COGNITIVE STUDIES | ÉTUDES COGNITIVES, 18
Warsaw 2018

Article No.: 1712
DOI: 10.11649/cs.1712

Citation: Benjamin, M. (2018). Inside Baseball: Coverage, quality, and culture in the Global WordNet. *Cognitive Studies | Études cognitives*, *2018*(18). https://doi.org/10.11649/cs.1712

MARTIN BENJAMIN

Kamusi Project International, Lausanne, Switzerland
martin@kamusi.org

INSIDE BASEBALL:[1] COVERAGE, QUALITY, AND CULTURE IN THE GLOBAL WORDNET

Abstract

The Global WordNet is succeeding in producing relatively open linguistic data that is coordinated to a degree among numerous languages. The project has grown organically, with no overall plan or direction. The result is a certain amount of incoherence in determining what items should be treated in wordnets, and how the various wordnets should aspire to consistent quality. Using the example of terms related to baseball, which constitute a non-trivial portion of the Princeton WordNet, this paper discusses problems of coverage selection both for English and for other languages, as well as methods to improve quality and depth through public review of current content, and contribution of missing terms and definitions. It is proposed that proper names be removed entirely from WordNet and treated as a separate project, and that individual languages produce annexes of indigenous concepts that can be readily considered within sister projects as a supplement to the Anglo–American weighting of the current endeavor. To produce a consistent product that transmits inter-intelligible understanding at a high level across languages, it is proposed that an open committee of interested stakeholders convene to consider the project's goals and develop a roadmap for how to achieve them.
Keywords: wordnet; lexicography; vocabulary; named entities; multilingual

1 Introduction

Baseball (00471613-n)[2] is "America's pastime",[3] and also an essential part of the cultures of Japan and many countries in Latin America. I hold cherished childhood memories of playing Little

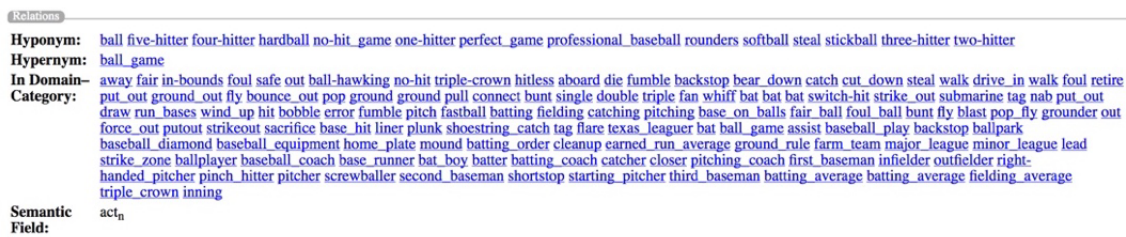
[1]The Oxford English Dictionary defines "inside baseball", n. 2.b., as "Details or information known to, and able to be appreciated and understood only by, aficionados or specialists; uninteresting technical details." A linkable definition on Wiktionary defines the term as "Matters of interest only to insiders": https://en.wiktionary.org/wiki/inside_baseball. Up-to-the-minute contemporary usage can be found at http://kamu.si/inside_baseball_tweets.

[2]All synset reference numbers are from Princeton WordNet 3.0, as provided by search results conducted on the Open Multilingual Wordnet (OMW) (http://compling.hss.ntu.edu.sg/omw/).

[3]Google results for "America's pastime": https://goo.gl/9XqoK2.

League (08231999-n), trading baseball cards (02799442-n), and going to the ballpark (02782778-n) to watch major league (08231499-n) ballgames (00471437-n). Now expatriate in Switzerland, I actively share this passion with my daughter, bringing her to a minor league (08231678-n) game on a visit to the States, filling my luggage with gloves (02800213-n) and wiffle balls (04584056-n), and going to the park to play catch (00458641-n) and practice batting (00126584-n) after school.

Wordnet (Fellbaum, 2008) is a massively important collaborative linguistic resource, with dozens of independently-produced open datasets that are a central part of my work to produce interlinked multilingual lexical resources. Most wordnets for individual languages are aligned to English synsets from the Princeton WordNet (PWN), making an excellent starting point for aligning the expression of concepts across languages. A big assist (00558008-n) in aligning wordnets has been the consistent LMF formatting (Vossen, Soria, & Monachini, 2013) provided by OMW for many languages (Bond & Foster, 2013). One feature of Kamusi,[4] the project I direct, is DUCKS (data unified conceptual knowledge sets), which involves aligning terms from various non-Wordnet bilingual datasets with the meanings defined for synsets in PWN.[5,6] Another feature (currently inactive for financial reasons) asks dictionary users to suggest a term in their language that is equivalent to a concept in the PWN-derived sense inventory, if their search reveals a null result. "Mauli"[7] is a forthcoming mobile app that will directly elicit terms from users for a raft of languages, matched to wordnet concepts and using PWN definitions whenever appropriate. The work that has been done and that continues on wordnets around the world is elemental to graphing a larger matrix of human expression (Benjamin, 2014), with the goal of creating a unified global linguistic data infrastructure.


Relations

Hyponym:	ball five-hitter four-hitter hardball no-hit_game one-hitter perfect_game professional_baseball rounders softball steal stickball three-hitter two-hitter
Hypernym:	ball_game
In Domain–Category:	away fair in-bounds foul safe out ball-hawking no-hit triple-crown hitless aboard die fumble backstop bear_down catch cut_down steal walk drive_in walk foul retire put_out ground_out fly bounce_out pop ground ground pull connect bunt single double triple fan whiff bat bat switch-hit strike_out submarine tag nab put_out draw run_bases wind_up hit bobble error fumble pitch fastball batting fielding catching pitching base_on_balls fair_ball foul_ball bunt fly blast pop_fly grounder out force_out putout strikeout sacrifice base_hit liner plunk shoestring_catch tag flare texas_leaguer bat ball_game assist baseball_play backstop ballpark baseball_diamond baseball_equipment home_plate mound batting_order cleanup earned_run_average ground_rule farm_team major_league minor_league lead strike_zone ballplayer baseball_coach base_runner bat_boy batter batting_coach catcher closer pitching_coach first_baseman infielder outfielder right-handed_pitcher pinch_hitter pitcher screwballer second_baseman shortstop starting_pitcher third_baseman batting_average batting_average fielding_average triple_crown inning
Semantic Field:	act_n

Figure 1: 140 terms with direct ontological relations to "baseball" (00471613-n) in PWN, as shown in OMW

These two things that I appreciate greatly, baseball and wordnet, combine in ways that demonstrate important challenges to the latter. I in no way wish to disparage either in this paper, nor to diminish the contributions of the game to the American English lexicon.[8] However, PWN contains a disproportionate volume of baseball-related terms.[9] At least 0.25% of terms are specifically re-

[4] `http://kamusi.org`, with mobile apps for iPhone (`https://bit.ly/kamusihere_ios`) and Android (`http://bit.ly/kamusihere_android`) that build on WordNet data across languages.

