wnlex2018 workshop:
wordnets as lexicographical resources

elexis european lexicographic infrastructure

CLARIN
Common Language Resources and Technology Infrastructure

Univerza v Ljubljani
FILOZOFSKA FAKULTETA

Motivation for this workshop

- **From dictionary to wordnet**

  The relation between mostly concept-based lexical-semantic networks (wordnets) and lemma-based lexical resources (dictionaries) has been explored so far mainly for wordnet-building purposes, and such projects and related issues are well documented.

- **From wordnet to dictionary**

  In spite of not being meant to serve lexicographical purposes, wordnets have become a de facto standard for the drafting of dictionary content. Experiences and related issues have just started to be systematically discussed.

- **Our Goal**

  A survey of solved and unsolved issues regarding wordnet-based lexicography

  - Data models and interoperability of lexical resources
  - Lexicographical processes, workflows

co-organized by:

wnlex speakers

- Andrea Bellandi
  - Institute for Computational Linguistics «A. Zampolli», Pisa, Italy
- Martin Benjamin
  - Kamusi Project (kamusi.org)
- John McCrae
  - National University of Ireland Galway / Ollscoil na hÉireann Gaillimh, Ireland
- Darja Fišer
  - University of Ljubljana, Slovenia
- Fahad Khan
  - Institute for Computational Linguistics «A. Zampolli», Pisa, Italy
- David Lindemann
  - Universität Hildesheim, Germany
- Maciej Piasecki
  - Wrocław University of Technology, Wrocław, Poland

wnlex registered participants

- *I am interested in the differences between human-oriented dictionaries and NLP-oriented lexical resources.*
- *I am working in the Elexis project and I am interested in defining extensions to the W3C OntoLex-Lemon model. I am also a user of WN, and main aim is to combine Ontolex and WN for allowing to attribute senses to morphological variants of lemmas, when this is needed.*
- *I would like to expand my knowledge in wordnets and learn how to incorporate the acquired knowledge in my current work with dictionaries. I am particularly interested in the linked-data qualities of wordnets and learning of ways to overcome the limited nature of lemma-based structures.*
- *Present my poster and discuss future research directions with invited speakers and other participants.*
- *I am interested in expanding my knowledge in Wordnets as lexical resources, and how to utilize related methods in the process of compiling a dictionary.*
- *Want to get some insights on lexicography and some basic knowledge, also regarding needs of users and struggles.*
- *Interested in wordnets as lexicographical resource as well as in data models for wordnet-like concept-based resources.*

Time Schedule

09:00 - 09:45	Use and Evaluation of Wordnets as Lexicographical Resources - David Lindemann
09:45 - 10:30	Lexicographic Perspective on Wordnet Interoperability in CLARIN - Darja Fišer and Maciej Piasecki
10:30 - 11:00	*Coffee break*
11:00 - 11:45	Representing WordNets with OntoLex and the Global Wordnet Formats - John McCrae
11:45 - 12:30	Linking Lexicographic Resources: The Opportunities and Challenges Offered by the Semantic Web - Fahad Khan & Andrea Bellandi
12:30 - 13:45	*Lunch Break*
13:45 - 14:15	Poster Session - Presentation of accepted posters
14:15 - 15:00	Wordnet as a crowd source for untreated languages, concepts, and data elements - Martin Benjamin
15:00 - 15:45	Corpus-based Wordnet Development and plWordNet as a Relational Semantic Dictionary - Maciej Piasecki
15:45 - 16:00	*Coffee break*
16:00 - 17:00	Wrap-up & Discussion - All participants

Introductory Speech: Use and Evaluation of Wordnets as Lexicographical Resources

David Lindemann
University of Hildesheim
david.lindemann@uni-hildesheim.de

Introductory Speech: Overview

- WordNet as lexicographical resource
  - Why WordNet?
- A lexicographers‘ view on wordnet data
  - Language related issues, English bias
  - Glosses / Definitions
  - Lexical-semantic relations
  - Translation equivalents
  - Sense granularity
  - Data Models
  - Wordnet in lexicographical workflows
- Summary
  - Open questions to work on

Why WordNet?

- Princeton WordNet = WordNet for English (Miller, Fellbaum et al.)
  - Many other wordnets with links to Princeton WN items, cf. OMWN (Bond et al.)
- De-facto standard for multilingual lexicography from scratch
  - Data model suitable for cross-language links at sense level
  - High rate of coverage of English standard lemma lists / English conceptualisations
  - High precision due to high amounts of manual lexicographical work
- Examples for multilingual e-dictionaries based on wordnet data
  - **BabelNet (+Wikipedia, etc.)** - (Navigli et al.)
  - **Kamusi (+crowdsourcing)** - (Benjamin)
  - OmniLexica
  - Langua.de
  - Memidex.com
  - Hyperdic.net
  - Lookwayup.com
  - ConceptNet

If you have a wordnet, use it!

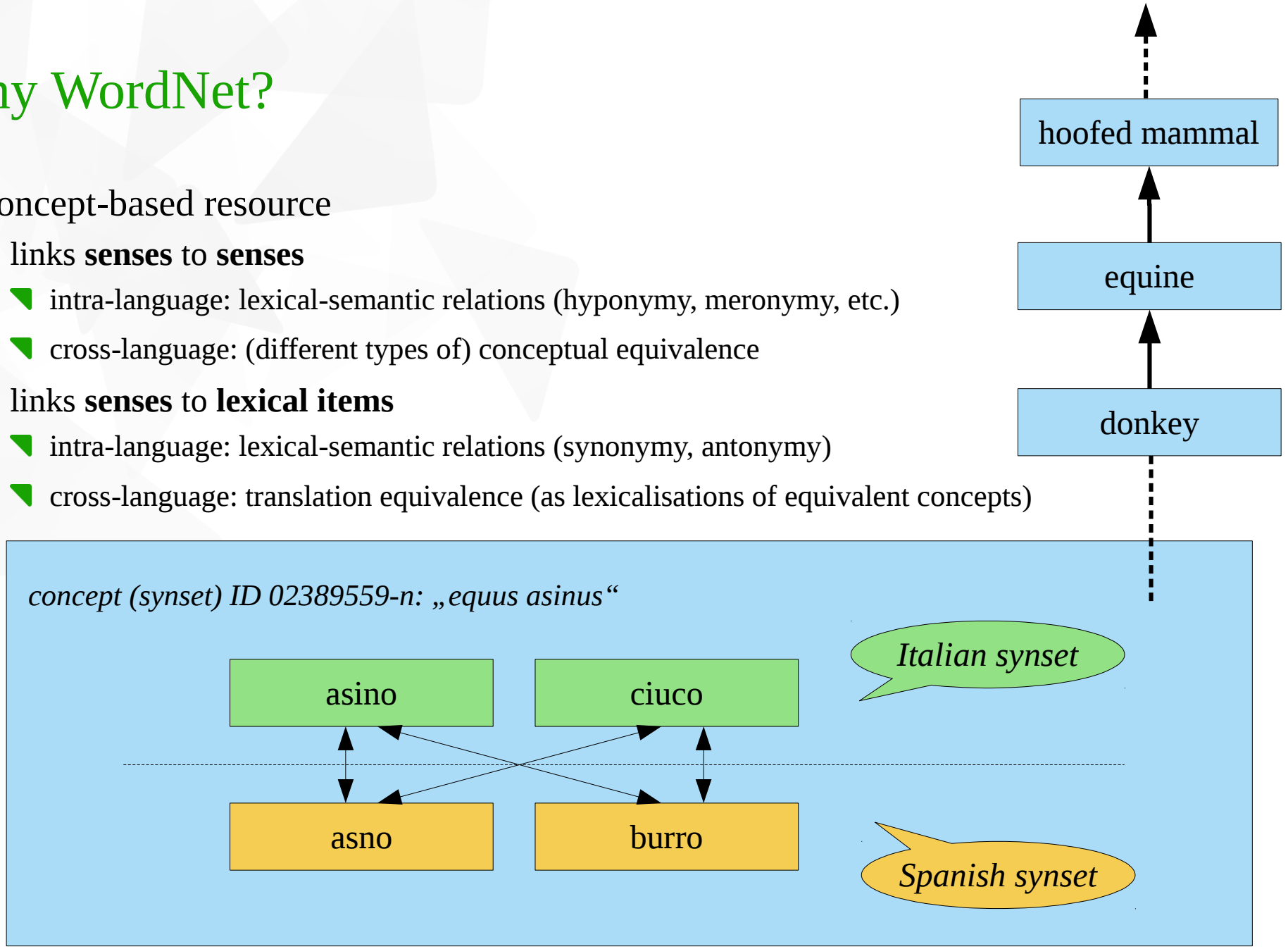
- **Example multilingual lexicography with Basque**

  Language resources for Basque: Somehow paradox situation

  - Lack of bilingual dictionaries (beyond ES, FR, EN, RU)
  - Availability of quite a large and precise, hand-crafted WordNet based on PWN synsets, and therefore aligned to a whole lot of languages
  - Research questions
    - Bilingual dictionary drafting using aligned wordnets - What does the lexicographer find?
    - Beyond synonymy and translation equivalence: What about all the other item types in a dictionary?
      - About the lemma: Phonetics, Morphology, Valency, Collocations,...
      - About the word sense: Other SemRels, Definitions, Example sentences
    - How can the Basque Wordnet benefit from wn-based lexicography?
      - Model for a bootstrapping loop

