ChatGPT

Cross-Linguistic WordNets: Aligning Synsets in English and Serbian

Introduction

WordNet is a lexical database that groups words into **synsets** (synonym sets) representing distinct concepts [1] [2]. Each synset is accompanied by a short definition or *gloss* and example, and synsets are interlinked by semantic relations (like hypernymy, meronymy, antonymy). Originally developed for English (the Princeton WordNet), this framework has been expanded to many languages, yielding a **global network of WordNets**. WordNets in different languages aim to represent the lexicon of each language while maintaining cross-linguistic alignment through shared concepts. This report examines how synsets and their definitions are created and aligned across languages – focusing on English and Serbian WordNets – and discusses the criteria for synset membership. We also explore how cross-linguistic semantic alignment is maintained or challenged, especially in a less-resourced language like Serbian, and provide specific examples comparing English and Serbian synsets. Finally, examples from Spanish, Dutch, Hindi, and other languages are included to illustrate broader multilingual WordNet dynamics.

Glosses and Definitions Across Languages

In a WordNet, each synset has a **gloss** – a textual definition (often with usage examples) that explicates the concept. In the English WordNet, glosses were originally written in English by lexicographers. For other languages, creating or aligning these definitions is a crucial step in WordNet development. A common approach in multilingual projects is to **translate the English glosses** into the target language. For example, the developers of the Spanish WordNet 3.0 explicitly state: *“We have translated the synsets and the glosses to Spanish”* [3], producing a fully bilingual resource. This means that for each English synset, the Spanish WordNet provides a corresponding Spanish gloss that aims to capture the same meaning. Alignment of glosses at the *word level* was done whenever possible [3] – ensuring that important content words in the definition correspond to equivalent words in the translated gloss. By translating glosses, the Spanish WordNet preserves the semantic content of the English definitions, facilitating direct comparisons and use in cross-lingual applications. Other languages have followed a similar “expand” strategy, translating Princeton WordNet’s glosses; for instance, the Serbian WordNet project initially borrowed many definitions from English WordNet or translated them via bilingual dictionaries [4].

However, translating glosses is not always straightforward. In some cases, the English gloss may need adaptation to be accurate and natural in the target language. The Serbian WordNet team found that *not even the glosses were exempt from revision* for quality purposes [5]. For example, an English gloss defined *mycoplasma* as *“the smallest self-reproducing prokaryote”*. Serbian experts noted that new scientific knowledge about life cycles required tweaking this definition. Consequently, *“the gloss for the synset in the Serbian WordNet has been changed”* to reflect updated understanding [6]. This illustrates that while many multilingual WordNets start with translated English glosses, they often **modify or rewrite definitions** to better suit the target language’s knowledge and usage. Glosses might be localized by consulting native dictionaries or encyclopedias. Indeed, in the Serbian WordNet development, lexicographers used resources like the Serbian translation of the *Cambridge Encyclopedia of Languages* and other specialized dictionaries to craft or verify glosses [7] [8] .

In summary, definitions across languages are largely aligned in content (anchored to the same underlying concept), but their formulation is adapted to each language. High-quality WordNets strive to provide idiomatic and precise glosses in the target tongue, rather than verbatim translations, especially when the English gloss would be unclear or factually debatable. By keeping glosses conceptually aligned (describing the same synset concept) yet linguistically natural for each language, multilingual WordNets ensure that users in each language can understand the synset meaning, and that the *semantic alignment* across languages remains intact via the shared underlying concept.

Synset Membership Criteria

A fundamental question in constructing any WordNet is deciding **which words belong together in a synset**. In Princeton WordNet, the guiding principle is that synset members are words (lemmas) with **essentially the same meaning, interchangeable in many contexts** [2] . In other words, if two words can be substituted for one another in a sentence without significantly changing the meaning or truth value, they can be considered synonyms and placed in the same synset. Additionally, all members of a synset share the same **part of speech** – an important constraint since cross-POS synonyms are generally not allowed [9] . For example, *happy* and *joyful* (both adjectives) can form a synset, but one would not place an adjective and a noun together, nor a verb and a noun, etc., as their syntactic categories differ [9] . This holds true across languages: each synset is POS-homogeneous, and the criteria of close semantic equivalence governs membership.

When expanding WordNet to other languages, the same criteria apply, but an extra step is needed: **finding the appropriate words in the target language** that express the concept of an existing synset. Using the “expand” approach (translating from English synsets), lexicographers identify one or more candidate words in the target language that correspond in meaning to the English synset gloss [10] . A word is included in a synset if it is judged to be a **true lexical equivalent** of the concept. Often, this involves consulting bilingual dictionaries, thesauri, and corpora. The Serbian WordNet team, for instance, validated potential Serbian synonyms by corpus frequency and multiple dictionary sources [11] [12] . If multiple Serbian words were possible translations for an English synset, they examined usage frequencies (even using Google hits) to decide which term is most prevalent and appropriate [11] . They also consulted linguists and domain experts via mailing lists for tricky cases [13] . This rigorous process ensures that the chosen words are indeed synonyms in Serbian that convey the same concept as the English synset.