[5] These sources are compared to Wordnet through games that show users a term and its sense definition or other known information in the new dataset, and ask them to identify concept matches from the sense definitions for the same English literal in the existing dataset.

[6] DUCKS will soon expand to a larger set of definition sources, beginning with the English Wiktionary, which will be merged with the PWN senses. Of 936,604 Wiktionary senses for the parts of speech included in Wordnet, 9,889 senses have been automatically identified as exact matches and about 807,950 as definite non-matches, leaving about 118,765 ambiguous items to be merged manually by game players.

[7] "Mauli" is a Hawaiian word that translates roughly as "the spirit of life".

[8] Indicative of the role of baseball in the USA, the Dickson Baseball Dictionary (2011) has 18,000 individual entries covering 10,000 terms, in a 1,000 page volume.

[9] WordNet's founder, George Miller, shared my fondness for the game, and enjoyed documenting its terms, according to personal communication with his academic successor Christiane Fellbaum. Producing a lexicographically perfect representation of the English language was not his original concern in the creation of WordNet, and he could not have foreseen the complications that would ensue when his project was extended to other uses in extra innings (15234212-n).

lated to baseball, with a ballpark (05126066-n) estimate as high as 0.5%, when all the ontological threads are followed (e.g. a "batter's box" (02810270-n) is a hyponym of "box" (02884607-n), which is a meronym of "diamond" (02780916-n), which is a meronym of "ballpark" (02782778-n), which is in the domain-category "baseball)"[10] and all of the terms pertaining to baseball that are not so marked are also discovered.[11] This high concentration on the vocabulary of one largely American pursuit provides a window to issues that affect both the English content of PWN, and the use of that resource as a foundation for generating linguistic data in other languages.

2 Issues for English

The terms in PWN were not chosen based on studied lexicographical criteria, and the definitions were not written to meet the scholarly standards of a dictionary. Nevertheless, the terms and definitions, as well as their synset memberships and ontological relationships, are more or less ossified.[12] These problems affect how well projects built on the WordNet foundations can fulfill their individual objectives. Problems of PWN as a data source for English should be considered separately from the problems that are introduced by using it as a cross-lingual resource for other languages.

2.1 Inventory of words and senses

The extensiveness of PWN's baseball vocabulary shines light on the absence of equally significant terms from other walks of life. "Football" (00468480-n) has about a third as many related terms as baseball, and conflates American football and soccer in a single definition. "Soccer" (00470966-n), the world's most popular sport, has a mere dozen relations. "Horse racing" (00450070-n) has only five.

Let us compare baseball and horse racing for a moment. Both are major sports in the US, and both have added significant contributions to the American idiom. From "across the board" to "on the nose" to "track record", American English is filled with terms related to raising and competing with horses. Many racing glossaries have been compiled (`https://goo.gl/vFLLFt`) that are similar in size to the baseball vocabulary in PWN. Yet, PWN provides the baseball sense of "closer" (09930257-n), "a relief pitcher who can protect a lead in the last inning (15255804-n) or two of the game", while overlooking the racing sense, "a horse who runs best in the latter part of the race, coming from off the pace".[13] These senses have a distinction between maintaining a lead or coming from behind, which is important in understanding financial or political news articles that use the word.[14] The inclusion of the baseball sense enriches PWN; the exclusion of the racing

[10] At least 235 synsets are linked ontologically at one level of remove or greater. These synsets contain 394 total terms, e.g. 10252921-n: southpaw, left-hander, left-handed pitcher, left hander, lefthander, lefty. However, many of those terms are different senses of the same literal, e.g. "bat" 00458456-n, "a turn trying to get a hit", 01413173-v, "strike with, or as if with a baseball bat", 01413561-v, "have a turn at bat", and 01413436-v "use a bat". (02806379-n, "a club used for hitting a ball in various games", is not counted because it is defined in a general sense, although also linked to baseball ontologically). The full PWN Wordnet Baseball Ontology dataset is available at `http://kamu.si/pwn_baseball_ontology` and `http://j.mp/2LLiVPZ`.

[11] Short of reading every definition in PWN, the number cannot be determined. Many defined baseball terms do not include the word "baseball" in their definition. Many terms, such as "ump" (10735984-n), "an official at a baseball game", have "baseball" in their definition, but do not link ontologically; in principle, these could be found by grepping the definition field in the database. Most difficult are terms such as "earned run" (00190040-n), "a run that was scored as a result of an error by the other team", or "down" (02061678-a), erroneously defined as "being put out by a strikeout", that are neither linked ontologically nor contain the word "baseball" within the definition.

[12] To suggest an improvement to the data for future generations of PWN, an in-the-know user who is registered at Github can go to `https://github.com/globalwordnet/english-wordnet/issues` and submit a bug report.

[13] DRF Glossary of Horse Racing Terms, `http://www1.drf.com/help/help_glossary.html`.

[14] As an example, the blurb for the pilot episode of the TV series "Suits" reads, "A "closer" for one of New York City's most successful law firms decides to hire an aloof genius who has passed the bar but never went to law school as his associate." (`https://www.imdb.com/title/tt1973786`). The series starred Meghan Markle, now the Duchess of Sussex. Her marriage to Prince Henry in 2018 was watched by tens of millions of people around the

term impoverishes it.