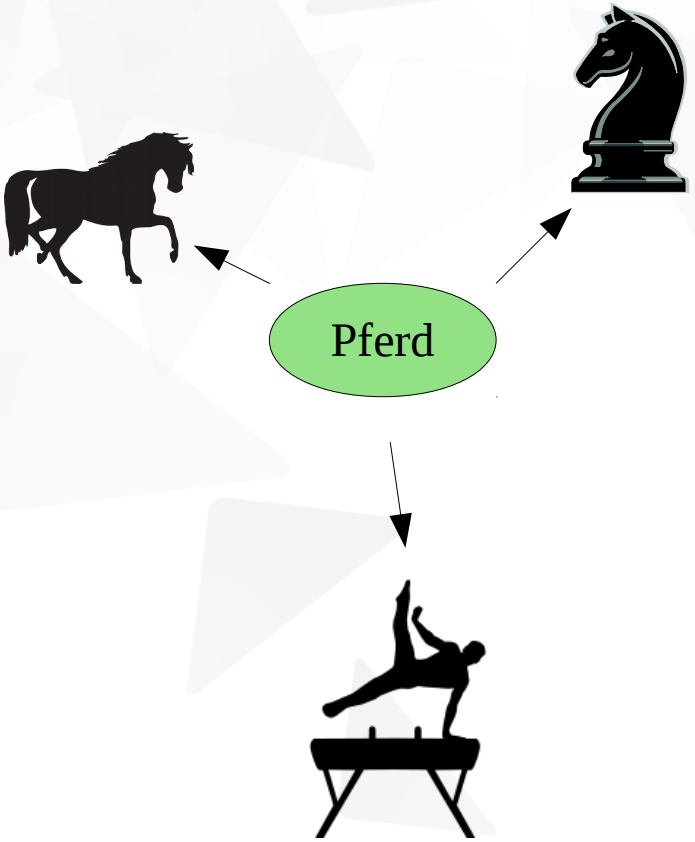
Why WordNet?

- Concept-based resource
  - links **senses** to **senses**
    - intra-language: lexical-semantic relations (hyponymy, meronymy, etc.)
    - cross-language: (different types of) conceptual equivalence
  - links **senses** to **lexical items**
    - intra-language: lexical-semantic relations (synonymy, antonymy)
    - cross-language: translation equivalence (as lexicalisations of equivalent concepts)

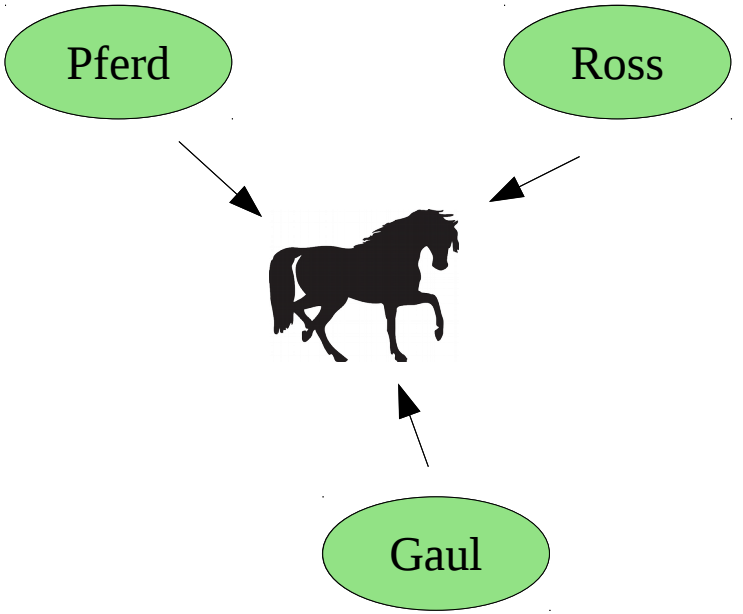


Open Multilingual WordNet (Bond et al.)

