ODENET
EIN DEUTSCHER BEITRAG ZUR „MULTILINGUAL OPEN WORDNET INITIATIVE“

Melanie Siegel

Agenda

- Projektidee und Projektziele
- Das Projekt OpenThesaurus
- Related Work und Optionen zur Umsetzung
- Arbeiten mit der deutschen Sprache
- Von OpenThesaurus-Synonymen zu WordNet Synsets, lexikalischen Einträgen und Relationen
- Weitere Entwicklungen
- Erste Zahlen
- Diskussion und Pläne

PROJEKTIDEE UND PROJEKTZIELE

- Francis Bond, am 2.2.2017:
  „Dear Melanie,
  I joined a CLARIN meeting where we discussed wordnets, and it was pretty clear that Erhard will not open up Germanet. So, shall we build a new one :-)?
  ...
  Then we would need some kind of wordnet by April."

- Entwicklungszeit des englischen WordNets: 30 Jahre
- Entwicklungszeit von GermaNet: 20 Jahre

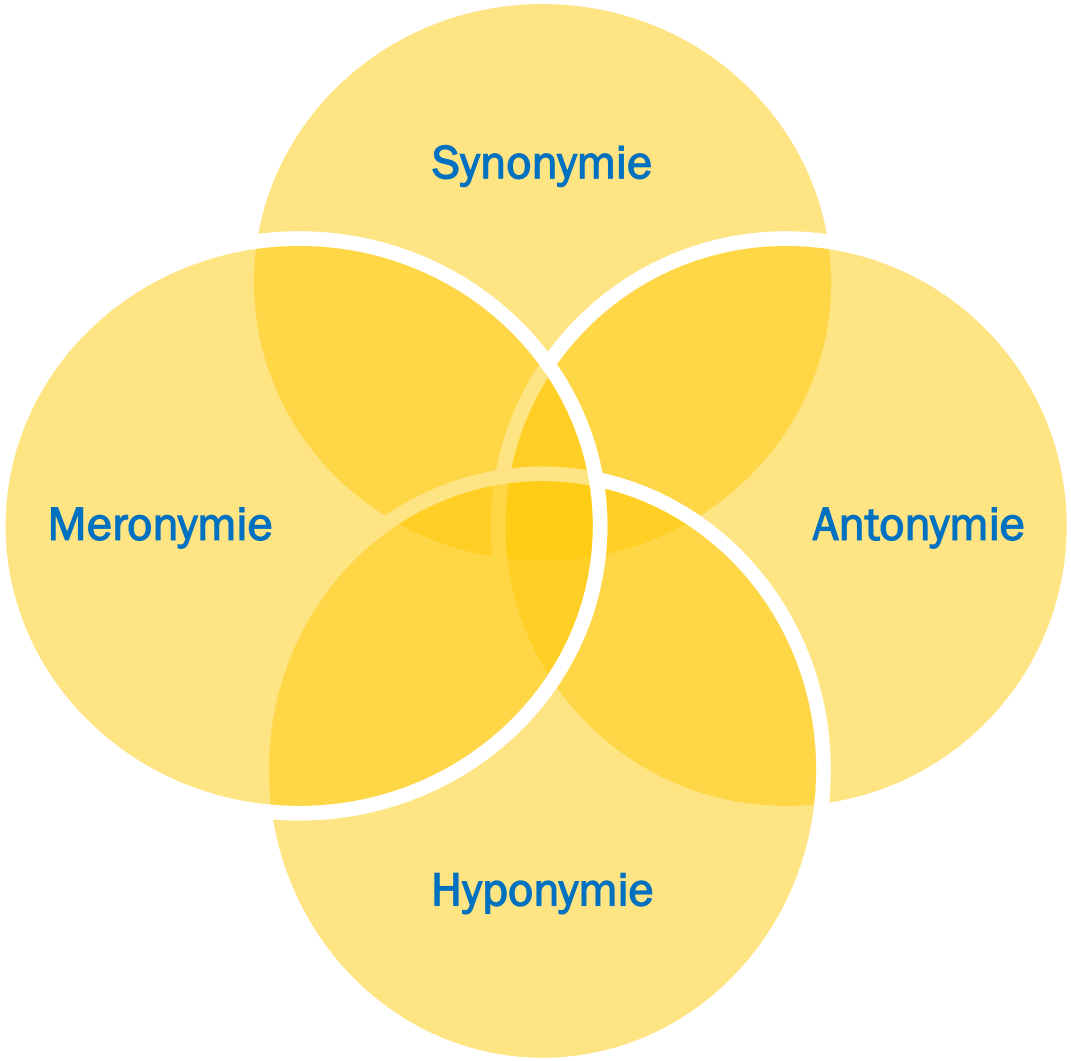
Was ist WordNet?

- 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*

Was ist WordNet?