Some additional criteria and practices include:
- **Register and style:** If the English synset contains both a formal and colloquial synonym (e.g., *automobile* vs. *car*), a target language might likewise include equivalents at different registers, but only if they are truly interchangeable in context. Dialectal variants can be included as separate synonyms if the WordNet aims for broad coverage (e.g., Spanish WordNet includes both *coche* (Spain) and *carro* (Latin America) for “car”) [14] .
- **Morphological variants:** WordNet tends to treat different morphological forms of the same root as separate lemmas if they are standard usage (e.g., *amoebic* vs *amebic* in English). In another language, such forms might collapse into one. For example, an English synset with five variants (*amoebic, amebic, ameban, amoeban, amoebous* – all variant forms meaning “caused by amoebas”) might correspond to just one adjective in Serbian (*amebni*), since those English terms are essentially spelling or morphological variants [15] . In Serbian WordNet, these would not be five separate synonyms – they would include only the single common term, ensuring the synset reflects actual synonymy rather than

orthographic variation.
- **Multi-word expressions:** If a language lacks a single-word equivalent for a concept, a multi-word expression might be included as a synset member. The criterion remains that the phrase functions as a **lexicalized unit** in that language. For instance, English has a single noun *uncle* for a concept that Serbian might only express as *majčin brat* ("mother's brother") or *očev brat* ("father's brother") if no umbrella term exists. Such phrases could be candidates for synset inclusion if they are commonly used as a unit. In practice, many multilingual WordNets prefer to include only true dictionary lemmata, not ad-hoc phrases, so a gap like "uncle" in Serbian might be handled via other means (see alignment section below). But for less formal concepts, multi-word synonyms are indeed used (e.g., the Dutch WordNet includes compound or phrasal synonyms when needed [16]).

In essence, a word is admitted into a synset if it **faithfully represents the same concept** defined by the synset, and is a synonym (or standard lexical variant) of the other members. The process in Serbian, Spanish, and other WordNets involves careful cross-linguistic mapping: not every English synset will have a direct equivalent word, and sometimes new synsets are created for language-specific concepts. The inclusion criteria must remain consistent to preserve the integrity of synonym sets – only true synonyms go together. This careful curation has a side effect: many synsets in languages like Serbian end up with fewer members on average than their English counterparts, since English (with its large vocabulary and dialectical variations) often has more synonyms per concept. Indeed, the Serbian WordNet averages about 1.68 words per synset, with most synsets having only one or two synonyms [17] , whereas English synsets often contain multiple synonyms for well-covered concepts. This reflects a genuine difference in available synonyms (and willingness to include variants) rather than a failure of lexicographers; it highlights how **lexical richness and redundancy differ by language**, even while the WordNet design principle of tight synonymy remains constant.

Cross-Linguistic Semantic Alignment: Maintenance and Challenges

A major goal of multilingual WordNets is to achieve **semantic alignment** across languages – meaning that there is a correspondence of synsets such that each concept is shared or linked between languages. In practice, this is often done by linking each non-English synset to an **Inter-Lingual Index (ILI)** or directly to an equivalent synset in the Princeton WordNet [10] . Two main strategies have been employed to build aligned WordNets:

- **Expand Approach:** Start with the Princeton WordNet as a blueprint, translate its synsets (and possibly glosses) into the target language, and copy the semantic relations from English WordNet [18] . This was the approach used in many projects (e.g., the Spanish WordNet, and initially the Serbian WordNet under the BalkaNet project). It guarantees a high degree of alignment because every synset in the target WordNet *originates* from an English synset. For instance, Serbian WordNet included around 5,381 synsets aligned with the common concept set defined in the BalkaNet project (which in turn maps to Princeton WordNet 2.0) [19] . All the semantic relations (hypernyms, meronyms, etc.) for those Serbian synsets were **inherited from English WordNet and then manually checked** [20] , ensuring that if "car" is a hyponym of "vehicle" in English, *automobil* is a hyponym of the Serbian equivalent of "vehicle", and so on. This top-down alignment makes cross-lingual correspondence straightforward, at the cost of sometimes forcing the target language into the English conceptual schema.