Lemma-oriented LR vs. Concept-oriented LR

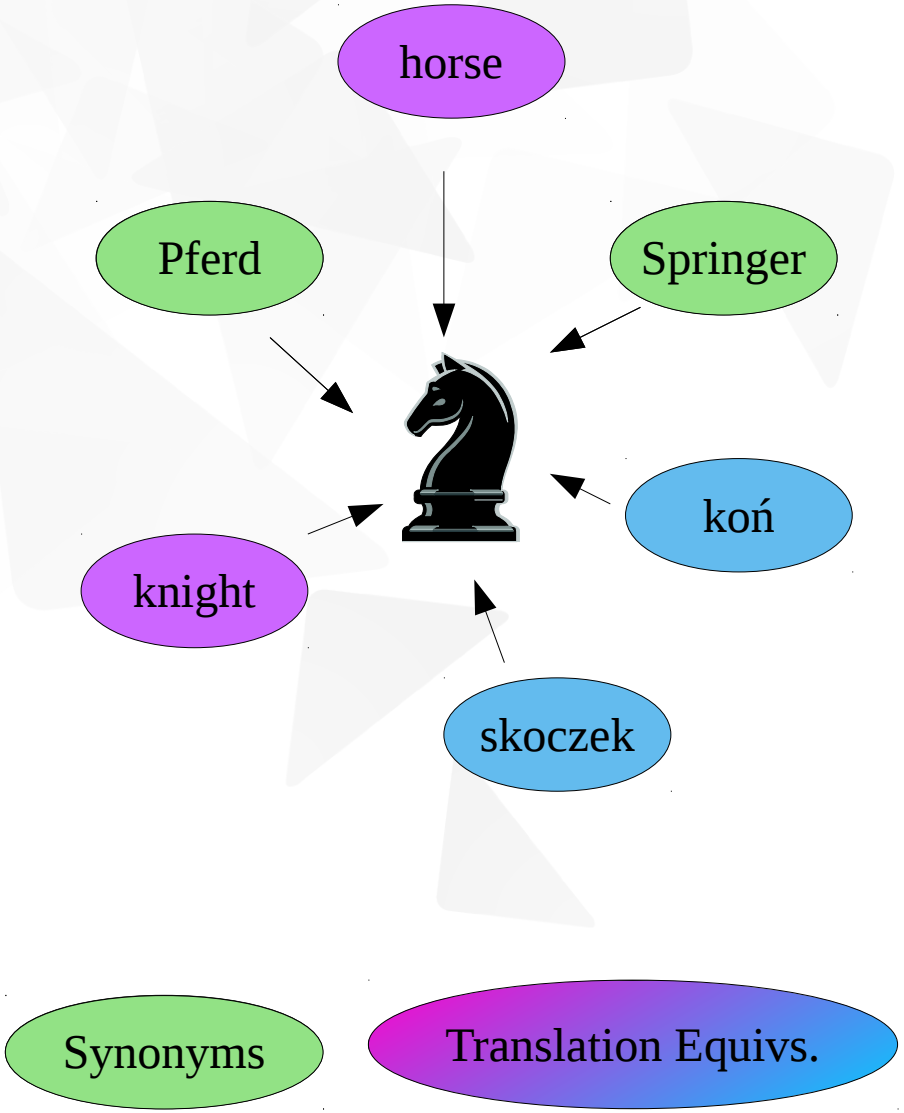


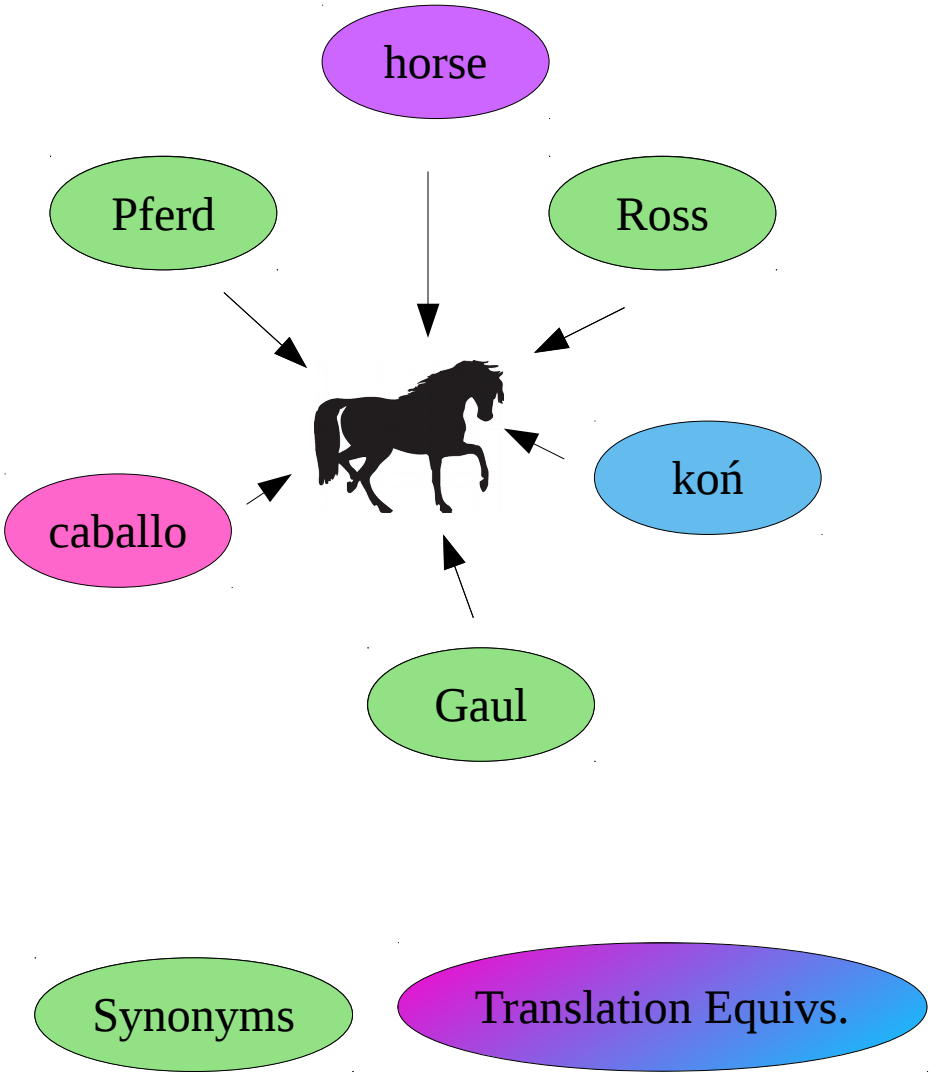
Polysemy: 3 word senses



Synonymy

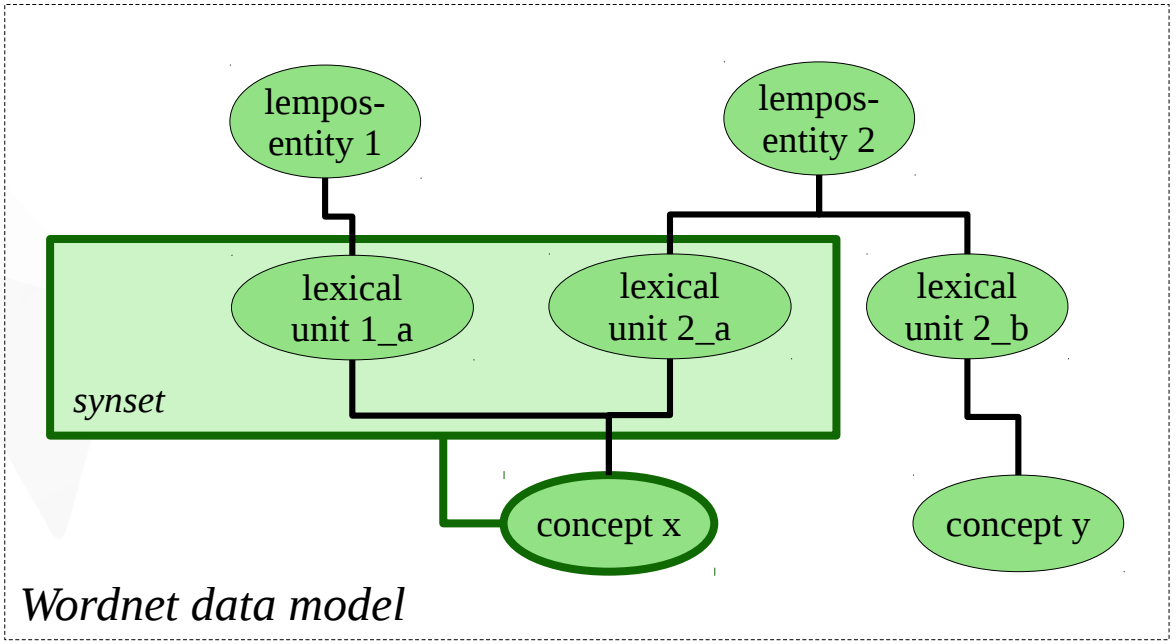
Concept-oriented LR





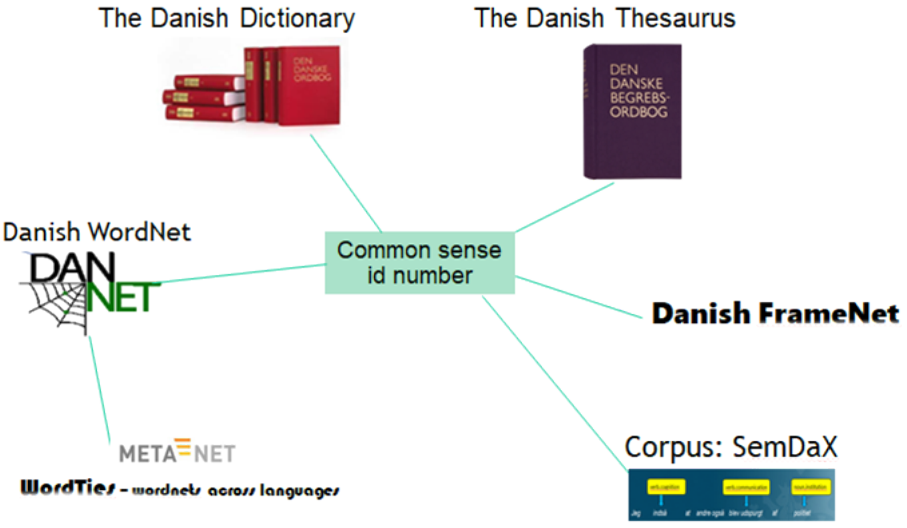
Why WordNet?

- Links to **other resources**
  - by lemma-sign (string) or lempos-entity (lemma with POS)
    - Always possible, but loses homograph / word sense disambiguation
  - by lexical item
    - by lemma_senseNr string (Bank_1 vs. Bank_2)
      - Princeton WN
    - by lexItemID
      - GermaNet data model
  - by sense
    - by senseID
      - Danish LR family
      - ILI: Open Multilingual WordNet
      - Global WordNet Grid / CILI (Bond, McCrae & Vossen 2016)



Wordnet data model

Linked Lexical Resources for Danish

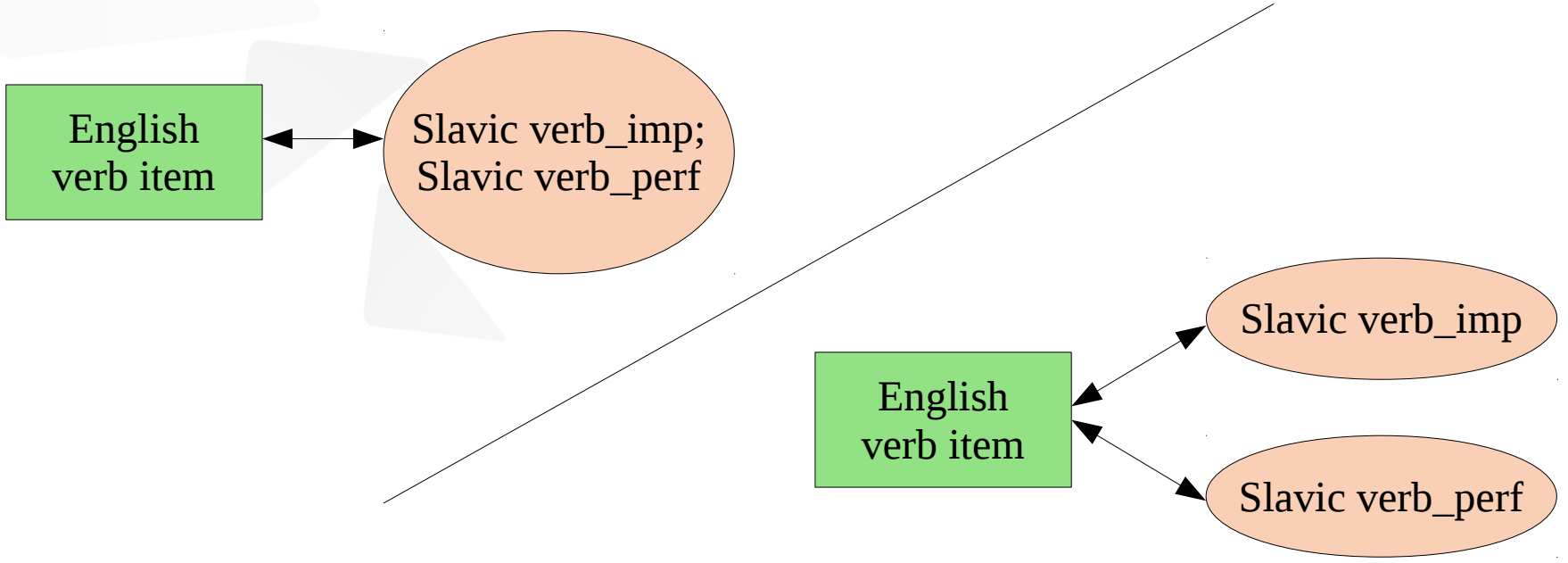


Wordnets in Lexicography: Some drawbacks & pitfalls

- English-bias
  - English-biased conceptualisation of our world
  - English-biased data model
- Glosses, Definitions
- Lexical-semantic relations
  - Relation fuzzyness
  - Granularity
- Translation equivalence
  - Relation fuzzyness
  - Errors of translations from PWN
- Sense granularity

