Corpus-based extraction of good example sentences with a high range of variation

Alexander Geyken, **Ulf Hamster**, Lothar Lemnitzer, Gregor Middell, Ji-Ung Lee**, Iryna Gurevych**

Berlin-Brandenburgische Akademie der Wissenschaften
**Technische Universität Darmstadt

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Outline

1. Introduction
2. Example **variation**
    - Motivation
    - Problem Statement
    - Adopted Solution
3. Train a **scoring** model
    - Interactive Learning approach
    - Best-Worst Scaling
4. Demo
    - Demo: train a scoring model
    - Demo: example variation

1. Introduction

Background

DWDS is a one-stop dictionary for contemporary German (spelling, grammar, etym, definition+ corpus examples, thesaurus)

DWDS consists to a large extent of legacy resources that are outdated or do not contain corpus examples at all.

150.000+ entries (WDG, Duden-99) have to be revised, including with corpus examples.

Corpus-base of dwds is very large (50b tokens: www.dwds.de/r)

- WDG: Wörterbuch der dt. Gegenwartssprache (1961-1977)
- Duden-99: Großes Wörterbuch der deutschen Sprache (1999)

DWDS www.dwds.de

Motivation

Starting point GDEX: “Automatically finding good dictionary examples in a corpus”. (Kilgarriff et al. 2008), Gute-Belege-Extraktor (Didakowski et al. 2012, adaptation for German)

Additional ideas to GDEX

- look vor variation of example sentences (in order to cover the different meanings)
- In addition to GDEX experiment with individualized model training by Active Learning approach

The implementation of both aspects dealt with in EVIDENCE project, joint initiative between BBAW and TU-Darmstadt

1. Example variation

Problem analysis

Observations

- Search results contain (near) duplicates
- ... refer to the meaning, event, linguistic concept (*semantics*)
- ... have similar sentence grammar (*syntax*)
- ... come from the same *sources* (author, book editions, time spans)

Causes

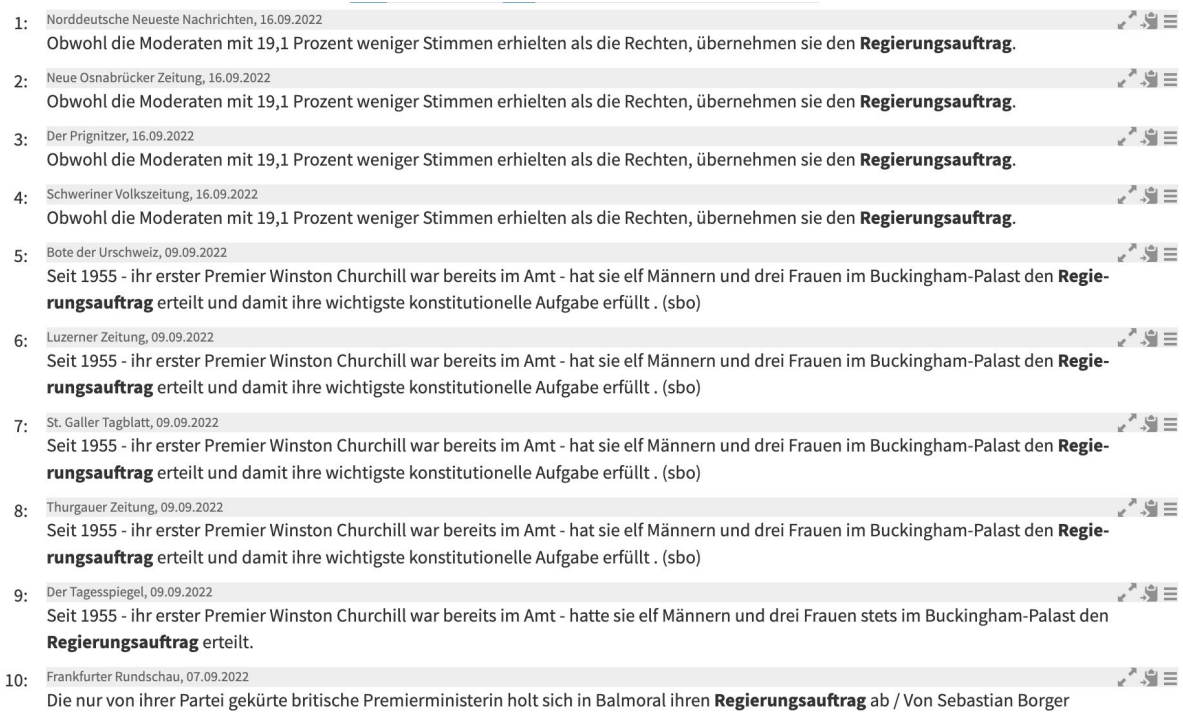
- Search results with an *equally high score* can refer to the same set of *evaluation criteria*.
- Scoring models evaluate single sentences *independently of each other.*

Example: Duplicates for the lemma “Regierungsauftrag”

Exact duplicates in multiple syndicated newspaper pages

Preferring different *sources* would not be sufficient.

