Connecting Multilingual Wordnets: Strategies for Improving ILI Classification in OdeNet

Johann Bergh
Lingolutions
Munich, Germany.
johann@lingolutions.com

PASSIVLINGO

Melanie Siegel
Darmstadt University
of Applied Sciences
melanie.siegel@h-da.de

h_da
darmstadt university
of applied sciences

Introduction: OdeNet

- Wordnet for German language
- Open-source license
- Automatically built 2017 on the basis of
  - OpenThesaurus data
  - NLP modules for German language, such as TextBlob and NLTK, as well as self-designed modules as for compound analysis
  - Google Translate
- Automatically and semi-automatically corrected and extended

Siegel, Melanie and Bond, Francis (2021). Compiling a German Wordnet from other Resources. In: Proceedings of the 11th International Global Wordnet Conference (GWC2021), pp. 192-198.

OdeNet corrections and extensions: first cycle

- Correction of POS
  - Multi-word lexemes and colloquial language words
  - Words in synsets that have different POS (extracted and corrected semi-manually)
  - Automatic correction of POS using information about German word endings

  → POS correct in ~93 %
- Adding hypernym relations
  - Implementing a compound analysis
  - Linking the compound to its head term as a hypernym

Further Problems in OdeNet

- Incorrect connections to OMW via ILI
- Duplicate assignment of ILI‘s to multiple synsets
- Inconsistent POS assignments

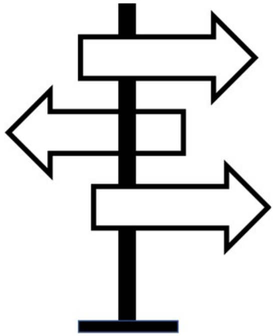
Problems in OdeNet: Incorrect / missing connections to OMW via ILI

- Ambiguity in synset translations: one to many
- German: *Unterlegscheibe*, English: *washer*

Name: washer
EWN ID: ewn-10788571-n
ILI: i94042
Definition: someone who washes things for a living

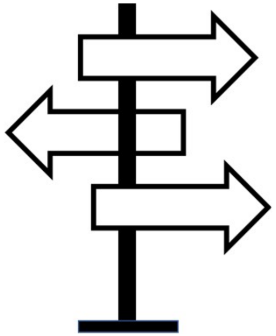
Name: washer
EWN ID: ewn-04562157-n
ILI: i60971
Definition: seal consisting of a flat disk placed to prevent leakage

Name: washer
EWN ID: ewn-04561970-n
ILI: i60970
Definition: a home appliance for washing clothes and linens automatically



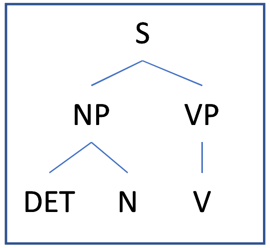
Problems in OdeNet: Duplicate assignment of ILIs to multiple synsets

- Ambiguity in synset translations: many to one
- Assignment of ILIs was not restricted to one-to-one
- Example:
  - odenet-4330-n [‘Anzahl’, ‘Zahl’] and odenet-688-n [‘Summe’, ‘Gesamtmenge’] had both ILI i35594
  - ILI i35594 in EWN: [‘measure’, ‘amount’, ‘quantity’]



Problems in OdeNet: Inconsistent POS assignments

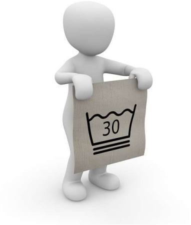
- Still ~7 % incorrect POS assignments
- Often multi-word lexemes
- For example:
  - [‘postmortal‘, ‘nach dem Tod‘, ‘post mortem‘] was categorized as POS "n", although it is POS "a".



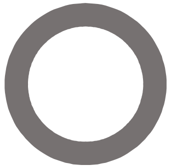
Idea: Improve automatic translation

- Translate from English to German (previously: German to English)
- Use DeepL for translation (previously: Google translate)
- Combine EWN synset elements with definition as translation input
  - i.e. add more context to the translation

Improving automatic translation: *Unterlegscheibe – washer*



- **EWN ID**: ewn-10788571-n
  ILI: i94042
  combination word and definition: *washer: someone who washes things for a living*
  automatic translation: *Wäscher: jemand, der beruflich Dinge wäscht*

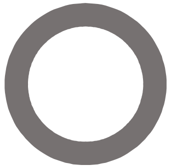


- **EWN ID**: ewn-04562157-n
  ILI: i60971
  combination word and definition: *washer: seal consisting of a flat disk placed to prevent leakage*
  automatic translation: *Unterlegscheibe: Dichtung, die aus einer flachen Scheibe besteht, um ein Auslaufen zu verhindern*