WordNet as lexicographical resource: Language related issues

- Language-related issues:
  - Princeton WordNet: A data model for English
  - Adaptation to features of other languages
    - Example: Aspect in Slavic languages like Slovene, Polish
      - English-biased data model leads to take verbs with different aspect/Aktionsart represented as synonyms
      - Adaptation of data model: One-to-many correspondances between verbs, equivalence typology
  - WordNet building: Translate Princeton WN vs. new, independent WN data model



Use of wordnet data in lexicography: WN Glosses

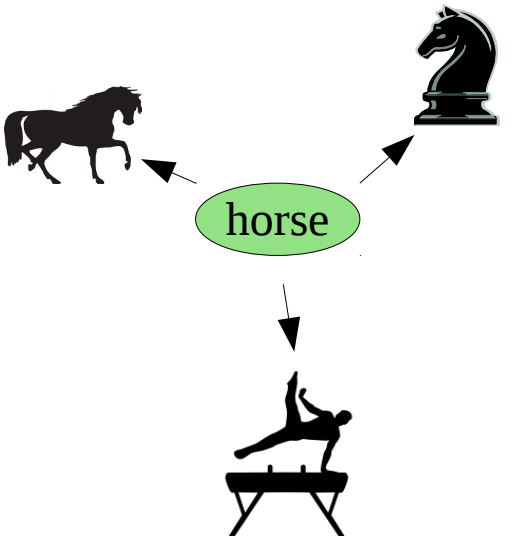
- Glosses, Definitions
  - Glosses: Hint for disambiguation for the human wordnet user
    - Just enough to be able to use as disambiguator
  - Definition in a language dictionary: Hint for the human dictionary user
    - Encyclopedic value as stand-alone text element
    - Bilingual dictionaries: Hints for word sense disambiguation in a foreign language
  - WN glosses as lexicographic definitions? cf. Benjamin 2016

WN lexical-semantic relations and Lexicography

horse #1
equine > odd-toed_ungulate > **ungulate >** placental > mammal > **vertebrate >** chordate > **animal**

horse #2
chessman > man > **game_equipment**

horse #3
gymnastic_apparatus > sports_equipment



- Hyperonymy-hyponymy:
  - Too fine-grained for a 1:1 use in a language dictionary?
  - More complete, more accurate than the information found in many dictionaries
    - (Pedersen et al. 2018: 103)

WN lexical-semantic relations and Lexicography

glass #1 {glass, drinking glass}
a container for holding liquids while drinking
> container > instrumentality

glass #2 {glass, glassful}
the quantity a glass will hold
> containerful > indefinite_quantity

snow #1 {snow, snowfall}
precipitation falling from clouds in the form of icy crystals
> precipitation > weather > athmospheric_phenomenon

snow #2 {snow}
a layer of snowflakes (white crystals of frozen water) covering the ground
> layer > region > location > object

- Metonomy:
  - More complete, more accurate than the information found in many dictionaries

WN lexical-semantic relations and Lexicography

- Synonymy
  - In wordnets: alternative lexicalisations for the same concept, interchangeable in a context
    - Quasi-synonymy sometimes represented as homonymy relation, then gloss concerning register
    - English {chalk, crank, glass, ice, methamphetamine, methamphetamine hydrochloride, Methedrine, meth, deoxyephedrine, chicken feed, shabu, trash}
    - English {policeman, officer, police officer} *a member of a police force*
    - English {cop, bull, copper, pig, fuzz} *uncomplimentary terms for a policeman*
    - Danish {betjent, funktionær, ordenshåndhæver, panser, politibetjent, strisser, strømer, tjenestemand}
  - In Lexicography: always quasi-synonymy (register, sociolect, dialect… pragmatics)
    - Thesauri (e.g. openthesaurus.de): Lexical items bear usage labels

Translation equivalence: Cross-language linking of items

- Interlingual indices
  - Open Multilingual WordNet (OMWN, cf. Bond & Foster 2013)
    - PWN synsets as pivot sense grid
  - Global WordNet Grid (Vossen, Bond & McCrae 2016)
    - English-independent sense repository
- Bilingual Dictionary Drafting using OMWN
  - Quantitative evaluation using source language lemma list as standard
  - Qualitative evaluation by human annotators: Adequateness as translation equivalent candidate
    - Do I want this candidate as it is to appear in my dictionary entry as an equivalent? OK
    - Is it an acceptable equivalent, but does it need some manual editing? FUZZY
    - Is this noise / an inadequately matched equivalent pair? FALSE

Translation equivalents extracted from WN: evaluation

- Lindemann et al. 2014: German-Basque
  - GermaNet v8 – BasqueWN v3: 21% recall, 83% precision
- Lindemann & Kliche 2017: Basque-English
  - BasqueWN v3 – Princeton WN v3: 31% recall, 89% precision
- Set of student assessments, BA course in computational lexicography, Hildesheim 2017
  - English – WOLF (FrenchWN): 58% precision
  - English – WONEF (FrenchWN): 74% precision
  - English – GermaNet v8: 87% precision
  - [English – BabelNet v3.7 German: 61% precision]

forget_Verb

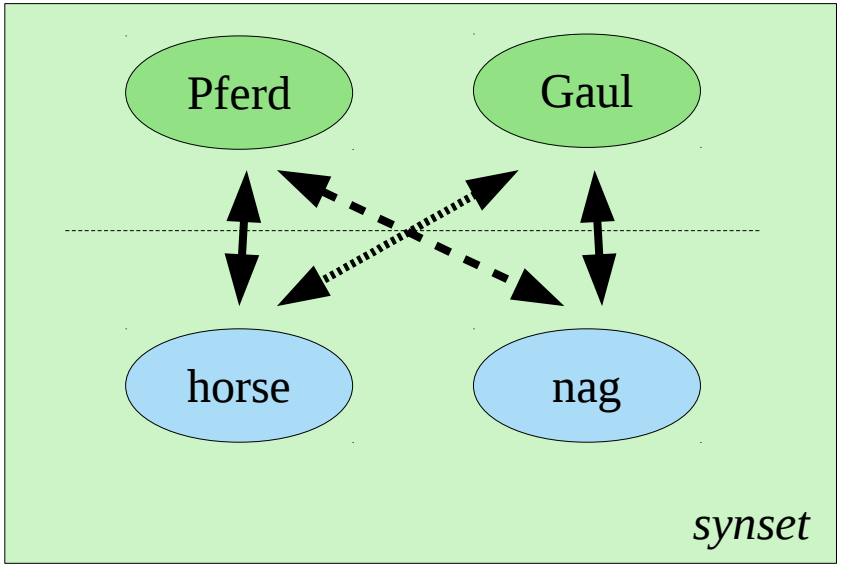
GermaNet8: **1** 30-00610167-v dismiss from the mind; stop remembering
German: unterdrücken (synonym) Equiv=OK

GermaNet8: **2** 30-00613018-v leave behind unintentionally
German: stehen lassen (near_synonym) Equiv=FUZZY
German: vergessen (synonym) Equiv=OK

GermaNet8: **3** 30-00614829-v forget to do something
German: verbummeln (near_synonym) Equiv=OK

Fuzzy equivalency (interlingual quasi-synonymy)

- More fine-grained evaluation of wordnet as multilingual lexicographical resource
- List of criteria for being represented in a more advanced wordnet data model
- 3 typologies of translation equivalence
  - Maks 2007: OMBI project (reverting bilingual dictionaries)
  - Adamska-Sałaciak 2010: Typology of interlingual equivalence
  - Rudnicka 2017: Features of a „super strong“ interlingual equivalence



OMBI (Maks 2007)

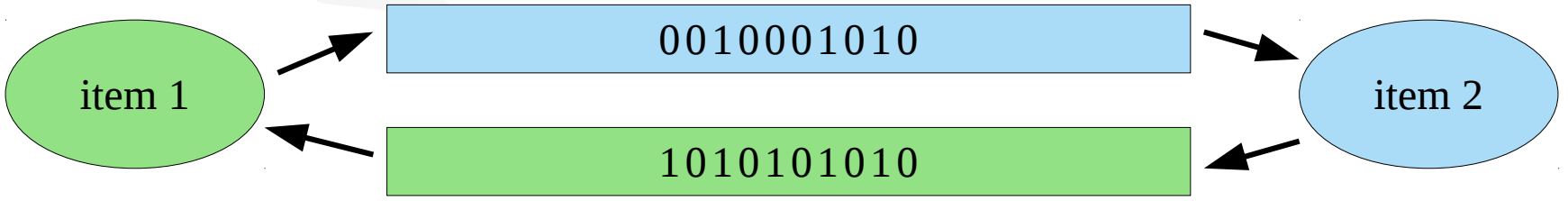
- *Contrasts in conceptual equivalence*
  - Hyponym
  - Hyperonym
  - Near Equivalent
  - Related
- *Pragmatic Contrasts*
  - Formal vs. neutral
  - Old-fashioned vs. neutral
- *Contrasts in degree of lexicalisation*
  (established lexical unit vs. explanatory equivalent)
  - Fully lexicalised
  - Semi-lexicalised
  - Non-lexicalised
- *Variant status*
  - Preferred synonym vs. term variant

