Open Multilingual Wordnet

Stand der Forschung

Erkenntnisse von der Global Wordnet Conference 2023

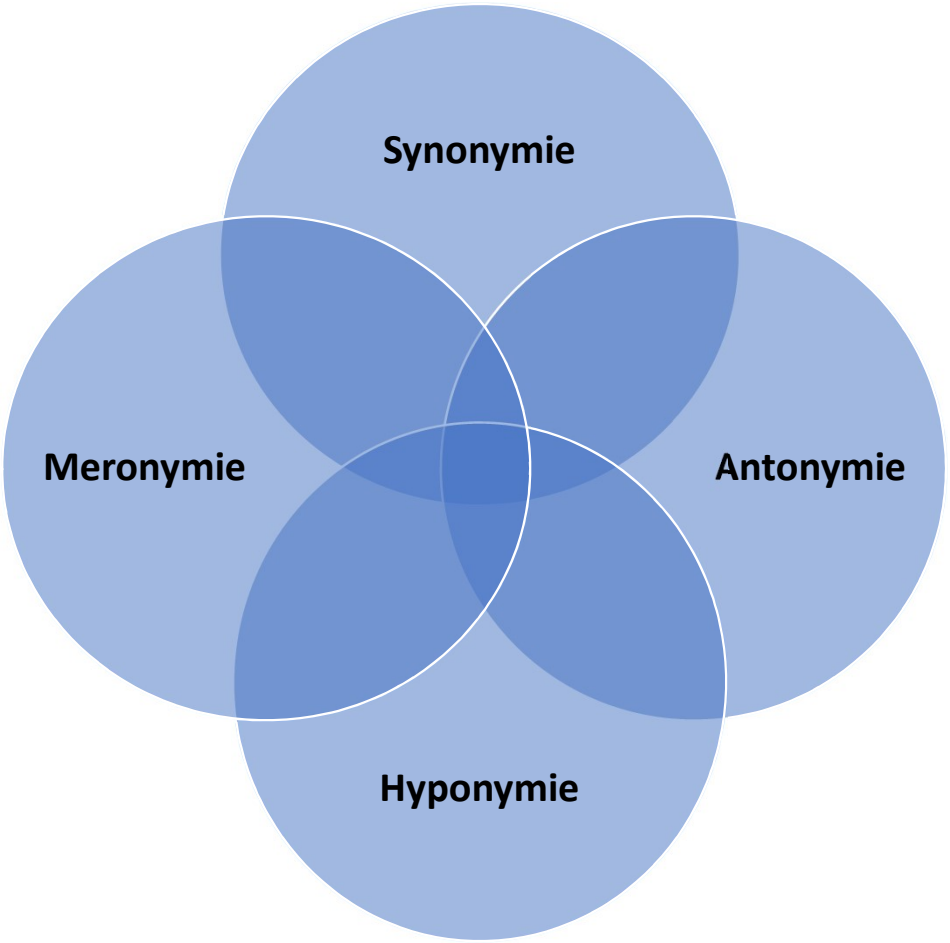
Wordnet

- Lexikalische Datenbank zunächst für Englisch
- Open-Source
- Entwickelt von der Princeton University (seit 1985)
- http://wordnet.princeton.edu/
- 4 semantische Teilnetze:
  - Nomen
  - Verben
  - Adjektive
  - Adverbien

Wordnet

- Organisation lexikalischen Wissens nach Wortbedeutungen
- Organisation in *synsets*
  - Mengen von Synonymen
  - Lesarten
  - Definitionen

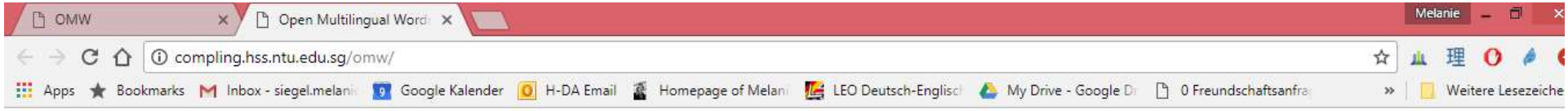
Lexikalische Beziehungen in Wordnet (Basis)



Nutzung von Wordnet

- Sprachtechnologie-Anwendungen
  - Lesarten-Disambiguierung
  - Informationserschließung und Informationsextraktion
  - Linguistische Annotation von Sprachdaten
  - Textklassifikation und automatische Textzusammenfassung
  - Entwicklung von Werkzeugen für Sprachanalyse und Sprachgenerierung
  - Maschinelle Übersetzung

Internationalisierung: Open Multilingual Wordnet



Open Multilingual Wordnet

This page provides access to open wordnets in a variety of languages, all linked to the Princeton Wordnet of English (PWN). The goal is to make it easy to use wordnets in multiple languages. The individual wordnets have been made by many different projects and vary greatly in size and accuracy. We have (i) extracted and normalized the data, (ii) linked it to Princeton WordNet 3.0 and (iii) put it in one place. The Open Multilingual Wordnet and its components are open: they can be freely used, modified, and shared by anyone for any purpose. There is a fuller list of wordnets at the Global Wordnet Association's Wordnets in the World page.

If you use these wordnets, please cite the original projects who created them (linked in Table 1), if you got value from this aggregation/normalization, please cite Bond and Paik (2012).