On the other hand, too many specialist terms would make PWN so unwieldy that the resource would become dysfunctional for users trying to sift through numerous esoteric senses. What is the boundary about whether to include the non-baseball sense of "batting" used by quilters (02810930-n), and the overlooked term "bearding" that describes when batting fibers migrate through surface material? Or the neglected sense of "first base" (03349771-n) that every teenager knows means a romantic kiss?[15] For that matter, is the baseball vocabulary incomplete, for instance omitting "sacrifice bunt" while including "sacrifice" (00130846-n), "bunt" (00128477-n), "fly" (00128638-n), and "sacrifice fly" (00130987-n), not to mention leaving out "grand slam"[16] and the play-on-words "salami" to signify a home run (00132355-n) with the bases (02797881-n) loaded? For the purposes of an unabridged dictionary, the more senses the merrier. For the purposes of a data source for natural language processing (NLP), there are cases where more might be too much — in practice, "salami" is so likely to refer to processed meat, except during a spoken game broadcast, that the baseball usage would merely add noise. There is an unstated difference between a wordnet and a terminology set, with wordnet aiming for general concepts and terminologies aiming to elucidate specific domains. Terminologies are a realm unto themselves, whether the focus of expensive translation efforts such as the 1.4 million multilingual entries in the InterActive Terminology for Europe (IATE) that includes many domains such as aviation,[17] or smaller monolingual efforts such as the stand-alone Aerofiles aviation glossary.[18] Subjectively speaking, though, I care much more that my pilot understands the aviation sense of "flare" that is not included in PWN, pitching the nose of the plane away from the ground just before landing, than the baseball sense (00150097-n) that is. Additionally, concepts like a landing flare that are excluded from PWN (and produced in six languages in IATE) are unlikely to be treated in other wordnets, meaning aspiring pilots for Finnair or Emirates or the Flight Simulator video game are unlikely to encounter them in their mother tongues.

While Section 3.1 proposes a solution for prioritizing the PWN terms that should be included in other wordnets, I do not have a good answer to the question of which senses of which terms should be included in PWN itself. However, we should be aware that the current selection is often arbitrary, including some terms that fall outside of general usage while excluding others that average users might wish to know. In the long run, perhaps some form of democratic selection could be devised, e.g. by tracking which sense users click for further inspection when confronted with polysemous search results. A process for culling extraneous items and seeking neglected pearls would require a programmatic decision by the wordnet community.

2.2 Bad definitions

Many definitions in PWN are adequate to guide a reader about the implications of the sense, but atrocious as lexicographic art. Because I know something about baseball, the definition "steal a base" tells me that "steal" (01111458-v) indicates a base runner (09841696-n) advancing around

world. (Henry is the brother of Prince William, second in line to the British throne, who has studied Swahili and has likely made use of Kamusi's Swahili resources.) The question of which sense of "closer" is intended in the show thus transforms from an academic distinction to one that touches the center of the cultural zeitgeist.

[15] Twitter results for "get to first base with" lead to innumerable examples of "first base" as a kiss: `https://twitter.com/search?q=%22get%20to%20first%20base%20with` , and Fast Company provides a photo: `https://www.fastcodesign.com/3050584/finally-you-can-get-to-first-base-with-your-coffee-cup`. Colloquially, the subsequent three bases indicate further advances in physical intimacy (Go Ask Alice!: `http://goaskalice.columbia.edu/answered-questions/definitions-bases-%E2%80%94-and-im-not-talking-baseball`).

[16] Grand slam "originated in the card game Bridge, referring to a player winning every trick. It carried over into baseball because it refers to a team scoring as many runs as possible in one at-bat", `http://m.mlb.com/glossary/standard-stats/grand-slam`. Video of a grand slam: `https://twitter.com/NCAAsoftball/status/959119271824674816`.

[17] `http://iate.europa.eu/about_IATE.html`

[18] A Glossary of Aviation Terms and Abbreviations: `http://www.aerofiles.com/glossary.html`.

the field independently of a hit (00043902-n) or walk (00127286-n),[19] but there is no way that a non-aficionado could garner a meaning. On the other hand, the definition for "strike" (00109414-n), "a pitch that the batter swings at and misses, or that the batter hits into foul territory, or that the batter does not swing at but the umpire judges to be in the area over home plate and between the batter's knees and shoulders", takes 44 words to describe what Merriam-Webster accomplishes in 18, "a pitched ball that is in the strike zone or is swung at and is not hit fair".[20] That a "baseball team" (08079319-n) is "a team that plays baseball" is a tautology, while non-fans who read that "retire" (01154175-v) means "cause to get out" will find that a mystery.

PWN has a particular problem with definitions that might suffice for some members of a synset, but not for all. A "baseball manager" (09841515-n) can be defined as "a coach of baseball players", but that definition fails for the affiliated "baseball coach". In such cases, either the definition for the entire synset needs to be rewritten, or consideration should be given to splitting the synset.

Bad PWN definitions leave their footprints throughout the Web and within NLP projects. Numerous projects have downloaded one or another generation of PWN as the starting (or complete) points of their own presentations. There is no systematic way for developers to know when changes have been made, and most developers treat the download as a one-off that they will never update. The PWN definition of "strike", for example, can be found in BabelNet,[21] vocabulary.com,[22] Lexipedia,[23] an Amharic dictionary,[24] and over 2400 more.[25] Many of these references will last forever, stored in the Internet Archive even if the original site goes down. The original PWN definitions therefore serve as a form of primary key across diverse datasets, that must be preserved in amber even if PWN itself eventually moves on.

Kamusi has enabled users to mark bad PWN definitions within DUCKS, and will extend that functionality to our main online and mobile search apps when we have time to program it.[26] We also have a working prototype for a game for users to suggest improved definitions (Benjamin, 2015, 2016), which will be activated in Facebook when coding labor is available; a player receives 10 points for writing a winning definition, or 1 point for voting for a winning definition submitted by someone else. The forthcoming pairing of senses from Wiktionary through DUCKS will provide an additional pool of definitions (Mrini & Benjamin, in press) that will often be better than PWN, since they have theoretically survived some level of public review.[27] In many cases, a better definition will be applied to only one or some members of a synset. Our intent is to mark a bad PWN definition as deprecated, but to continue to display it when it is matched to another language's wordnet that built upon that original indication. Additionally, if more than one definition is good, we will show multiple definitions for the same sense, such as displaying both PWN and Wiktionary renditions. When ready, improved definitions will be on offer to PWN and other wordnets.

[19] `http://m.mlb.com/glossary/standard-stats/stolen-base`

[20] `https://www.merriam-webster.com/dictionary/strike`

[21] `https://babelnet.org/synset?word=bn:00074671n&details=1&lang=EN&orig=strike`

[22] `https://www.vocabulary.com/dictionary/strike`

[23] `http://www.lexipedia.com/english/strike#nouns`

[24] `http://www.amharicpro.com/index.php?dr=101&searchkey=%E1%88%B0%E1%8A%95%E1%8B%9D%E1%88%AB`

[25] Google results for the baseball definition of "strike" from PWN: `http://j.mp/2Krdo02`.