Translation Equivalence (Adamska-Sałaciak 2010)

- Type C: Cognitive
  - (a.k.a. semantic, systemic, prototypical, conceptual, decontextualised, notional)
  - Has to be an established LU of TL > not always possible to provide
- Type E: Explanatory
  - (a.k.a. descriptive)
  - Always possible to provide
- Type T: Translational
  - (a.k.a. insertable, textual, contextual)
  - Adequate translation in context, word-level correspondance
- Type F: Functional
  - (a.k.a. situational, communicative, discourse, dynamic)
  - Adequate translation in context, without word-level correspondance

Rudnicka et al. 2017 and implications

- **Super-strong equivalence**:
  - i. identity in **grammatical category** (given from the synset mapping)
  - ii. identity in **number**
  - iii. identity in **sense** (synset (and lexical unit) relation structure and gloss)
  - iv. identity in **register**
  - v. identity in **countability**
  - vi. compatibility in (semantic) **gender** (if relevant/applicable)
  - vii. '**first choice**' equivalent: listed first in bilingual dictionaries
  - viii. **bidirectional**
  - ix. **high translation probability** if it appears in a parallel corpus
  - x. **unique** for a single lexical unit



Sense granularity

- WN sense clustering (creation of coarse senses): Several approaches
  - Surveys: Peters, Peters & Vossen 1998; Agirre & Lopez de Lacalle 2003 [senseval-2]
- The “autohyponymy“ problem (Pociello, Agirre & Aldezabal 2011)

Princeton WN 3.0

{celebration, festivity} (any festival or other celebration)
=> {merrymaking} (boisterous celebration)
=> {revel, revelry} (noisy partying)
=> {bout, spree} (a drunken revel)
=> {bender, bust} (an occasion for heavy drinking)
=> {carouse} (a merry drinking party)
=> {orgy} (a wild gathering involving drinking and promiscuity)
=> {whoopee} (noisy and boisterous revelry)

Basque WN 3.0

{festa, jai} (event or party organised to celebrate something)
=> {**parranda**} (boisterous celebration)
=> {**parranda**} (noisy partying)
=> {**parranda**} (a drunken revel)
=> {**parranda**} (an occasion for heavy drinking)
=> {**parranda**} (a merry drinking party)
=> {orgia} (a wild gathering involving drinking and promiscuity)
=> {**parranda**} (noisy and boisterous revelry)

WordNet sense clustering: Translation similarity

- Candidates for merging according to semantic distance calculated from cross-language lexicalization patterns
  - Resnik & Yarowski 2000
  - Chugur, Gonzalo & Verdejo 2002

Table 4. *Mapping between cross-linguistic sense labels and established lexicons*

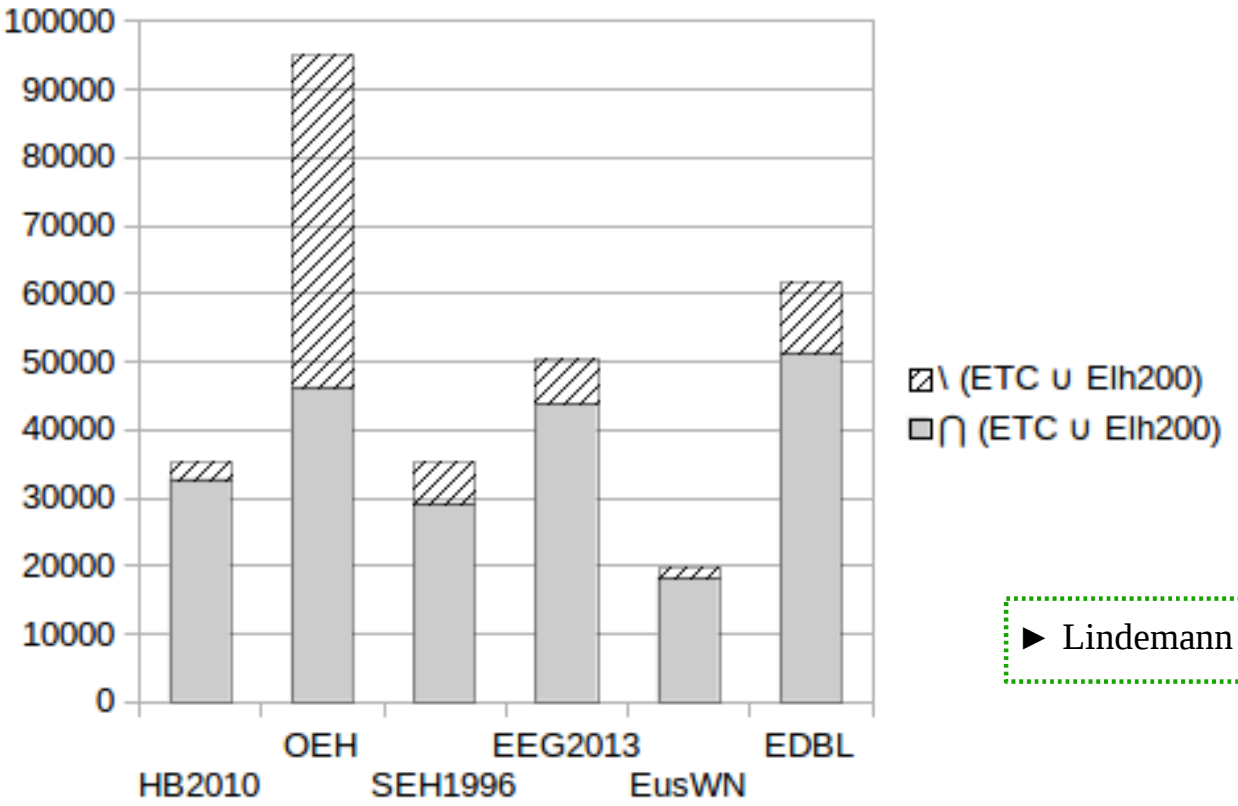
Target Word	WordNet Sense #	English description	Spanish	French	German	Italian	Japanese
interest (noun)	1	monetary (e.g. on loan)	interés, rédito	intérêt	Zinsen	interesse	rishi, risoku
	2	stake/share	interés, participación	intérêt participation	Anteil	interesse	riken
	3,4	intellectual curiosity	interés,	intérêt	Interesse	interesse	kanshin, kyōmi
	5	benefit, advantage	provecho, interés, beneficio	intérêt	Interesse	interesse	rieki
drug (noun)	1a	medicine	medicamento, droga	medicament	Medikament, Arzheimittel	medicina	kusuri
	1b	narcotic	narcótica droga	drogue	Drogue, Rauschgift	droga	mayaku
bank (noun)	1	shoreline	ribera, orilla	banc, rive	Ufer	sponda,riva	kishi
	2	embankment	loma, cuesta	talus, terasse	Erdwall	muccio	teibō
	3	financial inst.	banco	banque	Bank	banca	ginkō
	4	supply/reserve	banco	banque	Bank	banca	ginkō
	5	bank building	banco	banque	Bank	banca	ginkō
	6	array/row	hilera, batería	rang, batterie	Reihe	batteria	retsu
fire (t. verb)	1	dismiss from job	despedir, echar	renvoyer	feuern	licenziare	kubi ni shimasu
	2	arouse, provoke	excitar, enardecer	enflammer, animer	beflügeln entzünden	accendere infiammare	kōfun saseru
	4	discharge weapn	disparar	lâcher	abfeuern	sparare	happō s.
	5	bake pottery	cocer	cuire	brennen	cuocere	yaku

Basque Lexical Resources / A model for BDD

- Basque landscape of lexical resources
  - Institutions
    - Lexicography: Basque Language Academy
    - Lexicography: Basque Language Institute @ EHU
    - NLP: IXA CL-group @ EHU
    - Lexicography / NLP: Elhuyar
  - Scarcity of bilingual dictionaries
    - only ES, FR, EN, RU meet state of the art
  - State of the art NLP lexical resources
    - parameter files for spell checkers, taggers, RBMT engines
  - Basque WordNet, part of MCR
    - built by the ‘expand’ method
    - fully aligned to PWN 3.0
- Bilingual Dictionary Drafting (BDD)
  - Starting point: merged NLP lexicon and Wordnet (Lindemann & San Vicente 2016)

Basque Lexical Resources

- Corpus-based frequency lemma list for Basque
  - Lemmata extracted from ETC (Sarasola, Salaburu & Landa 2013), and Elh200 (Leturia 2014)
  - Comparison to 6 reference resources: 4 Dictionaries, Basque WN, 1 NLP lexicon



► Lindemann & San Vicente (2015)

Basque Dictionary Draft: (1) Homograph Level

```
<homograph homograph="aditu" corpus_counts="42042"/>
```

- Basic list of lemma-signs: 57.000
- 20+ occurrences in 200M-corpus *and* in 1+ reference resource
- Frequency data from Elh200 corpus