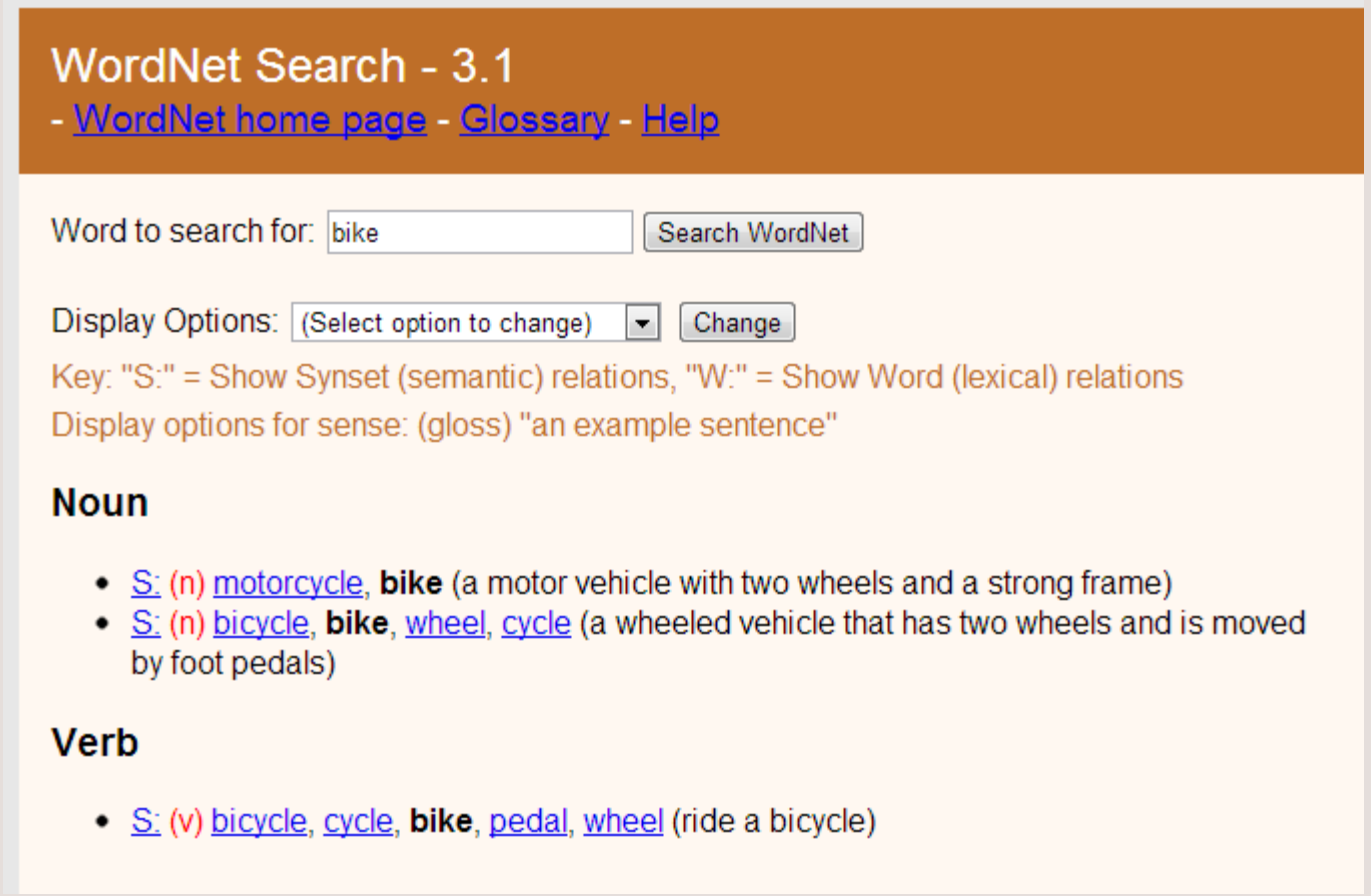
- Organisation lexikalischen Wissens nach Wortbedeutungen
- Organisation in *Synsets*
  - *Mengen von Synonymen*
  - *Lesarten*
  - *Definitionen*

Lexikalische Beziehungen in WordNet (Basis)



Wie wird WordNet genutzt?