You can access the wordnets through the (python) Natural Language Tool-Kit wordnet interface (NLTK).

We have an extended version with automatically extracted data for over a 150 languages from Wiktionary and the Unicode Common Locale Data Repository (Bond and Foster, 2013).

Documentation, News and Updates

Search

We have a simple search interface (search the extended wordnet). It uses the SQL database originally developed by the Japanese Wordnet.

Neue Wordnets für neue Sprachen

- Grundsätzliche Ansätze:
  - Von Hand
  - Expand: automatische Übersetzung aus dem Englischen
  - Merge: Nutzung eines monolingualen Lexikons und Überführung in das Wordnet-Format


gwc
2023
Donostia
HOME
CALL FOR PAPERS
IMPORTANT DATES
ORGANIZATION
PROGRAMME
SUBMISSION
INFORMATION
REGISTRATION

Korrekturen und Erweiterungen von Wordnets

Javier Álvez, Itziar Gonzalez-Dios and German Rigau. Towards Effective Correction Methods Using WordNet Meronymy Relations.

Eric Kafe. Mapping wordnets on the fly.

Kiril Simov and Petya Osenova. Recent Developments in BulTreeBank-WordNet (BTB-WN).

Marek Maziarz, Łukasz Grabowski, Tadeusz Piotrowski, Ewa Rudnicka and Maciej Piasecki. Lexicalized and non-lexicalized multi-word expressions in WordNet: a cross-encoder approach.

Francis Bond and Takyuki Kuribayashi. The Japanese Wordnet 2.0.

Ivelina Stoyanova and Svetlozara Leseva. Expanding the Conceptual Description of Verbs in WordNet with Semantic and Syntactic Information.

Ahti Lohk, Martin Rebane and Heili Orav. An Experiment: Finding Parents for Parentless Synsets by Means of CILI.

Svetla Koeva and Dimitar Hristov. Resolving Multiple Hypernymy.

Julie Kallini and Christiane Fellbaum. What to Make of make? Sense Distinctions for Light Verbs.

Erica Biagetti, Chiara Zanchi and Silvia Luraghi. Linking the Sanskrit WordNet to the Vedic Dependency Treebank: a pilot study.

Ramona Kuehn, Jelena Mitrović and Michael Granitzer. Hidden in Plain Sight: Can German Wiktionary and Wordnets Facilitate the Detection of Antithesis?

Julie Kallini and Christiane Fellbaum. What to Make of make? Sense Distinctions for Light Verbs

- Light Verbs:
  - make, have, get
  - sehr häufig und polysem
  - full, z.B. "make dinner", light, z.B. "make a request"
- Light Verbs in Wordnet:
  - viele Bedeutungen für diese Verben, die gruppiert werden sollten (z.B. 49 für "make")
  - keine Unterscheidung zwischen Vollverb "make" und Light Verb "make"
- Ziel: Gruppierung der Bedeutungen
- Idee: Gruppierung mit Word2Vec-Clustern der nominalen Komplemente
  - Basis: mit Dependenz annotierte Korpora (Treebanks)
  - Extraktion der nominalen Komplemente von "make"
  - Gruppierung der Komplemente in den Korpora mit Word2Vec
- Ergebnisse:
  - 493 Komplemente von make gefunden
  - 26 Cluster wurden daraus gebildet, z.B. [vodka, soup, wine, potato, food], [statement, announcement, speech]

Towards Effective Correction Methods Using WordNet Meronymy Relations

- Problem: Auch Wordnets, die in Handarbeit erstellt und gepflegt werden, enthalten Fehler
- Ziel: Automatisches Finden von Fehlern und Korrekturen mit möglichst wenig Handarbeit
- Methode:
  - Mapping zwischen Wordnet und SUMO (First-Order Logic Ontology), Cross-Check
  - Fokus auf Meronymie-Relationen in Wordnet und SUMO
  - Analyse von 200 Basis-Konzepten in Wordnet und ihr Mapping zu SUMO
  - Dann Übertragung der Korrekturen auf die Hyponyme
- Entdeckung von Diskrepanzen zwischen SUMO und Wordnet, Korrekturen

Eric Kafe. Mapping wordnets on the fly

- Problem:
  - Es existieren verschiedene Versionen des Princeton Wordnets, unter anderem OEWN
  - Word Senses können gelöscht oder hinzugefügt worden sein
  - Z.B.: zwischen Versionen 3.0 and 3.1 von PWN wurde “Pluto” verschoben von der griechischen Mythologie zur römischen.
  - In den Wordnets der anderen Sprachen wurden die Änderungen aber nicht nachvollzogen
- Methoden:
  - Implementierung von Mappings der Sense-Keys
- Ergebnisse:
  - 205 Sense-Keys stimmten zwischen PWN und OEWN nicht überein

Kiril Simov and Petya Osenova. Recent Developments in BulTreeBank-WordNet (BTB-WN).