[26] Without an operating budget, Kamusi Labs functions through volunteer student internships. Students from universities around the world are assigned to coding projects based on their interests and the relevant computer languages that they know. The laboratory whiteboard is maintained at `http://bit.ly/kamusilabs`. This defective scheme often means that programming proceeds in fits and starts, with no predictability about when new features or applications will be finalized.

[27] As an example, PWN gives a definition of "policewoman" (10449412-n) as "a woman policeman" that is problematic on several levels, whereas Wiktionary, `https://en.wiktionary.org/wiki/policewoman`, provides the perfectly satisfactory definition "A female police officer", for an entry that has more than 100 revisions since 2004. This is not to say that Wiktionary does not have many of its own weaknesses, such as offering "manager, 2. (baseball) The head coach" (`https://en.wiktionary.org/wiki/manager`), which Kamusi would separately treat as a bad definition and put to users to improve.

2.3 Named entities

An instance of "baseball coach" is the synset Stengel, Casey Stengel, Charles Dillon Stengel (11316429-n). In fact, Casey Stengel is the only instance of a baseball coach listed as such, rather than as a player, in PWN. Fourteen men are instances of "baseball player" (09835506-n), all heroes from days gone by. The variations of their names and nicknames constitute 45 literals, many of which have been diligently rendered in Malaysian, Thai, Romanian and Finnish. All the men are members of the National Baseball Hall of Fame (03810561-n) in Cooperstown (09118639-n), New York.

These men were great baseball icons, but awful wordnet subjects. None of their names are fixtures of American English, other than "the Babe",[28] which curiously is not included within the synset Ruth, Babe Ruth, George Herman Ruth, Sultan of Swat (11276100-n).[29] Barry Bonds and Mark McGwire have broken the all-time records of the selected 14, to enormous public excitement in the US, but have not broken into PWN. Conversely, lists are available from other sources that contain the names, teams, and playing years of all of the thousands of men who have ever played in the Major or Negro leagues.[30] An arbitrary set of fifteen names is not useful for English reference, for NLP, or for baseball fans. Similarly, either the National Baseball Hall of Fame should have company with other hyponyms of "Hall of Fame" (03479266-n), or it should not be included at all. Cooperstown has no more place in wordnet than any other burg of 2000 citizens and some incidental claim to fame.

Figure 2: Iconic button for the ERA (Equal Rights Amendment), a chief focus of the US women's movement in the 1970s that is not in PWN.

The baseball acronyms in PWN are similarly inappropriate for a general reference. True fans spend hours comparing players' RBIs (00190180-n) and ERAs (07261300-n), but fans of the major American motorsport NASCAR (not in PWN) will not find their DNAs. On the other hand, many acronyms, such as the genetics sense of DNA (14830364-n) are a legitimate part of the general vocabulary. NATO, the FBI, and the ABCs[31] are all included as they should be — but then, the AU (African Union) and MI6 (British security agency) should also rank. The space for acronyms should be broader than that for names, with a conference on the mound (03792334-n) about how to find the strike zone (08690974-n).

[28] Videos of "the Babe": `https://www.youtube.com/results?search_query=the+babe`.

[29] Yogi Berra (10848946-n), on the other hand, did coin many memorable aphorisms that are fixtures of American English, including, "It ain't over till it's over", "It's déjà vu all over again", and "When you come to a fork in the road, take it".

[30] A search on Amazon for "negro league books" lists 307 results, though not all of the results are unique or relevant matches to the query. `http://kamu.si/negro_league_amazon`. Four players who made the historic move from the Negro League to the Major League are included in PWN list of 14, but "Negro League" itself is absent from wordnet.

[31] The synset for ABCs (05872742-n) rudiment, first rudiment, first principle, alphabet, ABC, ABC's, ABCs, "the elementary stages of any subject", is a foul ball (00128091-n).

Baseball is just one example of how names are inappropriate within wordnet. PWN has too few names to be useful as a reference of American culture or the wider English-speaking world, and most decidedly does not feature named entities of significance elsewhere. There are 44 instances of "saint" in PWN, each with at least two literals in their synset. These saints are a smattering of the roughly 11,000 people canonized in Catholicism (Manning, 2013), with none included from other Christian denominations such as Eastern Orthodoxy, nor given as such from other religions such as Hinduism or Islam. Miller (2008b, p. 23) indicates that the issue of names was never given deep consideration: "No special attempt has been made to include proper nouns; on the other hand, since many common nouns once were names, no serious attempt has been made to exclude them." International wordnet teams exert unnecessary effort struggling to come up with local equivalents for people and places most have never heard of, or that belong to a hegemonic culture that has spent centuries overwriting local histories.

A solution would be to strip *almost all* proper nouns from wordnet, and establish a proper Names Net be set up in its place. One could argue that certain names that are linguistically embedded should remain in wordnet, such as the constellations in the Zodiac, or people or gods such as Sisyphus who have given rise to their own common words. Jesus's apostles would be good candidates to remain, as pillars of a religion that has adherents in every country, while Saint Vitus (11367725-n) does not belong. Santa Claus Santa Claus , Santa , Kriss Kringle , Father Christmas , Saint Nicholas , Saint Nick , St. Nick(10550673-n), not listed as an instance of a saint, should clearly feature in wordnet. "John Hancock" (06404907-n) must remain in the synset associated with "autograph", and probably also be added to the synset for a personal signature (06404582-n),[32] but should be removed as a historical figure *per se* (11027416-n). The slope is a little slippery, but not Sisyphean. Wikipedia addresses the problem by enabling the community to debate whether a person meets notability standards,[33] but that would be a very high-resource undertaking for GWN. Names Net would be a stand-alone repository of named entities — people, places, and organizations — culled from multiple sources, such as an available listing of every town in the world.[34] Data can be kept up to date by coordinating with sources such as the multilingual JRC-Names,[35] rather than the static collection of entities named in PWN. International projects would be encouraged to focus on names of local or regional relevance. Names Net would be a substantial project that would require thought and funding, but could make a contribution that the current instantiation of GWN does not.