Other examples:
- Clickbait Repostings,
- Errata in web articles,
- Book reprints



Korpustreffer für »Regierungsauftrag«, aus dem Korpus DWDS-Zeitungskorpus (ab 1945) des Digitalen Wörterbuchs der deutschen Sprache, Zeitraum 2021-2022.

Solution 1/4 - The MECE-Principle

MECE = "mutually exclusive and collectively exhaustive" (set theory)

“mutually exclusive” or disjunct

- The set of all selected sentence examples should have no or *minimal* ***intersections*** in terms of similarity in meaning (and grammar, and other criteria).

“collectively exhaustive”

- The ***union*** of all selected sentences should (ideally) *cover the entire range of* meanings (and grammars, and other criteria).

Solution 2/4
Quadratic Optimization as Search Filter

To sort the search results, determine weights w_i by maximizing the goodness score g_i of a sentence and minimizing the aggregated similarity matrix Q_{ij} between all sentence examples.

$$\min_{w_1,\ldots,w_N} \; -\lambda \cdot \underbrace{\sum_{i=1}^{N} w_i g_i}_{\text{total goodness}} + (1-\lambda) \underbrace{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} w_i Q_{i,j} w_j}}_{\text{total similarity}}$$

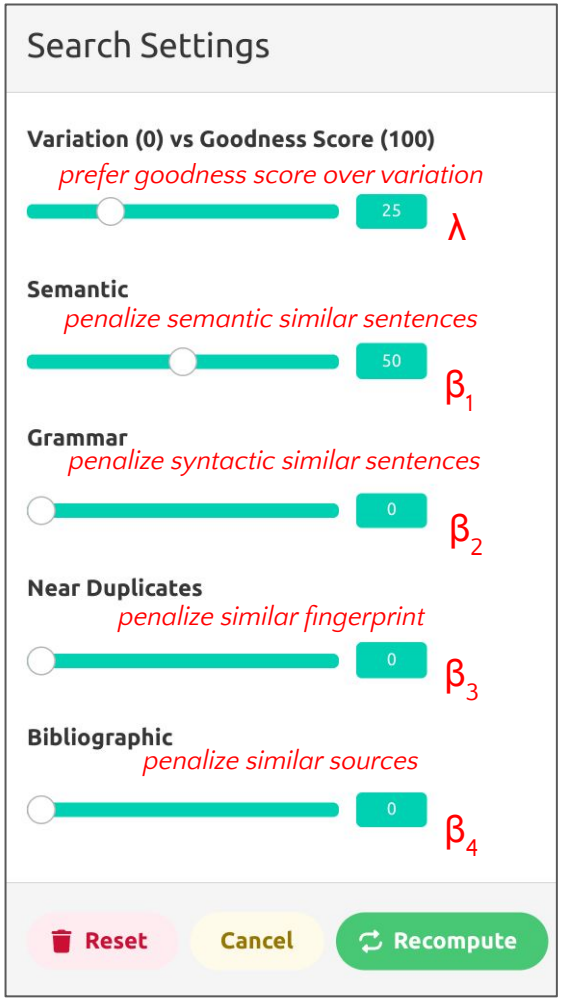
$$\text{s.t.} \quad \sum_{i=1}^{N} w_i = 1$$

$$w_i \geq 0 \quad \forall i$$

$$w_i \leq b \quad \forall i$$

$$Q_{i,j} = \sum_{k}^{M} \beta_k D_{i,j}^{(k)} \quad \forall i,j$$

e.g., $D_{ij}^{(semantic)}$ could be the cosine-similarities based on SBert representations



Implementation details:
Reverse Automatic Differentiation as approximation for quadratic optimization problems

$$\mathcal{L} = -\lambda \cdot \sum_{i=1}^{N} u_i g_i + (1-\lambda) \cdot \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} u_i Q_{i,j} u_j} + \alpha_1 \cdot \left(1 - \sum_{i=1}^{N} w_i\right)^2 + \alpha_2 \cdot \sum_{i=1}^{N} -\min(0, w_i) + \alpha_3 \cdot \sum_{i=1}^{N} -\min(0, b - w_i)$$

$$v_i = w_i - \min(0, w) \; \forall i$$

$$u_i = \frac{v_i}{\max(1e^8, \sum_{j=1}^{N} v_j)}$$

Why? Old SQP-Solver can process only few examples. Big-O complexity!

PyTorch and Tensorflow are “Reverse Automatic Differentiation” libraries.

We can reformulate an optimization problem as loss function.