- BTB-WN seit 2010
- Tool für die Arbeit mit BTB-WN
  - Präsentation von Information über Lemmata: Synsets, POS, Links zu OEWN, Hyponyme und Hyperonyme
- Manuelle Prüfungen und Korrekturen:
  - Definitionen
  - Links zu OEWN
  - Fehlende Bedeutungen
  - Relationen
  - Beispiele
- Ergebnisse:
  - 6.000 Lemmata im „Core Vocabulary“
  - 14.000 Synsets
- Anwendungen:
  - Verbindung zwischen lexikalischen Ressourcen
  - Plattform für die lexikalische Recherche
  - „Spiel“ für Sprachstudium

Marek Maziarz, Łukasz Grabowski, Tadeusz Piotrowski, Ewa Rudnicka and Maciej Piasecki. Lexicalized and non-lexicalized multi-word expressions in WordNet: a cross-encoder approach

- Problem:
  - Mehrwortlexeme (MWE) in NLP: es gibt sehr unterschiedliche (Idiome, Namen, feste Phrasen, Komposita, Kollokationen, u.a.)
  - Unterscheidung zwischen lexikalisierten und nicht lexikalisierten Mehrwortlexemen: Welche MWE sollen im Lexikon stehen?
- Methode:
  - Extraktion aller MWE aus PWN: 39.406, 86% davon Nomen
  - Zufällige Auswahl von ca. 10% der MWE
  - Prüfung in 6 verschiedenen Lexika des Englischen, ob es einen Eintrag dafür gibt
  - Training eines Transformers, um die anderen MWEs zu bewerten
- Ergebnisse:
  - 144 nicht lexikalisierte MWEs und 243 lexikalisierte MWEs aus dem Abgleich
  - Ca. 80% Korrektheit der Bewertung der automatischen Klassifikation

Francis Bond and Takyuki Kuribayashi. The Japanese Wordnet 2.0. - 1

- Orthographische Varianten
  - Hiragana, Kanji, Katakana, arabische Zahlen, lateinische Buchstaben
  - Displayform ist nicht vorgegeben, eigene Regeln erstellt
  - Manuelle Prüfung
  - Anwendungen:
    - Abdeckung des Lexikons auf Korpora
    - Hilfestellung für Japanisch-Lernende
- Frequenzen für Bedeutungen
  - Basierend auf Annotationen eines Korpus
  - Anwendungen:
    - Anzeige im OMW-Portal nach der Frequenz
    - WSD
- Grammatische Notizen
  - Z.B. Verbal Nouns

Francis Bond and Takyuki Kuribayashi. The Japanese Wordnet 2.0. - 2

- Neue Einträge:
  - Numeral Classifiers (gibt es nicht im Englischen)
  - Pronomen (nicht in PWN)
  - Exklamative (nicht in PWN)
  - Ausdrücke für Zeit/Datum (oft nicht ein Wort im Englischen)
  - Verwandtschaftsbezeichnungen (differenzierter als im Englischen)
  - Andere Konzepte, die es im Englischen nicht gibt, z.B.
    - Heiß – kalt – warm – kühl haben unterschiedliche Wörter, wenn es um das Gefühl oder die Berührung geht
    - Soba
    - Sunto (jährliche Aushandlung der Tarife)
- 770 neue Synsets

Ivelina Stoyanova and Svetlozara Leseva. Expanding the Conceptual Description of Verbs in WordNet with Semantic and Syntactic Information

- Ziel: Anreicherung von Wordnet mit Informationen aus FrameNet und VerbNet, auch sprachübergreifend
- FrameNet:
  - Subkategorisierungsrahmen von Verben (Agent, Theme, Location usw.)
- VerbNet:
  - Syntaktische und semantische Patterns von englischen Verben
- Ergebnisse:
  - Automatisches Mapping der Verbklassen von VerbNet auf Wordnet Synsets
  - Automatisches Mapping der Frames in FrameNet auf Wordnet Synsets
  - Manuelle Evaluation und Übernahme qualitätsgesicherter Information

Erica Biagetti, Chiara Zanchi and Silvia Luraghi. Linking the Sanskrit WordNet to the Vedic Dependency Treebank: a pilot study

- Sanskrit WordNet
- Nutzung einer Baumbank (syntaktische Information) für Argumentstrukturen von Verben
- Ergänzung dieser Information in das Wordnet

Ahti Lohk, Martin Rebane and Heili Orav. An Experiment: Finding Parents for Parentless Synsets by Means of CILI

- Problem: Wordnets enthalten Synsets ohne Hyperonym (z.B. OEWN 582 Fälle)
- Ziel: Automatisch potenzielle Hyperonyme für diese Synsets in mehreren Sprachen (8 Wordnets) finden
- Idee: Nutzung der ILIs, um sprachübergreifende Kandidaten zu finden

Ahti Lohk, Martin Rebane and Heili Orav. An Experiment: Finding Parents for Parentless Synsets by Means of CILI - 2

Z.B.:

WITHOUT PARENT

i29197 odenet-6447-v|['legen', 'stellen', 'tun', '...']

(OEWN equivalent: oewn-01496967-v|['pose', 'set', 'position', '...'])