- **Merge Approach:** Build the WordNet in the target language from the ground up (often using indigenous dictionaries and corpora) and then **map or merge** those synsets with Princeton WordNet via equivalence links [18] . This was done, for example, in the Hindi WordNet and the

IndoWordNet project. The Hindi WordNet team created synsets for Hindi using Hindi data, then later linked each Hindi synset to the closest matching English synset (Princeton WN 2.1) or marked where no equivalent exists [21] [22]. The merge approach can reveal concepts that are *missing* in Princeton WordNet or significantly different in the target language. It provides more freedom to include language-specific concepts from the start, and can thus highlight misalignments or gaps.

Regardless of approach, maintaining alignment means establishing an **equivalence relation** between synsets across languages [10]. In EuroWordNet and BalkaNet (projects involving Dutch, Spanish, Serbian, etc.), an Inter-Lingual-Index was used as a neutral mapping layer: each synset in each language was linked to an ILI entry (many of which corresponded to English synsets, plus some additional concepts) [23] [24]. This allowed cross-linguistic queries: a concept like *TREE* would have links from the English synset *{tree}*, the Dutch synset *{boom}*, the Spanish synset *{árbol}*, Serbian *{drvo}*, etc., all pointing to the same ILI concept ID. Alignment is thereby "maintained" by these links and by ensuring that the hierarchical relations inherited or mapped are consistent.

However, **challenges abound in maintaining semantic alignment**, especially for languages that differ significantly from English or have fewer resources:

- **Lexical Gaps and Surplus Concepts:** Languages often have concepts with no direct one-word equivalent in English, and vice versa. For example, Serbian (like many languages) has rich diminutives and augmentatives (e.g., a special word for "small house" vs "house") which English lacks as separate lexemes [25]. If Serbian WordNet were built from scratch, it might include synsets for these; but aligning to English WordNet (which has just one concept "house") forces a decision: either omit the diminutive concept or link it to the English parent concept with a note. Conversely, English has technical or cultural words that Serbian may lack. In early Serbian WordNet development, language-specific concepts that were not in Princeton WordNet were *not yet included* [26] – an artifact of the expand approach. This means some Serbian-specific synsets were absent to preserve alignment, an issue to be addressed in later expansions.

- **One-to-Many and Many-to-One Mappings:** A single concept in one language may correspond to multiple more specific concepts in another. This is commonly seen with kinship terms and other culturally specific domains. For instance, English has one synset {<u>uncle</u>} covering "the brother of one's father or mother (or the husband of one's aunt)." Serbian (and also Hindi, etc.) differentiate these relations: Serbian uses **ujak** for "maternal uncle" and **stric** for "paternal uncle." These are not interchangeable synonyms – they are separate words for different relations. Thus, there is no single Serbian synset equivalent to English *uncle*. In the Hindi WordNet, similarly, there are distinct synsets for maternal uncle, paternal uncle, etc. [27]. Maintaining alignment here is challenging: one approach is to link both *ujak* and *stric* synsets to the English *uncle* synset with a weaker equivalence (some projects mark them as "near equivalents" or link via a shared hypernym). Another approach is to introduce a new hypernymic concept in the Serbian WordNet for "parent's brother" and link that to *uncle*. The Hindi WordNet linkage project discussed using direct links where possible and **hypernym links when necessary** to handle such cases [28] [29]. The general solution is to not force a false synonymy (e.g., one would not lump *ujak* and *stric* into one Serbian synset, as that would violate the criteria of synonymy). Instead, alignment is maintained by mapping multiple target synsets to a single source concept in a structured way.

- **Cross-Part-of-Speech Mismatches:** Sometimes a concept is realized as a different part of speech in another language, which complicates direct synset alignment. We saw an example with *"peer"* (noun) in English vs *ravan* ("equal", adjective) in Serbian [30]. The English synset

{<u>peer</u>} “an equal in status” had no noun equivalent in Serbian, and the closest expression was an adjective meaning “equal (to someone)”. Such differences challenge the WordNet schema because synsets are typically monolingual and POS-specific. In Serbian WordNet, they either had to (a) omit the concept *peer*, (b) include *ravan* in an adjective synset linked to the noun concept (breaking strict POS parallelism), or (c) find a workaround (perhaps include a multiword noun phrase like *jednak po rangu* “one equal in rank”, if that were lexicalized). The expand approach revealed these tough spots, prompting the Serbian team to validate and adjust entries [31]. Each such case is handled individually, but it illustrates how **language-specific lexicalization patterns** (noun vs adjective, etc.) challenge direct alignment.

