The Latvian WordNet and Word Sense Disambiguation: Challenges and Findings

Ilze Lokmane[1], Laura Rituma[2], Madara Stāde[3], Agute Klints[4]

[1]University of Latvia, Department of Latvian and Baltic Studies, Visvalža 4a, Riga, LV-1050
[2]Institute of Mathematics and Computer Science, University of Latvia, Raina bulvaris 29, Riga, LV-1050
[3]University of Latvia, Department of Latvian and Baltic Studies, Visvalža 4a, Riga, LV-1050
[4]University of Latvia, Department of Latvian and Baltic Studies, Visvalža 4a, Riga, LV-1050
E-mail: ilze.lokmane@lu.lv, laura.rituma@lumii.lv, stade.madara@gmail.com, agute.klints@gmail.com

Abstract

The article addresses the issues of word sense disambiguation within the process of developing an electronic lexical semantic resource, the Latvian WordNet. Apart from word senses, the resource also contains semantic paradigmatic relations between these senses, and therefore sense granularity must align with the need for creating synonymous, hyponymic, meronymic and antonymic links between Latvian words, as well as external links with the Princeton WordNet.

The development of the Latvian WordNet started in 2020 and it is based on two sources: a summarising electronic dictionary Tēzaurs.lv and available corpora. Because the word senses listed in Tēzaurs.lv are not directly usable for the needs of computer linguistics due to a number of reasons, the developers of the Latvian WordNet checked and revised the senses manually based on corpus data. Thus, the work on distinguishing word senses serves two purposes: 1) creating a Latvian WordNet, and 2) improving the structure of existing entries in the dictionary Tēzaurs.lv.

The article primarily focuses on the elaboration of common criteria for distinguishing word senses. The analysis concentrates on verbs as these are the most complex part of speech from the point of view of making sense distinctions. The authors conclude that the process is based on a set of criteria that form a certain hierarchy depending on the semantic group of verbs, namely, syntactic distribution, semantic distribution, as well as the interrelation between the two, and semantic decomposition of senses. Particular attention is paid to the interrelations of superordinate senses and subsenses, from which it is possible to conclude that an absolutely uniform and consistent subsense distinction is not likely to be possible, and, therefore, in cases of uncertainty, decisions are made in favour of what is needed to develop the Latvian WordNet.

Keywords: word sense disambiguation; sense distinction; electronic lexical semantic resource; syntactic and semantic distribution; lexical decomposition

1. Introduction

The article focuses on major challenges and some preliminary findings in the field of word sense disambiguation with respect to the development of a Latvian WordNet[1], i.e. structured, machine-readable wide coverage inventory of word senses and semantic

[1] Project "Latvian WordNet and word sense disambiguation" No. LZP-2019/1- 0464

relations, such as synonymy, hyponymy, meronymy, antonymy and similar between these senses in Latvian. By word sense disambiguation we understand the task of determining which sense of a word is being used in a particular context (Jurafsky & Martin, 2020: 1). Therefore, finding criteria for deciding when different uses of a word should be represented as discrete senses is crucial. The aim of the project is to determine the senses of 5,000 commonly used Latvian words and to establish semantic links between them, but at the present stage approximately 150 words have been processed, most of which have multiple senses. The work is carried out using a specifically developed tool, which is described in more detail in Section 3.

The development of the Latvian WordNet began in 2020 and it is based on two sources: digital versions of pre-existing monolingual general and specialist dictionaries and available corpora.

An important Latvian lexical resource maintained by the Institute of Mathematics and Computer Science of the University of Latvia (IMCS UL) is Tēzaurs.lv (Spektors et al., 2016), which is a large (~ 378,000 entries in the last release in March, 2021) digital compilation of legacy dictionaries[2]. Our experience indicates that the word senses listed in Tēzaurs.lv are not directly usable for the purpose of computational linguistics due to issues with sense granularity and boundaries, as well as the outdated nature of many of the senses. Therefore, the word senses available in this resource are checked and revised using a corpus-based approach to determine if the senses are still currently relevant, whether any new senses have appeared or whether specific uses of a word demonstrate the validity of word sense distinction (based on a similar revising of sense distinctions and definitions in Estonian WordNet see Kerner, Orav & Parm, 2010).

The main source data for the lexical analysis is The Balanced Corpus of Modern Latvian (10 million tokens), which is also maintained by IMCS UL but has become the de-facto reference corpus for Latvian linguistic research (Levāne-Petrova, 2019). However, not all word senses can be found in the corpus, therefore other corpora are employed for identifying and illustrating less common or colloquial word senses: Corpus of the Saeima (the Parliament of Latvia) (Darģis et al., 2018), Latvian Blog Corpus 2015 (Laizāns, 2015), Latvian Web Corpus 2007 (Dzerins & Dzonsons, 2007) and CommonCrawl of Latvian 2020.

A corpus-based approach results in a better set of word senses than the commonly used alternative of directly mapping Princeton WordNet concepts to translations in the target language, which implicitly transfers the English linguistic patterns of many concepts that are often not a good match for the target language. While a corpus based approach requires more effort, we have chosen this to ensure the linguistic validity of the resulting resource.

[2] The total number of Tēzaurs.lv sources is 329.