POSSIBLE PARENT(S):

i30960 not in ODENET

PARENTS FROM OTHER WORDNET(S):

(i29197)->i30960 cow-01850315-v oewn-01854282-v|['displace', 'move']

(i29197)->i30960 enwn-ens-376187 oewn-01854282-v|['displace', 'move']

(i29197)->i30960 estwn-et-85-v oewn-01854282-v|['displace', 'move']

(i29197)->i30960 ewn-01854282-v oewn-01854282-v|['displace', 'move']

(i29197)->i30960 fiwn-01850315-v oewn-01854282-v|['displace', 'move']

(i29197)->i30960 lsg-01850315-v oewn-01854282-v|['displace', 'move']

(i29197)->i30960 oewn-01854282-v oewn-01854282-v|['displace', 'move']

(i29197)->i30960 slownet-eng-30-01850315-v oewn-01854282-v|['displace', 'move']

(i29197)->i30960 wnja-01850315-v oewn-01854282-v|['displace', 'move']

POSSIBLE GRANDPARENT(S):

i33830 not in ODENET

GRANDPARENTS FROM OTHER WORDNET(S):

(i30960)->i33830 estwn-et-128-v oewn-02424173-v|['act']

Svetla Koeva and Dimitar Hristov. Resolving Multiple Hypernymy

- Problem: 1.421 Synsets im PWN haben multiple Hyperonymie-Links
- Methode:
  - Visualisierung der Hierarchien
  - Manuelle Korrekturen
- Ergebnis:
  - Noun Graph ist jetzt ein Baum
  - multiple Hyperonyme nur noch in Ausnahmefällen
  - in diesen Ausnahmefällen muss geprüft werden, ob Polysemie vorliegt und PWN weiter korrigiert werden muss

Melanie Siegel and Johann Bergh. Connecting Multilingual WordNets: Strategies for Improving ILI Classification in OdeNet.

- Problem:
  - OdeNet ist automatisch erzeugt (Merge Approach) und enthält noch Fehler
  - Für die ILIs wurde automatische Übersetzung genutzt, die bei Ambiguitäten zu Fehlern führt
  - z.B. doppelte Zuweisung von ILIs zu Synsets
- Methode: Verbesserung der automatischen Übersetzung durch
  - Richtung Englisch -> Deutsch (DeepL)
  - Hinzufügung der Definitionen zur Übersetzung, um mehr Kontext zu haben
  - Klassifikationsfunktion mit Einbeziehung der Synonyme, falls immer noch Ambiguität besteht (Word2Vec-Modell)
  - englische Verben durch „to“ erweitern („run“ -> „to run“)
- Ergebnis:
  - keine doppelten ILIs mehr
  - 85% Korrektheit der ILIs (manuelle Evaluation)
  - Nebeneffekt Korrektur der POS: 99% korrekt

Ramona Kuehn, Jelena Mitrović and Michael Granitzer. Hidden in Plain Sight: Can German Wiktionary and Wordnets Facilitate the Detection of Antithesis?

- Problem:
  - Antonym-Relationen fehlen in vielen Wordnets
  - sie werden aber gebraucht, um Antithesen zu erkennen
- Methode:
  - reguläre Ausdrücke, um Antonyme aus Wiktionary zu extrahieren
  - OdeNet enthält Antonym-Relationen für 3.049 Wörter
- Ergebnis:
  - Antonyme für 45.499 Wörter
  - (ich habe bereits angefragt, ob wir diese in OdeNet aufnehmen können)

Information aus Deep Learning-Sprachmodellen für Wordnet

Konrad Wojtasik, Arkadiusz Janz and Maciej Piasecki. Wordnet for Data Augmentation in Pretraining for Encoder-Decoder Architecture.

Hugo Gonçalo Oliveira. Studying the Acquisition of WordNet Relations from Pretrained Masked Language Models for Portuguese.

Konrad Wojtasik, Arkadiusz Janz and Maciej Piasecki. Wordnet for Data Augmentation in Pretraining for Encoder-Decoder Architecture.

- Problem:
  - Das polnische Wordnet hat in 40% der Synsets keine Definition und in 60% keine Beispiele
  - Definitionen ambiger Wörter sind oft abhängig vom Kontext
- Idee: Nutzung von großen Sprachmodellen wie GPT und T5, um Definitionen zu generieren
- Methode: Fine-Tuning von T5 (PL) mit Bedeutungen und Beispielen für Wörter, aus Treebanks

Hugo Gonçalo Oliveira. Studying the Acquisition of WordNet Relations from Pretrained Masked Language Models for Portuguese

- Problem: weitere Relationen zwischen Synsets sollen im portugiesischen Wordnet eingepflegt werden
- Idee: Nutzung von vortrainierten Sprachmodellen (BERT) für neue Relationen (Synonym, Antonym, Hyponym, Hyperonym)
- Methode:
  - Anfragen in der Art von “a x2 is a type of x1”
  - Z.B. “a dog is a type of [MASK]” (Ergebnis: “animal”)
- Ergebnisse:
  - Monolinguale Sprachmodelle funktionieren am besten
  - Bessere Ergebnisse für Nomen als für Verben
  - Komplette Automatisierung ist noch zu risikoreich

Wordnets für weitere Sprachen

Francis Bond and Takyuki Kuribayashi. The Japanese Wordnet 2.0.

Peteris Paikens, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde and Laine Strankale. Latvian WordNet.

Luis Chiruzzo, Marvin Agüero-Torales, Aldo Alvarez and Yliana Rodríguez. Initial Experiments for Building a Guarani WordNet.