- **Cultural Vocabulary Mismatch:** Domains like flora/fauna, tools, musical instruments, food, and kinship often reflect local culture. Less-resourced languages might have rich terminology in some areas that English glosses over, or vice versa. The Hindi WordNet linkage identified categories where English and Indian languages diverge a lot: *kinship*, *musical instruments*, *kitchen utensils*, *tools*, *species of plants/animals*, *grains*, etc. [32] [33]. For example, Hindi has a plethora of names for specific musical instruments and cooking utensils unique to South Asian culture. One example: Hindi **तबला (tablā)** – a particular type of twin hand-drum – has no exact English counterpart [34]. English WordNet might have a general synset {<u>drum</u>} and maybe an entry for *tabla* as a loanword, but generally “tabla” is absent as a concept in PWN. In the Hindi WordNet, *tablā* is a synset on its own; when linking to English, the best one can do is link it as a hyponym or near-equivalent of “drum” [29]. Similarly, in the domain of utensils, Hindi distinguishes a *large ladle* (**करछा (karaćhā)**) and a *small ladle* (**करछी (karaćhī)**) as separate items [35] [36]. English WordNet has just {<u>ladle</u>} with no size distinction – a single concept. The Hindi synsets for big-ladle and small-ladle both have to align to the one English *ladle* synset (one perhaps as primary equivalent and the other noted as a specific subset). As the Hindi-English linkage paper notes, *“English…use[s] a single term, ladle, which is not size-specific,”* whereas Hindi needs two terms [36]. Serbian and other languages also face this issue (though perhaps to a lesser degree than Hindi): e.g., Serbian has both **kašika** (generic “spoon”) and **kašičica** (diminutive “teaspoon”) as separate everyday words – English has *spoon* but differentiates by context or compound “teaspoon” (which *is* a separate synset in PWN, fortunately). The alignment challenge is ensuring these pairs map correctly (teaspoon to kašičica, spoon to kašika, etc.).

- **Resource Limitations and Quality Control:** In less-resourced languages like Serbian, building a WordNet is often hampered by lack of large lexical databases and sense-tagged corpora. This makes semantic validation harder – it’s challenging to verify if two Serbian words are truly synonyms or just loosely related. The Serbian team tackled this by using the expand approach (thus leveraging English structure) and then performing **corpus-based validation** [37]. They computed frequency and co-occurrence statistics from a Serbian corpus to see if the introduced synsets made sense in usage [37]. Nonetheless, some misalignments can slip in, especially if an English concept is forced where a natural gap exists. The less-resourced nature also means that the Serbian WordNet initially did not include purely Serbian concepts outside the English inventory [26], potentially under-representing the Serbian lexicon’s richness (this was a conscious decision to focus on the common cross-lingual core first). Over time, efforts like **BalkaNet** and the Global WordNet Grid encourage adding those language-specific synsets (marked as such and linked via generalized relations) once the core alignment is done [38] [39].

In spite of these challenges, multilingual WordNets have developed techniques to **maintain alignment**. The use of **Common Base Concepts (CBC)** – a set of high-level concepts shared by many languages – helped ensure that each language WordNet covers an overlapping core vocabulary [40] [41]. Serbian was involved in the BalkaNet project, which extended the base concept set to 4,689 fundamental synsets common to Balkan languages (aligned to PWN 2.0) [42]. By concentrating on this shared core and linking everything to the ILI, the Serbian and English WordNets achieve a high degree of semantic overlap. As multilingual WordNets grow, they continue to update mappings (for example, linking to newer Princeton WordNet versions or to the Open Multilingual WordNet indices) to keep alignment current [43] [44].

To summarize, cross-linguistic alignment is maintained through systematic linking (expand/merge strategies, ILIs, base concept frameworks) and **manual curation** to handle mismatches. It is challenged by language-specific phenomena – but these very differences are informative. Projects like IndoWordNet explicitly note that linking different languages' WordNets "pay particular attention to language specific phenomena" [45], ensuring that while alignment is a goal, it should not come at the cost of misrepresenting a language's lexical structure. Instead, alignment sometimes requires creative solutions (like mapping via hypernyms) and accepting that the network of concepts won't be a perfect one-to-one mirror across tongues.

English vs. Serbian Synset Examples

To concretely see how synsets function across English and Serbian, let's compare a few specific examples. These illustrate differences in synonym count, part-of-speech alignment, and lexicalization between the two WordNets:

Example 1: Cross-POS Equivalent – *Peer* (English) vs *ravan* (Serbian)
- **English:** *peer* (noun), gloss: "a person who is of equal status or age" [30]. Synonyms in the English synset: *peer* (this sense doesn't have a close synonym like *equal* as a noun in common use, so it might be a single-member synset).
- **Serbian:** The concept of "equal (person of same status)" is expressed by the adjective **ravan** (literal gloss: "equal, level") used in constructions like *on je njemu ravan* ("he is equal to him") [46]. Serbian WordNet does not have a noun for "peer" – instead, the adjective *ravan* is used in that meaning. This Serbian synset for *ravan* is aligned to the English {peer} concept, despite the POS mismatch [30]. No other Serbian synonyms apply here. This example shows how a concept can be realized differently – a noun in English vs an adjective in Serbian – complicating direct synset matching.