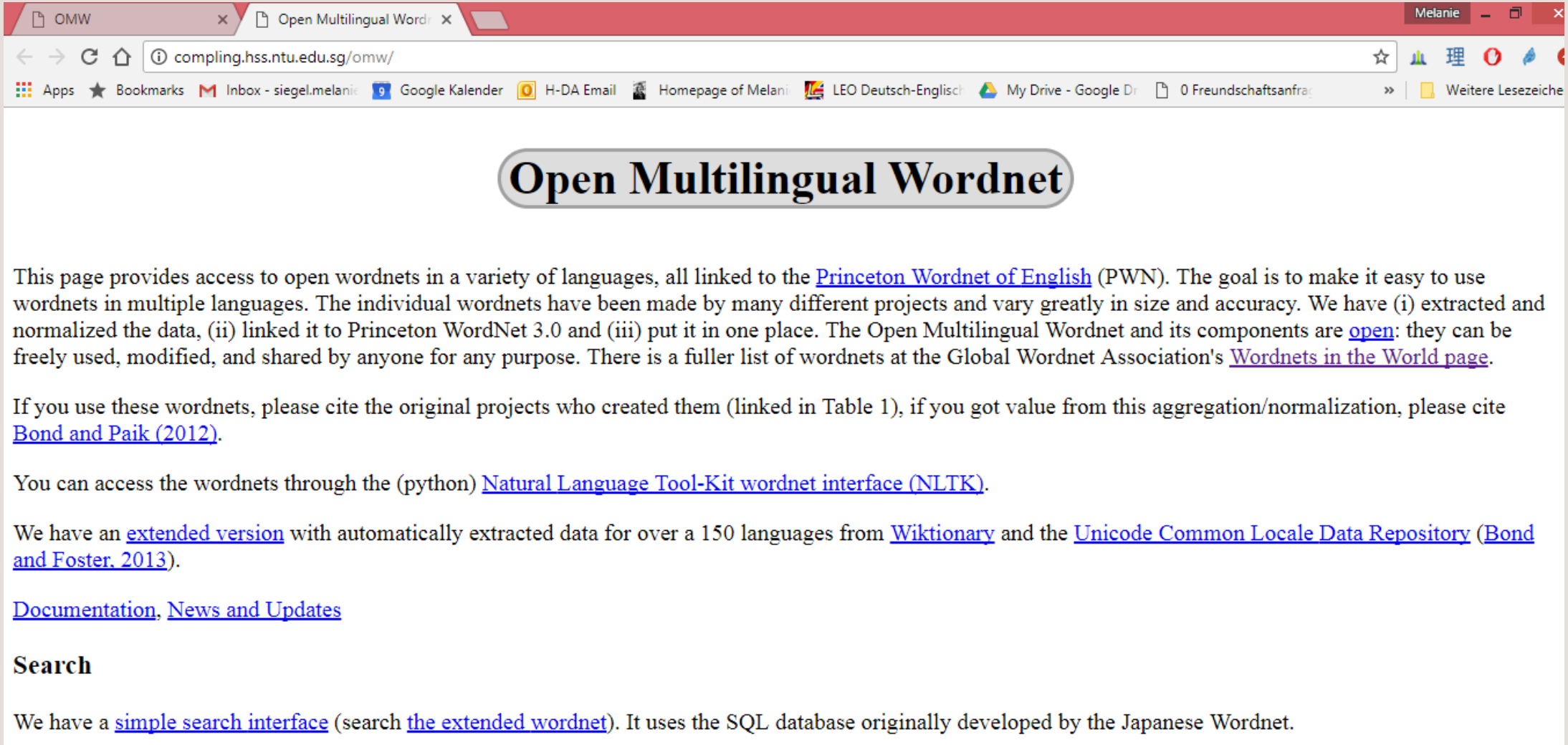
- Sprachtechnologie-Anwendungen
  - *Lesarten-Disambiguierung*
  - *Informationserschließung und Informationsextraktion*
  - *Linguistische Annotation von Sprachdaten*
  - *Textklassifikation und automatische Textzusammenfassung*
  - *Entwicklung von Werkzeugen für Sprachanalyse und Sprachgenerierung*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


WordNet Search - 3.1
- WordNet home page - Glossary - Help
Word to search for: bike
Search WordNet
Display Options: (Select option to change)
Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"
Noun
S: (n) motorcycle, bike (a motor vehicle with two wheels and a strong frame)
S: (n) bicycle, bike, wheel, cycle (a wheeled vehicle that has two wheels and is moved by foot pedals)
Verb
S: (v) bicycle, cycle, bike, pedal, wheel (ride a bicycle)

XML-Format, Zugriff über Python/NLTK möglich

- <LexicalEntry id ='w35249'>
  <Lemma writtenForm='bike' partOfSpeech='v'/>
  <Sense id='w35249_01935476-v' synset='eng-10-01935476-v'/>
  </LexicalEntry>
- <Synset id='eng-10-01935476-v'>
  <Definition>ride a bicycle</Definition>
  <SynsetRelation target='eng-10-01955984-v' relType='hypernym'/>
  <SynsetRelation target='eng-10-01935846-v' relType='hyponym'/>
  <SynsetRelation target='eng-10-01935953-v' relType='hyponym'/>
  </Synset>

Internationalisierung: Open Multilingual WordNet (OMW)



WordNets in OMW

Wordnet	Lang	Synsets	Words	Senses
Albanet	als	4,675	5,988	9,599
Arabic WordNet (AWN v2)	arb	9,916	17,785	37,335
BulTreeBank Wordnet (BTB-WN)	bul	4,959	6,72	8,936
Chinese Open Wordnet	cmn	42,312	61,533	79,809
Chinese Wordnet (Taiwan)	qcn	4,913	3,206	8,069
DanNet	dan	4,476	4,468	5,859
Greek Wordnet	ell	18,049	18,227	24,106
Princeton WordNet	eng	117,659	148,73	206,978
Persian Wordnet	fas	17,759	17,56	30,461
FinnWordNet	fin	116,763	129,839	189,227
WOLF (Wordnet Libre du Français)	fra	59,091	55,373	102,671
Hebrew Wordnet	heb	5,448	5,325	6,872
Croatian Wordnet	hrv	23,12	29,008	47,9
IceWordNet	isl	4,951	11,504	16,004
MultiWordNet	ita	35,001	41,855	63,133
ItalWordnet	ita	15,563	19,221	24,135
Japanese Wordnet	jpn	57,184	91,964	158,069
Multilingual Central Repository	cat	45,826	46,531	70,622
Multilingual Central Repository	eus	29,413	26,24	48,934
Multilingual Central Repository	glg	19,312	23,124	27,138
Multilingual Central Repository	spa	38,512	36,681	57,764
Wordnet Bahasa	ind	38,085	36,954	106,688
Wordnet Bahasa	zsm	36,911	33,932	105,028
Open Dutch WordNet	nld	30,177	43,077	60,259
Norwegian Wordnet	nno	3,671	3,387	4,762
Norwegian Wordnet	nob	4,455	4,186	5,586
plWordNet	pol	33,826	45,387	52,378
OpenWN-PT	por	43,895	54,071	74,012
Romanian Wordnet	ron	56,026	49,987	84,638
Lithuanian WordNet	lit	9,462	11,395	16,032
Slovak WordNet	slk	18,507	29,15	44,029
sloWNet	slv	42,583	40,233	70,947
Swedish (SALDO)	swe	6,796	5,824	6,904
Thai Wordnet	tha	73,35	82,504	95,517