- **EWN ID**: ewn-04561970-n
  ILI: i60970
  combination word and definition: *washer: a home appliance for washing clothes and linens automatically*
  automatic translation: *Waschmaschine: ein Haushaltsgerät zum automatischen Waschen von Kleidung und Wäsche*

Improving automatic translation: *Unterlegscheibe – washer*

- **EWN ID**: ewn-10788571-n
  ILI: i94042
  combination word and definition: *washer: someone who washes things for a living*
  automatic translation: *Wäscher: jemand, der beruflich Dinge wäscht*

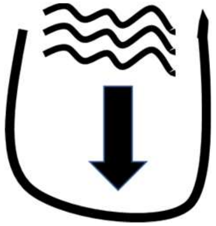


- **EWN ID**: ewn-04562157-n
  ILI: i60971
  combination word and definition: *washer: seal consisting of a flat disk placed to prevent leakage*
  automatic translation: *Unterlegscheibe: Dichtung, die aus einer flachen Scheibe besteht, um ein Auslaufen zu verhindern*

- **EWN ID**: ewn-04561970-n
  ILI: i60970
  combination word and definition: *washer: a home appliance for washing clothes and linens automatically*
  automatic translation: *Waschmaschine: ein Haushaltsgerät zum automatischen Waschen von Kleidung und Wäsche*

Still there is ambiguity: One to many

- There is still ambiguity, as for example:
  - ILI: i66412
    combination word and definition: *depth: the intellectual ability to penetrate deeply into ideas*
    automatic translation: *Tiefe: die intellektuelle Fähigkeit, tief in Ideen einzudringen*
- Lemma *Tiefe* in OdeNet:
  - odenet-847-n: [’Tiefe’, ’Tiefsinn’]
  - odenet-6615-n [’Abgrund’, ’Tiefe’, ’Schlund’, ’Holle’]
  - odenet-16328-n [’Tiefe’, ’Teufe’]



(it should be odenet-847-n)

Still there is ambiguity: Many to many

- More than one EWN synset can match a single OdeNet synset, e.g.
- **ILI**: i6124
  combination word and definition: *ethic: the principles of right and wrong that are accepted by an individual or a social group*
  automatic translation: *Ethik: die Grundsätze des Richtigen und Falschen, die von einem Individuum oder einer sozialen Gruppe akzeptiert werden*
- **ILI**: i68929
  combination word and definition: *ethics: the philosophical study of moral values and rules*
  automatic translation: *Ethik: das philosophische Studium der moralischen Werte und Regeln*
- *odenet-10-n* [’Sittlichkeit’, ’Wertvorstellungen’, ’Wertmaßstäbe’, ’Wertesystem’, ’Moral’, ’Moralvorstellungen’, **’Ethik**’, ’sittliche Werte’, ’moralische Werte’]
- *odenet-4879-n* [**’Ethik**’, ’Morallehre’, ’Sittenlehre’, ’Tugendlehre’]

Dealing with ambiguity: classification function

- OdeNet is very synonym rich
- Therefore, we use the synonyms in combination with a Word2Vec model
- We extract the definition part of the translated lemma and definition
- The content words in this translation are added to a vector v_1
- All synonyms are added to a vector v_2

Dealing with ambiguity: classification function

- For each value in v_1 and v_2 a similarity value is computed
- These values are summed and normalised to a value between 0 and 1
- This is the weighted value for the candidate synset in OdeNet competing for the ILI in a specific EWN synset

