



# A machine learning method based on word profiles for semi-automatic update of polysemous dictionary entries in legacy dictionaries

Alexander Geyken – BBAW

Christian Poelitz, Thomas Bartz (TU Dortmund)

ENel

Wien, 12-02-2015

# Outline

1. DWDS Project: Background
2. GdEx + WSD
  - a. Motivation
  - b. Resources
  - c. Method
  - d. Experiment
  - e. Results

(A) More than 450 000 dictionary entries

→ Grimm, Etymological Dictionary (Pfeifer), DWDS-Sync (i.e. WDG+Duden) OpenThesaurus, GermaNet

(B) Corpora (4 billion tokens)

→ Reference corpus (20th c.), Deutsches Textarchiv (17-19th c.), newspapers (Bild, SZ, Welt, Spiegel, Zeit...), CMC-corpora

(C) Word statistics

→ Time lines, word sketches





**DWDS-Wörterbuch**

**Mission**  
 fem., -, -en  
 Herkunft: Latein

- ehrentvoller Auftrag, verpflichtende Aufgabe**  
 eine hohe, historische, nationale, humanistische, kulturelle Mission haben
- von einer Regierung mit besonderem (diplomatischem) Auftrag ins Ausland entsandte Person, Gruppe von Personen**  
 eine Mission nach K senden
- meist unter der Leitung eines Gesandten oder Geschäftsträgers stehende diplomatische Vertretung eines Staates im Ausland (und deren Sitz)**  
 die Errichtung einer Mission in Warschau
- Religion die (Äußere) Mission** Verbreitung einer religiösen Lehre unter Andersgläubigen durch Missionare ohne Plural  
 Mission treiben

Dazu:  
 Version: 0.4.23 | Quelle: WDG | Artikeltyp: Vollartikel Kompakt | Details

**Etymologisches Wörterbuch**

**Mission**, fern: Missionar, missionieren

**Mission** f. 'Sendung, Botschaft, Aufgabe, Auftrag' wird im 16. Jh. aus lat. *missio* (Gen. *missionis*) 'das Absenden, Entlassung' entlehnt, einem Verbalsubstantiv zu lat. *mittere* (*missum*) 'schicken, senden'. Der zunächst dem lat. Wörtinhalt folgende Gebrauch nimmt zusätzlich im 17. Jh., kirchensprachlichem *missio* entsprechend, die Bedeutung 'Ausendung im Namen Christi; Bekehrung der Heiden' an. Den besondere zwischenstaatliche Beziehungen bezeichnenden Sinn 'diplomatischer Auftrag, zu bestimmten Aufgaben entsandte Personengruppe, Gesandtschaft (eines Staates)' vermittelt im 18. und 19. Jh. frz. *mission*. Aus den genannten Verwendungsweisen entwickelt der allgemeine Sprachgebrauch die Bedeutung 'gottgewollter bzw. ernster, bedeutsamer Auftrag, Sendung, Bestimmung'. – **Missionar** m. 'beauftragter Verkünder der christlichen Religion bei Ungläubigen' (17. Jh.), *nlät. missionarius*. **missionieren** Vb. 'zur christlichen Religion bekehren' (19. Jh.).

Version: 1.0.87 | © Dr. Wolfgang Pfeifer

**OpenThesaurus**

**Synonymgruppen für Mission**

Synonymgruppe: Auslandsvertretung, Botschaft, Gesandtschaft, Konsulat, Mission, diplomatische Vertretung, ständige Vertretung

Version: 2014-07-07 | Quelle: OpenThesaurus

**DWDS-Wortprofil 3.0**

Abfragewort:  Vergleichswort:

Substantiv logDice 20

Überblick zu 'Mission'

Arbeiterklasse Arbeiterwohlfahrt benannte Caritas  
 Diakonie diplomatischen erfüllen Evangelisation  
 Fehlschlag Friedenseinsätze geheimer heikle  
 Kampfeinsatz Kampfeinsätze Mission Sonde  
 Sondergesandten Twister Ökumene

'Mission' hat Attribut  
 'Mission' hat Genitivattribut  
 Koordination mit 'Mission'