(1) Homograph, (2) Syntactical Entity

```
<homograph homograph="aditu" corpus_counts="42042">
    <ADI lemma="aditu" pos="ADI_SIN" corpus_counts="18989"/>
    <IZE lemma="aditu" pos="IZE_ARR" corpus_counts="13945"/>
    <ADJ lemma= "aditu" pos="ADJ_ARR" corpus_counts="5486"/>
</homograph>
```

- Syntactical Entities (lempos-entities) from Elh200 corpus
- Corpus pos-tagged with EusTagger, based on EDBL data
- Frequency data for each lempos-entity
  - interesting for lexicographer
  - interesting for dictionary user

(1) Homograph, (2) Syntactical Entity, (3) Sense

```
<homograph homograph="aditu" corpus_counts="42042">
    <ADI lemma="aditu" pos="ADI_SIN" corpus_counts="18989">
        <sense synset="30-00588888-v" equivs="understand"/>
        <sense synset="30-02169702-v" equivs="hear"/>
        <sense synset="30-02571901-v" equivs="heed mind listen"/>
    </ADI>
    <IZE lemma="aditu" pos="IZE_ARR" corpus_counts="13945">
        <sense synset="30-09617867-n" equivs="expert"/>
        <sense synset="30-10557854-n" equivs="scholar scholarly_person bookman"/>
    </IZE>
    <ADJ lemma= "aditu" pos="ADJ_ARR" corpus_counts="5486">
        <sense synset="30-02226162-a" equivs="adept expert skillful"/>
    </ADJ>
</homograph>
```

- Word senses from EusWN (Basque WordNet)
- Linking of senses to syntactical entities (as child elements)

Drafted Basque dictionary content

	Corpus-based SE	SE with one or more EusWN Word senses	Total EusWN Word senses	Polysemy ratio	SE present in corpus but not in EusWN	SE present in EusWN but not found in corpus
Verbs	4,151	1,636	6,567	2.01	2,515	279
Common Nouns	23,921	15,193	30,613	4.01	8,728	3,479
Proper Nouns	2,443	132	153	1.16	2,311	60
Adjectives	6,147	50	141	2.82	6,097	8
Adverbs	1,556	0	0	0.00	1,556	0
Total	**38,218**	**17,011**	**37,474**	**2.20**	**21,207**	**3,826**

Dictionary Draft SE Gap Detection: semi-automatic

```
<homograph homograph="aditu" corpus_counts="42042">
    <ADI lemma="aditu" pos="ADI_SIN" corpus_counts="18989">
        <sense synset="30-00588888-v" equivs="understand"/>
        <sense synset="30-02169702-v" equivs="hear"/>
        <sense synset="30-02571901-v" equivs="heed mind listen"/>
    </ADI>
    <IZE lemma="aditu" pos="IZE_ARR" corpus_counts="13945">
        <sense synset="30-09617867-n" equivs="expert"/>
        <sense synset="30-10557854-n" equivs="scholar scholarly_person bookman"/>
    </IZE>
    <ADJ lemma="aditu" pos="ADJ_ARR" corpus_counts="5486">
        <sense synset="30-02226162-a" equivs="adept expert skillful"/>
    </ADJ>
</homograph>
```

- Blank SE (present in EDBL, not in EusWN):

  Find corresponding synset in Princeton WordNet, copy ID

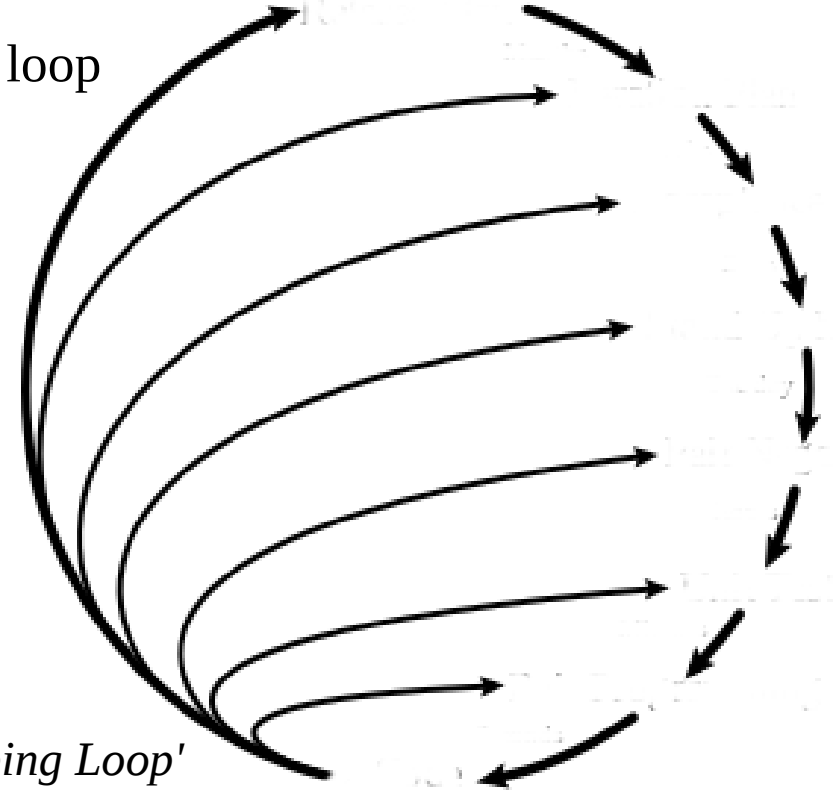
Dictionary Draft Sense Gap Detection: Manual work!

EusWN Lexical Unit	Definition EN	EusWN 3.0 synset	EN synset	CAT synset
adar_1	*one of the bony outgrowths on the heads of certain ungulates*	adar_1	horn_2	banya_1
adar_2	*a railway line connected to a trunk line*	adar_2	branch_line_1 spur_track_1 spur_5	enforcall_1 forcall_1
adar_3	*a warning signal that is a loud wailing sound*	adar_3, sirena_2 turuta_5	siren_3	
adar_4	*a local branch of some fraternity or association*	adar_4	chapter_3	capítol_2
adar_5	*a division of a stem, or secondary stem arising from the main stem of a plant*	adar_5 abar_2 besanga_1 beso_12	branch_2	branca_1 branc_1
adar_6	*an alarm device that makes a loud warning sound*	sirena_4 adar_6 turuta_6	horn_9	
adar_7	*a device used for easing the foot into a shoe*	zapata_sartzeko_1	shoehorn_1	calçador_1

Manual postediting of WordNet-based dictionary drafts

- Crowdsourcing (as in kamusi.org)
  - User is prompted to fill lexical gaps in his language's WN (wich is aligned to other WNs)
    - Language community empowerment (kamusi.org)
    - Alignment at concept (word sense) level from the very beginning
- Concepts new to the multilingual WN: Global WN Grid
- Lexicographical workflows for a bootstrapping loop
  - Manual editing of bilingual dictionary drafts
  - Reuse of hand-validated data for upgrading the original resources
  - Planned project: New series of bilingual dictionaries with Basque

Every single bit of manual work,
every gap that is filled,
every sense that is split,
every link that is set,
every error that is found,
shall allow to upgrade both
EDBL and Basque WordNet.



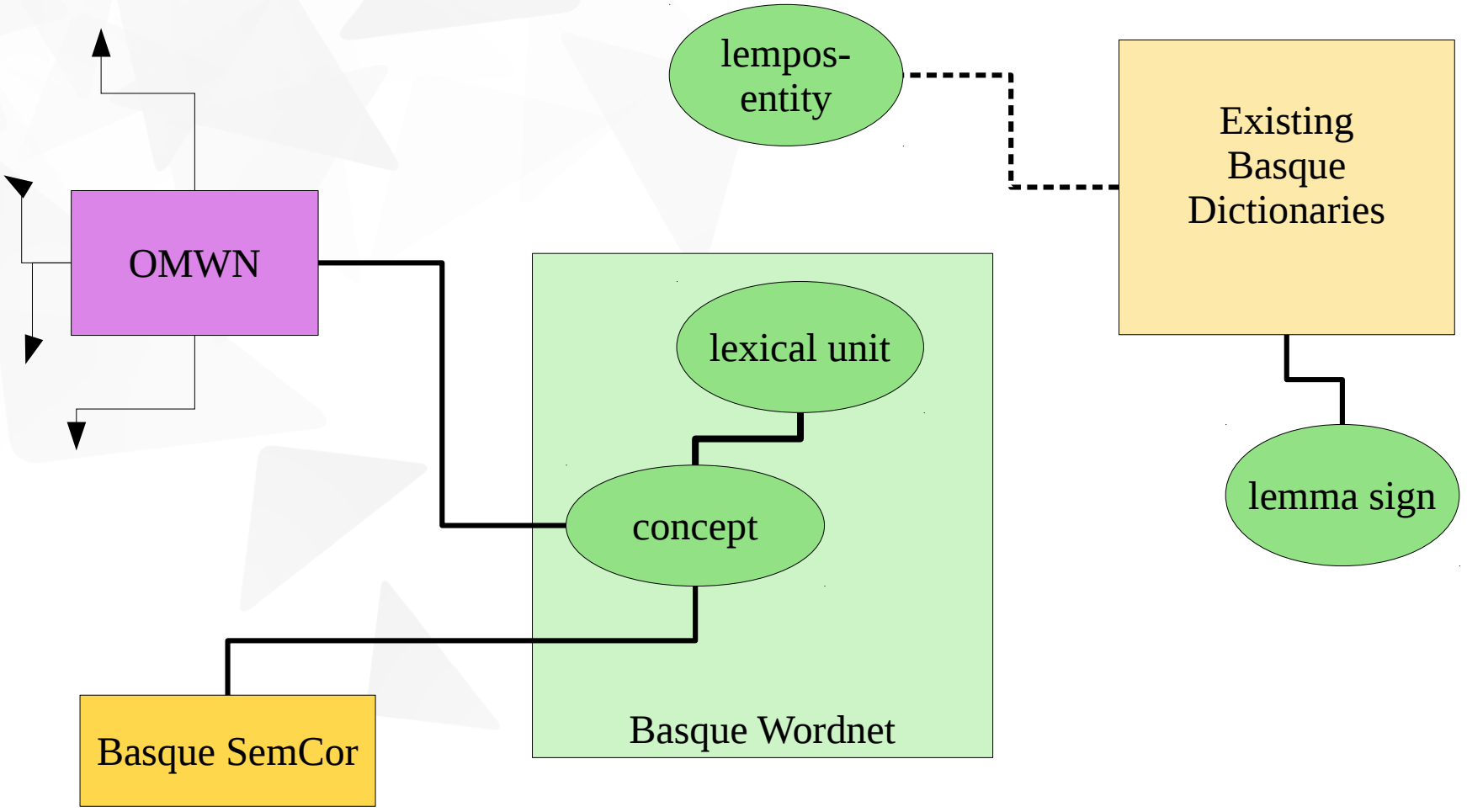
'Bootstrapping Loop'