2.4 Frozen vocabulary

The list of baseball heroes included in PWN, suspended in time, points to the additional problem that WordNet marks a particular era in the ever-changing history of the English language. PWN obtained its words from a variety of lists (Miller, 2008a). About 50,000 words were distilled from the Brown Corpus,[36] about a million words of edited English prose printed in the United States during the calendar year 1961. An unspecified number of additional words came from a wordlist compiled by Fred Chang at the Naval Personnel Research and Development Center in the 1980s that has left no evident fossil trail of publications or citations.[37] About 10,000 more terms were included

[32] A good dictionary usage example to associate John Hancock, autograph, and signature comes from this baseball-related tweet: "'I stole my signature from John Hancock. I had to have a good autograph for baseball.' — Jarrett" `https://twitter.com/emilycdohogne/status/464457481209786370`. Searching Twitter for "John Hancock" and "baseball" leads to a rich vein of tweets about signed baseballs and other memorabilia.

[33] Category: Wikipedia notability guidelines: `https://en.wikipedia.org/wiki/Category:Wikipedia_notability_guidelines`

[34] National Geospatial Intelligence Agency: `http://geonames.nga.mil/gns/html/namefiles.html`

[35] European Commission >EU Science Hub >Language Technology Resources >JRC-Names: `https://ec.europa.eu/jrc/en/language-technologies/jrc-names`

[36] Francis, W. and Kucera, H. (1964; revised 1979). Brown Corpus Manual. Providence, RI, Department of Linguistics, Brown University: `http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM` (Francis & Kucera, 1979).

[37] A university librarian and a program officer at the Office of Naval Research joined in the search, but we went down swinging (00571444-n).

by comparison to the Comlex Syntax set (Grishman, Macleod, & Meyers, 1994), which itself was a distillation of headwords from the 1980 version of the Oxford Advanced Learner's Dictionary (Hornby, 1980). Other terms came through reference to books of synonyms. The provenance of the baseball vocabulary cannot be reconstructed via the PWN website or published bibliographies, but the large bulk of terms overall came from sources published between 1961 and 1986.

This is not to say that WordNet is completely frozen in time. The PWN site has a link to suggest missing terms via email, but this is not a loudly trumpeted feature — version 3.0[38] had just 62 more synsets than version 2.1.[39] Four late arrivals that could only have entered the lexicon from the late 1990s are Viagra (04218383-n), Cialis (04383537-n), Levitra (04521428-n) and their short-lived competitor Vasomax (03922561-n). A selection of technology terms from the 1990s are included, such as Java (06901053-n), html (06788262-n), Google (06578905-n), and browser (06571301-n), but other terms from that decade or later ("app", "Apple", "smart phone", "javascript") are absent. By and large, newer terms do not enter wordnet no matter how prevalent they become in English, so that a term such as "low-carb diet", a major fad in the aughts,[40] does not appear alongside "low-fat diet" (07564101-n) or "low-salt diet" (07564292-n). Older terms that largely evaporated from popularity before the PWN catchment period might make it into the pool, an example being "avoirdupois" (04999401-n) that had its heyday in the 19th century.[41] However, though in 1866 John Langdon Down fully described the syndrome that bears his name, pre-natal screening did not begin until the 1980s, and the rocketing of "Down syndrome" into common use starting in that decade[42] did not earn it entry to PWN. Without belaboring the point, medical concerns that came into public consciousness after the 1980s are probably not in wordnet; AIDS (14127782-n), mad cow disease (14261846-n), SIDS (14310292-n) and Ebola (14135623-n) made the cut, but "Zika", and "chikungunya" did not. Fellbaum (2016) delivered a workshop presentation that lays out the many challenges confronting additions to PWN, suggesting that corpus-based frequency analysis and crowdsourcing could provide some railings. A systematic method to incorporate new terms has not yet been introduced.

3 Issues for other language wordnets

Most wordnets have been built using the "extend" approach that seeks translations from the PWN master set (Vossen, 1998). This makes a great deal of sense as a starting point, because the labor of identifying many items of significance to people worldwide has already been accomplished. All people sleep, all people have noses in their DNA, and all languages have terms for such concepts. However, many other concepts are not universal. Most African languages do not have words for *winter* or *subway*, because neither describe things experienced by most people on the continent. Japanese, Korean, the Chinese of Taiwan, and the Spanish of Latin America are replete with baseball terms, the rest of the world not so much. Issues regarding both language and culture present a number of challenges to the Global WordNet.

[38] `https://wordnet.princeton.edu/documentation/wnstats7wn`

[39] `https://wordnet.princeton.edu/documentation/21-wnstats7wn`

[40] Neither "aughts" nor 2000s are in wordnet, though the former gained some currency as the appellation for the latter (Mead, 2010). Decade names in PWN are worth investigating as a small tangent. Some are in (1750s, 1850s, 1950s), some are not (1810s, 1910s, seemingly anything from 1740 or earlier). The 1880s and 1980s both have the literal "eighties" in their synsets, and the 1890s and 1990s are both the "nineties", but all other included decades from the nineteenth century are only shown with digits. Decade names are gimmes (a term that has entered the popular lexicon, but not PWN, through golf: `https://www.dictionary.com/browse/gimme`) for other languages, but only about 11 have taken on some of them. Dutch and Slovene usually misrepresent the way decades are described in their language, only giving a number like "1820". Neither the 1390s of the Islamic calendar nor the 5730s of the Hebrew calendar, contemporaneous with the 1970s, are mentioned, though the calendar month names for both of those systems are present. "Twentieth century" is the only century name that is included. No other calendrical systems are evident.

[41] Google Books Ngram Viewer for "avoirdupoids": `http://j.mp/2zpHSdW`

[42] Google Books Ngram Viewer for "Down syndrome": `http://j.mp/2L3jX5w`

3.1 Which PWN senses to cover

Which concepts to cover has been only a limited focus of discussion within the wordnet community for years (Rodríguez et al., 1998). In particular, a credible "core"[43] set of 5000 base concepts has been derived that distills many of the more universal aspects of the human experience, with synsets selected by comparing a frequency analysis of the British National Corpus with PWN, and then manually deciding which senses had the most universal relevance (Boyd-Graber, Fellbaum, Osherson, & Schapire, 2006). *Train* is included but *subway* is not. *Winter* is included as a noun, but not its less common usage as a verb. *Baseball* is not included at all. However, six concepts from from the baseball domain do remain in the core: *hit* (00043902-n), *base on balls* (00127286-n), *catch* (01082454-v), *right field* (04091839-n), *lead* (08592165-n), and *inning* (15255804-n). The proceedings of the nine Global WordNet (GWN) conferences since 2002 contain a variety of discussions about how to treat concepts that are missing from PWN, whether those ideas should be merged in from other languages (section 3.4) or from projects for specific domains such as theology (Slaughter, Wang, Morgado da Costa, & Bond, 2018). However, unless I have missed something, no papers have been written along the lines of, "What is a bullpen (02917964-n) and should we balk (00107279-n) at having it in our wordnet?". PWN is by and large considered a fixed inventory to which concepts could potentially be added, or optionally ignored, sometimes repaired, but never deleted.

