

# 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

# 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DWDS consists to a large extent of legacy resources that are outdated or do not contain corpus examples at all.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](http://www.dwds.de/r))

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



www.dwds.de

## Hürde, die engl: 'hurdle' or met. 'obstacle'

**Grammatik** Substantiv (Femininum) · Genitiv Singular: **Hürde** · Nominativ Plural: **Hürden**  
**Aussprache** **h** [ˈhʏrɔ] **de**  
**Worttrennung** Hür-de  
**Wortbildung** mit ›Hürde‹ als Erstglied: ↗ Hürdenlauf ... **7 weitere** · mit ›Hürde‹ als Letztglied: ↗ 5-%-Hürde ... **7 weitere**  
**Mehrwortausdrücke** ↗ die Hürden liegen hoch

### Bedeutungsübersicht

1. (geflochtenes) Gestell zum Einzäunen eines Platzes für das Herdenvieh, besonders für Schafe  
schützend eingezäunter Platz für das Herdenvieh, Pferch
  2. Gestell zum Aufbewahren oder Trocknen, Horde
  3. [Sport] Hindernis, über das bei einem Wettlauf die Läufer, beim Pferderennen die Pferde springen müssen  
[übertragen] (eine Hürde nehmen) ein Hindernis, eine Schwierigkeit überwinden
3. **Sport** Hindernis, über das bei einem Wettlauf die Läufer, beim Pferderennen die Pferde springen müssen
- BEISPIELE:**  
die **Hürden** nehmen, überspringen  
die Läuferin siegte über 100-Meter **Hürden** (= beim Hürdenlauf auf einer Strecke von 80 Metern)
- ✓ **übertragen** (eine **Hürde** nehmen) ein Hindernis, eine Schwierigkeit überwinden
- BEISPIEL:**  
mit dem erfolgreichen Examen hatte er die letzte **Hürde** seines Studiums genommen

# 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 for 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

- 1: Norddeutsche Neueste Nachrichten, 16.09.2022  
Obwohl die Moderaten mit 19,1 Prozent weniger Stimmen erhielten als die Rechten, übernehmen sie den **Regierungsauftrag**.
- 2: Neue Osnabrücker Zeitung, 16.09.2022  
Obwohl die Moderaten mit 19,1 Prozent weniger Stimmen erhielten als die Rechten, übernehmen sie den **Regierungsauftrag**.
- 3: Der Prignitzer, 16.09.2022  
Obwohl die Moderaten mit 19,1 Prozent weniger Stimmen erhielten als die Rechten, übernehmen sie den **Regierungsauftrag**.
- 4: Schweriner Volkszeitung, 16.09.2022  
Obwohl die Moderaten mit 19,1 Prozent weniger Stimmen erhielten als die Rechten, übernehmen sie den **Regierungsauftrag**.
- 5: Bote der Urschweiz, 09.09.2022  
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hat sie elf Männern und drei Frauen im Buckingham-Palast den **Regierungsauftrag** erteilt und damit ihre wichtigste konstitutionelle Aufgabe erfüllt . (sbo)
- 6: Luzerner Zeitung, 09.09.2022  
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hat sie elf Männern und drei Frauen im Buckingham-Palast den **Regierungsauftrag** erteilt und damit ihre wichtigste konstitutionelle Aufgabe erfüllt . (sbo)
- 7: St. Galler Tagblatt, 09.09.2022  
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hat sie elf Männern und drei Frauen im Buckingham-Palast den **Regierungsauftrag** erteilt und damit ihre wichtigste konstitutionelle Aufgabe erfüllt . (sbo)
- 8: Thurgauer Zeitung, 09.09.2022  
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hat sie elf Männern und drei Frauen im Buckingham-Palast den **Regierungsauftrag** erteilt und damit ihre wichtigste konstitutionelle Aufgabe erfüllt . (sbo)
- 9: Der Tagesspiegel, 09.09.2022  
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hatte sie elf Männern und drei Frauen stets im Buckingham-Palast den **Regierungsauftrag** erteilt.
- 10: Frankfurter Rundschau, 07.09.2022  
Die nur von ihrer Partei gekürte britische Premierministerin holt sich in Balmoral ihren **Regierungsauftrag** ab / Von Sebastian Borger

*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.

