Lexemes in Wikidata

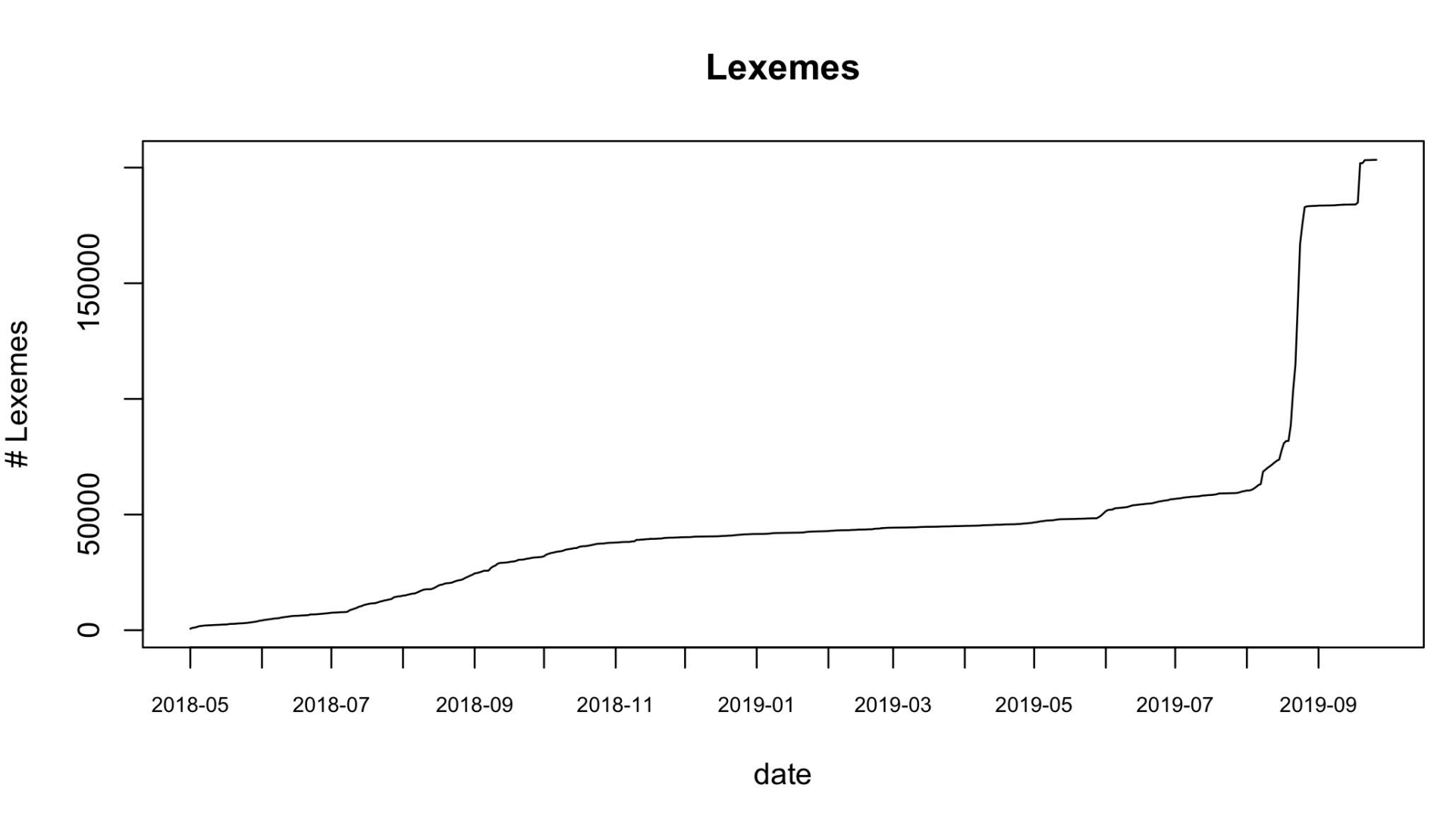
WikidataCon 2019, Berlin