Example 2: Indefinite Kind – "some sort of"
- **English:** *sort* (noun) in the sense of *"some sort of X"*, gloss: "a kind or type (in expressions like 'sort of dessert')". In WordNet 3.0, one synset for *sort* has this meaning of an unspecified kind. Synonyms include *kind* (in certain uses) or phrases like *"sort of"*.
- **Serbian:** The equivalent idea is not a noun at all, but an **indefinite pronoun/adjective** *nekakav* (literally "some-kind-of") [47]. In Serbian WordNet, *nekakav* would carry the meaning of an unspecified type, matching the usage of English "sort" in phrases. Again, no true Serbian noun corresponds directly. The English synset {sort (as in "some sort of")} is thus linked to a Serbian synset {nekakav}, different syntactic category. This was identified as a problem during development, as the *most natural translation* of the example *"she served a creamy sort of dessert thing"* is *"poslužila je nekakvo kremasto zaslađenē"* [48], where *nekakvo* ("some kind of") appears [48]. The Serbian WordNet includes *nekakav* to cover this concept, ensuring the synset alignment, even though on the surface an English noun maps to a Serbian indefinite adjective.

Example 3: Number of Synonyms – "Automobile"
- **English:** The synset for the concept *automobile* (car) is famously rich in synonyms: **{car, auto, automobile, machine, motorcar}** all appear as nouns for "four-wheeled motor vehicle" [49]. (Note: *machine* is an archaic informal synonym for car, included in WordNet; *auto* is colloquial, *motorcar* somewhat dated, etc.) This reflects English's large vocabulary and dialectal variants.

- **Serbian:** The primary words for this concept are **{automobil, auto}**. *Automobil* is the standard Serbian word for car (borrowed from French/International), and *auto* is the everyday colloquial form (a truncation) – these two are synonyms and both are included. Unlike English, Serbian does not have a whole slew of different words for “car”; even the word *mašina* (“machine”) is **not** used to mean a car. So the Serbian synset has essentially 2 members (possibly a third if one counts *kola*, see note). *(Kola*, literally “wheels”, is a common slang for car in Serbian; it’s plural in form. It may or may not appear in Serbian WordNet – if included, it would be another synonym in that synset.) _In any case, the English synset has five literals versus Serbian’s two. Both synsets are aligned to the same concept (vehicle for personal transport) and share the gloss equivalent to “a motor vehicle with four wheels, used for transportation of people”. Spanish and Dutch WordNets likewise include multiple synonyms: Spanish has {automóvil, auto, coche, carro, máquina}, covering regional terms_ [50] _, and Dutch has {auto, wagen, automobiel} etc. This example highlights that synset sizes can vary by language_*, even though the core concept is the same. Each language includes synonyms that are actually used in that language. English, with its mix of Germanic and Latinate words and global variants, often lists more synonyms. A less-resourced or more homogeneous language like Serbian lists fewer – not due to a flaw, but because it genuinely has fewer commonly used synonyms for that concept.

The table below summarizes some of the above comparisons plus additional examples, aligning English synsets with their Serbian counterparts:

Concept (Gloss)	English Synset (words)	Serbian Synset (words)
Peer “person of equal status/rank”	*{peer}* (noun)	*{ravan}* (adj., “equal” in status) [30]
“Some sort” unspecified kind/type	*{sort}* (noun, in phrase “some sort of”)	*{nekakav}* (indefinite adj., “some kind of”) [47]
Automobile (car) “motorized road vehicle”	*{car, auto, automobile, motorcar, machine}* [49]	*{automobil, auto}* (also *kola* informally)
Amoebic “related to amoeba (disease)”	*{amoebic, amebic, ameban, amoeban, amoebous}* [15]	*{amebni}* (one term, covers all)
Anemic “lacking power (figurative)”	*{anemic, anaemic}* (American/ British spelling)	*{anemičan}* (single term, spelling difference not applicable)

Table 1: Example synsets in English vs Serbian, showing differences in synonym count and part-of-speech. (Serbian terms in curly braces are synset members equivalent to the English concept.) The “peer” and “some sort” examples illustrate how Serbian uses a different POS to express the concept. The “automobile” and “amoebic” examples show English having multiple synonyms (including variant forms) where Serbian has fewer. Such comparisons demonstrate the careful alignment effort: even when a one-to-one match of words isn’t possible, the WordNets align the *concept*. The Serbian WordNet might have a note or a cross-reference for cases like *ravan* = *peer* to explain usage. Overall, English and Serbian synsets cover the same ground but with language-specific expression.

Multilingual Perspectives: Spanish, Dutch, Hindi, etc.