Aslı Kuzgun, Oğuz Kerem Yıldız and Olcay Taner Yildiz. A CCGbank for Turkish: From Dependency to CCG.

Thierry Declerck, Thomas Troelsgård and Sussi Olsen. Towards a RDF Representation of the Infrastructure consisting in using WordNet(s) as a conceptual Interlingua between multilingual Sign Language Datasets.

Laurette Marais and Laurette Pretorius. Extending the usage of adjectives in the Zulu AfWN.

Luis Morgado Da Costa, František Kratochvíl, George Saad, Benidiktus Delpada, Daniel Simon Lanma, Francis Bond and Natálie Wolfová. Linking SIL Semantic Domains to Wordnet and Expanding the Abui Wordnet through Rapid Words Collection Methodology

Fahad Khan, John P. McCrae, Francisco Javier Minaya Gómez, Rafael Cruz González and Javier E. Díaz-Vera. Some Considerations in the Construction of a Historical Language WordNet.

Peteris Paikens, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde and Laine Strankale. Latvian WordNet

- Problem: Wordnet für Lettisch
- Methode:
  - Merge-Ansatz, basierend auf existierenden Lexikon-Ressourcen für Lettisch und häufigen Wörtern des Lettischen (in einem Text-Korpus)
  - Extraktion der häufigsten Wörter
  - Nutzung eines Thesaurus für diese Wörter
  - Manuelle Bearbeitung von immer drei Annotator*innen
  - Hinzufügen von ILIs
  - Hinzufügen von Relationen (Synonym, Hyponym, Hyperonym, Antonym)
- Ergebnis:
  - 7609 Wörter, 6515 Synsets
  - Fälle, in denen die Sprachen differieren, z.B. anderes Bedeutungsspektrum (“Geschwister” für “Schwestern” und “Brüder” gibt es nicht im Lettischen)

Luis Chiruzzo, Marvin Agüero-Torales, Aldo Alvarez and Yliana Rodríguez. Initial Experiments for Building a Guarani WordNet

- Guarani: indigene südamerikanische Sprache mit wenigen Sprachressourcen
- Methode:
  - Expand Approach
  - Übersetzung von Synsets (bzw. von Lemmata in Synsets)
  - Basierend auf bilingualen Lexika Spanisch – Guarani (Englisch gibt es nicht)
  - Strategien für Polysemien, z.B. die Schnittmenge der Lemmata in zwei Synsets in beiden Sprachen
- Ergebnisse:
  - 6.519 Synsets
  - Precision 0,613 (muss weiter manuell evaluiert werden)

Fahad Khan, John P. McCrae, Francisco Javier Minaya Gómez, Rafael Cruz González and Javier E. Díaz-Vera. Some Considerations in the Construction of a Historical Language WordNet

- Old English WordNet
- manuelle Erstellung von Terminologie, die Emotionen betrifft
- Basis: Spreadsheets mit emotionalen Termen aus einem anderen Projekt
- Anbindung an OEWN-Synsets
- Erweiterung der Informationen in Wordnet:
  - Etymologie als Relation zwischen lexikalischen Elementen
  - grammatisches Geschlecht als Attribut in lexikalischen Elementen
  - zusätzliche Definition für Senses

Thierry Declerck, Thomas Troelsgård and Sussi Olsen. Towards a RDF Representation of the Infrastructure consisting in using WordNet(s) as a conceptual Interlingua between multilingual Sign Language Datasets.

- Ziel: Lexikon von Gebärdensprachen (Deutsch und Griechisch) im Linked-Data-Format (OntoLex-Lemon)
- Methode:
  - Wordnets können in OntoLex-Lemon repräsentiert werden
  - Gebärden sind als Videos repräsentiert
  - Synsets für Gebärden, Link zu OMW (via ILI)
  - Zusammenführung verschiedener Ressourcen

Laurette Marais and Laurette Pretorius. Extending the usage of adjectives in the Zulu AfWN

- African languages Wordnet (AfWN) for Zulu (ZWN) wurde mit Expand-Approach entwickelt (übersetzt)
- Problem: Adjektive im Englischen werden durch andere Konstruktionen in Zulu übersetzt: Verben, Copula, Possessive, mit morphologischen Veränderungen
- Methode:
  - Nutzung einer Grammatik für Zulu
  - Hinzufügung linguistischer Information, z.B. über Nullpronomina

Luis Morgado Da Costa, František Kratochvíl, George Saad, Benidiktus Delpada, Daniel Simon Lanma, Francis Bond and Natálie Wolfová. Linking SIL Semantic Domains to Wordnet and Expanding the Abui Wordnet through Rapid Words Collection Methodology

- Ziele:
  - Für linguistische Feldforschung Wordnet als interessante Ressource vorstellen
  - Abui Wordnet erweitern
- Abui Wordnet:
  - Expansion Approach, seit 2022
  - 1.475 Synsets
  - Manuell geprüft von Muttersprachler
- Methode:
  - Rapid Word Collection Workshops:
    - 10 Tage, 25 Personen, 17.000 Einträge, Übersetzungen in Indonesisch, Malay oder Englisch (aber nicht gleichzeitig)
    - Außerdem Links zu SIL Semantic Domains, einer linguistischen Ontologie mit vielen Sprachen
  - Mapping der SIL Semantic Domains auf Wordnet
- Ergebnisse:
  - 10.780 neue Eintragskandidaten, die nun noch manuell geprüft werden
  - Manuelle Evaluation zeigt ca. 90% Korrektheit

Wordnets für andere Ressourcen

Bolette Pedersen, Sanni Nimb, Nathalie Sørensen, Sussi Olsen, Ida Flörke and Thomas Troelsgård. Reusing the Danish WordNet for a New Central Word Register for Danish - a Project Report.