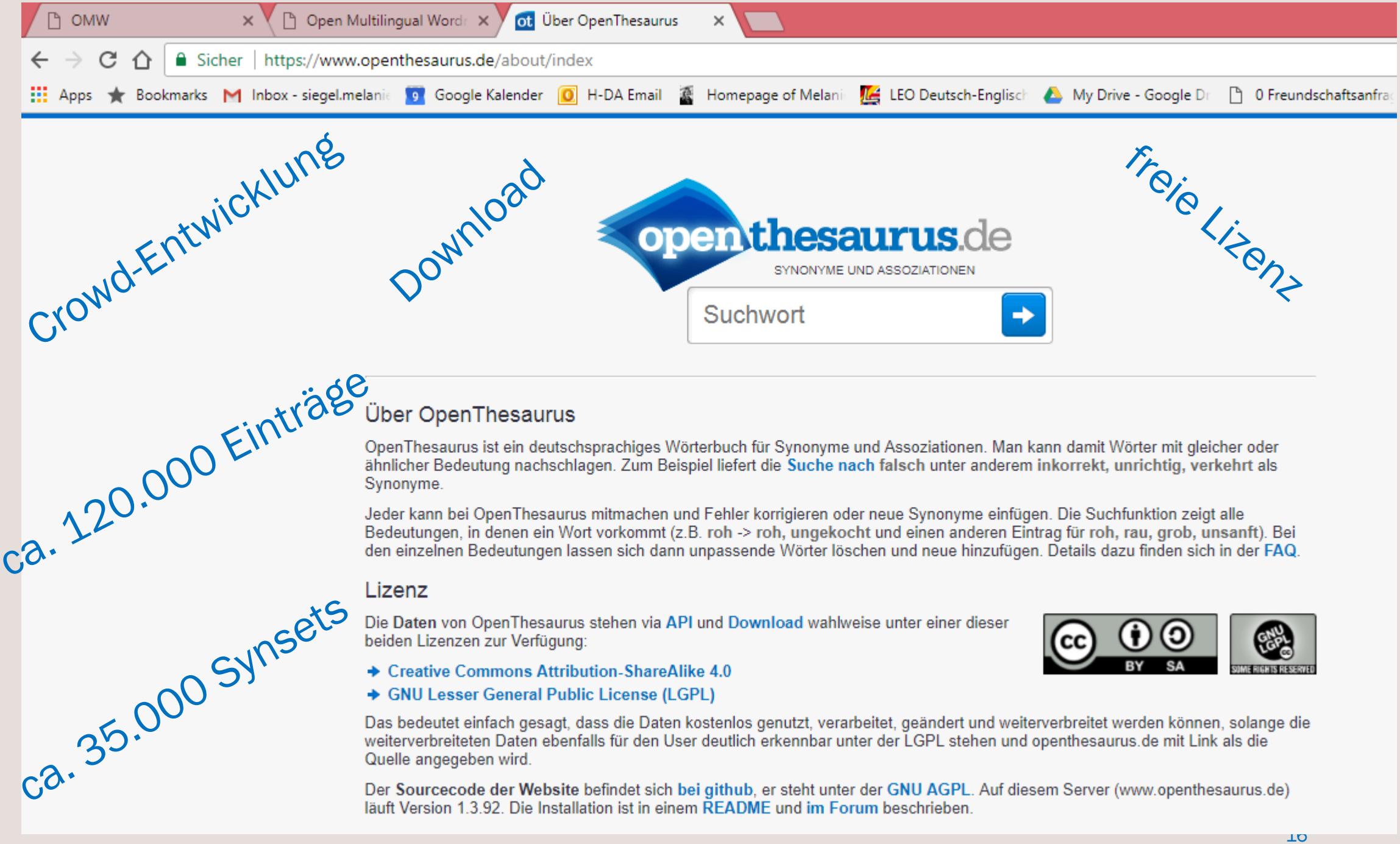
WordNets in (Weiter-)Entwicklung in OMW

- English — Princeton Wordnet 3.0 (3.0)
- German — Offenes Deutsches WordNet (1.2)
- Hindi — Hindi Wordnet (1.4)
- Marathi — Marathi Wordnet (1.4)
- Polish — plWordnet (Słowosieć) (3.0)
- Sanskrit — Sanskrit Wordnet (1.4)