Image Source: Wikimedia Commons

Application example: WordNet/BabelNet bootstrapping for EUS-SLO

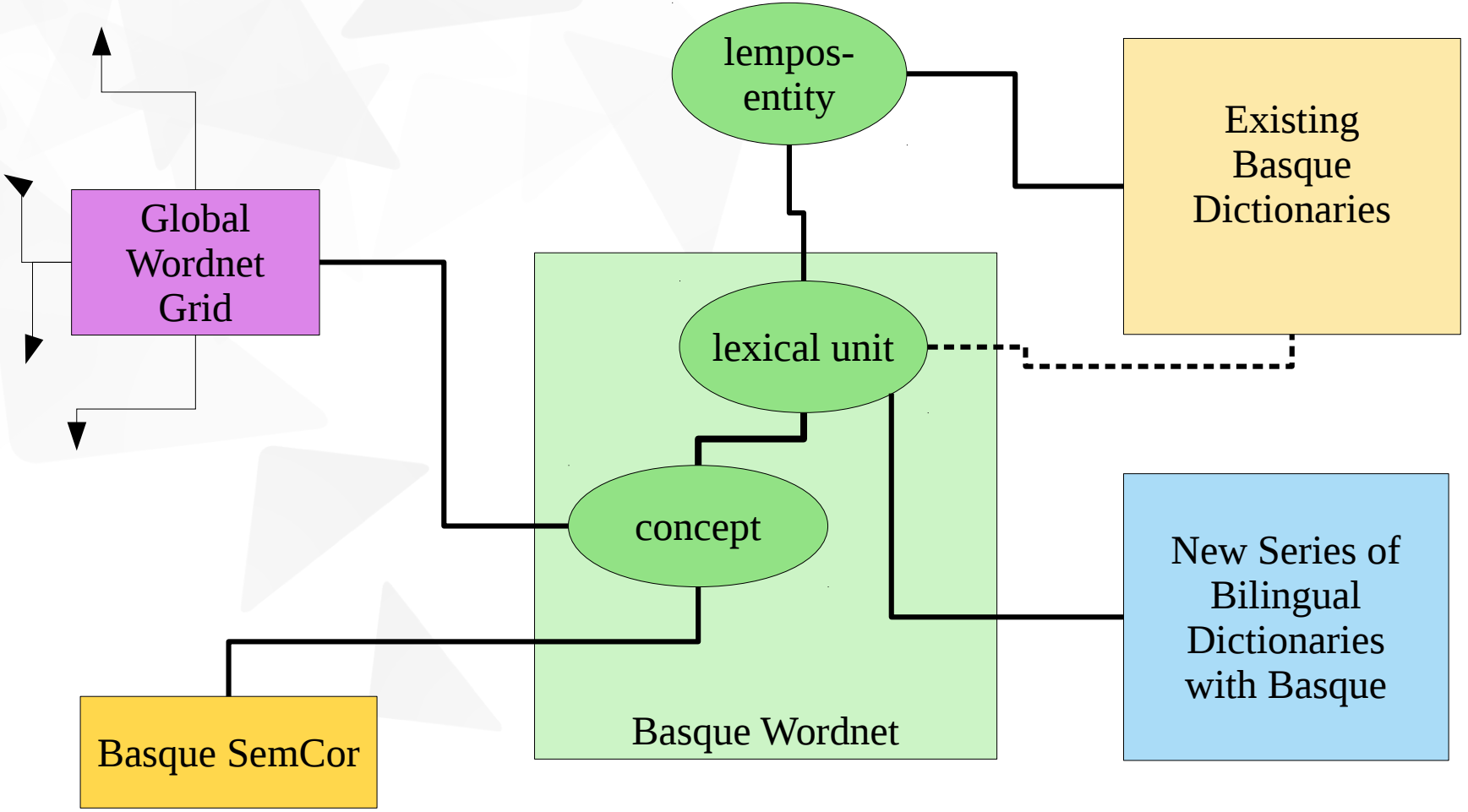
- Basque (EUS) – Slovene (SLO): A totally uncovered pair of 'smaller' languages
- Quantitative Evaluation
  - Recall: Synsets that contain 1+ Basque standard headword and 1+ Slovene item
    - EusWN / SloWNet 20% (66% of 30%)
    - BabelNet 31% (78% of 40%)
  - Recall on 5,000 most frequent Basque headwords (BabelNet): 74% (3,707)
  - Recall on 20,000 most frequent Basque headwords (BabelNet): 53% (10,549)
- Qualitative Evaluation
  - Precision: Unknown. EN-SL precision to be measured first.

Basque Lexical Resources today



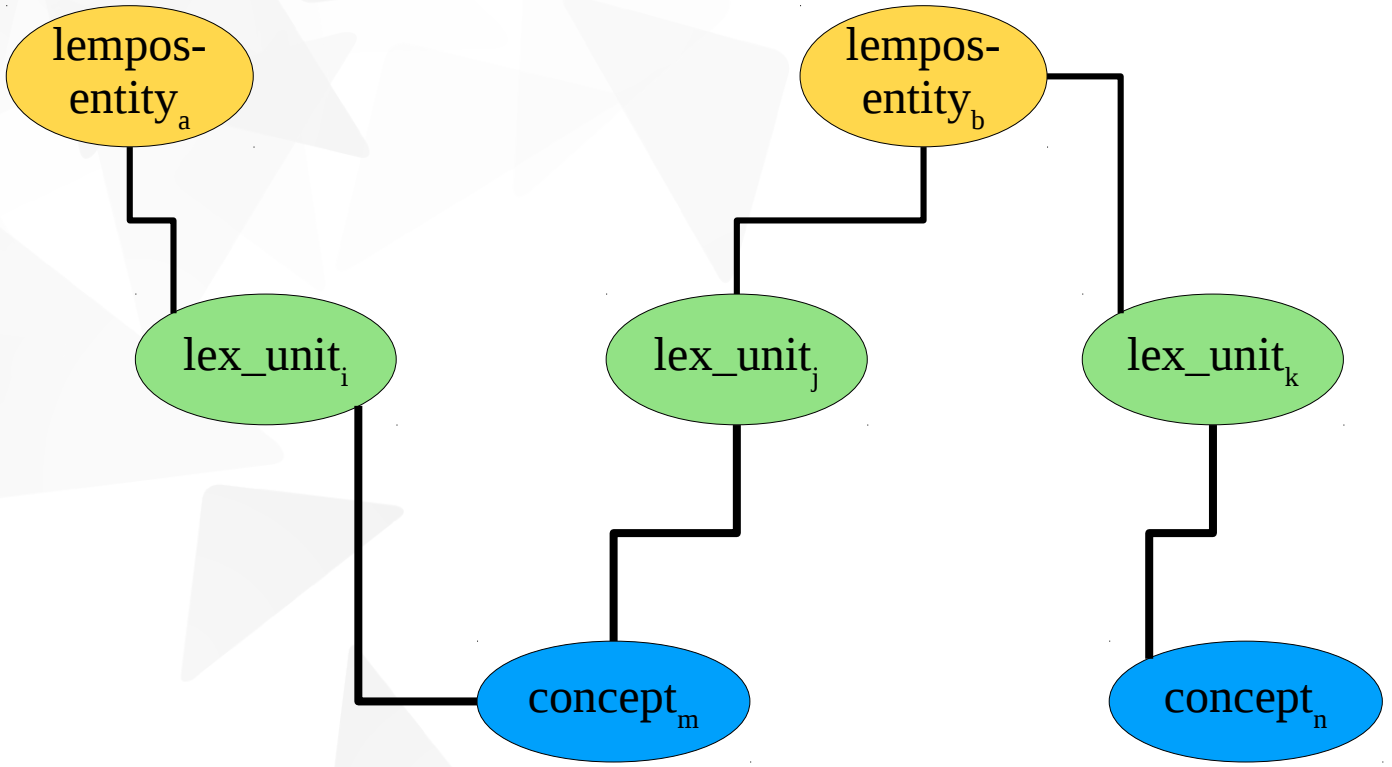
lemma sign: lemma string without POS and sense disambiguation
lempos-entity: lemma-sign with POS, all word senses
lexical unit: lemma-sign with POS and unique word sense