Alexandre Rademaker, Abhishek Basu and Rajkiran Veluri. Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus.

Joanna Sio and Luis Morgado Da Costa. The Open Cantonese Sense-Tagged Corpus.

Wiktor Walentynowicz and Maciej Piasecki. Wordnet-oriented recognition of derivational relations.

Elena Zotova, Montse Cuadros and German Rigau. Towards the integration of WordNet into ClinIDMap.

Arkadiusz Janz and Marek Maziarz. Data Augmentation Method for Boosting Multilingual Word Sense Disambiguation.

Bolette Pedersen, Sanni Nimb, Nathalie Sørensen, Sussi Olsen, Ida Flörke and Thomas Troelsgård. Reusing the Danish WordNet for a New Central Word Register for Danish - a Project Report

- Aufbau eines zentralen dänischen Lexikons als Open-Source-Ressource für NLP/KI-Anwendungen
- Übertragung von Informationen des dänischen Wordnets(DanNet) in die neue Ressource
- Teilweise Vereinfachung, wie Weglassen seltener Bedeutungen
  - Z.B. „Hühnerbrust“ nur als FOOD, nicht als Teil von ANIMAL
  - Dagegen z.B. „Konstruktion“ als PROCESS und RESULT
- Korrekturen werden wieder in DanNet eingebaut

Alexandre Rademaker, Abhishek Basu and Rajkiran Veluri. Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus

- Problem:
  - “Princeton Annotated Gloss Corpus”: Textkorpus mit Sense-annotierten Wörtern nach PWN
  - Noch nicht alle Wörter im Korpus sind annotiert
  - Definitionen in PWN enthalten ambige Wörter, diese müssen ebenfalls Sense-annotiert werden
- Methode:
  - Parsing der Definitionen mit ERG (HPSG für Englisch), als Ergebnis semantische Repräsentationen (oft mehrere), die gewichtet sind
- Ergebnisse:
  - Inkonsistenzen in PWN wurden entdeckt
  - Korpus wurde vollständig annotiert

Joanna Sio and Luis Morgado Da Costa. The Open Cantonese Sense-Tagged Corpus

- Kantonesisches Wordnet
  - seit 2019 komplett von Hand erstellt, Expand Approach und dann erweitert
  - 5.252 Lemmata in 16.336 Synsets
- Cantonese Wordnet Corpus
  - Korpus von Beispielen für Verb-Bedeutungen nach Wordnet
  - Problem Wortsegmentierung: Was ist ein Wort?
  - Annotationstool IMI
  - Aktuell 300 Sätze
  - 709 neue Bedeutungen (Senses) für Wordnet

Sana Ghanem, Mustafa Jarrar, Radi Jarrar and Ibrahim Bounhas. A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms

- Problem:
  - Aufstellung von Synonym-Listen
  - Synonymie wird in unterschiedlichen Kontexten (Ontologie, Wordnet etc.) unterschiedlich verstanden – Fuzzy Konzept
- Methode:
  - 500 Synsets aus dem arabischen Wordnet
  - Annotiert von vier Linguist*innen
  - Skala von 1 (keine semantische Relation) bis 100 (totale Synonymie)
  - → Fuzzy-Werte für Synonymie
  - Training eines Algorithmus für die Erkennung von Synonymen
  - Extraktion von potentiellen Synonymen aus Wörterbüchern
  - Anwendung des Algorithmus

Wiktor Walentynowicz and Maciej Piasecki. Wordnet-oriented recognition of derivational relations

- Problem:
  - Derivationen in Polnisch, z.B. dom (Haus) – domek (Häuschen)
- Ziel:
  - Automatische Erkennung von derivationalen Relationen in der Sprache
- Methode:
  - Training auf plWordNet-Relationen (derivationale)
  - Bag of Character Vector (jeder Buchstabe ist eine Stelle im Vector)
  - FastText Vector (non-supervised Learning auf Textkorpora, N-Gramme)
- Ergebnis:
  - F1 um die 80%
  - Problem nicht ausgewogene Daten

Elena Zotova, Montse Cuadros and German Rigau. Towards the integration of WordNet into ClinIDMap.