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

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

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

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

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

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

### Search Settings

**Variation (0) vs Goodness Score (100)**  
*prefer goodness score over variation*

25  $\lambda$

**Semantic**  
*penalize semantic similar sentences*

50  $\beta_1$

**Grammar**  
*penalize syntactic similar sentences*

0  $\beta_2$

**Near Duplicates**  
*penalize similar fingerprint*

0  $\beta_3$

**Bibliographic**  
*penalize similar sources*

0  $\beta_4$

# Implementation details:

## Reverse Automatic Differentiation as approximation for quadratic optimization problems

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

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

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

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 regularization penalties.

*Quadratic Optimization with Tensorflow in Python*

[https://github.com/satzbeleg/keras-quadopt/blob/main/keras\\_quadopt/problem.py](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  $D_{ij}$  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 trunkit.
- the dependency tree is decomposed into partial trees (treesimi pypi package).
- partial trees are serialized for MinHash (datasketch pypi package).
- the jaccard similarity  $D_{ij}$  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.

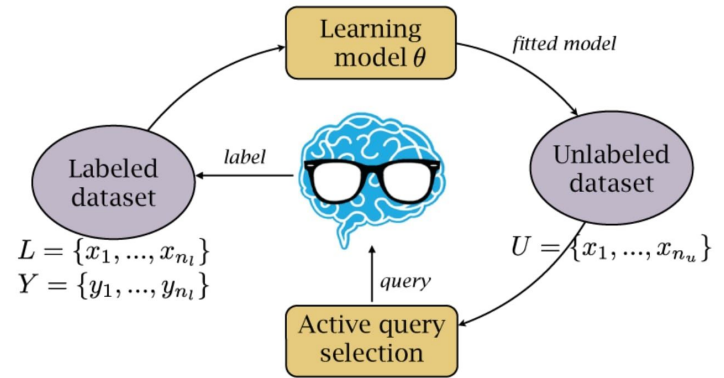
Train a scoring model  
with Interactive Learning

# 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)

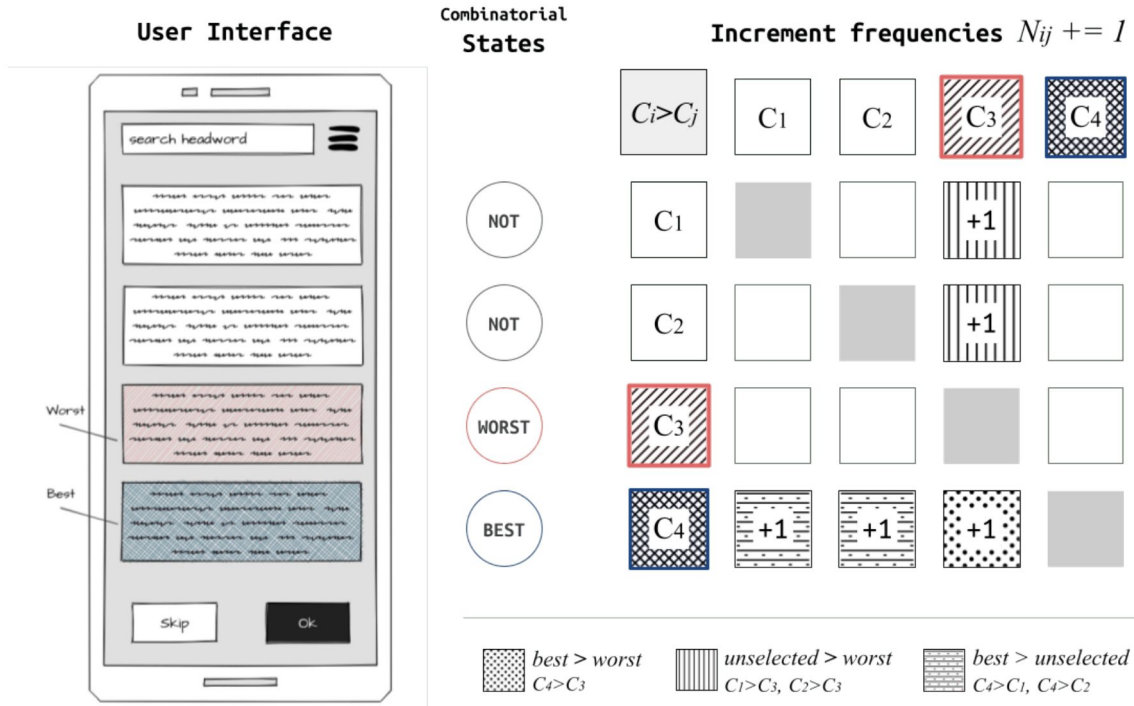
1. User annotates (**few!**) examples
2. (Warm-Start) Model retraining (few! epochs)
3. Model predicts scores for step 4.
4. Sampling of new examples for step 1.