In addition, such an approach would meet the needs of both WordNet development and the improvement of word sense inventory of Tēzaurs.lv. Therefore, the development of the Latvian WordNet is primarily a linguistic (lexicographic) challenge, as the separation of senses is performed by manually aligning corpus evidence with lexicographic data.

2. Problematic Issues of Distinguishing Word Senses

Before describing the process of WordNet creation and the criteria for distinguishing senses, we would like to point out the main issues that arose in this process.

First of all, the process of word sense distinction is a complicated task in itself. We tend to agree with cognitive linguists that the question of how many senses the word has may not have a clear-cut answer. There is always a question whether two different uses of a word exemplify two separate senses, or contextual modulations of the same sense (Taylor, 2009: 144). Some linguists even claim that a word has just a single abstract meaning which is instantiated in a range of sometimes very different usage situations (Taylor, 2009: 147–148).

Therefore, the word sense system is not a stationary and entirely fixed one, and semantic derivation is an active and ongoing process. It could be said that the range of word meanings is continuous and diffuse, and the fixation of individual meanings is linked to a certain degree of schematisation. The concept of polysemy, on the other hand, is based on the idea of discreteness of lexical meanings and, as a consequence, researchers and lexicographers, in particular, try to discern strict boundaries around what is in fact an unclear grey area.

Therefore, lexicographic resources display a considerable variation in the number of word senses. Even though overall coverage of the senses is the same, dictionaries may have differently clustered senses and subsenses, with the same semantic space merged and split in various ways. For example, metaphoric and metonymic meaning extensions are not always set apart as distinct meanings. In addition, it is possible to use certain words creatively in new contexts, and it is not easy to determine whether it illustrates an already existing meaning or is considered an individual metaphorical or metonymic use and, hence, does not require including in the dictionary.

Thus, the question of what marks the point when a meaning should be regarded as a distinct sense or subsense and included in a dictionary is probably one of the most difficult issues of lexicographic work. As Allen (1999: 61) states, lexicographers can be divided into two broad categories - ‘lumpers’ and ‘splitters’: “The ‘lumpers’ like to lump meanings together and leave the extraction of the nuance of meaning that corresponds to a particular context to the user, whereas the ‘splitters’ prefer to enumerate differences of meaning in more detail; the distinction corresponds to that between summarizing and analyzing.” Furthermore, Jackson (2002: 89) admits, that

"most dictionaries tend to be of the 'splitting' type, though different dictionaries do not necessarily agree on where to make the splits between senses." This is also fully applicable to existing dictionaries of the Latvian language.

In our opinion and given the point of view of the user of the dictionary, it is better to list fewer senses, thus making the entry more transparent and reader-friendly. A lexicographer is able to discern between slight nuances of meaning, whereas an everyday user outside the realm of linguistics might find it difficult to grasp the difference between word senses, especially if they are accompanied by long and complex definitions. Initially, we planned to generalise the division of word senses in the dictionary and make it less detailed, but over the course of the work it became clear that a general division is not always entirely useful for WordNet purposes. In addition, the legacy of Tēzaurs.lv had to be treated with great care in order not to erase the dialectal, terminological and other word senses included there, even if modern language corpora do not contain examples of their use. Therefore, the corpus-based approach applies only to a certain part of the word senses.

Therefore, and **secondly**, in the revision of word senses a compromise was necessary between two extremes: an excessively generalised or fine-grained division of word senses. The need for a more detailed division arises in cases when synonymous, hyponymic and other semantic relations between senses are formed, as well as during the formation of external links with the Princeton WordNet. Our definitive solution for cases of ambiguity aligns with the needs of WordNet: word senses are identified in more detail when a sense and subsense form individual synonymic or other semantic links to a sense or subsense of another word.

Thirdly, despite the substantial semantic differences between various parts of speech and separate semantic groups within a part of speech, the selected approach to word sense distinction should be as consistent as possible. The defined criteria and their application are described in more detail in Section 4.

Fourthly, we encountered the problem of defining and dividing superordinate senses and subsenses. In such cases, it was noticeably more difficult to identify a consistent solution that would be equally applicable to words in all semantic groups, therefore defining subsenses is the most subjective step in the WordNet creation process and requires a more detailed explanation.

Latvian lexicographers have so far avoided studying the theoretical problems of word subsense, so the division found in the Latvian language dictionaries is inconsistent and intuitive. Semanticists, on the other hand, do not examine the problem of separating superordinate senses from subsenses and regard it as a topic pertaining more to lexicography. The basis for identifying a subsense is usually more detailed semantic differences attributable to the same sense, as well as grammatical and functional features of the word (LLVV, 1972: 11). They are as follows:

1) A subsense can differ from a superordinate sense by a certain semantic component. For example, the verb *uztvert* (*to catch*) has a sense 'to grasp' with a subsense 'to grasp and deflect'[3], therefore the semantic component 'to deflect' is added.

2) A subsense can differ from a superordinate sense by semantic distribution, namely, the semantic roles of the participants of the situation or the semantic groups they pertain to. For example, the verb *rakt* (*to dig*), has a sense 'to impale and move soil or dirt with a shovel', which indicates a person as the agent, whereas the subsense reveals other possible agents, such as equipment or animals. The semantics of the instrument is also different: humans dig with a shovel, while animals dig by using their muzzle or limbs.