- Problem:
  - Es gibt Ontologien für medizinische Konzepte (z.B. ClinIDMap), die verknüpft werden sollen: Interoperabilität zwischen Ressourcen
  - Es existieren nach diesen Ontologien annotierte Korpora, allerdings meist nur für Englisch
- Methode:
  - Wikidata-Items mit SPARQL zugreifen, die einen Link zu Wordnet-Synsets haben
  - Wordnet Domains (eine Erweiterung von Wordnet) nutzen, um medizinische Konzepte zu finden
  - mit ClinIDMap annotierte Korpora mit Wordnet annotieren und so die Label miteinander verbinden
  - Übertragung (über ILI) auf viele Sprachen
  - Übertragung der ILIs auf annotierte Korpora
- Ergebnis:
  - 5 medizinische spanische Korpora sind mit Wordnet-Synsets annotiert

Arkadiusz Janz and Marek Maziarz. Data Augmentation Method for Boosting Multilingual Word Sense Disambiguation

- Problem:
  - Word Sense Disambiguation
  - Ansätze mit neuronalen Netzen funktionieren nicht gut für Sprachen mit wenigen Ressourcen
  - Neuronale Netze benötigen dafür große Korpora, die mit Bedeutungen annotiert sind
- Idee:
  - Nutzung des polnischen Wordnets, das mit dem englischen stark verbunden ist (über ILI)
  - Nutzung der Beispiele und Definitionen im polnischen Wordnet zum Training

Wordnet-Information für die Analyse von Deep Learning-Sprachmodellen

Filip Klubička and John Kelleher. Probing Taxonomic and Thematic Embeddings for Taxonomic Information.

Wondimagegnhue Tufa, Lisa Beinborn and Piek Vossen. A WordNet View on Crosslingual Transformers.

Filip Klubička and John Kelleher. Probing Taxonomic and Thematic Embeddings for Taxonomic Information

- Ausgangsproblem:
  - Neuronale Netze mit Word Embeddings sind der aktuelle Standard bei NLP-Aufgaben
  - Aber welche Information verbirgt sich in den Vektoren?
- Ziel:
  - Herausfinden, inwieweit taxonomische Information in den Vektoren kodiert ist.
- Aufgabe:
  - von zwei Wörtern herausfinden, was das Hyperonym und was das Hyponym ist.
- Testdatensatz kommt direkt aus Wordnet:
  - Hyperonym - Hyponym - Paare
- Ergebnisse:
  - Embeddings enthalten taxonomische Information
  - Man kann Word2Vec-Modelle auf Wordnet-Daten trainieren, die dann noch mehr taxonomische Information haben

Wondimagegnhue Tufa, Lisa Beinborn and Piek Vossen. A WordNet View on Crosslingual Transformers

- Wordnet repräsentiert Relationen zwischen Wörtern und Konzepten
- Transformer repräsentieren Wörter im Kontext, lassen aber die Konzepte implizit
  - Basis ist die durchschnittliche Wahrscheinlichkeit, sodass seltenere Lesarten von Wörtern eher vernachlässigt werden. Z.B.: "star"
- Neuere Modelle (BERT et al.) repräsentieren Wörter auch abhängig vom Kontext.
  - Es gäbe also zwei verschiedene Vektoren für "star", abhängig vom Kontext.
- Sprachübergreifende Modelle (XPTLMs) wie XLM-RoBERTa sind komplexer, denn dort gibt es auch noch sprachübergreifende Ambiguität.
  - Die sprachübergreifenden Modelle nutzen ein Vokabular für alle Sprachen.
  - "star" im Niederländischen bedeutet z.B. "unflexibel".
- Da die meisten Wordnets mit der expand-Methode aufgebaut worden sind, gibt es dort wenig sprachübergreifende Ambiguität.
  - XPTLMs eröffnen neue Möglichkeiten für Wordnets, die auf expand basieren: Neue Konzepte
- Frage: Wie erfassen Sprachmodelle die Relation zwischen Wort und Konzept im Fall von Polysemie?
  - Experimente mit Englisch, Deutsch und Niederländisch, anschließen Arabisch und Amharisch
- Experiment:
  - Datenbasis: XLEnt Datensatz (El-Kishky et al., 2021) mit 160 Millionen mit Englisch alignierten NE-Paaren in 120 Sprachen
  - Auswahl von 21 NEs: Tokens in allen drei Sprachen, polyseme und nicht ambige Beispiele, z.B. „Mercury“ (Ort, Organisation und auch Person)
- Ergebnis: Die Einbeziehung von mehr Sprachen bei der Desambiguierung ist effektiv. Wenn man mit mehr Sprachen trainiert, wird der Fundus an Konzepten erweitert, was mit der expand-Methode in Wordnet vergleichbar ist.

Oscar Sainz, Oier Lopez de Lacalle, Eneko Agirre and German Rigau. What do Language Models know about word senses? Zero-Shot WSD with Language Models and Domain Inventories.

- Problem: WSD
  - Z.B. „The medicine can only be obtained with a prescription."
  - "prescription" gehört zu vier verschiedenen Synsets
- Methode:
  - Nutzung von Inferenz:
    - Z.B. „Two men on bicycles competing in a race" → "People are riding bikes."
  - Training von Transformern auf verfügbaren Inferenz-Datensätzen
- Ergebnis:
  - Language Models haben eine gewisse Vorstellung von Bedeutungen

Tools für die Wordnet-Entwicklung

Oğuzhan Kuyrukçu, Ezgi Sanıyar and Olcay Taner Yildiz. StarNet: A WordNet Editor Interface.

Francis Bond, Luis Morgado da Costa, Ewa Rudnicka, Alexandre Rademaker and Michael Wayne Goodman. Documenting the Open Multilingual Wordnet.