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

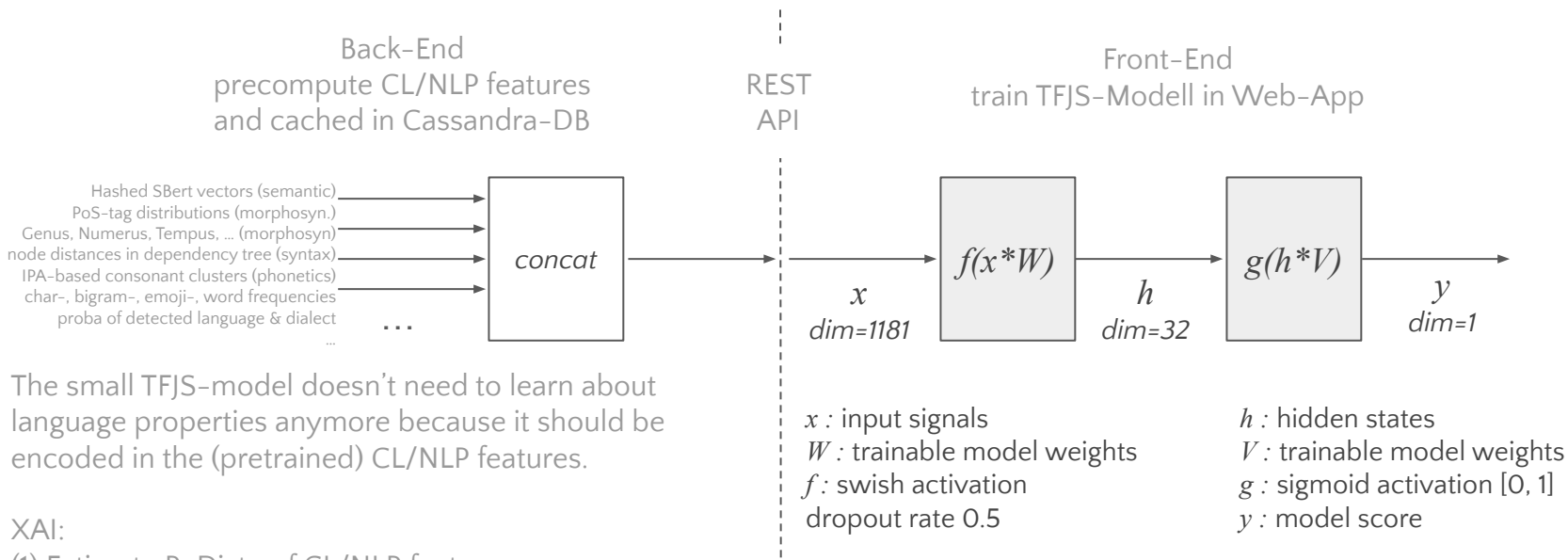
Software repositories for the EVIDENCE project:

<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.