3) A subsense can differ from a superordinate sense by syntactic distribution, for example, the superordinate sense can have transitive and subsense intransitive properties or vice versa. The creators of the Latvian WordNet believe that the use of a transitive verb without a direct object should not be considered as a subsense if the object can be understood from context or situation or if it is so general that it is not necessary to be named. For example, the word *dzert* (*to drink*) has a transitive superordinate sense 'to imbibe and swallow (a liquid)' and an intransitive subsense, e.g. *Dzert gribi?* 'Do you want a drink?'[4] Only if the sense of a verb that is being used in its intransitive use is joined by a new semantic component is there a basis for defining a subsense, as is demonstrated by the verb *lasīt* (*to read*), which has the transitive superordinate sense 'to take in a written text' and an intransitive subsense, which has the added semantic element of 'being able to'.

4) Cases of diathesis demonstrate the interrelation of semantic and syntactic distribution. Here, a situation is illustrated by the same verb from different points of view. The participants in the situation remain the same, but their syntactic status is changed. For example, the act of digging involves both the agent (*cilvēks rok* 'a person is digging') and the instrument (*rakt ar lāpstu* 'to dig with a shovel'), as well as patients of different kinds: that, which is moved (*rakt zemi* 'to dig soil') and that, which is created (*rakt bedri* 'to dig a hole'). Syntactically, only one of them can be realised at a time, but the situation as a whole does not change. The instrument can also be used as a subject (*lāpsta labi rok* 'this shovel digs well', *ekskavators rok* 'the excavator is digging'). Various cases of diathesis have been extensively examined in semantic studies (Paducheva, 2004: 51–79), as well as divided into types, which differ slightly in each respective language. In other semantic theories such extensions of a certain verb have been described as metonymic (Pustejovsky, 1998: 31–33), whereas in cognitive semantics this process is called profiling (Saeed, 2000: 328–330).

[3] All sense definitions referred to in this article are taken from Tēzaurs.lv.

[4] All examples of word usage are taken from Latvian language corpora.

Therefore, it can be concluded that to a certain extent subsenses can illustrate the continuity of lexical semantics of words and the gradual transition from one sense to another. It can be seen further in the paper that subsenses can be distinguished on the same principles as superordinate senses (see Section 4).

Fifthly, an optimal definition (sometimes called a gloss) of sense is necessary, as the definition method of a word sense can affect the entire system of word senses. For instance, a more general definition may lead to two or more senses being combined whereas specific definitions allow the contrary, i.e. splitting a sense into separate senses or subsenses.

Different forms of definition are appropriate to different types of words (Jackson, 2002: 94). Practical lexicography offers three main methods of defining sense: definition by synonym, definition by periphrasis and a scientific definition. Each of the listed methods has its advantages and disadvantages, which we will examine in more detail.

In the process of developing the Latvian WordNet, definition by synonym has been one of the most useful methods, as it facilitates finding synonym links between senses of various words, e.g. *domāt* (*to think*), the third sense of which is 'to care for'. However, taking into account the revelation of the lexical semantics of a word, this approach to definition also has notable disadvantages. Firstly, there is a risk of circularity (Jackson 2002: 94), secondly, by using a synonym, the meaning is essentially left unexplained, and thirdly, not all senses have synonyms. Moreover, the synonym used in the definition could have multiple senses as well.

Definition by periphrasis, unlike definition by synonymy, attempts to determine the semantic components that form the sense, e.g. *skriet* (*to run*) – 'to move steadily by springing steps, so that both feet occasionally leave the ground at the same time at each step'. For this method it is important to find the essential features, i.e. those that distinguish the realia from others, and not to include irrelevant information. The number of specific features should be sufficient (Zuicena, 2010: 370) and the words used in the definition should be simpler than the word that is being defined (Jackson, 2002: 93). Therefore, this method is similar to lexical decomposition. However, this approach has certain limitations: the first is that the proportion of words which lend themselves to this sort of analysis is relatively restricted; the second is that the analysis leaves much semantic knowledge unaccounted for (Cruse, 2004: 242). In practical lexicography the periphrastic definition method is often used intuitively, thus it is not always sufficiently accurate and is used mostly in cases when there are no synonyms.

A scientific approach or at least elements of it are sometimes used to define sense, such as the noun *bullis* (*a bull*) – 'a male representative of hollow-horned or antlered ruminants'. There are reasonable objections to this type of explanation, namely, that the definition of a scientific concept is not part of ordinary linguistic competence (Goddard, 1998: 28). However, it should be kept in mind that language users may have certain (albeit rudimentary) scientific knowledge of specific realia. Although

explanatory dictionaries are not encyclopediae, there is no strict boundary between the meaning of a word and the knowledge of certain realia.

It should also be noted that the definition of a word sense often requires information on typical distribution. It is mostly used in verb definitions, e.g. *čivināt* (*to twitter*) – 'to make short, rhythmic chirping noises (about birds)'. When defining verbs of certain semantic groups, it is even impossible to do without this approach. For example, specific senses of sound verbs cannot be fully revealed either by synonymy or periphrasis. Definitions can also be supplemented by elements typical of the referent, introduced by the adverb *parasti* (*typically, usually*) (Jackson, 2003: 95), e.g. *glāze* (*a glass*) – 'a small (usually cylindrical) drinking container without a handle made of glass or other material'.