The observations above are echoed when we look at other languages' WordNets, each bringing its own flavor to the multilingual network. Here we highlight a few points with Spanish, Dutch, and Hindi to contrast with English (and Serbian):

- **Spanish:** As an Indo-European language with a long literary tradition, Spanish has a well-developed WordNet (initially through EuroWordNet and later projects). Spanish WordNet closely followed the expand model, translating English synsets and glosses [3]. Many concepts align neatly since Spanish and English share a lot of common concepts (though from different roots at times). One interesting aspect is dealing with **regional synonyms**. Spanish is spoken in many countries, and terms can vary (e.g., *coche* vs *carro* for "car"). The Spanish WordNet chose to include multiple variants in synsets to cover these: for "car", synonyms include *automóvil* (formal), *coche* (Spain), *carro* (Latin America), *auto* (short form), even *máquina* (Cuba, slang for car) [50]. This makes some Spanish synsets quite large – reflecting dialect diversity rather than purely semantic nuance. English similarly has regional terms (e.g., *truck* vs *lorry* are separate synsets for US vs UK because they actually refer to the same concept but WordNet splits them by region). Spanish decided to merge regional variants when meaning is the same. Another Spanish-English difference is polysemy vs monosemy: sometimes Spanish has a single word where English has two. For instance, *reloj* in Spanish means both "clock" and "watch". In WordNet, there are separate synsets for *wristwatch* and *clock* (timepiece vs wall clock), so Spanish WordNet has to handle *reloj* mapping to two synsets (perhaps as *reloj (de pulsera)* vs *reloj (de pared)* with qualifiers). This is a *one-to-many* mapping issue similar to what we discussed for kinship in Serbian/Hindi, but here within Spanish itself one word corresponds to two English synsets. Spanish lexicographers resolved this by treating the senses of *reloj* separately, effectively creating two entries for *reloj* (one in each synset). The glosses in Spanish make the distinction clear, and thus alignment is preserved: one Spanish synset = one English synset, even if a form is identical, the sense is separated.

- **Dutch:** Dutch was one of the original languages in EuroWordNet, and its WordNet (sometimes called Cornetto for a later version) likewise aligns with Princeton WordNet. Dutch, being Germanic, shares many basic concepts with English and often has straightforward equivalents (one could almost translate WordNet directly in a large portion of the vocabulary). However, Dutch has its own set of synonyms and some unique cultural terms. For example, Dutch has two common words for "bike" – *fiets* and *rijwiel* – where English mainly has *bicycle* (and *bike*). Dutch WordNet includes both as synonyms for the concept "bicycle". In general, Dutch synsets tend to have fewer members than English; where English might have a cluster of near-synonyms, Dutch often uses one. But there are cases of Dutch having synonyms where English doesn't. One noted case in lexical studies is the many Dutch diminutive forms that carry particular connotations (though WordNet usually doesn't list every diminutive as a separate lemma unless the meaning shifts). The **structural alignment** in EuroWordNet was aided by a top-ontology and base concept list [40] [41], meaning Dutch was built to mirror the semantic hierarchy of English to a large extent. Still, differences like separable verbs (a peculiarity of Dutch) required some adaptation – e.g., the English concept "to set up (an appointment)" is a single verb, whereas Dutch might have a separable verb *afspreken* ("to agree/arrange") that's one word. These are not synonyms issues per se, but syntactic representation differences that the Dutch WordNet handles internally. In multilingual alignment, Dutch did not present the degree of conceptual mismatch that Hindi did; rather, the focus was on ensuring each language's wordnet remains **autonomous yet interlinked** [51]. Dutch added some concepts not in English, such as culturally specific terms, into the ILI with a new ID (so as not to lose them). Overall, Dutch and English

WordNets are highly compatible, requiring relatively fewer special mapping maneuvers compared to languages from other families.