The two issues to highlight here are stopping and keeping on going. Many wordnet teams have stopped work at or near the edges of the core. Icelandic, for example, completed 4,951 synsets, skipping all of the baseball terms and 43 others. This limited concept set is useful as a bilingual pocket reference; when IceWordNet[44] has been imported to Kamusi, the data will be adequate to help a Basque speaker to change planes at the Reykjavik airport using our mobile app, without the need to speak English. Yet a lexicon of 5000 concepts remains a toy when it comes to serious research or use in NLP. When the Basque speaker arrives in New York, she is going to need to find the subway. Furthermore, she might wish to take in a ballgame, perhaps even a subway series,[45] but the wordnet will not help her because the Basque group (sensibly) omitted most baseball terms. Though she might well speak Spanish, a language with millions of passionate baseball fans, the Spanish wordnet was compiled in soccer-mad Iberia instead of the Caribbean basin and was thus also flummoxed by the domain. On the other hand, the teams producing Basque or Castellano should not feel the need to expend time figuring out terms of no relevance to their lives, nor feel they have struck out (01509280-v) by not translating all the terms from an abstruse and idiosyncratic foreign inventory. A partial solution that we will soon launch experimentally is to elicit missing equivalents directly from knowledgeable members of the public, enabling Cubans and Dominicans to contribute Spanish terms for their beloved baseball where the wordnet team in Spain left off, but this is a haphazard approach to gathering extensive, consistent data among dozens of languages.

Again, baseball is not the problem. The problem is that there is a certain randomness to wordnet, both in the terms covered in PWN and in the terms that other languages choose to treat from the PWN set, that makes for a hit-or-miss (01924803-a) user experience. If the goal is a consistent product that transmits inter-intelligible understanding at a high level across languages, a more coherent strategy for concepts above the core should be pursued. Kamusi plans a data-driven solution after all available wordnets have been imported to our system. The number of other languages that have produced equivalents for each synset will be measured and reported, on the premise that the concepts with the highest batting average (13817872-n and thus 13818143-n) have the most universal resonance and will appear in the most wordnets. The logic of this assumption is not airtight, notably because wordnets based on the core or on computation do

[43] `http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt`

[44] `http://www.malfong.is/index.php?lang=en&pg=icewordnet`

[45] "Subway series" is not defined in PWN, but has about 380,000 Google search results. The concept is discussed at `https://www.sportingcharts.com/dictionary/mlb/subway-series.aspx`

not necessarily involve prioritizations by speakers of the languages in question, but should give an overall approximation of human evaluation. Notably, most languages choose to ignore most baseball terms. Japanese appropriately delves deep into the repertoire, and non-baseball cultures Finnish and Thai, and to a lesser extent French and Romanian, also frequently provide dubious equivalents (such as "joc de baseball" for "baseball play" (00564177-n) in Romanian that better translates as a game of baseball from first pitch to last out (00129527-n) than as a short rehearsed sequence of actions), but the extent to which other languages shun the baseball vocabulary speaks loudly to its usefulness in GWN. On the other hand, equivalents for "soccer" have been provided in at least three dozen wordnets, making it more popular than air, food, or water. Although not a formal selection process, the choices actually made by individual wordnet teams — dozens including "fly" the insect (02190166-n) versus a handful pursing "fly" the baseball hit high into the air (00128638-n) — will serve as votes for the PWN terms of global significance. We will offer the earned run average (07261300-n) rankings to the wordnet community as open data, where they could be used by individual projects to expand their vocabulary of concepts that are revealed as desirable within a unified global wordnet. On the part of Kamusi, the rankings will be used to prioritize concepts for languages that do not yet have wordnets, as well as missing concepts in existing wordnets, in the effort to build or improve dictionaries for those languages; a result will be de facto wordnets for those new languages, with the selection of terms weighted a priori for those already favored by existing languages within GWN.

3.2 Mistranslations

Though Mickey Mantle (11155196-n) retired from baseball in 1968, he still holds several major league records. He also, according to WOLF (French wordnet), would be called "manteau" in French, a translation of the mantle (03719343-n) over a fireplace. If Mantle grounded (01406356-v and 01406512-v) a ball so that it rolled on the surface instead of flying through the air, he probably did not think of it as a shipwreck. The Finnish wordnet team tried translating both of those synsets, guessing or computing the first time that it was "haaksirikkoutua" or "ajaa matalikolle" (something to do with washing upon the shoals) and the second time that it was "lyödä maapallo" (crashing to earth). As a consequence, these Finnish terms are inextricably associated with "ground", without context, in numerous downstream projects,[46] with the real likelihood of being plugged into a translation that should use the electrical sense (01292534-v). When another wordnet team makes a similarly wrong translation, such as the term "roletear" apparently invented for grounder in the Spanish wordnet, completely false equivalencies ripple into the firmament of "fact".[47]

The baseball vocabulary is not unusual in the extent to which English terms are whiffed (01409888-v) in other wordnets. Arabic, for example, gives terms for the idea of "passing sentence upon" in its translation of *judge* (00672433-v) in the sense of estimating, the Finns mistakenly translate John Hancock as "Matti Meikäläinen" (which actually matches to the non-PWN "Joe Schmoe" in English) and Romanians start quibbling with translations the moment they encounter the data. Mistakes are inevitable. However, baseball shows how the methods used to construct many wordnets may have exacerbated the problem; either the human translators were forced to guess about items for which they would need advanced or specialist English, or machine methods found false positives due to shared spellings with more common concepts.

In the first pass, Kamusi is just as vulnerable to perpetrating mistranslations as are the other projects that build on wordnet data. The solution is manual correction of errors. This is something we plan to implement within Kamusi as soon as possible, with corrections available to feed back to the original groups. Users will be challenged to fix mistakes they encounter, with the chance to mark bad entries and propose alternatives. The intent is for users to have fun and feel rewarded

[46] Google results for "lyödä maapallo" : `http://j.mp/2C9oez0`
[47] Glosbe: `http://j.mp/2Egoq5z`; BabelNet: `http://j.mp/2Egoja9`

for improving the resource for their language, while requiring contributions to pass through a validation process that ensures accuracy. Whether this approach is successful depends on our ability to attract a crowd.