It should also be taken into account that there is no universal principle or method for defining the senses of words of all semantic groups and parts of speech. For example, distribution is more important for defining the semantics of verbs than it is for nouns. Polyvalent verbs are more effectively defined by describing their distribution (e.g. the meaning of the verb *īrēt* (*to rent*) can be revealed by listing *who, what, to whom, for how long and for what payment*), whereas in case of verbs with zero valency, e.g. *snigt* (*to snow*) the distribution analysis yields little information and other methods should be employed.

And lastly, certain problems are also caused by the separation of distinct word senses and multi-word expressions. However, this topic deserves separate research, therefore it is not examined in this article.

3. Lexicographic Infrastructure and Tools

The software infrastructure for this work is based on the existing tools for maintaining the Tēzaurs.lv lexicographic platform which was already used for maintenance of structured data for entries, glosses, word senses and usage examples. As we wanted to base the Latvian WordNet on the existing Tēzaurs.lv word sense data where possible, we chose to extend the Tēzaurs.lv editor tools with the required functionality instead of managing the WordNet data in a separate existing tool (for example, WordNet Loom and DebVisDic). This choice adds certain complexity due to need to balance the requirements (for example, for the word sense granularity) of the WordNet project with the expectations of generic dictionary users of Tēzaurs.lv, as they would see the same word senses, but it also has the potential to make the resulting resource more accessible to a wider general audience, which would be less likely to use separate tools for browsing WordNet data. The choice of integration also means that all work on improving word sense definitions and usage examples improves the general dictionary data.

The technical platform for the Tēzaurs.lv lexicographical database is built as JavaScript (Vue.js) web interface to a custom PostgreSQL database for the lexical data.

In order to manage WordNet data, we extended the Tēzaurs.lv database and tools with support for managing synsets and semantic links (including external links to the Princeton WordNet), as well as streamlining functionality for mapping corpus examples to specific word senses and subsenses (see Figure 1). The data is developed in an internal environment with quarterly releases of new data versions to the general public on the Tēzaurs.lv online platform. At project milestones, we plan to release the WordNet data along with the Tēzaurs.lv lexical database in machine-readable structured format.

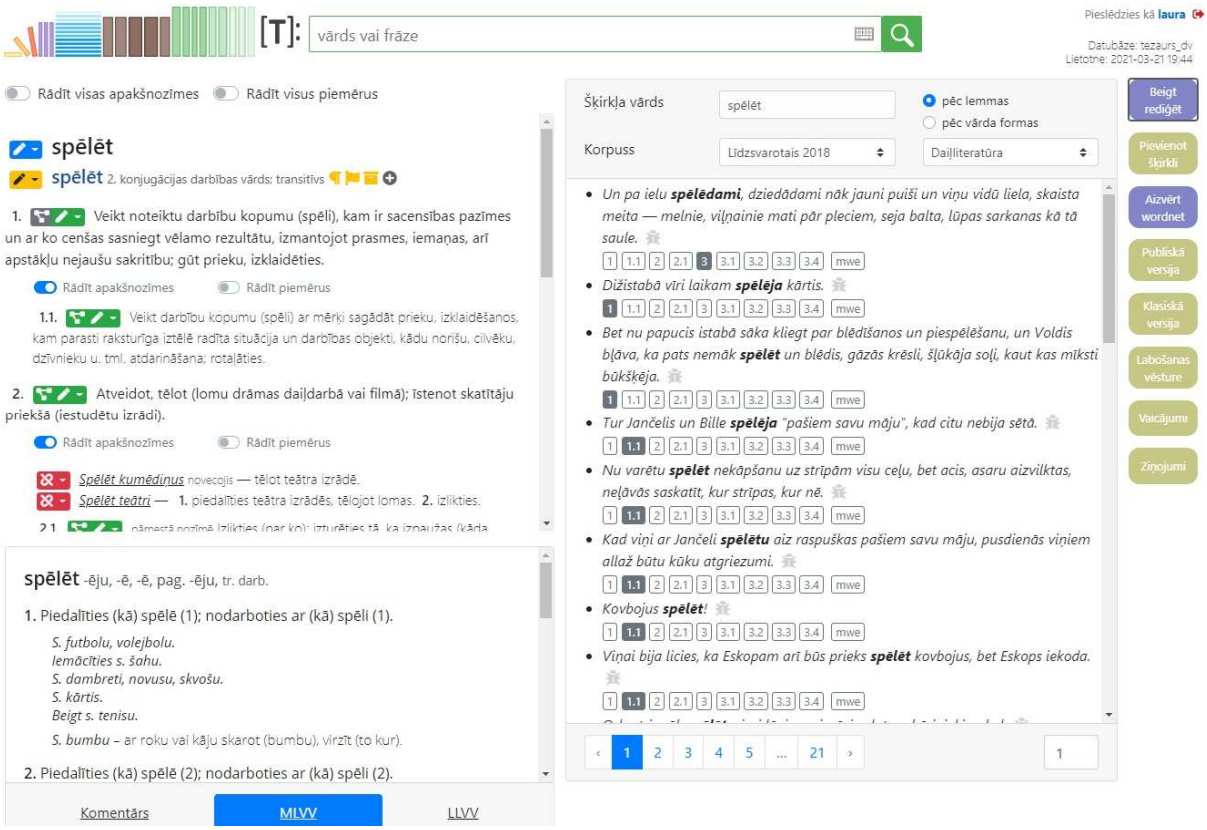

Figure 1. The sense editing and example selection function view in the tool. The left side shows senses and subsenses listed in Tēzaurs.lv and two other dictionaries for comparing differences. The right side shows all the examples with the corresponding lemma in the selected corpus; each example can be marked with the matching word sense number.