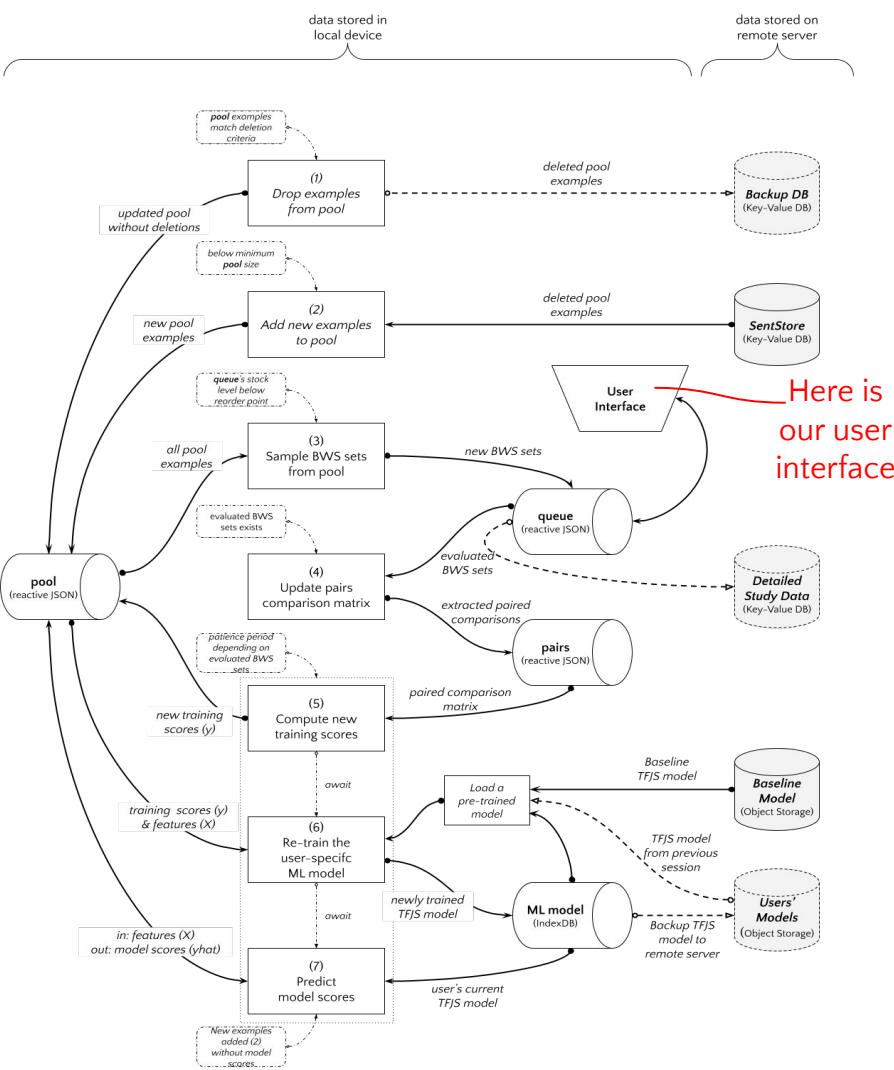
XAI:

- (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.

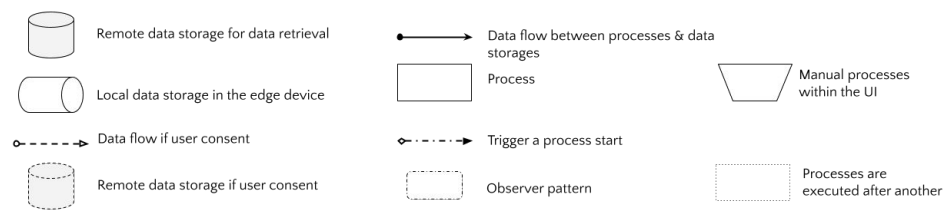
Feature Engineering: <https://github.com/satzbeleg/evidence-features/tree/main#features-overview>

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 model-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

# Käse (en: cheese) 1/2 – sorted by goodness ( $\lambda=100\%$ )

Top-Scorers:  
cheese as food

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

*w<sub>0</sub>: 0.0073 | Schweizer, E. u. a.: Abendmahl. In: Gallig, 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<sub>1</sub>: 0.0073 | Röhrich, Lutz: Käse. In: Lexikon der sprichwörtlichen Redensarten [Elektronische Ressource], Berlin: Directmedia Publ. 2000 [1994], S. 27405*

– Käse und Butter, Lapšin.

*w<sub>2</sub>: 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<sub>3</sub>: 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<sub>4</sub>: 0.0067 | Alexander Granach, Da geht ein Mensch: Leck: btb Verlag 2007, S. 307*

# Käse (en: cheese) 1/2 – sorted by goodness ( $\lambda=100\%$ )

Top-Scorers:  
cheese as food

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

*w<sub>0</sub>: 0.0073 | Schweizer, E. u. a.: Abendmahl. In: Gallig, 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<sub>1</sub>: 0.0073 | Röhrich, Lutz: Käse. In: Lexikon der sprichwörtlichen Redensarten [Elektronische Ressource], Berlin: Directmedia Publ. 2000 [1994], S. 27405*

– Käse und Butter, Lapšin.

*w<sub>2</sub>: 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<sub>3</sub>: 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<sub>4</sub>: 0.0067 | Alexander Granach, Da geht ein Mensch: Leck: btb Verlag 2007, S. 307*

# Käse (cheese) 2/2 - different semantics ( $\lambda=0\%$ , $\beta_1=100\%$ )

Schweizer Käse in MWA

Die eiskalten Pole des Roten Planeten sind löchrig wie Schweizer Käse.