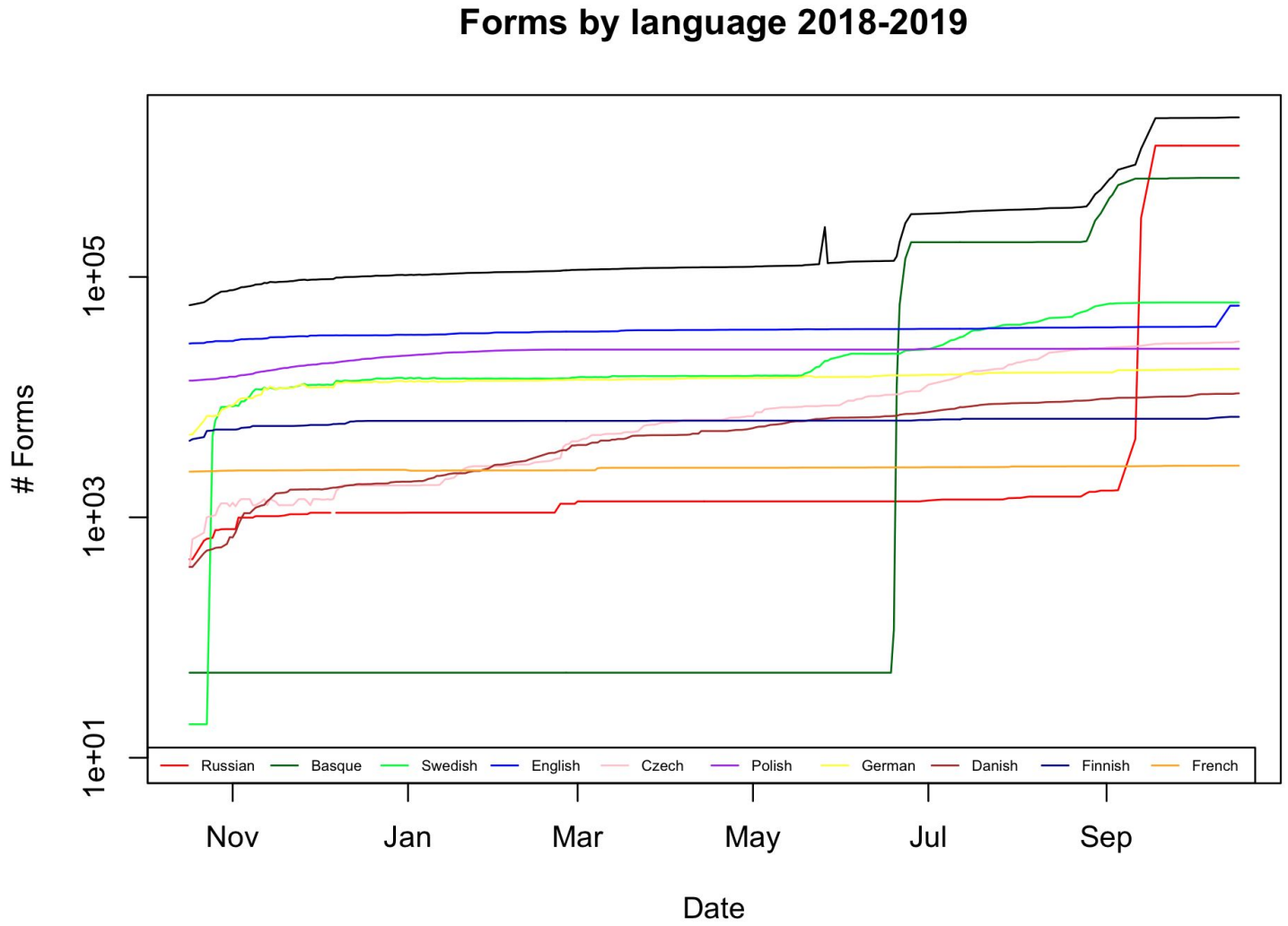
Arthur Smith (American Physical Society) &

Alicia Fagerving (Wikimedia Sverige)