The workflow consists of the following steps: 1) editing entries by modifying word senses, their order and definitions and adding new entries and senses, 2) browsing through various examples from different corpora and adding them to word senses or multi-word expressions in an entry (10–30 examples for each sense), 3) creating synsets between separate meanings of various words, 4) creating various types of links between synsets, 5) linking Latvian meanings/synsets with those of the Princeton WordNet (see Figure 2).

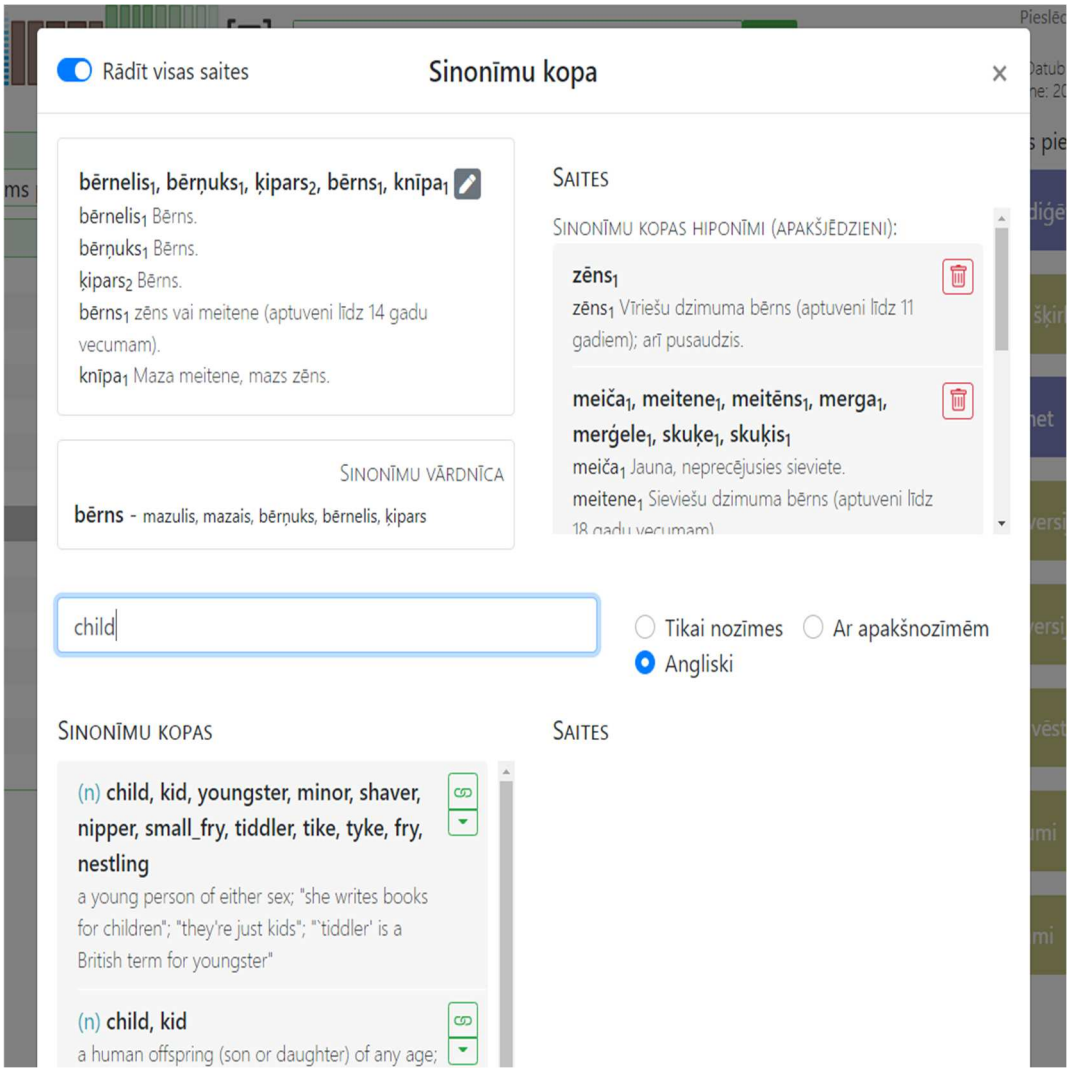

Figure 2. The window for creating synsets and semantic links; the process of reviewing word senses for "child". The synset with all the synonyms included is shown at the upper part of the window. Below the synset there are synonym suggestions from the dictionary of synonyms. The search window is in the middle, where the developer can search for word senses on Tēzaurs.lv or corresponding English synsets on the Princeton WordNet. All links added to the synset are displayed on the right side.