*w<sub>7</sub>: 0.0135 | Michael Remke, Erwischt!, in: Bild 15.03.2000, S. 7*

Käse ('food')

Der Therapeut drückt mit einer Hand auf den schulterhoch ausgestreckten Arm des Patienten, und der hält dabei gleichzeitig ein Stück Käse in seinem Mund.

*w<sub>26</sub>: 0.0134 | Jörg Zittlau, Handauflegen soll verborgene Allergien aufspüren, in: DIE WELT 16.02.2002, S. TV6*

Eine Frau, die sich makrobiotisch ernährte, litt an Übelkeit, bis sie sich erlaubte, Käse zu essen.

*w<sub>66</sub>: 0.0134 | Wilberg, Gerlinde M.: Zeit für uns, München: Frauenbuchverl. 1979, S. 25*

Die angewandten Tests reagierten auch auf Käse positiv; dies sei die einzig denkbare Erklärung für das Ergebnis.

*w<sub>6</sub>: 0.0134 | o.A., Etikettenschwindel bei Rindfleisch-Wurst, in: Süddeutsche Zeitung 28.12.2000, S. M 2*

Käse ('nonsense')

Auf den Käse haben sie mir noch nicht geantwortet.

*w<sub>85</sub>: 0.0133 | Brief von Ernst G. an Irene G. vom 26.01.1943, Feldpost-Archive mkb-fp-0270*

# digitalisieren (digitize) 1/2 – goodness vs. semantics & grammar ( $\lambda=25\%$ , $\beta_1=50\%$ , $\beta_2=50\%$ )

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<sub>13</sub>: 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<sub>0</sub>: 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<sub>1</sub>: 0.0184 | Die Zeit, 03.09.1998, Nr. 37*

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

*w<sub>24</sub>: 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<sub>48</sub>: 0.0183 | o.A., Der Computer, der aufs Wort gehorcht, in: Süddeutsche Zeitung 16.02.1995, S. 24*

# digitalisieren 2/2 – variation of sources ( $\lambda=25\%$ , $\beta_1=50\%$ , $\beta_2=50\%$ , $\beta_4=100\%$ )

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<sub>3</sub>: 0.0188 | Die Zeit, 28.06.1996, Nr. 27*

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

*w<sub>45</sub>: 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<sub>6</sub>: 0.0184 | konkret, 1996*

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

*w<sub>59</sub>: 0.0183 | 37*

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

*w<sub>2</sub>: 0.0183 | o.A., NEWS-Gute-Schlechte, in: Bild 23.03.2000, S. 4*

# Blau (en: blue) 1/2 – semantics ( $\lambda=0\%$ , $\beta_1=100\%$ )

‘color’

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

*w<sub>47</sub>: 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<sub>48</sub>: 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<sub>19</sub>: 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<sub>96</sub>: 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<sub>32</sub>: 0.0133 | Jürgen Wenzel, Luxus-Luder fährt Amok auf'm Kudamm, in: Bild 16.08.2005, S. 5*



# blau 2/2 – more grammar variation ( $\lambda=0\%$ , $\beta_1=50\%$ , $\beta_2=50\%$ )

previous hits

Er lächelt, mit der blauen Mütze, seinem roten Halstuch.

*w<sub>97</sub>: 0.0135 | o. A.: Reportage vom Großen Preis von Deutschland auf dem Nürburgring, 17.07.1932*

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

*w<sub>19</sub>: 0.0135 | o.A., 214 Hamburger Singles zum Verlieben, in: Bild 24.01.2006, S. 1*

Beim "Blauen Engel" war das natürlich überdimensional.

*w<sub>143</sub>: 0.0133 | Der Spiegel, 29.03.1993*

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

*w<sub>96</sub>: 0.0133 | Hollemann, Günter: Ein Dioden-gepumpter Nd:YAG Laser für ein Indium-Frequenznormal, Garching bei München: Max-Planck-Inst. für Quantenoptik 1993, S. 54*

new hits

Die physikalisch unterschiedlichen Umfelder bewirken unterschiedliche Verschiebungen der Farbe der physikalisch identischen blauen Testfelder.

*w<sub>22</sub>: 0.0133 | Hoffmann, K.-P. u. Wehrhahn, Christian: Zentrale Sehsysteme. In: Dudel, Josef u. a. (Hgg.) Neurowissenschaft, Berlin: Springer 1996, S. 424*