The optimization constraints become regulariztion penalties.

Quadratic Optimization with Tensorflow in Python
https://github.com/satzbeleg/keras-quadopt/blob/main/keras_quadopt/problem.py

Refactored Code in TFJS
https://github.com/satzbeleg/evidence-app/blob/main/src/components/variation/quadopt.js

tl;dr
=> Old slow numerical optimization problems
can be reformulated and solved with faster modern software

Solution 3/4 - Similarity Metrics

We are using off-the-shelf algorithms, e.g., SBert (LM), Datasketch (MinHash)

What does ***semantic similarity*** mean here?

- for each sentence example, the representation vector is computed with SBert (Contextual Sentence Embeddings),
- and the cosine-similarity $\boldsymbol{D_{ij}}$ for each pair of sentence examples i and j is computed.

How is ***syntactic similarity*** determined?

- for each sentence example, the dependency grammar tree is computed with trankit.
- the dependency tree is decomposed into partial trees (treesimi pypi package).
- partial trees are serialized for MinHash (datasketch pypi package).
- the jaccard similarity $\boldsymbol{D_{ij}}$ for each pair of MinHashes i and j is computed.

And how are ***(near) duplicates*** and ***similar citations*** detected?

- decompose the texts into shingles (kshingle pypi package).
- use shingles to generate a MinHash, and compute jaccard similarities.

Solution 4/4 - Goodness Scores

How is the goodness score g_i of the i-th sentence example computed?

- In our case, we implemented an Interactive Learning model
  - Users rank with Best-Worst Scaling UI to produce training scores.
  - An individualized TensorFlowJS-Modell is trained in the App/Browser directly.
  - The local TFJS-model predicts g_i
- In general, **any other Scoring-Modell can be deployed** variation search filter
  - For example, precompute scores g_i with GDEX (Kilgarriff et al. 2008), and retrieve from backend.

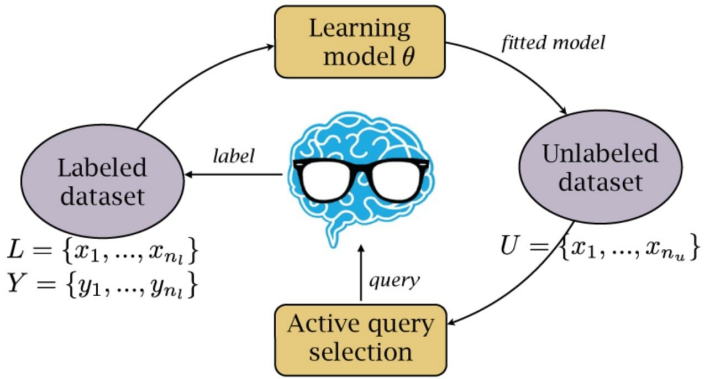
Train a scoring model with Interactive Learning

Interactive Learning approach

Synonym [ml.]: Human-in-the-loop AI, Cooperative Learning, Interactive Learning

Type of ML: “Semi-Supervised Learning” with “Incomplete Supervision” (Zhou, 2018)

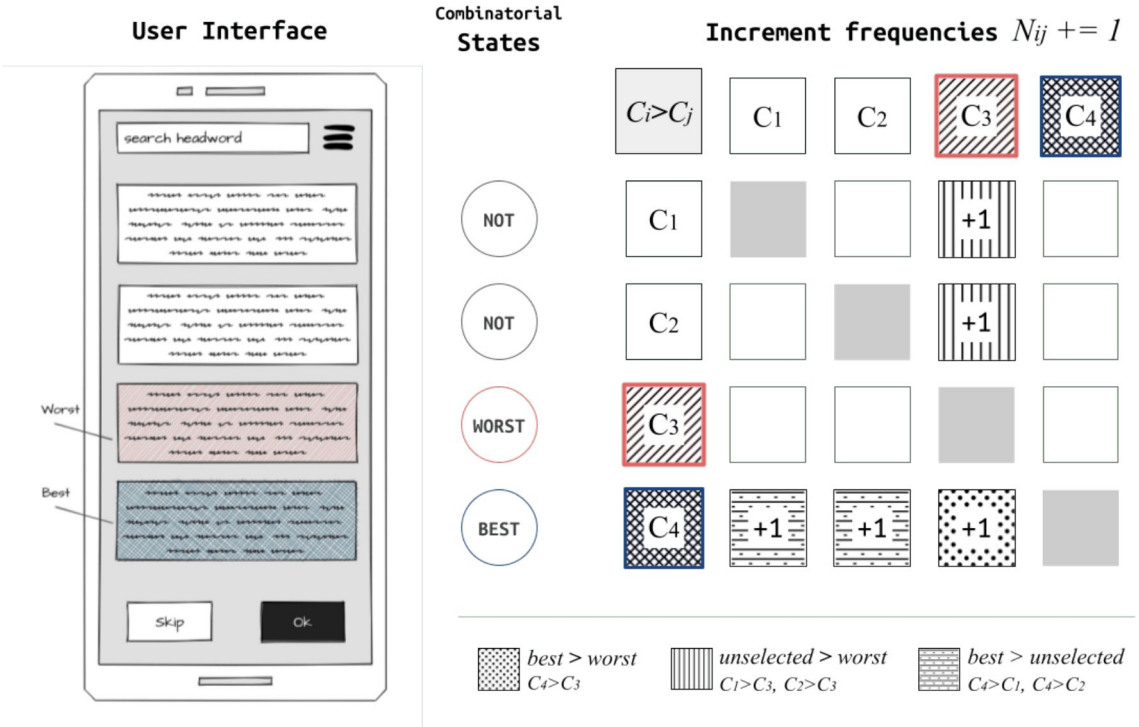
1. User annotates (**few**!) examples
2. (Warm-Start) Model retraining (few! epochs)
3. Model predicts scores for step 4.
4. Sampling of new examples for step 1.