From the WordNet perspective the main motivation of selecting a substantial quantity of examples from corpora is to use them as training data for supervised machine learning in developing a Word Sense Disambiguation system. As the usage examples are searched in corpus, the selected wordform/inflection is annotated with the manually chosen word sense identifier, forming a sense-annotated corpus. The review of examples also helps to ensure that the chosen word sense split is based on actual usage, and a manually chosen subset of most representative examples are also used in the public Tēzaurs.lv version to aid dictionary readers by illustrating the differences between specific word senses, in contrast to the earlier approach of Tēzaurs.lv which used automatically selected corpus examples for the whole entry, without explicit linking to word senses.

4. Word Sense and Subsense Distinction Criteria and their Applications

As mentioned before, the criteria of distinguishing word senses can differ depending on various parts of speech and even semantic groups (e.g. sound and directional verbs). The approach chosen in the development of the Latvian WordNet is based on word sense separation by a set of features. As verbs may be considered the most challenging part of speech with respect to deciding how many discrete senses a word has, we will examine this part of speech by concentrating on criteria which have proved to be useful.

Latvian is a highly inflected language, and thus the **syntactic distribution** of the verbs, namely, valency frame (arguments and their coding), has to be taken into account first of all (on the implementation of valency models of verbs in Polish WordNet see Dziob & Piasecki, 2018). The syntactic distribution shows what syntactic constructions a word is a part of, e.g. whether it has a direct or indirect object, certain adverbial modifiers, etc. Syntactic distribution can be particularly important when separating the senses of highly desemanticised and grammaticalised verbs. For example, the verb *būt* (*to be*) has a meaning 'to be situated', which becomes clear in a construction involving adverbials of place, e.g. *Visapkārt mājai ir priedes* 'There are pine trees all around the house', whereas the meaning 'to belong' can be understood in a construction containing the dative of possession: *Tev būs tieši tāda māja* 'A house just like this will someday belong to you'.

The role of syntactic distribution in word sense distinction can also be illustrated by the verb of cognition *domāt* (*to think*). For example, the distribution of the sense 'to consider' is typically associated with an object clause introduced by conjunction *ka* (*that*) (*Domāju, ka tas nav godīgi pret auto izmantotājiem* 'I think that it isn't fair to car users') or deicitc adverbs *tā* (*thus, this way*) and *tāpat* (*in the same way, similarly*) (*Tā jau es domāju* 'That's what I thought'), whereas the sense 'to envisage, to get ready' is demonstrated when combined with infinitive: *Ko tu domā darīt ar tiem?* 'What are you thinking of doing with them?'.

Although syntactic distribution could be considered a fairly objective criterion in distinguishing word senses, it should be noted that sometimes two different senses can be used in the same syntactic construction. For example, the verb *domāt* (*to think*) in combination with a prepositional phrase can represent both the basic sense of 'to think' (*Es nezinu, par ko domāja viņš* 'I don't know what he was thinking about'), as well as the secondary sense of 'to care for' (*Katrs īpašnieks sāktu domāt tikai par savu peļņu* 'Each owner would start to think only of their own profit'). The latter sense can be identified based on the semantics of the object – the desirable things that are obtained through effort (e.g., *profit*). Therefore, it is not surprising that in some instances of word use there is ambiguity between these two senses, e.g. *Par to viņiem nav jādomā* 'They don't have to think about it'.

Interestingly, in Latvian the verb *domāt* (*to think*) has two senses that mostly materialise in one grammatical form, namely, the past passive participle. The first one, meaning 'to be meant for a certain purpose' is used in combination with adverbials of purpose (*Bibliotēka domāta ne tikai lasīšanai, bet arī sarunām* 'The library is meant not only for reading but also for having talks'), whereas the second sense 'to understand by' is used with a prepositional phrase (*Ar meitām un dēliem ir domāti vecāku miesīgie pēcnācēji* 'One's direct descendants are understood by the terms 'daughters and sons'').

Secondly, **semantic distribution** including the **semantic roles** and **semantic features** of the arguments has proved to be useful. The semantic distribution of verbs includes the semantic roles of the participant (e.g. agent, patient, experiencer, beneficiary, addressee, instrument) and general or more specific semantic features (e.g. *animate / inanimate, abstract / concrete, countable / uncountable*).

The main problem associated with this method is that it is not clearly defined which semantic roles or characteristics are sufficiently important to be taken into account in the process of word sense distinction, e.g. whether the semantic opposition *human / other living beings* always enables one to fully differentiate between senses or not. Traditionally, in Latvian lexicography the verbs of motion, like *iet* (*to go*), *skriet* (*to run*) and so on, have different senses based on whether the action is performed by a human or animal, however, the developers of WordNet have chosen to overlook this in favour of a view that the nature of direction is not greatly changed by this. In this case, the animacy / inanimacy of the subject is a much more important characteristic. For example, in the basic sense of the verb *skriet* (*to run*) the subject is animate, whereas in derived senses it is an inanimate object (*Pa lāstekām uz leju skrien ūdens pilītes* 'Water droplets are running down the icicles'), physical phenomenon (*Uguns skrien uz priekšu* 'Fire is running forward') or phenomenon related to the subjective perception of humans (*Laiks skrēja nemanot* 'The time ran by unnoticed'; *Domas skrēja ātri* 'Thoughts ran through (one's) head'). In this case, the process of word sense distinction is based on the semantic groups of subjects, which can be viewed as a justified approach, given that significant features of the action directly depend on the subject: physical movement through space with or without legs, or movement through time or mental space. In contrast, the sense distinction process for the verb *mainīties* (*to change*) is not based on the animateness of the subject, even though it can relate to both animate subjects (*Nemaz neesi pa šiem gadiem mainījies* 'You haven't changed a bit over these years'), as well as inanimate ones (*Tomēr beidzamjā laikā situācija ir mainījusies* 'However, in recent times the situation has changed'). In our view, the process of change is a very general one and is not affected by the animateness of the subject.