Scenario: Basque Lexical Resources



lempos-entity: lemma-sign with POS, all word senses
lexical unit: lemma-sign with POS and unique word sense

Data modeling



Three entities to link item types to: lempos entity, lexical unit, concept

- Abkürzungsangabe (AbkA)
- Abkürzungsauflösungsangabe (AbkAA)
- Akzentsilbenangabe (AkSA)
- Antonymenangabe (AntA)
- Ausspracheangabe (AusA)
- Belegbeispielangabe (BBeiA)
- Bedeutungsangabe (BA)
- Bedeutungsparaphrasenangabe (BPA)
- Beispielangabe (BeiA)
- Beispielgruppenangabe (BeigA)
- Belegangabe (BelA)
- Belegstellenangabe (BStA)
- Datierungsangabe (DatA)
- Deklinationsangabe (DekA)
- Diminutivangabe (DimA)
- Diminutivgruppenangabe (DimgA)
- Etymologieangabe (EtyA)
- Fachgebietsangabe (FGA)
- Formvariantenangabe (FVA)
- Genusangabe (GA)
- Graduierungsangabe (GradA)
- Graduierungsbeschränkungsangabe (GradbA)
- Grammatikangabe (GrA)
- Hinweisangabe (HinA)
- Häufigkeitsangabe (HA)
- Kompetenzbeispielangabe (KBeiA)
- Kompositagruppenangabe (KompgA)
- Kompositumangabe (KompA)
- Konjugationsangabe (KonjA)
- Lemmazeichengestaltangabe (LZGA)
- Markierungsangabe (MarkA)
- Phrasemangabe (PhrasA)
- Pluralbildungsangabe (PlbA)
- Pluraletantumangabe (PltA)
- Polysemieangabe (PA)
- Pragmatische Angabe (PragA)
- Pragmatisch-semantische Angabe (PragsemA)
- Quellenangabe (QuA)
- Rechtschreibangabe (RA)
- Rektionsangabe (RekA)
- Satzmusterangabe (SmA)
- Schreibungsangabe (SchrA)
- Silbenangabe (SA)
- Silbentrennungsangabe (STrA)
- Singularbildungsangabe (SgbA)
- Singularetantumangabe (SgtA)
- Sprachenidentifizierungsangabe (SpIA)
- Sprachenvergleichsangabe (SpVA)
- Sprichwortangabe (SprichwA)
- Stilschichtangabe (StilA)
- Symptomwertangabe (SympA)
- Synonymenangabe (SynA)
- Themaangabe (ThA)
- Umlautangabe (UmlA)
- Verbvalenzangabe (VVA)
- Verweisangabe (VerwA)
- Vokalquantitätsangabe (VQA)
- Wortartenangabe (WAA)
- Wortakzentangabe (WAkA)
- Wortäquivalentangabe (WÄA)
- Wortformenangabe (WFA)
- Zeichengestaltangabe (ZGA)

Linguistische Angaben nach Wiegand 1989

Treatment of Homonymy

Noun

- S: (n) **tear**, teardrop (a drop of the clear salty saline solutio… glands) *"his story brought tears to her eyes"*
- S: (n) rip, rent, snag, split, **tear** (an opening made forcibly as … *was a rip in his pants"; "she had snags in her stockings"*
- S: (n) bust, **tear**, binge, bout (an occasion for excessive eating or … *on a bust that lasted three days"*
- S: (n) **tear** (the act of tearing) *"he took the manuscript in both hands and gave … mighty tear"*

Princeton WordNet 3.1, noun "tear"

…ilding or structure *They tore …*
hospital and built some offices.
tear sth off to quickly remove your clo…
He tore off his shirt and jumped into the stream.
tear sth up to tear paper into a lot of small pieces *He tore up her photograph.*

tear[2] /teəʳ/ *noun* [C] a hole in a piece of cloth, paper, etc where it has been torn

tear[3] /tɪəʳ/ *noun* [C] a drop of water that comes from your eye when you cry *Suddenly he* ***burst into tears*** (= started crying). ○ *I was* ***in tears*** (= crying) *by the end of the film.* ● **tearful** *adjective* crying *a tearful goodbye* ● **tearfully** *adverb* ➲See also: in floods (**flood**[2]) of tears.

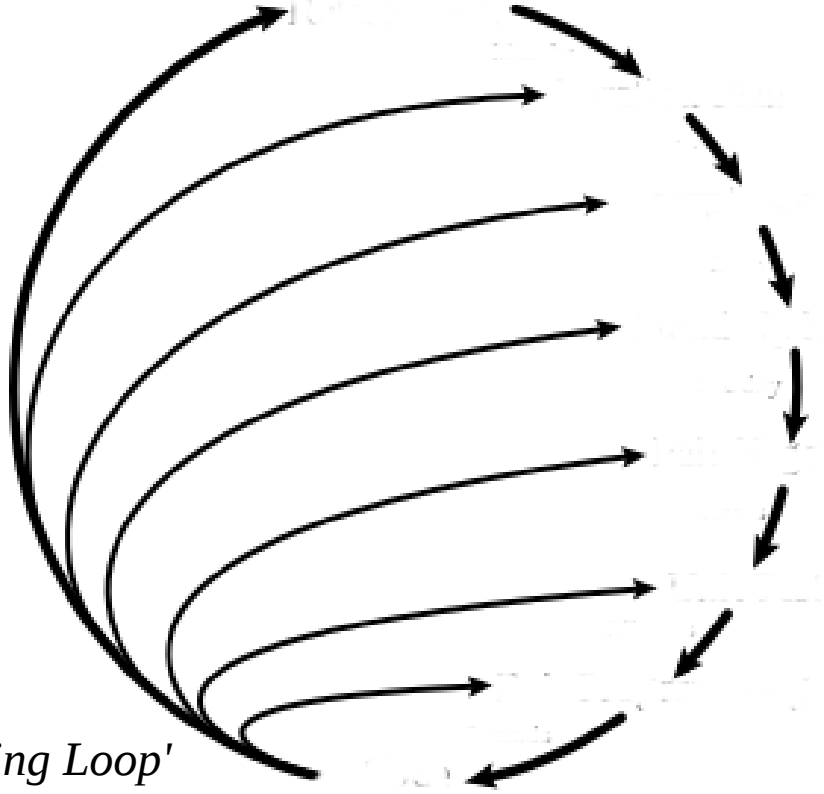
'tear ˌgas *noun* [U] a gas that makes pe… eyes hurt, used by the police or ar… …olent crowds

Cambridge Learners' Dic., noun "tear"

```
<homograph homograph="tear">
   <entity pos="noun" phon="/tiə/" equiv="Träne"/>
   <entity pos="noun" phon="/teə/" equiv="Riss"/>
   <entity pos="verb" phon="/tiə/" equiv="tränen"/>
   <entity pos="verb" phon="/teə/" equiv="reißen"/>
</homograph>
```

Workflow proposal for Basque

- Bilingual Dictionary Draft for Basque-English including sense-to-sense mappings
  - Encouraging recall and precision rates; can be applied to other language pairs
- Preliminaries for a research project
  - Bilingual Dictionary Drafts for many uncovered language pairs
  - Data model that allows
    - Manual and semi-automated (bulk) editing
    - Edition of e-dictionaries including more item types
    - Retro-updating of original resources: 'Bootstrapping Loop'
  - Engagement of lexicographers for editing 'their' language pair
  - Edition of a new series of bilingual dictionaries with Basque



'Bootstrapping Loop'

Image Source: Wikimedia Commons

Summary: Some open questions

- Multilingual WordNet: Data modeling
  - Types of translation equivalence
    - Representation of relations between synsets / between lexical units
  - Inclusion of / linking to more lexicographic item types
  - Homonymy vs. Polysemy
  - Interoperability with existing standards
- Linking of lexical resources of different shape
  - lemma-based resources, lemma-based links
  - concept-based resources, concept-based links
- Evaluation of automatically built resources
- Definition of lexicographic workflows
  - Hand-crafted edits / upgrades of wordnet-dictionaries
  - Tutorials / best practice guidelines

Thank you for your attention

david.lindemann@uni-hildesheim.de

Please find the bibliography at:
https://www.zotero.org/groups/2164775/wnlex/items/collectionKey/BH2Y7CYQ