Projektziele und Projektidee

- Ziel:
  - *offenes deutsche WordNet mit einer ersten Version in 3 Monaten erstellen*
- Idee:
  - *automatische Kompilierung aus existierender Ressource, Anreicherung mit linguistischer Information und Anbindung an OMW durch Übersetzung*

DAS PROJEKT OPENTHESAURUS


OMW
Open Multilingual Word
Über OpenThesaurus
Sicher
https://www.openthesaurus.de/about/index
Apps
Bookmarks
Inbox - siegel.melani
Google Kalender
H-DA Email
Homepage of Melani
LEO Deutsch-Englisch
My Drive - Google Dr
0 Freundschaftsanfra
Crowd-Entwicklung
Download
freie Lizenz
openthesaurus.de
SYNONYME UND ASSOZIATIONEN
Suchwort
ca. 120.000 Einträge
ca. 35.000 Synsets
Über OpenThesaurus
OpenThesaurus ist ein deutschsprachiges Wörterbuch für Synonyme und Assoziationen. Man kann damit Wörter mit gleicher oder ähnlicher Bedeutung nachschlagen. Zum Beispiel liefert die Suche nach falsch unter anderem inkorrekt, unrichtig, verkehrt als Synonyme.
Jeder kann bei OpenThesaurus mitmachen und Fehler korrigieren oder neue Synonyme einfügen. Die Suchfunktion zeigt alle Bedeutungen, in denen ein Wort vorkommt (z.B. roh -> roh, ungekocht und einen anderen Eintrag für roh, rau, grob, unsanft). Bei den einzelnen Bedeutungen lassen sich dann unpassende Wörter löschen und neue hinzufügen. Details dazu finden sich in der FAQ.
Lizenz
Die Daten von OpenThesaurus stehen via API und Download wahlweise unter einer dieser beiden Lizenzen zur Verfügung:
Creative Commons Attribution-ShareAlike 4.0
GNU Lesser General Public License (LGPL)
Das bedeutet einfach gesagt, dass die Daten kostenlos genutzt, verarbeitet, geändert und weiterverbreitet werden können, solange die weiterverbreiteten Daten ebenfalls für den User deutlich erkennbar unter der LGPL stehen und openthesaurus.de mit Link als die Quelle angegeben wird.
Der Sourcecode der Website befindet sich bei github, er steht unter der GNU AGPL. Auf diesem Server (www.openthesaurus.de) läuft Version 1.3.92. Die Installation ist in einem README und im Forum beschrieben.

RELATED WORK UND OPTIONEN ZUR UMSETZUNG

Manuelle Entwicklung

- PWN (Englisch) (Fellbaum 1998)
- GermaNet (Deutsch) (Hamp et al. 1997)
- viele Jahre Entwicklungszeit
- präzise und verlässliche Einträge
- muss bei Anwendung auf eine Domäne dennoch weiter angepasst werden

Expand Approach

- Japanese WordNet (Isahara et al. 2008)
- ausgehend von Synsets im englischen PWN
- Lexikoneinträge zu diesen Synsets erzeugen
- manuelle Prüfung
- Unterschiede in der Konzeptstruktur zwischen Japanisch und Englisch → nicht alle Synsets konnten übertragen werden

Merge Approach

- Russian WordNet (Alexeyevsky and Temchenko 2016)
- Monolinguales Lexikon mit Wortdefinitionen als Basis
- Nutzung der Definitionsstruktur für Hyponymie-Beziehungen:
  - *WORD:HYPERNYM …*

Crowdsourcing

- YARN (Braslavski et al. 2016)
- Aufbau eines großen russischen Thesaurus durch Crowdsourcing
- direkt im WordNet-Format
- keine Verlinkung zu OMW

Umsetzung in OdeNet

- Nutzung der Crowdsourcing-Ressource OpenThesaurus
- Umformatierung in WordNet-Format
- Anreicherung mit linguistischer Information
- Verbindung zu OMW-Konzepten
- Nutzung/Übertragung der Hierarchie-Informationen in OMW-Konzepten

ARBEITEN MIT DER DEUTSCHEN SPRACHE

Lexikalische Ambiguität im Deutschen

- viele Beispiele (wie in allen Sprachen), z.B. „Mutter“ (Person, Schraube), „umfahren“
- häufig nicht parallel zum Englischen
- meistens innerhalb der (groben) syntaktischen Kategorie (anders als Englisch)
  - *Ausnahme z.B. „verlegen“ (V + ADJ)*

Mehrwortlexeme

- Support Verb Constructions, z.B. „Abschied nehmen“, „in Rechnung stellen“
- Reflexiv-Konstruktionen, z.B. „sich waschen“
- Idiome, z.B. „das geht auf keine Kuhhaut“
  - *Bestimmung der syntaktischen Kategorie ist schwierig*

Komposita

- Produktiv und komplex
  - *Donaudampfschifffahrtskapitänsmütze*
- Ambig
  - *Barbezahlung*
  - *Glücksautomaten*
- Im regulären Fall Hyponymie-Beziehung zum Kopf
  - *Wassereis - Eiswasser*