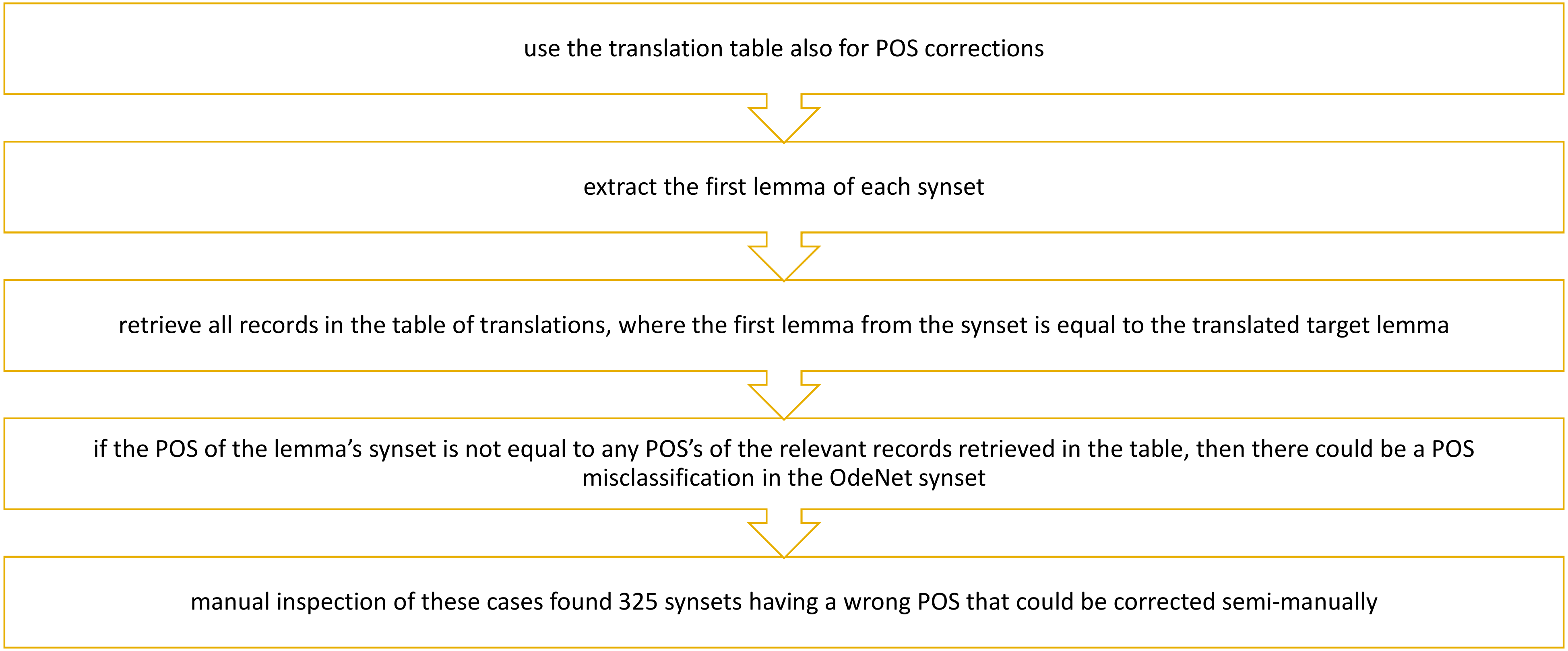
$$f(v_1, v_2) = \frac{\sum_i \sum_j dist(v_{1i}, v_{2j})}{|v_1| \times |v_2|}$$

Optimising machine translation by pre-processing verbs

- Problem: many English words are ambiguous between noun and verb
  - such as *search*
- Translation results improve, when adding *to* in front of English verbs

search → to search

Correct POS classification in OdeNet



use the translation table also for POS corrections

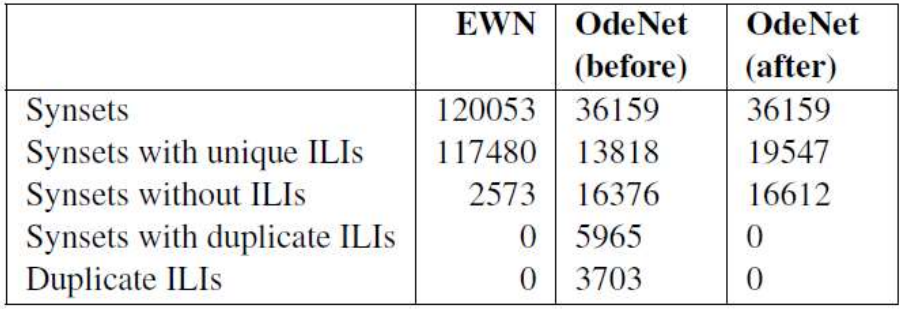
extract the first lemma of each synset

retrieve all records in the table of translations, where the first lemma from the synset is equal to the translated target lemma

if the POS of the lemma’s synset is not equal to any POS’s of the relevant records retrieved in the table, then there could be a POS misclassification in the OdeNet synset

manual inspection of these cases found 325 synsets having a wrong POS that could be corrected semi-manually

Results



	EWN	OdeNet (before)	OdeNet (after)
Synsets	120053	36159	36159
Synsets with unique ILIs	117480	13818	19547
Synsets without ILIs	2573	16376	16612
Synsets with duplicate ILIs	0	5965	0
Duplicate ILIs	0	3703	0

- Complete elimination of duplicate ILIs
- Much better linking of OdeNet synsets to OMW
  - Experiment with 100 examples: 59 had an ILI, 9 of these were wrong, 85% correct
- Identified and corrected the POS entries of 325 synsets