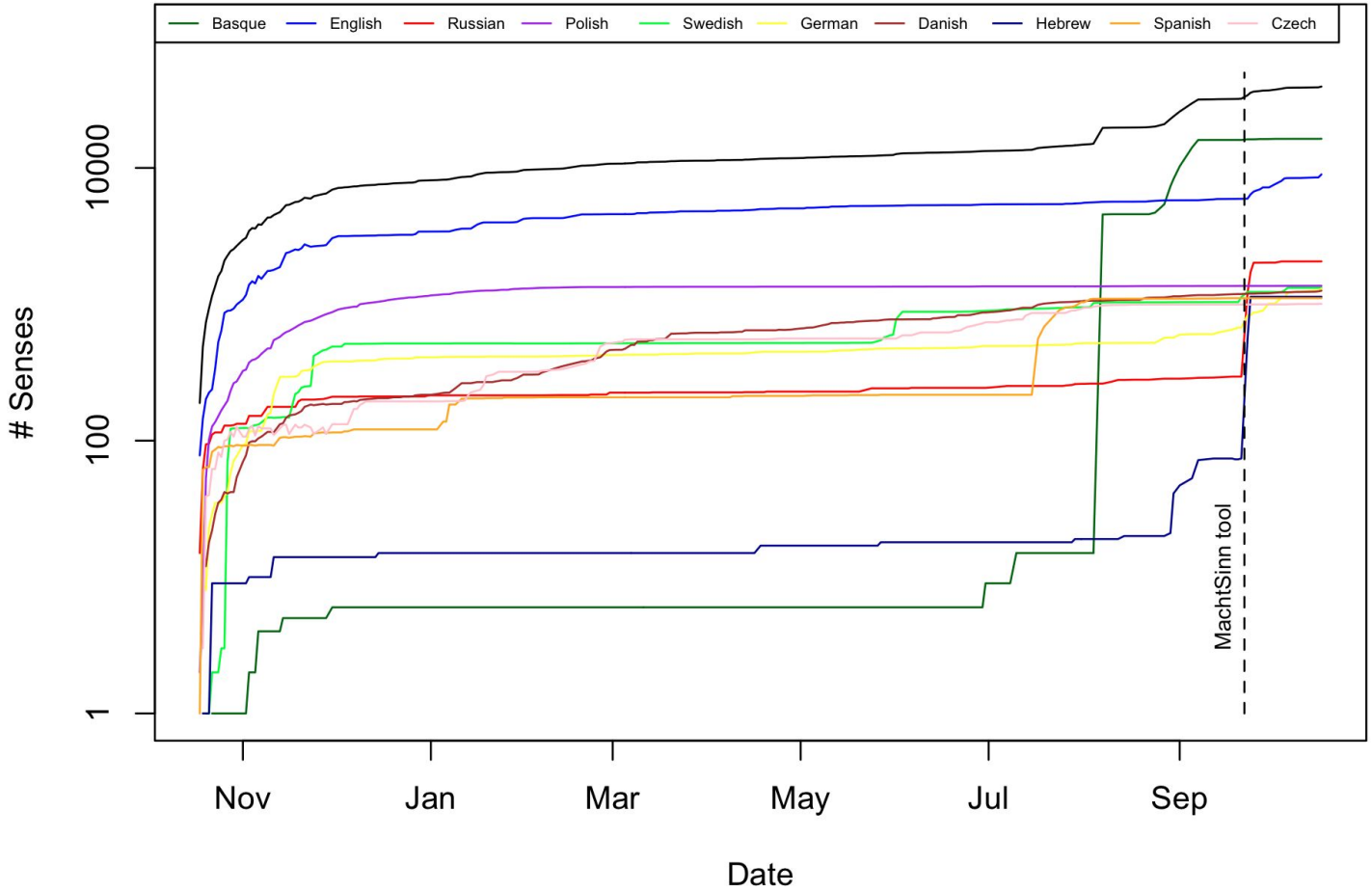
Agenda

1. What this is all about – lexeme basics
   a. https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation
2. Where we are now – some statistics
3. Properties used with lexemes, forms and senses
4. Tools
5. Bots
6. Discussion...


Lexemes
Lexemes
0
50000
150000
2018-05
2018-07
2018-09
2018-11
2019-01
2019-03
2019-05
2019-07
2019-09
date


Forms by language 2018-2019
Forms
1e+01
1e+03
1e+05
Russian
Basque
Swedish
English
Czech
Polish
German
Danish
Finnish
French
Nov
Jan
Mar
May
Jul
Sep
Date

Senses by language 2018-2019



Common properties directly on lexemes

- P5911 – “inflection class” (over 200,000 times)
- P5185 – “grammatical gender” (over 100,000)
- P1552 – “has quality” (over 100,000) – generally Q51927539 (inanimate) or Q51927507 (animate)
- P5187 – “word stem” (73702)
- P6838 – “Elhuyar Dictionary ID” (for Basque - 10280)
- P5238 – “combines” (4646)
- P31 – “instance of” (4485)
- P5831 – “usage example” (3531)
- P5191 – “derived from” (2935)

Common properties on forms

- Varies greatly by language!
- Very few on Basque lexeme forms
- P7243 – “pronunciation” (ru)
- P5279 – “hyphenation” (ru, cz)
- P898 – “IPA transcription” (cz, en, ru)
- P443 – “pronunciation audio” (cz, en, ru, sv)
- P6712 – “precedes word-initial” (en)
- P2440 – “transliteration” (ru)

Common properties on senses

- P5137 “item for this sense” (15637 times)
- P5972 “translation” (2322)
- P18 “image” (1367)
- P5973 “synonym” (983)
- P31 “instance of” (795)
- P6271 “demonym of” (774)
- P5974 “antonym” (178)
- P5323 “attested in” (88)
- P5831 “usage example” (73)
- P6191 “language style” (68)
- P5185 “grammatical gender” (26)