Source: Z.-H. Zhou, “A brief introduction to weakly supervised learning,” National Science Review, vol. 5, no. 1, pp. 44–53, Jan. 2018, doi: 10.1093/nsr/nwx106

Diagram: https://images.deepai.org/glossary-terms/46a6b355896c490cae75d2a0a15d4f65/active-learning.jpg

Extract pairs comparisons with Best-Worst-Scaling (BWS)



2 clicks (worst, best) result in

5 pair comparisons (1x direct, 4x implicit)

N_{ij} is a big & growing, sparse matrix

Derive training scores s_i from N_{ij}

Why?
Cold Start problem
Generate training data fast

Demo how to count frequencies: https://github.com/satzbeleg/bwsample/blob/main/docs/count.ipynb
Algorithms to derive scores from paired comparisons: https://doi.org/10.31219/osf.io/ev7fw

Demo

Demo: Train a scoring model

Corpus: approx 1 Mrd. tokens (SZ, Bild, NOZ, political speeches)

1. Login: https://evidence.bbaw.de/#/auth/login
2. Go to “Settings”: https://evidence.bbaw.de/#/settings
   a. Set “Sampling Sentences from Pool” to “semantic-similar” (=> reduce pool size!)
   b. Set “Maximum Pool Size” to 100 - 150
   c. Set “Re-Train patiences” to 5
3. Go to “Ranking”: https://evidence.bbaw.de/#/bestworst4
4. Search for a lemma
5. Start ranking BWS sets
6. Observe “Training Loss” (lower-right corner of the screen with red/yell/green)
7. Click on “Rankings” button (upper-left corner of the screen)

Demo: Search for varied examples

1. Go to “Variation”: https://evidence.bbaw.de/#/variation2
2. Search for a lemma
3. Case 1: Disable similarity penalties (Sort only by Goodness Score)
4. Case 2: Enable “semantic” penalty

Quellen - Literatur

[1] Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P., 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. Presented at the Proceedings of the 13th EURALEX International Congress, pp. 425–432.

[2] Didakowski, J., Lemnitzer, L., Geyken, A., 2012. Automatic example sentence extraction for a contemporary German dictionary, in: Proceedings of the 15th EURALEX International Congress. Presented at the EURALEX 2012, Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo,Norway, pp. 343–349.

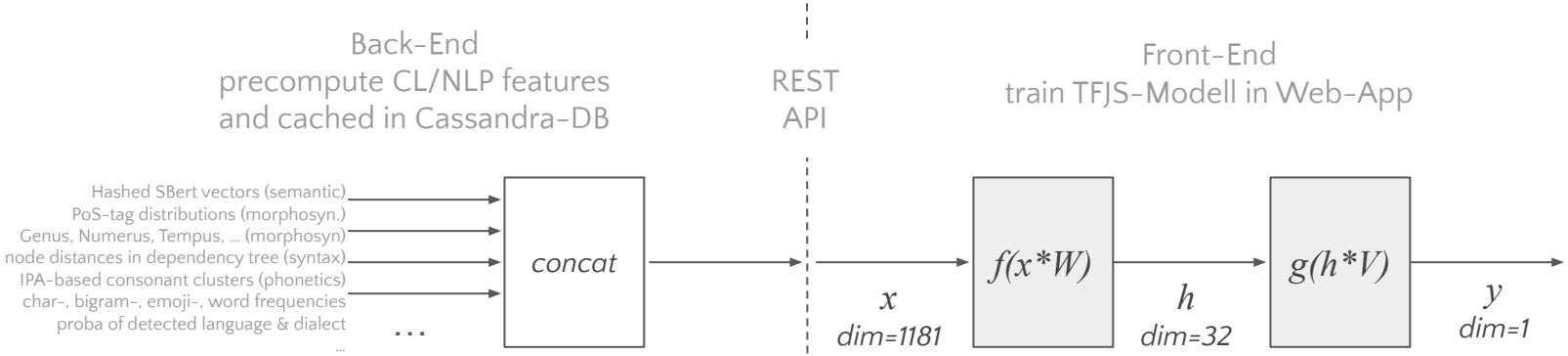
Quellen – Software

Software repositories for the EVIDENCE project:

https://github.com/satzbeleg

Appendix

Implementation details:
Precompute CL/NLP features & Keep the TFJS-Model simple



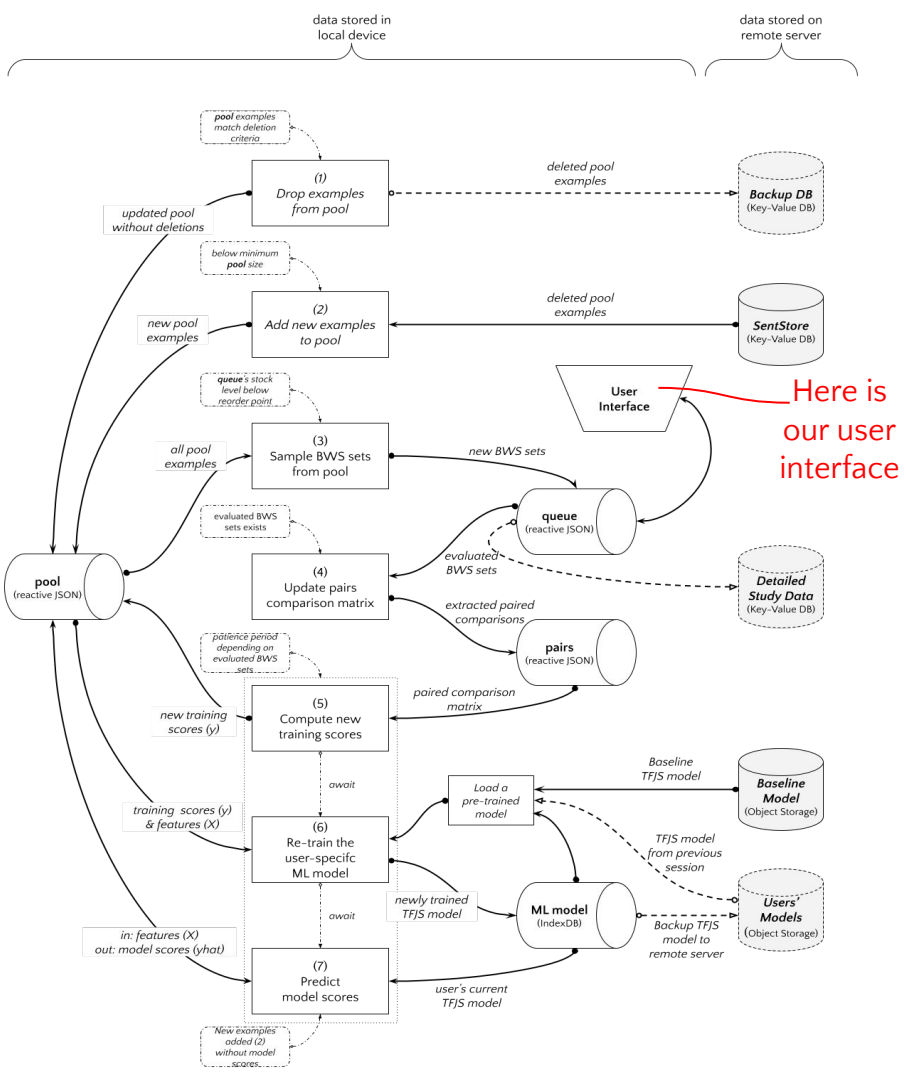
The small TFJS-model doesn't need to learn about language properties anymore because it should be encoded in the (pretrained) CL/NLP features.

XAI:
(1) Estimate Pr.Distr. of CL/NLP features,
(2) Run MC-Simulation on each group of CL/NLP feats.,
(3) Measure the sensitivity on model score.