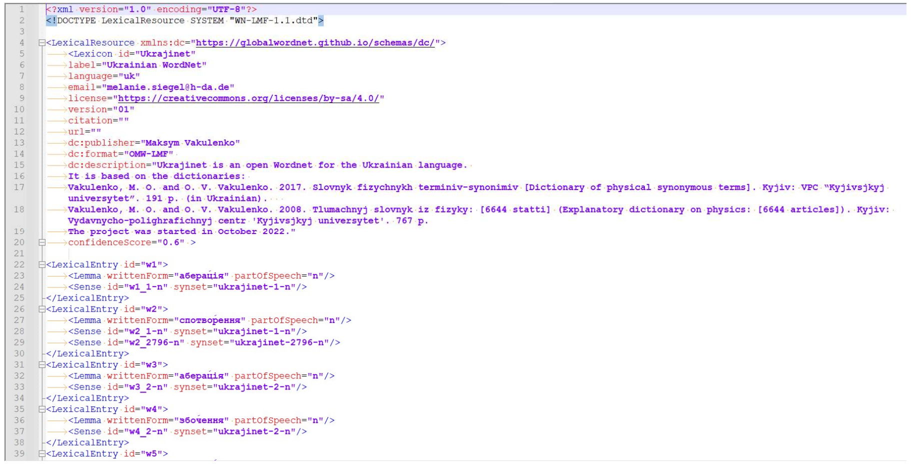
→ POS 99 % correct

Concluding Remarks

- OdeNet: open-source wordnet
  - Automatically compiled from OpenThesaurus
  - connected to the multilingual wordnets in the OMW initiative by machine-translating synsets
- Problems:
  - Machine translation was partly incorrect, mainly because translation context was missing
  - Duplicate interlingual indicators (ILIs) were assigned
  - POS information was not always correct
- Solution:
  - matching ILIs to OdeNet synsets, taking the English definitions into account as context
  - ILI classification weight function for desambiguation

Next steps

- Find more ways to further correct ILI information
- Link more OdeNet entries with OMW via ILI
- Apply the methods to build a wordnet for another language: Ukrainian



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LexicalResource SYSTEM "WN-LMF-1.1.dtd">

<LexicalResource xmlns:dc="https://globalwordnet.github.io/schemas/dc/">
    <Lexicon id="Ukrajinet"
    label="Ukrainian WordNet"
    language="uk"
    email="melanie.siegel@h-da.de"
    license="https://creativecommons.org/licenses/by-sa/4.0/"
    version="01"
    citation=""
    url=""
    dc:publisher="Maksym Vakulenko"
    dc:format="OMW-LMF"
    dc:description="Ukrajinet is an open Wordnet for the Ukrainian language.
    It is based on the dictionaries:
    Vakulenko, M. O. and O. V. Vakulenko. 2017. Slovnyk fizychnykh terminiv-synonimiv [Dictionary of physical synonymous terms]. Kyjiv: VPC “Kyjivsjkyj
    universytet”. 191 p. (in Ukrainian).
    Vakulenko, M. O. and O. V. Vakulenko. 2008. Tlumachnyj slovnyk iz fizyky: [6644 statti] (Explanatory dictionary on physics: [6644 articles]). Kyjiv:
    Vydavnycho-polighrafichnyj centr 'Kyjivsjkyj universytet'. 767 p.
    The project was started in October 2022."
    confidenceScore="0.6" >

<LexicalEntry id="w1">
    <Lemma writtenForm="аберація" partOfSpeech="n"/>
    <Sense id="w1_1-n" synset="ukrajinet-1-n"/>
</LexicalEntry>
<LexicalEntry id="w2">
    <Lemma writtenForm="спотворення" partOfSpeech="n"/>
    <Sense id="w2_1-n" synset="ukrajinet-1-n"/>
    <Sense id="w2_2796-n" synset="ukrajinet-2796-n"/>
</LexicalEntry>
<LexicalEntry id="w3">
    <Lemma writtenForm="аберація" partOfSpeech="n"/>
    <Sense id="w3_2-n" synset="ukrajinet-2-n"/>
</LexicalEntry>
<LexicalEntry id="w4">
    <Lemma writtenForm="збочення" partOfSpeech="n"/>
    <Sense id="w4_2-n" synset="ukrajinet-2-n"/>
</LexicalEntry>
<LexicalEntry id="w5">
```

h_da
darmstadt university
of applied sciences

OdeNet

Thank you for your attention!

Questions?

Melanie Siegel (melanie.siegel@h-da.de)
Johann Bergh (johann@lingolutions.com)

PASSIVLINGO