VON OPENTHESAURUS-SYNONYMEN ZU WORDNET SYNSETS, LEXIKALISCHEN EINTRÄGEN UND RELATIONEN

Format der OpenThesaurus-Exportdatei

- knüpfen;spinnen;wirken;weben;flechten
- Flügel;Tragfläche;Flugzeugflügel
- Mobilität;Unabhängigkeit;Beweglichkeit
- Geflunker;Erfindung;Jägerlatein;Windei;Schwindelei;Lügengeschichte;Ammenmärchen;Aufschneiderei;Münchhausiade;Münchhauseniade;Bluff;Flunkerei;Räuberpistole;Seemannsgarn;Erdichtung;Märchen;Anglerlatein;Schwindel
- Paralogismus;Fehlannahme;non sequitur;Missverständnis;Irrglaube;Denkfehler;klarer Fall von denkste;falsche Annahme;Trugschluss;Fehlschluss;Irrtum
- Hausschwein;Borstenvieh;Wutz;Schwein
- Streubreite;Streuung;Varianz;Standardabweichung
- einwandern;zuwandern;immigrieren

Ziel: drei Lexikoneinträge und ein Synset-Eintrag

```
<LexicalEntry id="w39185">
   <Lemma writtenForm="Mobilität" partOfSpeech="n"/>
   <Sense id="w39185_9784-n" synset="odenet-9784-n"></Sense>
</LexicalEntry>

<LexicalEntry id="w35624">
   <Lemma writtenForm="Unabhängigkeit" partOfSpeech="n"/>
   <Sense id="w35624_9784-n" synset="odenet-9784-n"></Sense>
</LexicalEntry>

<LexicalEntry id="w33556">
   <Lemma writtenForm="Beweglichkeit" partOfSpeech="n"/>
   <Sense id="w33556__9784-n" synset="odenet-9784-n"></Sense>
</LexicalEntry>


<Synset id="odenet-9784-n" ili="i62097“ partOfSpeech="n"
            dc:description="the quality of moving freely">
   <SynsetRelation target='odenet-23172-n' relType='hypernym'/>
</Synset>
```

Synset id="odenet-9784-n"

- Input-Datei als csv einlesen und jedem Synset eine eindeutige ID zuweisen
  - *([Mobilität, Unabhängigkeit, Beweglichkeit], odenet-9784-n)*

partOfSpeech="n"

- Python-Library TextBlob_DE
- POS für das erste mögliche Lexem
- Abbildung der TextBlob-POS auf n – v - a
- Bei Mehrwortlexemen POS des letzten Wortes
  - *z.B. „sich vergnügen“*

ili="i62097“

- Übersetzung der Lexeme im Synset mit google translate (Python-Anbindung goslate)

  >>> print(gs.translate('Mobilität, Unabhängigkeit, Beweglichkeit', 'en'))

  Mobility, independence, mobility
  - *Vorteil der statistischen Übersetzung: andere Wörter bilden den Kontext*
- Auswahl der Übersetzung:
  - *die, die mehrfach vorkommt*
  - *die erste, die ich im englischen WN finde*
- Nachschlagen der ili im englischen WN:
  - *ili für die erste Bedeutung des Wortes*
- Knapp 20.000 Synsets konnten so eine ili bekommen

dc:description="the quality of moving freely"

- Wenn es eine ili gibt:
- englische Definition mit dem NLTK-Interface zugreifen und aufnehmen

<SynsetRelation target='odenet-23172-n, relType='hypernym'/>

- Wenn es eine ili gibt:
- Im englischen Synset:

  <SynsetRelation target='eng-10-04723816-n' relType='hypernym'/>
- Relation mit dem NLTK-Interface zugreifen
- über ili die deutsche Entsprechung finden
- die OdeNet-Synset-ID für die Relation eintragen

Bedeutungen für die lexikalischen Einträge zusammenführen und Word-IDs vergeben

- ("Beweglichkeit","odenet-8203-n),
  ("Beweglichkeit","odenet-9784-n“),
  ("Beweglichkeit","odenet-11420-n"),
  ("Beweglichkeit","odenet-19087-n")

- <LexicalEntry id="w33556">
  <Lemma writtenForm="Beweglichkeit" partOfSpeech="n"/>
  <Sense id="w33556_8203-n" synset="odenet-8203-n"></Sense>
  <Sense id="w33556_9784-n" synset="odenet-9784-n"></Sense>
  <Sense id="w33556_11420-n" synset="odenet-11420-n"></Sense>
  <Sense id="w33556_19087-n" synset="odenet-19087-n"></Sense>
  </LexicalEntry>