- **Hindi (and other Indo-Aryan languages):** The Hindi WordNet offers a strong contrast as discussed earlier. Built via the merge approach, it has revealed significant **cross-linguistic alignment challenges**. Hindi (and related languages like Marathi, Bengali, etc. in IndoWordNet) often have multiple specific terms where English has one general term. We saw kinship and utensil examples. Another striking example is with **species and gender**: Hindi (like many languages) has different words for male vs female animals (e.g., *Sher* = male tiger, *Sherni* = tigress), whereas English WordNet might represent “tiger” as a single concept with perhaps a pointer to “tigress” as a female variant. In linking, they had to decide whether to map both *sher* and *sherni* to “tiger” or treat *sherni* as a hyponym (female tiger). Typically, *tigress* is in English WordNet as a hyponym of *tiger*, so the structure can be preserved: *Sherni* maps to *tigress*, *Sher* to *tiger* (if *tiger* synset is treated as generic or male? WordNet may have “tiger” glossed as species, and “tigress” as female tiger). The Hindi WordNet team reports many such cases where a direct one-to-one equivalence wasn’t available, requiring careful linking strategies [52] [53]. They implemented a combination of **direct matches and hypernym matches** to maximize linkage [28]. If a Hindi synset couldn’t find an exact English counterpart, they linked it to the nearest more general concept in WordNet. For instance, many unique tools and grains in Hindi were linked to a broader English category (Hindi *ankusii* – a coconut meat remover tool – has no specific English name, so it might link under the synset for “kitchen tool” or “utensil”) [54] [55]. This way, the Hindi-specific concept is not lost; it’s connected at a higher level. Maintaining semantic alignment in such scenarios meant sometimes accepting *partial* equivalence. The Global WordNet Association now encourages using **link types** (exact, narrow, broad equivalence) to annotate these alignments rather than forcing everything into exact matches [56] [57].

- **Other Languages:** Many other languages have their WordNets (over 200 languages have some WordNet project [58]). Each offers learning opportunities. For example, the **Finnish WordNet** might illustrate how compound-rich languages handle multi-word concepts (likely by including many compound words as separate entries). The **Arabic WordNet** dealt with a root-based morphology and figured out how to represent synonyms in different dialects. **Basque WordNet** had to create many terms from scratch for concepts that lacked Basque words, sometimes borrowing or coinage. All these WordNets align to Princeton WordNet via projects like the Open Multilingual WordNet (which links dozens of languages to English WordNet 3.0) [44]. Meanwhile, **multilingual projects like BabelNet** take a different approach by blending WordNet with Wikipedia to cover named entities and encyclopedic concepts, creating a massively interconnected ontology (but BabelNet goes beyond the strict synonymy principle of WordNet).

The key observation is that **cross-linguistic alignment is a balancing act**: each language’s WordNet must be true to its own lexicon (so that native speakers find it usable and accurate), yet it should map onto a shared conceptual space so that we can translate or transfer knowledge across languages. English, Serbian, Spanish, Dutch, Hindi, etc., each contribute their perspective. Alignment is maintained through a combination of structural planning (shared base concepts, ILIs, expansion strategies) and painstaking manual or semi-automatic reconciliation of differences. Challenges like missing equivalents or multiple mappings are tackled with creative data modeling (e.g., adding new synsets for local concepts, using hypernym links, or grouping similar foreign concepts under one umbrella). This ensures that, for example, a query for a concept like *“rainbow”* can retrieve the Italian *arcobaleno*, the Hindi *Indradhanush*, the Spanish *arcoíris*, the Serbian *duga*, and the Dutch *regenboog* – all through their link to the English *{rainbow}* synset. The synsets function as interlingual anchors in a vast semantic web.

Conclusion

WordNets across languages demonstrate a remarkable blend of **universal ontology and linguistic individuality**. On one hand, they share a common inventory of concepts (thanks to projects like EuroWordNet, BalkaNet, IndoWordNet, and the Open Multilingual WordNet), aligning thousands of synsets via interlingual indices so that a "synset" essentially represents the same concept in many tongues. On the other hand, the realization of each concept – in terms of synonyms and definitions – is tailored to each language's lexical repertoire and cultural context.

We saw how **glosses** are translated and refined: definitions are kept aligned in meaning but worded in each language's natural way [3] [6]. We explored **criteria for synset membership**: the necessity of true synonymy and equivalent meaning, which guides lexicographers in deciding which words in Serbian or Spanish can join a given synset that originated in English [2]. We discussed how each language may have more or fewer synonyms for a concept, reflecting genuine lexical richness rather than inconsistency – English often has dense synsets (due to historical layers of vocabulary), whereas Serbian's synsets are sparser on average [17], but in both cases the included words are interchangeable equivalents in context.

Maintaining **cross-linguistic semantic alignment** is both a design goal and a challenge. Alignment is facilitated by linking structures (like the Princeton WordNet as a hub, or the Inter-Lingual Index) and by collaborative selection of base concept sets common to many languages [40] [41]. Yet, alignment is challenged whenever one language doesn't lexicalize a concept the same way as another – whether due to cultural specificity (e.g., Hindi's many kinship terms [59]), grammatical differences (Serbian's use of aspect, or cross-POS mappings like *nekakav* for "some sort" [47]), or simply gaps in one language's lexicon. We saw that Serbian, as a less-resourced language, relied initially on the English structure, which caused some strain where Serbian had to stretch or adjust to fit English concepts [25]. Over time, the strategy has been to enrich Serbian WordNet with its own nuances (for instance, adding any missing Serbian-specific concepts later, and fine-tuning glosses and synonyms for naturalness [5]). In the broader context, languages like Spanish and Dutch, being well-resourced, achieved full WordNets relatively early and aligned them closely with English (with minor regional adaptations), whereas languages like Hindi (and other Indian languages) built robust WordNets that then illuminated where alignment needed careful handling (often informing improvements to the interlingual tools and methodologies).