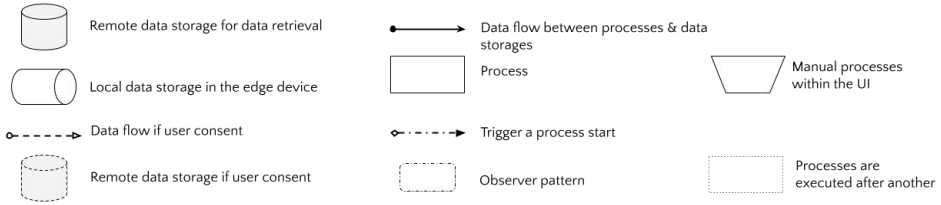
x : input signals
W : trainable model weights
f : swish activation
dropout rate 0.5

h : hidden states
V : trainable model weights
g : sigmoid activation [0, 1]
y : model score

Feature Engineering: https://github.com/satzbeleg/evidence-features/tree/main#features-overview
TFJS Model: https://github.com/satzbeleg/evidence-app/blob/main/src/components/bestworst/interactivity.js#L930
How to compress SBert features by 94 to 99%: https://arxiv.org/abs/2304.02481



Here is our user interface



Implementation Details

- processes are triggered by events (e.g., observed counter exceeds threshold)
- processes work on data, and run in parallel in the background
- data is stored in "reactive" variables (JS) which show the changes to the UI when background processes are done
- syncing local reactive variables with backed-server happens in the background too (e.g., offline-first principle if the internet connection is interrupted)

Demo (offline)

Example Käse (en: cheese)

Käse, der

Grammatik	Substantiv (Maskulinum) · Genitiv Singular: **Käses** · Nominativ Plural: **Käse**
Aussprache	
Worttrennung	Kä-se
Wortbildung	mit ›Käse‹ als Erstglied: ↗ Käseauflauf ... 46 weitere · mit ›Käse‹ als Letztglied: ↗ Analogkäse ... 40 weitere · mit ›Käse‹ als Binnenglied: ↗ Dreikäsehoch
Mehrwortausdrücke	↗ Tilsiter Käse · ↗ der Käse ist gegessen

1. ‘food’
2. [pej. ‘nonsense’

Bedeutungsübersicht

1. aus der Milch durch Säuerung oder Lab gewonnenes, gelblich-weißes, fetthaltiges und eiweißhaltiges, streichfestes oder schnittfestes Erzeugnis, das als Nahrungsmittel dient
2. [salopp, abwertend, übertragen] törichtes, dummes Geschwätz, Unsinn, Unfug

www.dwds.de/wb/käse

Käse (en: cheese) 1/2 - sorted by goodness (λ=100%)

Top-Scorers:
cheese as food

Außer Brot und Wein wurden noch lange Öl (zur Taufe?), Käse, Oliven, Erstlingsfrüchte (S. 115 ed. Hauler; Klauser- Rech, RAC II, 500), Blumen u. a. »geopfert«, zum Unterhalt von Armen und Klerikern.

w_0: 0.0073 | Schweizer, E. u. a.: Abendmahl. In: Galling, Kurt (Hg.), Die Religion in Geschichte und Gegenwart, Berlin: Directmedia Publ. 2000 [1957], S. 220

(Kaum) drei Käse hoch sein: (noch) ganz klein sein, spöttisch vor allem von einem kleinen Gernegroß gesagt, einem (Drei-)Käsehoch; schon 1767 im 'Versuch eines bremisch-niedersächsischen Wörterbuchs' (Band 2, S. 762):» Een Junge twe Kese hoog: ein kleiner kurzer Junge «; im niederdeutschen Raum machte man früher auf allen Höfen Käse nach Art der (Holländer)

w_1: 0.0073 | Röhrich, Lutz: Käse. In: Lexikon der sprichwörtlichen Redensarten [Elektronische Ressource], Berlin: Directmedia Publ. 2000 [1994], S. 27405

– Käse und Butter, Lapšin.

w_2: 0.0067 | Schlögel, Karl: Petersburg, München Wien: Carl Hanser Verlag 2002, S. 223

Dieser Erstlingskäse ist ein vollfetter Käse von erstklassigem Geschmack und hohem Nährwert.

w_3: 0.0067 | Die Landfrau, 24.01.1925

Die anderen, die vom Jungen beliefert worden waren, gaben uns einen Trostschluck und einen Happen Käse und neckten uns dafür.

w_4: 0.0067 | Alexander Granach, Da geht ein Mensch: Leck: btb Verlag 2007, S. 307

Käse (en: cheese) 1/2 - sorted by goodness (λ=100%)

Top-Scorers:
cheese as food

Außer Brot und Wein wurden noch lange Öl (zur Taufe?), Käse, Oliven, Erstlingsfrüchte (S. 115 ed. Hauler; Klauser- Rech, RAC II, 500), Blumen u. a. »geopfert«, zum Unterhalt von Armen und Klerikern.

w_0: 0.0073 | Schweizer, E. u. a.: Abendmahl. In: Galling, Kurt (Hg.), Die Religion in Geschichte und Gegenwart, Berlin: Directmedia Publ. 2000 [1957], S. 220