A more interesting situation is presented by transitive verbs, where the semantic features and semantic roles of not only the subject but also the object can be crucial. Besides a direct object in the accusative, the verb *dot* (*to give*) takes an indirect object in the dative as well. It is also important to note that the direct object can have a wide spectrum of meaning, from a real object to abstract states, conditions etc. The position

of the subject can be occupied not only by people or a group of people but also, for example, by circumstances. That is, everything that can serve as the basis for someone receiving something. So, the act of giving is interpreted very broadly as a causal relationship. Due to the previously examined semantic features, the verb *dot* (*to give*) is an often used one and has a wide distribution. This is also one of the verbs which tend to grammaticalise in many languages (Heine & Kuteva, 2002: 149–155), meaning that the semantics of the verb itself often play a fairly insignificant role in the semantics of phrases.

Word sense distinction for the verb *dot* (*to give*) is mainly based on the semantics of the object: it can be an inanimate object (*Nu tad dod to grozu un desmit santīmus šurp* 'Then give me the basket and 10 santims'), a state or a circumstance (*Nolēmām dot iespēju jaunam censonim* 'We decided to give the new contestant a chance'), information (*Norādes dot jau es varu* 'At least I can give directions'). At the same time, the structure of senses of this verb effectively demonstrates the interaction of grammatical and semantic criteria, for example, with the word sense 'to procure, to provide (conditions)', which has two subsenses. The first one, 'to have by birth', is usually realised through the passive participle in the past tense (*Viņam no dabas ir daudz dots* 'He was already given much from birth'), whereas the second subsense 'to let' is demonstrated through a syntactic construction with the infinitive (*Dodiet man arī pamēģināt!* 'Let me try!').

Thirdly, the differences in syntactic and / or semantic distribution are often combined with differences in **semantic components.** According to lexical decomposition theory, a word's sense may be broken down into smaller semantic components or features. As Cruse (2004: 235) states, "it is probably true to say that virtually every attempt to explicate a rich word meaning ends up by giving some sort of breakdown into simpler semantic components". In some cases, the semantic components that the meaning is composed of are the only criterion that delimits senses. For example, the verb *dot* (*to give*) has the sense of 'to allow to use (something) or take into possession', the semantic elements of which differ from the basic sense: instead of the physical act of giving, it describes the act of giving permission, even though the semantic type of the object is the same (*Ķeizars došot zemi* 'They say the Emperor will give land'). Semantic components influence, for example, the metaphorical subsense 'to pretend' of the verb *spēlēt* (*to play*): *Viņš spēlē gudrinieku* 'He's playing the smart guy'.

The method of semantic decomposition is more relevant in the analysis of monovalent or zero-valent verbs. However, it is also associated with the following problems.

1) It is problematic to define the semantic components, as they can have various degrees of generalisation. Semantic components can be identified best by comparing, for example, the senses of two words or the use of one word in different contexts.

2) The naming of semantic components can also be quite problematic, as words of natural language need to be used and the choice of words will affect the identification of semantic components as well. One attempt at solving this problem is by choosing a

limited number of words, which are used to explain the meaning of other words (see, for example, Wierzbicka, 1996; Goddard, 1998). However, there is no such inventory of semantic components fit for explaining all words of a language, and it is unlikely it could exist, or it would otherwise be too vast for convenient use.

3) The number of semantic components is not finite; in practice, each researcher puts forward a set of semantic components corresponding to the purpose of his research. However, in the work of a lexicographer and also in the development of electronic resources, such an approach would not present a solution, as the entire vocabulary of a language would have to be covered.

4) Even if a detailed decomposition or a word sense is possible, it is not possible to determine specifically how many and which semantic components must differ in order to register different word senses in a dictionary. In this case, a consistent solution is not possible, and the work of the lexicographer, as a rule, involves the use of intuition to determine which semantic components are sufficiently important for their change to create a new sense. If each case of a single differing semantic component was considered a new word sense, the resulting division of senses would be too exhaustive. Therefore, this criterion is usually applied in combination with the syntactic and semantic distribution, which was mentioned earlier.

