Inside Baseball:[1]

Coverage, quality, and culture in the Global WordNet

Martin Benjamin

LSIR, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
martin.benjamin@epfl.ch

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 position 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

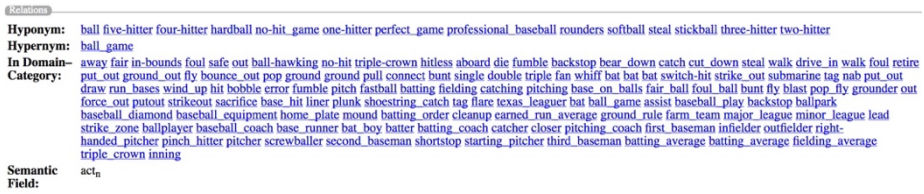
Baseball (00471613-n)[2] is "America's pastime" (https://goo.gl/9XqoK2), and also an essential part of the cultures of Japan and many countries in Latin America. I hold cherished childhood memories of playing Little League (08231999-n), trading baseball cards (02799442-n), and going to the ballpark (02782778-n) 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

[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

[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/)

(04584056-n), and going to the park to play catch (00458641-n) and practice batting (00126584-n) after school.

Wordnet [1] 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. One feature of Kamusi,[3] the project I direct, is DUCKS (data unified conceptual knowledge sets), which involves aligning terms from various bilingual datasets with the meanings defined for synsets in PWN.[4,5] 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. The work that has been done and that continues on Wordnets around the world is elemental to graphing a larger matrix of human expression [2], with the goal of creating a unified global


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 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

linguistic data infrastructure.

These two things that I appreciate greatly, baseball and Wordnet, combine in ways that demonstrate important challenges to the latter. PWN contains a significant number of baseball-related term,[6] constituting at least 0.25% of synsets and, ballpark (05126066-n), perhaps as high as 0.5%.[7] I in no way wish to disparage either in this

[3] http://kamusi.org, with mobile apps for iPhone (https://is.gd/PDXJ15) and Android (https://is.gd/IyODZl) that build on WordNet data across languages.

[4] 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.

[5] 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.

[6] 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 later innings.

[7] Short of reading every definition in PWN, the number cannot be determined because many defined baseball terms do not include the word "baseball" in their definition and are not linked

paper, nor to diminish the contributions of the game to the American English lexicon.[8] This high concentration on the vocabulary of one largely American pursuit, though, 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. 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", 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".[9] 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. The inclusion of the baseball sense enriches PWN; the exclusion of the racing 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 sense of "batting" used by quilters (02810930-n), and the overlooked term "bearding" that describes when batting

ontologically, an example being "wild pitch" (00109892-n), defined as, "an errant pitch that the catcher cannot be expected to catch and that allows a base runner to advance a base".

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

fibers migrate through surface material? 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" and "salami" to signify a home run (00132355-n) with the bases 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 that the baseball usage would merely add noise. I do not have a good answer to the question of which senses of which terms should be included in PWN. 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, "steal a base" tells me which sense of "steal" (01111458-v) is indicated, 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".[10] 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.

We have enabled users to mark bad PWN definitions within DUCKS for almost a year, and will extend that functionality to our main online and mobile search apps after we have installed user authentication. We have designed a game for users to suggest improved definitions [3, 4], which we also hope to activate within Facebook after authentication is resolved; 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 [5] that will often be better than PWN, since they have theoretically survived

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

some level of public review.[11] 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 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.

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", which curiously is not included within the synset {Ruth, Babe Ruth, George Herman Ruth, Sultan of Swat} (11276100-n).[12] 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. 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.

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. At the same time, international teams exert unnecessary effort struggling to come up with local equivalents for people and places most have never heard of.

A solution would be to strip *all* proper nouns from Wordnet, and establish a proper Names Net be set up in its place. Names Net would be a repository of named entities – people, places, and organizations - culled from multiple sources, such as an available

---

[11] 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 errors, which Kamusi is seeking to address separately.

[12] Yogi Berra (10848946-n) coined many memorable aphorisms, 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".

listing of every town in the world.[13] Data can be kept up to date by coordinating with sources such as the multilingual JRC-Names,[14] 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 does not.

3 Issues for other language Wordnets

Most Wordnets have been built using the "extend" approach that seeks translations from the PWN master set. 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, 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, 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.

3.1 Which PWN senses to cover

Which concepts to cover has been a focus of discussion within the Wordnet community for years, so I will not try to rehash old ground. In particular, a credible set of 5000 base concepts has been derived that distills many of the more universal aspects of the human experience. *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 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; within a few weeks of this writing, when IceWordNet 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, but the Wordnet will not help her because the Basque group (sensibly) omitted most

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

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

baseball terms and, though she might well speak Spanish, that Wordnet was compiled in Iberia instead of the Caribbean 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 terms directly from knowledgeable members of the public, 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 that 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.

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. If Mantle grounded (01406356-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, though the Finns who translated the concept as "haaksirikkoutua" guessed or computed that the term had something to do with washing upon the shoals.

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, 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.

The solution is manual correction of errors. This is something we plan to implement within Kamusi, hopefully before the end of 2017, 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 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. 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 divine the meaning of a concept based on the way it is described in PWN. For example, 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 a player having their turn 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 fixed, 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 in that domain for consideration by the teams developing wordnets for six South African languages. Twelve terms are in the domain for cricket (00476389-n), a sport followed by hundreds of millions who speak the 18 languages in the IndoWordNet. The issue of missing concepts is again not new to the Global Wordnet. Several projects have gone to some extent to include local terms, notably many new synsets developed in Polish [6] and unique expression of body parts in South African languages [7]. 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. A useful procedure would be 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.

Figure 2: Woman wearing a protective head cushion

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 roadmap 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

1. Fellbaum, C.: WordNet: An Electronic Lexical Database. Cambridge, MA, MIT Press (2008)
2. Benjamin, M.: Molecular Lexicography: A Lexical Data Model for Human Language Technology. https://kamusi.org/molecular_lexicography (2014)
3. Benjamin, M.: Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography. AsiaLex 2015, Hong Kong (2015)
4. Benjamin M.: Problems and Procedures to Make Wordnet Data (Retro)Fit for a Multilingual Dictionary. Global WordNet Conference 2016, Bucharest, Romania (2016)
5. Mrini, K. and Benjamin, M.: Linking the English Wiktionary: a Source for New Multilingual Data for Kamusi and WordNet, Language Technology: Special Issue on Linking, Integrating and Extending Wordnets (2017 Forthcoming)
6. Piasecki,M., Szpakowicz, S., Maziarz, M., and Rudnicka, E.: plWordNet 3.0 -- Almost There. Global WordNet Conference 2016, Bucharest, Romania (2016)
7. Mojapelo, M. Semantics of body parts in African WordNet: a case of Northern Sotho. Global WordNet Conference 2016, Bucharest, Romania (2016)