(Kaum) drei Käse hoch sein: (noch) ganz klein sein, spöttisch vor allem von einem kleinen Gernegroß gesagt, einem (Drei-)Käsehoch; schon 1767 im 'Versuch eines bremisch-niedersächsischen Wörterbuchs' (Band 2, S. 762):» Een Junge twe Kese hoog: ein kleiner kurzer Junge «; im niederdeutschen Raum machte man früher auf allen Höfen Käse nach Art der (Holländer)

w_1: 0.0073 | Röhrich, Lutz: Käse. In: Lexikon der sprichwörtlichen Redensarten [Elektronische Ressource], Berlin: Directmedia Publ. 2000 [1994], S. 27405

– Käse und Butter, Lapšin.

w_2: 0.0067 | Schlögel, Karl: Petersburg, München Wien: Carl Hanser Verlag 2002, S. 223

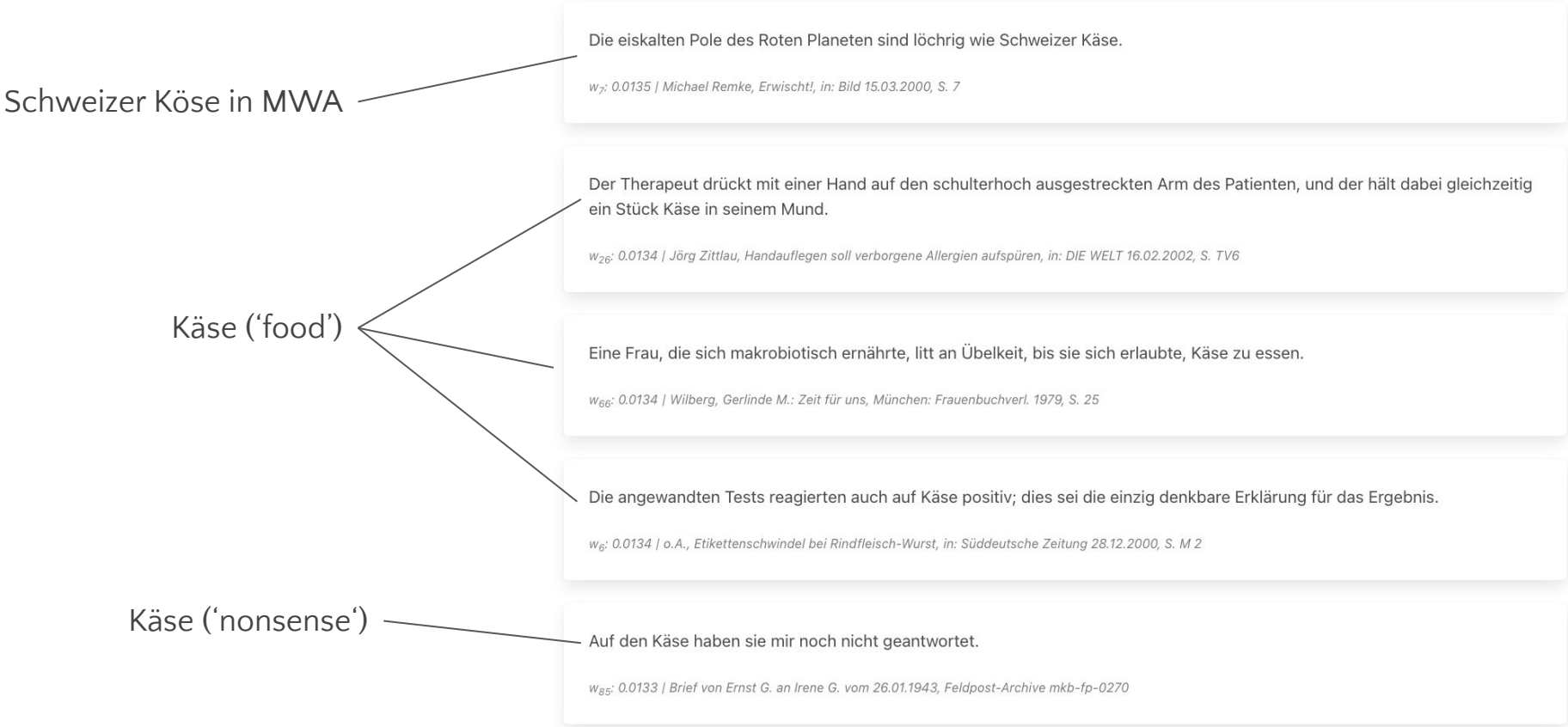
Dieser Erstlingskäse ist ein vollfetter Käse von erstklassigem Geschmack und hohem Nährwert.

w_3: 0.0067 | Die Landfrau, 24.01.1925

Die anderen, die vom Jungen beliefert worden waren, gaben uns einen Trostschluck und einen Happen Käse und neckten uns dafür.

w_4: 0.0067 | Alexander Granach, Da geht ein Mensch: Leck: btb Verlag 2007, S. 307