WEITERE ENTWICKLUNGEN

POS-Korrekturen

- Problem: Syntaktische Kategorie (POS) in Mehrwortlexemen ist schwer zu bestimmen
  - *z.B. : „zur selben Zeit“*
- Evaluation: Nur 77% Korrektheit
- Test: Synsets mit Lexikoneinträgen, die unterschiedliche POS haben
- Manuelle Korrektur (mit Hilfsmechanismen)
- Erneute Evaluation: 90 % Korrektheit
- weitere Ideen:
  - *Suche nach Lexemen, die auf –ung, -heit, -keit enden, aber kein Nomen sind*
  - *Suche nach Lexemen, die auf –lich enden, aber kein Adjektiv sind*
  - *Suche nach großgeschriebenen Lexemen, die kein Nomen sind*
  - *Suche nach Lexemen mit ili, bei denen das englische Synset ein anderes POS hat*

Domänenarbeiten

- Domänen „Lageberichte großer Unternehmen“ und „Projektmanagement“:
  - *Aufnahme neuer Lexeme*
  - *Ergänzungen für existierende Lexeme:*
    - deutsche Definitionen
    - ili
    - Relationen

Annotation von Wörtern aus dem Grundwortschatz Deutsch

- Wörter von hier: https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Grundwortschatz
- Wörter, die noch nicht in Odenet sind, aufgenommen (außer Funktionswörtern)
- Wörter automatisch annotiert:
- <LexicalEntry id="w178" **dc:type="basic_German"** confidenceScore="1.0">
  <Lemma writtenForm="wahr" partOfSpeech="a"/>
  <Sense id="w178_35-a" synset="odenet-35-a"></Sense>
  <Sense id="w178_6674-a" synset="odenet-6674-a„></Sense>
  <Sense id="w178_9353-a" synset="odenet-9353-a"></Sense>
  </LexicalEntry>
- Begonnen, die dazugehörigen Synsets zu erweitern

Deutsche Komposita für die Hyponymie-Relation

- Reguläre Komposita haben eine Hyponymie-Beziehung zum Kopf
  - *Eiswasser*
  - *Wassereis*
- Der Großteil der deutschen Komposita sind regulär
- Idee: Ausnutzung für das Einfügen von Hyponymie-Beziehungen in Synsets

Deutsche Komposita für die Hyponymie-Relation

- Implementierung einer einfachen Kompositaanalyse für deutsche Nomen
  - *Basis: Nomenliste aus dem TIGER-Korpus*
  - *früher implementierte Lemmatisierungskomponente*
  - *Python-Implementierung*
- Analyse aller Lemmata in OdeNet, Auflistung aller Komposita und ihrer Teile
- Suche nach Synset-IDs für den jeweils letzten Teil (Kopf)
- wenn das Kompositum und auch der Kopf nur eine Bedeutung in OdeNet haben,
  - *Hinzufügung einer Hyponymie-Relation*

Deutsche Komposita für die Hyponymie-Relation - Ausnahmen

- systematische Untersuchung von Ausnahmefällen
  - *Fachterminus ist synonym zu Terminus → keine Eintragung, wenn die IDs gleich sind*
  - *Zauberei ['Zauber', 'Ei'] --> alle Relationen zu "Ei" gestrichen*
  - *ist Nichtfiktion Fiktion? (eher Antonym)*
  - *Nichtraucher ist antonym zu Raucher, aber Nichteisenmetall ist ein Metall → alle mit „Nicht“ durchgesehen*
- unklare Fälle (ausgelassen)
  - *ist eine Pseudo-Dokumentation eine Dokumentation?*
  - *ist eine Scheinschwangerschaft eine Schwangerschaft? --> Scheinehe? Scheinfirma?*
  - *ist Fruchtfleisch Fleisch?*

Deutsche Komposita für die Hyponymie-Relation – Ergebnisse

- ca. 5.000 neue Hyponymie-Relationen eingefügt
- z.B.:
  - *Musiktheaterstück* *['Musik', 'Theaterstück']*
    - ['Singspiel', 'Musiktheaterstück', 'Musical']
    - ('hypernym', 'odenet-3461-n')
  - *Theaterstück ('odenet-3461-n‘)*
    - ['Schauspiel', 'Stück', 'Theaterstück', 'Spiel', 'Bühnenstück', 'Drama', 'Repertoirestück']
  - *Künstlerwerkstatt['Künstler', 'Werkstatt']*
    - ['Studio', 'Atelier', 'Künstlerwerkstatt']
    - ('hypernym', 'odenet-7839-n')
  - *Werkstatt ('odenet-7839-n')*
    - ['Betrieb', 'Werkstatt']

ERSTE ZAHLEN

Größe

- 120.000 Lexikoneinträge
- 36.225 Synsets
- 19.870 Verknüpfungen zum multilingualen WordNet

Relation	Anzahl
Hypernyme	9.907
Hyponyme	1.601
Holonyme	748
Meronyme	371
Antonyme	2.653

Erste Evaluation

- Je 30 Verben, Nomen, Adjektive, zufällig ausgewählt
- POS-Zuordnung korrekt in 90 % der Fälle (nach der ersten Runde von Korrekturen)
- 61 von 90 Synsets hatten ili-Links
  - *davon waren 87% korrekt*
- 23 von 90 Synsets (26%) hatten Relationen (vor der Kompositaanalyse)
  - *davon waren 3 falsch (87 % korrekt)*