Version: 3.0 Einstellungen

**'DWB (1854-1961)**

Kein Eintrag vorhanden

Version: @rev145

**Deutsches Textarchiv (wöchentlich aktualisiert)**

Treffer: 1112

KWIC Datum Datum Zufällig Links Rechts

- 1913 "riester geweiht, arbeitet an vMissionen, kehrte 1854 nach Eun
- 1913 t stets an den Werken der JnnMission beteiligt und das Martha-
- 1913 ), war 1888-93 VereinsgeistlichMission in Braunschweig u. ist sei
- 1913 Er ging 1875 in politischerMission nach London, wurde auch
- 1913 m ihre Feder u. Kenntnisse gaMissionen zu stellen.
- 1913 farrers, der 1869 als DirektorMission nach Leipzig berufen war;
- 1913 1889 Hilfsgeistlicher beim VereMission und Lehrer an der Dumas
- 1913 n zu Anfang des Jahres 1852 iMission.
- 1913 ungen aus der südafrikanischMission, 1891. - Erntekranz ( Per
- 1913 sonders auf dem Gebiete der irMission tätig und schrieb vorwieg
- 1913 - MagdalenesMission ( E. ), 1905.
- 1913 er noch unmittelbar vor dem dMission von seiten des Königs zur
- 1913 r Vaterstatt, um er sich auch dMission lehnte betätigte.

**DIE ZEIT**

Treffer: 8330

KWIC Datum Datum Zufällig Links Rechts

- 2014 )hen sagt, klingt es wie eine poMission.
- 2014 t in diesen Tagen, dann kann dMission über 14. fast 15 Jahre jet
- 2014 ) der Leyen sagte, dass sie diMission auch nach so langer Zeit i
- 2014 SeineMission habe er direkt von Allah a
- 2014 insatz von dem Sie reden, warMission der OSZE.
- 2014 gerübergreifend " im März eineMission ihrer Organisation beschl
- 2014 ngig davon, mit Kiew eine bilMission in Marsch setzt, unter Lei
- 2014 in die Brüsseler RegulierungsMission geworden, seit er nicht m
- 2014 loh habe eine Mission, handle im Interesse me
- 2014 entbehren vom Volk an die Mission, es zu befreien.
- 2014 zeit ist zwar zu Ende, aber meMission ist noch nicht fertig", sag
- 2014 Er glaubt, dass er eine höhereMission erfüllt, indem er Russlanc
- 2014 enik: Mir sind dankbar für dieMission der OSZE in unserem L ar

**Kernkorpus 20**

Treffer: 2128, davon anzeigbar: 1895

KWIC Datum Datum Zufällig Links Rechts

- 1999 DieMission blieb erfolglos, obwohl Be
- 1999 d der Griechen bester Redner Mission.
- 1999 heint das unruhliche Ende dMission.
- 1999 hten, sowenig sahen sich jeneMission. In erheblich größerem
- 1999 and malten, war Teil ihrer sozMission.
- 1999 und zwei Fanfarekapellen stMission startete.
- 1999 dem eher im Gegenteil als MonMission des Kapitalismus, von t
- 1999 auch die Vorstellung von der dMission eigens vom Weltgeist ges
- 1999 ten, oder in Bereung ihrer SMissionen beschenken, kann der
- 1999 in geheimerMission!
- 1999 wie die Jungfrau ihre historisMission verletzt, als sie sich ganz
- 1999 Zu ihrer besonderenMission wird die Menschlichkeit.
- 1999 nd - mehrere Firmen bereiten eMissionen vor, und das Gros der
- 1999 - begonnen, nachdem die beiMission geschickt hatten.
- 1999 ind Persönlichkeiten einsetzenMission haben und das Persona
- 1999 atfinden, da der Graf in diplomMission andertrags schon abreise

Version: 1.1 Optionen

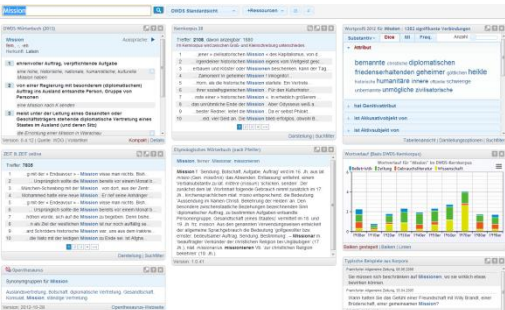
**Wortverlauf (Basis DWDS-Kernkorpus)**

Wortverlauf für "Mission" im DWDS-Kernkorpus

● Selbsteristik ● Zeitung ● Gebrauchsliteratur ● Wissenschaft


Jahr	Selbsteristik	Zeitung	Gebrauchsliteratur	Wissenschaft
1900	100	50	20	10
1910	120	60	30	15
1920	150	80	40	20
1930	180	100	50	25
1940	200	120	60	30
1950	220	140	70	35
1960	250	160	80	40
1970	280	180	90	45
1980	300	200	100	50
1990	320	220	110	55
1999	350	250	120	60

# Panel: DWDS-synch dictionary



## DWDS-Wörterbuch (2013)

**Mission**  
fem., -, -en  
Herkunft: Latein

Aussprache: 