Käse (cheese) 2/2 – different semantics (**λ=0%, β_1=100%**)



digitalisieren (digitize) 1/2 – goodness vs. semantics & grammar (λ=25%, β_1=50%, β_2=50%)

Too many examples from the same source (Die ZEIT)

Der Protest mit Massen-Mails sollte das Rathaus blockieren, in dem alle Entscheidungsprozesse digitalisiert sind.

w_{13}: 0.0184 | Die Zeit, 29.10.1998, Nr. 45

Die gemeinnützige Stiftung, deren 35-Millionen-Dollar-Haushalt von der amerikanischen Regierung mitfinanziert wird, digitalisiert und katalogisiert die Videobänder zu einer »digitalen Bibliothek«.

w_0: 0.0184 | konkret, 1996

Wird Digital-Fernsehen vor allem mit Bezahlfernsehen identifiziert, so dürfte es schwer werden, das soeben von der Bundesregierung proklamierte Ziel zu erreichen und bis zum Jahr 2010 die Übertragungstechnik komplett zu digitalisieren.

w_1: 0.0184 | Die Zeit, 03.09.1998, Nr. 37

Das Sterben wird also zügig digitalisiert, das Leben nach dem Tod nicht minder.

w_{24}: 0.0183 | Die Zeit, 03.05.1996, Nr. 19

Ähnlich wie bei der Aufnahme einer Compact Disc (CD) wird das analoge Sprachsignal zunächst digitalisiert.

w_{48}: 0.0183 | o.A., Der Computer, der aufs Wort gehorcht, in: Süddeutsche Zeitung 16.02.1995, S. 24

digitalisieren 2/2 – variation of sources (λ=25%, β_1=50%, β_2=50%, **β_4=100%**)

Sources (ZEIT, C't, konkret, ZEIT, Bild)

Vor allem sollten wir nach den Menschen fragen, die darüber entscheiden, welche Information digitalisiert wird und welche nicht; ob uns in der Flut der Netzinformationen noch Zeit zur Erinnerung und Gedächtnis genug bleibt, uns an anderes zu erinnern als an die perfekte Beherrschung des Netzes und seiner Chancen.

w_3: 0.0188 | Die Zeit, 28.06.1996, Nr. 27

Dabei besteht das geringste Problem darin, alle Bilder zu digitalisieren.

w_{45}: 0.0185 | C't, 2000, Nr. 8

Die gemeinnützige Stiftung, deren 35-Millionen-Dollar-Haushalt von der amerikanischen Regierung mitfinanziert wird, digitalisiert und katalogisiert die Videobänder zu einer »digitalen Bibliothek«.

w_0: 0.0184 | konkret, 1996

Und was wäre, wenn man jedes der schätzungsweise 60 Millionen existierenden Bücher nur einmal digitalisieren würde?

w_{59}: 0.0183 | 37

Die weltberühmte Gutenbergbibel (aus dem Jahr 1456) gibt es jetzt digitalisiert.

w_{21}: 0.0183 | o.A., NEWS-Gute-Schlechte, in: Bild 23.03.2000, S. 4

Blau (en: blue) 1/2 – semantics (λ=0%, β_1=100%)

‘color’

Die Farbe der rohen Bohne ist grau, blau, grün oder gelb.

w_{47}: 0.0135 | Kölling, Alfred: Fachbuch für Kellner, Leipzig: Fachbuchverl. VEB 1962 [1956], S. 112

Die Farbe ist je nach Grund blau, weiß, gelb, gold, auch mit andersfarbigem Rand, Gold setzt sich durch.

w_{48}: 0.0134 | Schiller, G.: Nimbus. In: Die Religion in Geschichte und Gegenwart, Berlin: Directmedia Publ. 2000 [1960], S. 13604

Die hübsche Nina (22) träumt von einem Mann mit blauen Augen.

w_{19}: 0.0134 | o.A., 214 Hamburger Singles zum Verlieben, in: Bild 24.01.2006, S. 1

Es wurde mit einem Gitterspektrometer die Luftwellenlänge der zweiten Harmonischen im blauen Spektralbereich gemessen.

w_{96}: 0.0134 | Hollemann, Günter: Ein Dioden-gepumpter Nd:YAG Laser für ein Indium-Frequenznormal, Garching bei München: Max-Planck-Inst. für Quantenoptik 1993, S. 54

‘Black eye’

(discolored flesh around the eye resulting from a blow)

Doch diesmal kommt sie nicht mit einem blauen Auge davon.

w_{32}: 0.0133 | Jürgen Wenzel, Luxus-Luder fährt Amok auf'm Kudamm, in: Bild 16.08.2005, S. 5

blau 2/2 – more grammar variation (λ=0%, β_1=50%, β_2=50%)