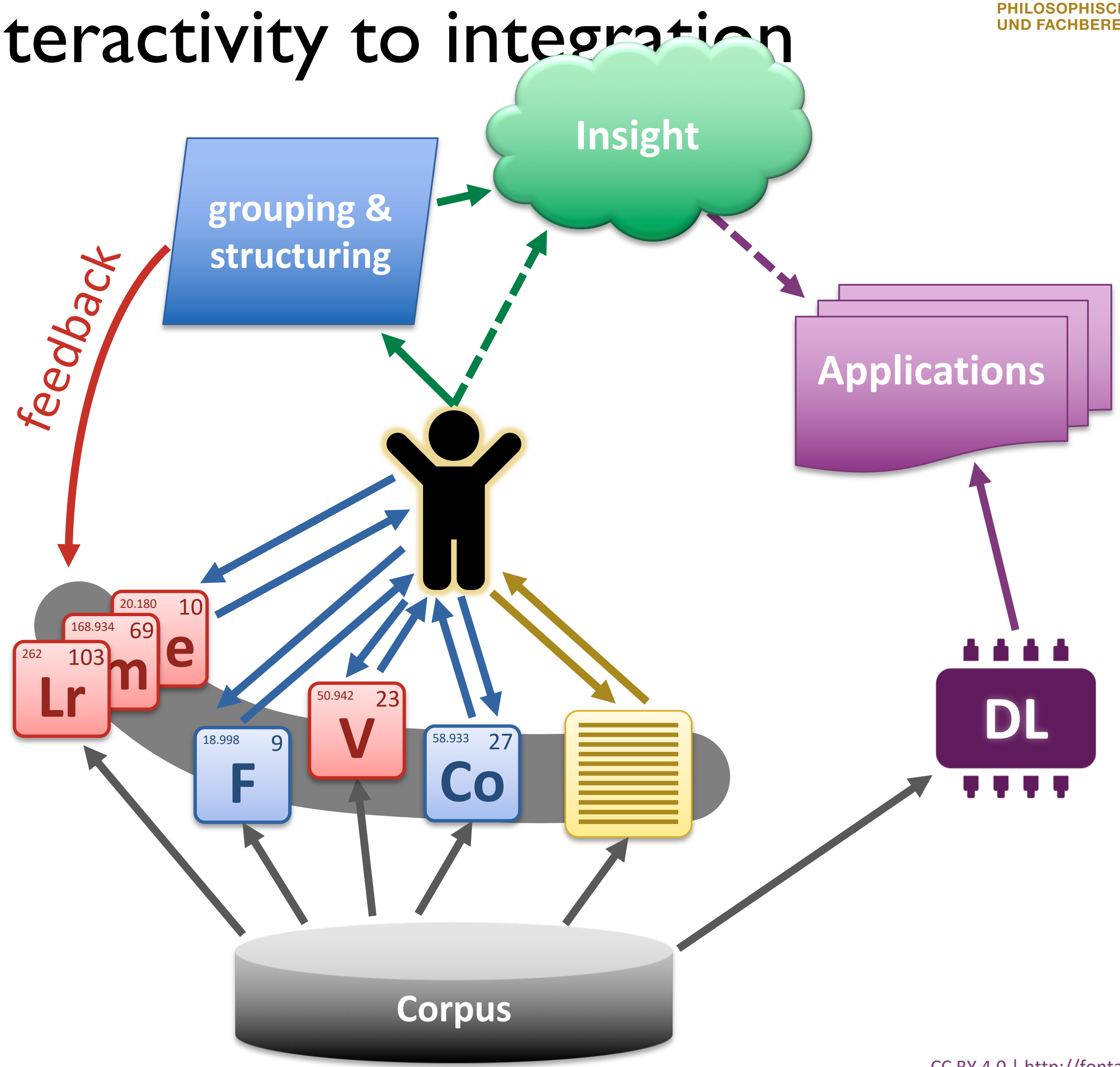
PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

From interactivity to integration



CC BY 4.0 | http://fontawesome.com/

From interactivity to integration



CC BY 4.0 | http://fontawesome.com/

MMDA: next steps

- Testing, improved usability, performance tuning
- Additional quantitative data & visualization
  - sentiment analysis
  - temporal distribution
  - communities in Twitter network
- Lexical items as unit of analysis
  - lemmas, NER, multiword units (EN compounds), …
- Enhanced interactivity & feedback
  - re-organize semantic map by dragging items
  - discoursemes attract semantically similar items and secondary collocations

MMDA: the future

- Contextualizing discoursemes
  - goal: disambiguation of lexical items in discoursemes
  - syntactic function, sentiment of context, metadata, …
  - n-grams and discourseme collocates
  - context-sensitive embeddings
  - mutual disambiguation in constellation (= discursive pos.)
- Automatic detection of discursive positions
  - high accuracy based on contextualized discoursemes
  - allows quantitative analysis of the spread of discursive positions across networks, media and languages
- Multi-corpus / multi-lingual MMDAs

THANK YOU

MMDA toolkit available on-line now

https://geuselambix.phil.uni-erlangen.de/

(ask us for a demo account)