Oğuzhan Kuyrukçu, Ezgi Sanıyar and Olcay Taner Yildiz. StarNet: A WordNet Editor Interface

Figure 1: Interface of Literal Matcher with the synonyms of *red* in Turkish

Figure 2: Interface of Synset Matcher; the first and final match example of red

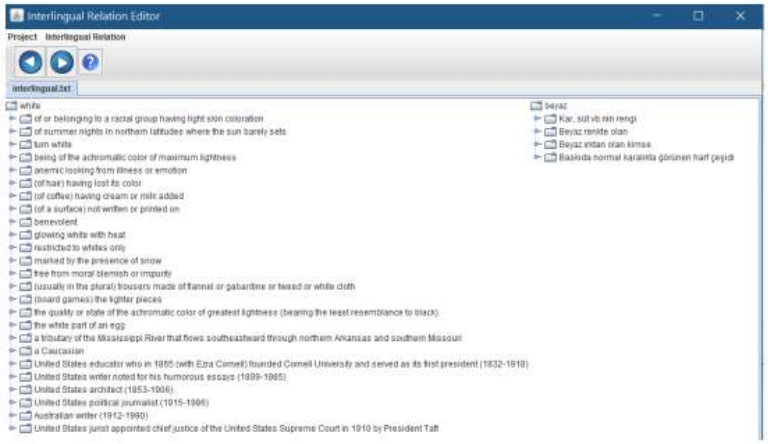


Figure 3: Interface of Interlingual Matcher with the English-Turkish synsets of white

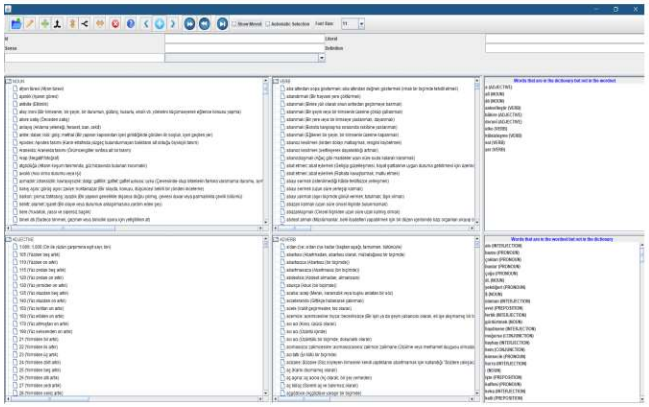


Figure 4: Interface of Hypernym Editor

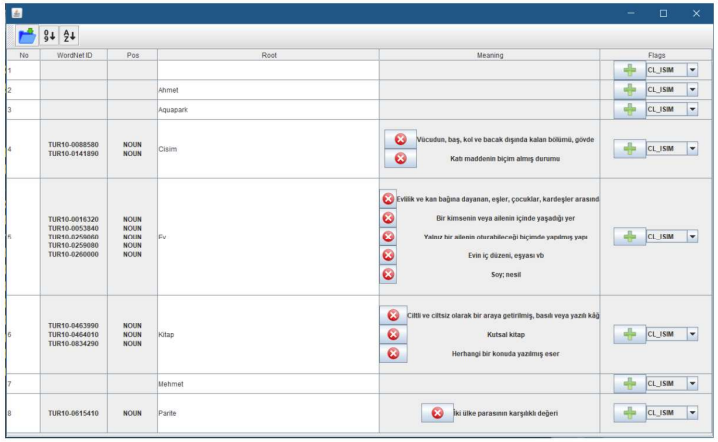


Figure 6: Interface of Dictionary Editor

Francis Bond, Luis Morgado da Costa, Ewa Rudnicka, Alexandre Rademaker and Michael Wayne Goodman. Documenting the Open Multilingual Wordnet.

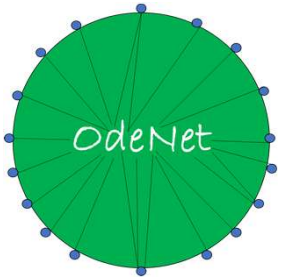
- Open Multilingual Wordnet:
  - Plattform für Wordnets verschiedener Sprachen
  - gemeinsames Format
  - Software, mit der man die Wordnets zugreifen kann
  - offene Lizenzen
- Probleme:
  - Inkonsistenzen in der Benennung von Relationen und POS Tags
  - Veraltete Dokumentationen
- Methode:
  - Dokumentation der semantischen Relationen
  - Dokumentation der OMW Strukturen (Relationen, POS, Definitionen, Beispiele, …)
  - OMW Interface-Dokumentation
- Ergebnis:
  - https://omwn.org/docs.html

Zusammenfassung

- Wordnet wird seit vielen Jahren gepflegt und erweitert
- Multilingualität ist ein wesentliches Prinzip
- Aktuelle Arbeiten in den Bereichen:
  - Korrekturen und Erweiterungen
  - Information aus Deep Learning-Sprachmodellen für Wordnet
  - Wordnets für weitere Sprachen
  - Wordnets für andere Ressourcen
  - Wordnet-Information für die Analyse von Deep Learning-Sprachmodellen

Wordnet in Darmstadt

- OdeNet: offenes deutsches Wordnet
- Ukrajinet: Wordnet für die ukrainische Sprache



Vielen Dank für die Aufmerksamkeit!

melanie.siegel@h-da.de