3.3 Missing definitions

Few wordnets produced definitions in their own languages. As a result, terms are assumed to mean whatever they are said to mean in the English definition of the English terms in the synset. For example, Finnish "laiton", an adjective meaning "illegal" or "illegitimate", can only be understood in reference to one of the 18 English definitions for which it is said to be equivalent, including the nouns "strike" in baseball and "no ball" in cricket. As discussed above, many of the English definitions are problematic to begin with. When one factors in the semantic drift induced by inexact equivalents between languages or by human or machine mistranslations, a lot of uncertainty can result about how closely the terms match across languages. Own-language definitions provide clarity, and are essential for speakers of a language who do not also happen to speak English well enough to devine the meaning of a concept based on the way it is described in PWN. A Finnish definition for "polttaa", which they use to translate *whiff* (01409888-v), would show that the terms do not align very closely, but that the Finnish term does fit correctly with the broader concept of players having their turns terminated and leaving the field. On the other hand, definitions are difficult to write and take a lot of time.

As with missing words, Kamusi has a system for users to provide missing definitions, using the same core as the system for English improvements discussed above. The components have been programmed, but actually administering the system requires management resources that have not been found. As with terms that are added or repaired, definitions gathered through this method will be available for their original groups to use or improve.

3.4 Missing indigenous concepts

Rugby (00470966-n) is popular in South Africa, but only eight terms are directly in that domain category for consideration by the teams developing wordnets for six South African languages. Cricket is followed by hundreds of millions who speak the 18 languages in the IndoWordNet, but only twelve terms are in the domain category for the sport (00476389-n), along with a smattering of entries that are not directly connected.[48] The Basque sport of jai alai (00480366-n) appears in name only, with no entries for player positions, equipment, or anything else that is unique to the game. The issue of missing concepts is again not new to the GWN. Several projects have gone to some extent to include local terms, such as the unique expressions of body parts in South African languages (Mojapelo, 2016). Notably, Polish was built from scratch using Polish corpora and subsequently manually mapped to English, producing 8000 new noun lemmas (9000 noun lexical units) that do not occur in PWN (Piasecki, Szpakowicz, Maziarz, & Rudnicka, 2016). What has not yet happened is for the English side of those synsets to be recirculated for consideration by other language groups, which would expand the global set to cross-border concepts that did not make it into PWN. For example, we elicited words from the crowd in more than 30 languages, including Portuguese and Australian languages, for the head cushion worn by African women for carrying heavy loads, "inkatha" in Zulu, which does not exist as an English term.

Certainly, cricket terms developed in India would have an appreciative audience in South Africa, and other branches of knowledge could be addressed by diving into a specialist's or aficionado's domain of interest — boating, birding, basketry. Finding untreated indigenous concepts from the general vocabulary is a much more difficult task.

Lexicographers of well-documented languages have recourse to corpora, which could be readily mined and matched against the term inventory in a given language's wordnet. Comparing existing

[48] `http://www.cfilt.iitb.ac.in/indowordnet/index.jsp`

Figure 3: Woman wearing a protective head cushion.

wordnets to corpora would start with Sketch Engine,[49] to generate a list of candidate terms in situ and subtract the items that appear in both the wordnet and the corpus. Further substantial human review would be required to determine the remaining terms that are headwords that should enter a wordnet. Many of these revealed terms will in fact have matches in PWN, which could be discovered by manually searching PWN for an English equivalent. For example, were an analysis of a Spanish corpus to reveal that "gráfico de sectores" did not appear in the Spanish wordnet, a person could look up "pie chart", choose the appropriate PWN entry, and mark the two as equivalent. The DUCKS process does this within Kamusi, though the base of eligible concepts is expanding beyond those demarcated in PWN. Once the matching process is complete, the leftover items are candidates for membership in the indigenous lexicon. Such terms may in fact have direct equivalents in English or other languages, but the local language would be the starting point for the discovery of the concept set for an expanded GWN.

Languages without corpora, about 6900 of the 7000 spoken globally, present an even more difficult challenge. Hundreds have dictionaries that can be mined in ways similar to corpora. Doing so is time-consuming and arduous work. We experimented for the Fula language with a dictionary (Osborn, Dwyer, & Donohoe, 1993) that is more amenable to parsing than most, and were able to find likely matches to PWN in about 72% of cases. This does not mean, however, that the remaining 28% of terms are strictly indigenous. For example, the source dictionary has the term "annduɗo" that is described in English as "person who is knowledgeable, wise", whereas PWN has "wise man" that is arguably the same concept, and was missed by our automated process (Mrini & Benjamin, 2017). As with corpus mining, additional human review can line up some dictionary terms with PWN, and highlight the remainders as indigenous. The efficacy is limited, though; the Fula dictionary only contained 11,000 headwords, which is large for a non-market language. For languages that do not have any digitized resources to use as a starting point, Kamusi or other tools to bring in equivalents to PWN and other global data are inherently ill-suited to uncovering indigenous terms. Languages that use the Rapid Word Collection method[50] could align to the Global WordNet through DUCKS and then contribute their non-matching concepts as indigenous, but success would require a number of planets (funding, intellectual property, personnel) to come into alignment first.

When a set of indigenous terms can be produced for a language, a useful procedure would be

[49] `https://www.sketchengine.eu/`
[50] `http://rapidwords.net/`

to publish indigenous terms explicitly as annexes — a Polish annex, a Dutch annex, a Zulu annex — that other groups could then work through to the extent they find relevant. An annex system could also be a reasonable solution to domain vocabularies that are underrepresented in PWN, such as cricket or theology.

4 Conclusion

Constituting as many as one in 200 terms in the overall data, baseball probably has an outsized influence in PWN relative to its place within English, and is certainly disproportional as a focus for the vocabulary of most other languages. Its special place in wordnet provides a lens to examine issues of larger importance to the project, namely questions of the range and selection of concepts and vocabulary, as well as issues of quality and cultural bias. Many of these problems have been raised before. However, the way forward is at an impasse, first because changing the existing trunk could cause problems for the many projects that branch off it, and second because there is no central organization that exists to debate such matters and arrive at amenable solutions. Unlike baseball, wordnet has no rulebook and no governing board. Perhaps this is an acceptable condition, with each project continuing on its own track and then coordinating informally at the biennial GWN conferences. If there is a hunger for a wordnet that has systematic integration across languages, though, it might be time to convene an organizing committee that can home in on the goals and boundaries of the overall project, and develop a general game plan for how to get there — or, shall we say, step up to the plate (03528901-n) and knock an outside (00023655-a) curve (00107875-n) in a blast (00128867-n) out of the park (02782778-n).