DISKUSSION UND PLÄNE

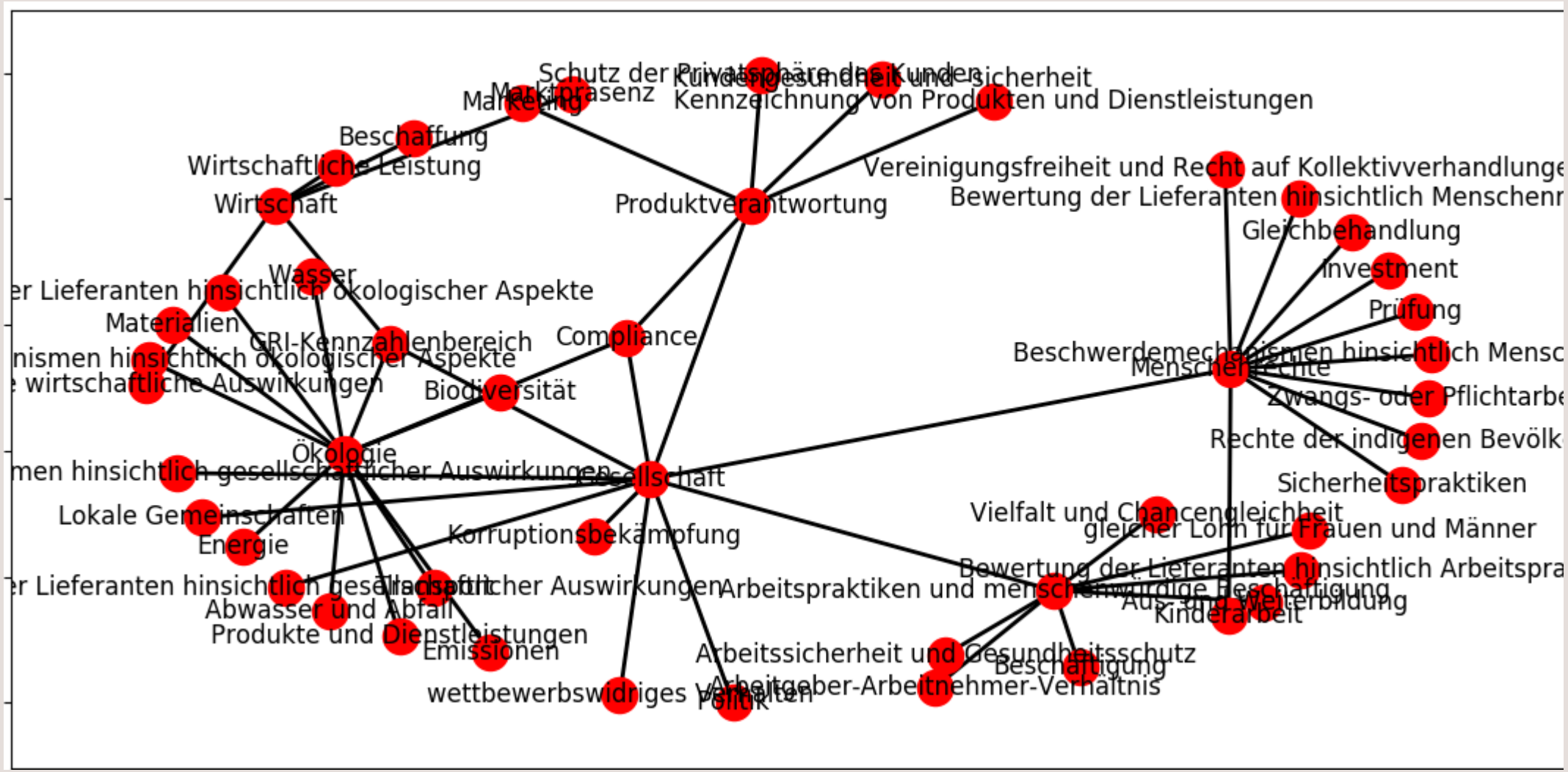
Fazit

- OMW bietet eine große Chance, lexikalische Ressourcen in vielen Sprachen zu erstellen
  - *open-source*
  - *standardisiertes Format*
  - *standardisierte Schnittstelle*
- Es ist möglich, mit existierenden Ressourcen, mit Sprachtechnologie und in kurzer Zeit ein neues WordNet zu erstellen
  - *akzeptable Qualität*
  - *hohe Abdeckung*
  - *Erweiterbarkeit*
  - *Zugreifbarkeit*

Pläne

- Einbettung in NLTK
- weitere Korrekturen und Ergänzungen
  - *POS*
  - *deutsche Definitionen durch automatische Übersetzung der englischen*
  - *deutsche Definitionen aus anderen Quellen (Wikitionary?)*
  - *automatisches Einfügen von Rück-Relationen (z.B. Hyponym – Hyperonym, Antonymie)*
  - *Hierarchie-Relationen durch Verarbeitung von Definitionssätzen*
  - *Visualisierungen*
- weitere Domänen
  - *Sentimentanalyse: Annotation von Lexemen mit Polarität*
  - *Leichte Sprache: Annotation von Lexemen mit Komplexitätsinformation*
  - *Geschäftsberichte: spezielle Bedeutungen, Relationen und Definitionen in der Domäne*

Domäne Geschäftsberichte



Vielen Dank für die Aufmerksamkeit!