In conclusion, WordNets and synsets **function as a bridge between languages**: each language's WordNet is an autonomous lexical network reflecting its unique semantic structure, yet through synset alignment, they all plug into a global lexical *grid*. Definitions (glosses) provide a human-readable alignment of meaning, and the interlingual links provide a machine-readable alignment of concepts. Criteria for synset membership ensure that within each language the integrity of meaning is preserved – a synonym set in Serbian is just as semantically coherent as one in English [2]. Cross-linguistic alignment is maintained by those equivalence links and common frameworks, even as it is tested by linguistic diversity.

By examining examples across English, Serbian, Spanish, Dutch, Hindi and others, we observe a consistent theme: **the concept is the pivot**. Each language may encode the concept differently – one word, multiple words, different parts of speech, or multiple near-synonyms – but through the WordNet architecture, these all coalesce around a shared synset ID. The multilingual WordNet enterprise thus underscores both the universality of many human concepts and the beautifully varied ways in which languages express them. Researchers and lexicographers continue to refine this alignment, adding new languages and updating existing ones (the Global WordNet Association lists over 200 WordNets [58], and efforts are ongoing to link even sign languages via ILIs [60]). Each addition enriches the network and sometimes challenges it, leading to better criteria and tools for alignment. In practical terms, these multilingual WordNets enable cross-language information retrieval, machine translation disambiguation, and comparative linguistic studies, all grounded in the humble synset – a concept with many names.

Sources: The analysis above draws on WordNet project descriptions and research papers, including the development of Serbian WordNet [30] [47] [6] , the expansion of Spanish WordNet [3] , general principles from the Global WordNet Association [18] [2] , and detailed examples from Hindi-English alignment studies [59] [36] , among others. Each cited source is indicated in-text with a reference number corresponding to the list below for further reading.

---

[1] [9] wordnetcode.princeton.edu
https://wordnetcode.princeton.edu/5papers.pdf

[2] [10] [43] [44] [60] (PDF) WordNet, EuroWordNet and Global WordNet
https://www.researchgate.net/publication/228527736_WordNet_EuroWordNet_and_Global_WordNet

[3] (PDF) The Spanish Version of WordNet 3.0
https://www.academia.edu/965601/The_Spanish_Version_of_WordNet_3_0

[4] [PDF] Creating a Synthetic Evaluation Dataset for the Serbian SentiWordNet
https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/download/2024.24.1.3_en/716/

[5] [6] [7] [8] [11] [12] [13] [15] infoteka.bg.ac.rs
http://infoteka.bg.ac.rs/pdf/Eng/2008/INFOTHECA_IX_1-2_May2008_59a-78a.pdf

[14] Word embeddings : Una guía fácil de entender - Pangeanic Blog
https://blog.pangeanic.com/es/word-embeddings-una-guia-facil-de-entender

[16] [PDF] The Dutch Wordnet - Fon.Hum.Uva.Nl.
https://www.fon.hum.uva.nl/paul/papers/1999-uva-VossenBloksmaBoersma.pdf

[17] [19] [20] [25] [26] [30] [31] [37] [46] [47] [48] RJIST-MATF.dvi
http://poincare.matf.bg.ac.rs/~cvetana/biblio/RJIST-MATF.pdf

[18] [38] [39] [40] [41] [42] globalwordnet.github.io
https://globalwordnet.github.io/resources/gwa-base-concepts

[21] [22] [27] [28] [29] [32] [33] [34] [35] [36] [45] [52] [53] [54] [55] [56] [57] [59] Instructions for ACL-IJCNLP 09 Proceedings
https://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/04_iwn_Hindi%20to%20English%20Wordnet%20Linkage%20-%20Challenges%20and%20Solutions.pdf

[23] [51] (PDF) EuroWordNet: A Multilingual Database of Autonomous and ...
https://www.researchgate.net/publication/230585377_EuroWordNet_A_Multilingual_Database_of_Autonomous_and_Language-specific_Wordnets_Connected_via_an_Inter-Lingual-Index

[24] [PDF] EuroWordNet - VU Research Portal
https://research.vu.nl/files/74104709/VossenMulti1

[49] WordNet - EcuRed
https://www.ecured.cu/WordNet

[50] Sample usage for wordnet - NLTK
https://www.nltk.org/howto/wordnet.html

[58] Language - Global WordNet Association
https://globalwordnet.github.io/resources/wordnets-in-the-world