References

Benjamin, M. (2014). Molecular lexicography: A lexical data model for Human Language Technology. Retrieved March 2, 2018, from `https://kamusi.org/molecular_lexicography`

Benjamin, M. (2015). Crowdsourcing microdata for cost-effective and reliable lexicography. In *Proceedings of AsiaLex 2015 Hong Kong* (pp. 213–221).

Benjamin, M. (2016). Problems and procedures to make Wordnet Data (Retro)Fit for a multilingual dictionary. In V. Barbu Mititelu, C. Forascu, C. Fellbaum, & P. Vossen (Eds.), *Proceedings of the Eighth Global WordNet Conference* (pp. 27–33). Retrieved from `http://jiangbian.me/papers/2016/gwc2016.pdf`

Bond, F., & Foster, R. (2013). Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL–2013* (pp. 1352–1362). Sofia: Association for Computational Linguistics (ACL).

Boyd-Graber, J., Fellbaum, C., Osherson, D., & Schapire, R. (2006). Adding dense, weighted connections to WordNet. In P. Sojka, K.-S. Choi, C. Fellbaum, & P. Vossen (Eds.), *GWC 2006: Third International WordNet Conference, GWC 2006 Jeju Island, Korea, January 22–26, 2006: Proceedings* (pp. 29–35). Retrieved from `http://semanticweb.kaist.ac.kr/conference/gwc/pdf2006/gwc06.pdf`

Fellbaum, C. (Ed.) (2008). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Fellbaum, C. (2016). *How and when to add a new concept and how to define it.* Paper presented at Workshop on the Collaborative Interlingual Index, Global WordNet Conference 2016, Bucharest, Romania.

Francis, W., & Kucera, H. (1979). *Brown Corpus Manual.* Providence, RI: Department of Linguistics, Brown University. Retrieved from `http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM`

Grishman, R., Macleod, C., & Meyers, A. (1994). Comlex Syntax: Building a computational lexicon. In *COLING '94 Proceedings of the 15th conference on Computational linguistics* (Vol. 1, pp. 268–272). `https://doi.org/10.3115/991886.991931`

Hornby, A. S. (Ed.). (1980). *Oxford advanced learner's dictionary of current English.* Oxford: Oxford University Press.

Manning, K. (2013, November). How many saints are there? *US Catholic*, *78*(11), 46. Retrieved March 2, 2018, from `http://www.uscatholic.org/articles/201310/how-many-saints-are-there-28027`

Mead, R. (2010, January 4). What do you call it? *The New Yorker*. Retrieved March 2, 2018, from https://www.newyorker.com/magazine/2010/01/04/what-do-you-call-it

Miller, G. (2008a). Forward. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. xv–xxii). Cambridge, MA: MIT Press.

Miller, G. (2008b). Nouns in Wordnet. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 23–46). Cambridge, MA: MIT Press.

Mojapelo, M. (2016). Semantics of body parts in African WordNet: A case of Northern Sotho. In V. Barbu Mititelu, C. Forascu, C. Fellbaum, & P. Vossen (Eds.), *Proceedings of the Eighth Global WordNet Conference* (pp. 233–241). Retrieved from http://jiangbian.me/papers/2016/gwc2016.pdf

Mrini, K., & Benjamin, M. (2017). Towards Producing Human-Validated Translation Resources for the Fula language through WordNet Linking. In *The Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)* (pp. 58–64). Varna: RANLP. https://doi.org/10.26615/978-954-452-042-7_008

Mrini, K., & Benjamin, M. (in press). Linking the English Wiktionary: A source for new multilingual data for Kamusi and WordNet. *Linguistic Issues in Language Technology: Special Issue on Linking, Integrating and Extending Wordnets*.

Osborn, D., Dwyer, D., & Donohoe, J. (1993). *A Fulfulde (Maasina)–English–French Lexicon: A root-based compilation drawn from extant sources followed by English-Fulfulde and French-Fulfulde listings*. East Lansing: Michigan State University Press.

Piasecki, M., Szpakowicz, S., Maziarz, M., & Rudnicka, E. (2016). plWordNet 3.0 — Almost there. In V. Barbu Mititelu, C. Forascu, C. Fellbaum, & P. Vossen (Eds.), *Proceedings of the Eighth Global WordNet Conference* (pp. 290–299). Retrieved from http://jiangbian.me/papers/2016/gwc2016.pdf

Rodríguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., & Roventini, A. (1998). The top–down strategy for building EuroWordNet: Vocabulary, base concepts, and top ontology. In P. Vossen (Ed.), *EuroWordNet: A multilingual database with lexical semantic networks* (pp. 45–80). Dordrecht: Springer. https://doi.org/10.1007/978-94-017-1491-4_3

Slaughter, L., Wang, W., Morgado da Costa, L., & Bond, F. (2018). Enhancing the Collaborative Interlingual Index for Digital Humanities: Cross-linguistic analysis in the domain of theology. In F. Bond, C. Fellbaum, & P. Vossen (Eds.), *Proceedings of the 9th Global Wordnet Conference, Singapore, 8–12 January 2018*. Global Wordnet Association.

Vossen, P. (1998). Introduction to EuroWordNet. In P. Vossen (Ed.), *EuroWordNet: A multilingual database with lexical semantic networks* (pp. 1–17). Dordrecht: Springer. https://doi.org/10.1007/978-94-017-1491-4_1

Vossen, P., Soria, C., & Monachini, M. (2013). Wordnet-LMF: A standard representation for multilingual Wordnets. In G. Francopoulo & P. Paroubek (Eds.), *LMF Lexical Markup Framework* (pp. 51–66). Hoboken, NJ: Hermess/Lavoisier. https://doi.org/10.1002/9781118712696.ch4

A conference version of this work was financed by the Distributed Systems Information Laboratory (LSIR) at the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. The work undertaken to revise and extend the paper for publication was funded at the author's own expense.

The author declares that he has no competing interests.

Publisher: Institute of Slavic Studies, Polish Academy of Sciences, University of Silesia & The Slavic Foundation