Template:Lexicographical properties

v · t · e

Lexicographical properties	
General	item for this sense (P5137) · grammatical gender (P5185) · conjugation class (P5186) · word stem (P5187) · derived from (P5191) (mode of derivation (P5886), derived from form (P5548), derived from sense (P5980)) · Wikidata property example for lexemes (P5192) · Wikidata property example for forms (P5193) · officialized by (P5194) · attested in (P5323) · stroke count (P5205) · combines (P5238) · hyphenation (P5279) · auxiliary verb (P5401) · homograph lexeme (P5402) · Han character in this lexeme (P5425) · Japanese pitch accent type (P5426) · valency (P5526) · requires grammatical feature (P5713) · usage example (P5831) (demonstrates form (P5830), demonstrates sense (P6072)) · inflection class (P5911) · root (P5920) · creates lexeme type (P5923) · translation (P5972) · synonym (P5973) · antonym (P5974) · troponym of (P5975) · false friend (P5976) · Wikidata property example for senses (P5977) · classifier (P5978) · location of sense usage (P6084) · language style (P6191) · collective noun for animals (P6571)
Languages	has grammatical gender (P5109) · has grammatical person (P5110) · has conjugation class (P5206) · has grammatical mood (P3161) · has grammatical case (P2989) · has tense (P3103) · has inflection class (P5913)
Phonetics	pronunciation audio (P443) · IPA transcription (P898) · X-SAMPA Code (P2859) · Soundex (P3878) · Cologne phonetics (P3879) · pronunciation variety (P5237) · Slavic phonetic alphabet (P5276) · pronunciation (P7243)
Other properties useful in lexicography	image (P18) · described by source (P1343) · quote (P1683)
Dictionaries and databases	OED Online ID (P5275) · SJP Online ID (P5274) · SGJP Online ID (P5455) · Doroszewski Online ID (P5497) · Kopaliński Online ID (P5533) · WSO Online ID (P5627) · WSJP ID (P5720) · Dobry słownik ID (P5793) · Vocabolario Treccani ID (P5844) · SPXVI ID (P5877) · SJPXVII ID (P5876) · Uralonet ID (P5902) · Álgu ID (P5903) · Oqaasileriffik online dictionary ID (P5912) · Ġabra lexeme ID (P5928) · Oudnederlands Woordenboek GTB ID (P5937) · Vroegmiddelnederlands Woordenboek GTB ID (P5938) · Middelnederlandsch Woordenboek GTB ID (P5939) · DanNet 2.2 word ID (P6140) · Bantu Lexical Reconstructions ID (P6168) · described at URL (P973) · search formatter URL (P4354) · formatter URL (P1630)
Values for property instance of or has quality of the lexeme	plurale tantum/collective noun/singulare tantum · inanimate/animate · imperfective aspect/perfective aspect · reconstructed word · acronym
Values for property instance of or has quality of the form	obsolete form · depreciative form · rare form · potential form · non-depreciative form · vocalic form · non-vocalic form · colloquial form · strong form · weak form · incorrect form · former form
Values for property instance of or has quality of the sense	old sense · colloquial sense · archaism · rare sense · humorous sense · euphemism · vulgarism · pejorative
Sandboxes	bar/sandbox (L123) · Sandbox-Lexeme (P5188) · Sandbox-Form (P5189) · Sandbox-Sense (P5979)

Wikidata:Lexicographical data · Examples & resources · Property proposals · Create a new Lexeme · Wikidata Lexeme Forms

Tools for editing & exploring lexemes

- Wikidata Lexeme Forms - https://tools.wmflabs.org/lexeme-forms/
- Ordia - https://tools.wmflabs.org/ordia/
- Hauki - https://tools.wmflabs.org/hauki/
- MachtSinn - https://tools.wmflabs.org/machtsinn/
- Lexeme Senses - https://tools.wmflabs.org/lexeme-senses/
- Wikidata Wordmap - https://esterpantaleo.github.io/wikidata_wordmap/
- Others? (more listed at https://www.wikidata.org/wiki/Wikidata:Tools/Lexicographical_data)

Lexeme Bots

- November 2018: MewBot (Northern Sami lemmas, lexemes only)
- December 2018: Lingua Libre Bot (add pronunciation audio to lexeme forms)
- August 2019: YurikBot (Russian lexemes and forms)
- August? 2019: Elhuyar Fundazioa bot (Basque lexemes and forms)
- October 2019: SixTwoEightBot (English adverbs – lexemes and forms)
- October 2019: Uzielbot (Hebrew lexemes and forms)

Workshop Discussion

- Review criteria for bots?
- Acceptable (CC-0) sources for sense data?
- Missing properties?
- Anything you hate about the lexicographical namespace? What could be fixed?
- Other questions?