### 1 ehrenvoller Auftrag, verpflichtende Aufgabe

*eine hohe, historische, nationale, humanistische, kulturelle Mission haben*

### 2 von einer Regierung mit besonderem (diplomatischem) Auftrag ins Ausland entsandte Person, Gruppe von Personen

*eine Mission nach K senden*

### 3 meist unter der Leitung eines Gesandten oder Geschäftsträgers stehende diplomatische Vertretung eines Staates im Ausland (und deren Sitz)

*die Errichtung einer Mission in Warschau*

## 2.a Motivation

- maintenance of examples in legacy dictionaries (replace outdated examples, add more examples, more genres)
- polysemous entry: map the example sentences to the appropriated dictionary senses by using ML techniques
- specific motivation: 50,000 “skeleton entries” from Duden-99 are to be integrated into DWDS-system.
- However: Duden does not grant the right to publish the dictionary examples

## 2b. Resources WDG (Duden)

- a large monolingual legacy dictionary for German (WDG, 1964-1977)
  - WDG: for each sense: definition, synonyms, constructed examples and (possibly) outdated corpus examples;
  - 90.000 full entries (235.000 usage examples/patterns)
- a sketch-engine like database of German collocations (Wordprofile, 12 million collocations, based on DWDS-corpus (2G tok; logDice > 0))
- GdEx sentences extracted from DWDS corpus (Kilgarriff 2008, Didakowski 2012, for German)

## 2c. Method (1/3)

- step 1: select keywords in dictionary
- step 2: compute GdEx for each keyword
- step 3: Init Knowledgebase: extract+lemmatize (A,N,V) for each sense of each keyword
- step 4: Enrich knowledgebase
  - compute set of all collocations for all (A,N,V)
  - (compute synsets, hyp(o|-er)nyms with GermaNet)
- step 5: sense annotation by hand for each GdEx
- step 6: determine best sense (LESK, MEC)



## 2c. Method (2/3)

### step 6: determine best sense (LESK, MEC)

- LESK1: the max. number of intersecting words
- LESK2: the max. of sum of logDices
- MEC: a Maximum Entropy Classifier (Nigam 1999) to learn the correct mapping between example sentence to its sense number based on hand labeled sense annotations.
  - MEC estimates the probability of a sense for a given example
  - (the sense with the max. probability is then selected)

## 2c. Method (3/3) - Maximum Entropy Classifier

- Formally, the probability of a sense  $s$  for a given example  $e$  is defined as  $p(s|e) = \frac{e \cdot w}{Z}$  with a feature vector, a weight vector  $w$  and the normalization constant  $Z$ .
- The features are extracted from the Word Profiles of (A,N,V) in  $e$  and  $s$ . Each feature is the sum of the weights of WordProfile for (A,N,V) in  $e$  and  $s$
- We find the optimal weights by maximizing the joint probability over a training set of sentences with hand labeled senses.

- step 1: select keywords
- 100 highly polysemous words from WDG (75 nouns, 25 verbs) from the WDG with a total of 857 senses (314 main senses).
  - nouns: Schloss (2), Mutter(2) ... Grund (5),... Satz(8)
  - verbs: kosten (2), scheinen (2), ... anstellen(6),  
anschließen (6)

# Experiment

Example: w= Grund (ground, reason...)

sense: **I.1.**: "Boden"

sense: **I.1.a)** has 1 def. and 3 examples:

- Erdboden
- fetter, magerer, lehmiger, trockener, sandiger, steiniger, unebener Grund
- auf festem, schlüpfrigem Grund stehen
- etw. bis auf den Grund(völlig) zerstören

# Experiment

## step 3: select GdEx

- step 3a: extract content words + lemmatize
- step 3b: hand-labelling with WDG senses
- 20 examples for each keyword => 2000 examples
- 2 annotators (inter annotator agreement: kappa = 0.88 for main senses).

# Experiment

- step 4 : compute enriched knowledge base (with collocations)
- step 5: LESK1, LESK2, ML-method (training set on 50 keywords/1000 examples), evaluation on the remaining 50 keywords/1000 examples

# Preliminary Results

- Baseline 11.67% (100/8.57)
- LESK1: 31.17%; LESK2 ~ 32 %
- MEC: 43% for verbs, and 52% for nouns
- homographs are recognized with much higher precision (up to 90%) than entries with a fine-grained sense distinction.
- The lower accuracy for verbs: WDG frequently uses only placeholders (such as s.o., sth.) in its sense descriptions.

# Future work

- Add (synset/hyp(o|er)nym information from GermaNet
- use phrases as features
- WDG has few corpus examples, and if so, with very short context
  - ⇒ Apply method on WDG+ (i.e. WDG + 10GdEx-examples for a given key-word); apply algorithm with the remaining 10GdEx.
- Apply method to “legacy”-Duden-99