And lastly, the difference in semantic components can be indicated by the possibility to replace one word with various **synonyms** in different contexts. As substitution with a synonym is a traditional and widely used method of explaining meaning, it can also be used in word sense or subsense distinction. For example, the word *spēlēt* (*to play*) can be substituted by verb *atskaņot* (*to perform*) in connection with music or a piece of music (*spēlēt / atskaņot skaņdarbu, mūziku* 'to play / perform music, a piece of music'), but not in connection with a musical instrument (*spēlēt vijoli* 'to play the violin', but not *atskaņot vijoli* 'to perform the violin'). That is a sufficient basis for a subsense 'to use (a musical instrument) to create sound' to be established. This subsense is also the only one that forms hyponymic relationships with words *trinkšķināt* (*to fiddle*), *čīgāt* (*to saw*), as well as other words for playing musical instruments. The synonyms used in the definitions of word meanings can directly refer to synsets, but it should be noted that synonymy is essentially a relative concept, as the meanings of words can be more or less synonymous and they can have more or less in common.

5. Conclusions

The division of a word's lexical semantics into separate senses may vary depending on the purpose. The aim of word sense distinction in the context of development of WordNet is to obtain such a degree of word sense granularity that would allow to create synonymous, hyponymic, meronymic and antonymic links between word senses and subsenses and at the same time be transparent and easily perceived by any user of the Tēzaurs.lv electronic dictionary, including language learners. In cases of uncertainty, the decision is made in favour of what is needed to develop the Latvian WordNet.

The procedure of distinguishing word senses is based on a set of specific criteria, which are not equally substantial but jointly form a certain hierarchy. However, not all semantic groups demonstrate this hierarchy in the same way. In the sense distinction of polyvalent verbs syntactic distribution (syntactic functions of arguments and ways of coding) and semantic distribution (semantic roles of arguments and general or more specific semantic features) are more important, with semantic components and the possible replacement by a synonym playing a secondary role.

Although the concept of subsense has not been clearly defined yet, in the process of developing the Latvian WordNet the separation of senses and subsenses of verbs has proven necessary. Mostly, a subsense is a way of displaying metonymic (and less often metaphorical) shifts, which cannot be given the status of a separate sense. Regarding verbs, a subsense is most often distinguished by the semantic group of the subject or object. However, it should be emphasised that a consistent solution to subsense distinction is not likely, as it is not possible to determine exactly how large or significant the differences should be in order to consider them as a sign of a separate sense. The authors of the project have tried to formulate the superordinate sense in a sufficiently broad manner for it also to include subsenses. In cases when such an approach was not possible, a subsense was converted into an independent sense. In the formation of synsets and semantic links between word senses, the subsenses listed in the Latvian WordNet function in the same way as superordinate senses: they can form synsets or other semantic relations with other word senses.

Further work on the development of the Latvian WordNet will show whether the selected criteria for word sense distinction will prove useful for automatic word sense disambiguation and linking the Latvian WordNet with the Princeton WordNet. However, the authors are confident that the results of the chosen approach of manually processing the data are of a high quality and will serve as a valuable contribution to the development of lexicography and semantics of the Latvian language.

Acknowledgements

This research work was supported by the Latvian Council of Science, project "Latvian WordNet and word sense disambiguation", project No. LZP-2019/1- 0464.

References

Allen, R. (1999). Lumping and splitting. *English Today*, 15 (4), pp. 61–63.

Cruse, A. (2004). *Meaning in language. An introduction to semantics and pragmatics.* Oxford: Oxford UP.

Darģis, R., Auziņa, I., Bojārs, U., Paikens, P., Znotiņš, A. (2018). Annotation of the Corpus of the Saeima with Multilingual Standards. *Proceedings of the 2018 ParlaCLARIN Workshop.*

Dzerins, J. & Dzonsons, K. (2007). Harvesting national language text corpora from the

Web. *Proceedings of the 3rd Baltic Conference on Human Language Technologies (Baltic HLT).*

Dziob, A. & Piasecki, M. (2018). Implementation of the Verb Model in plWordNet 4.0. Available at: https://www.aclweb.org/anthology/2018.gwc-1.14.pdf

Goddard, C. (1998). *Semantic analysis. A practical introduction.* Oxford & New York: Oxford UP.

Heine, B. & Kuteva, T. (2002). *World lexicon of grammaticalization.* Cambridge: Cambridge UP.

Jackson, H. (2002). *Lexicography. An introduction.* London & New York: Routledge.

Jurafsky, D. & Martin, J.H. (2020). Word Senses and WordNet. In: *Speeeh and Language Processing* (3rded.draft). Available at: https://web.stanford.edu/~jurafsky/slp3/

Kerner, K., Orav, H., & Parm, S. (2010). Growth and revision of Estonian wordnet. *Principles, Construction and Application of Multilingual Wordnets*, pp. 198–202.

Laizāns, M. (2015). *Latviešu valodas korpusa izveide no emuāru tekstiem. Bakalaura darbs.* (*Creation of a Latvian Language corpus from blog posts. Bachelor thesis.*) Rīga: Latvijas Universitāte.

Levane-Petrova, K. (2019). LVK2018: Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos (LVK2018: The Balanced Corpus of Modern Latvian and its role in grammar studies) *Language: Meaning and Form 10*, pp. 131–146.

LLVV: *Latviešu literārās valodas vārdnīca. (Dictionary of Standard Latvian)* 1.–8. (1972–1996). Rīga: Zinātne.

Paducheva, E. (2004). *Dinamicheskie modeli v semantike leksiki.* (*Dynamic models in lexical semantics*) Moskva: Yazyki slavyanskoj kul'tury.

Pustejovsky, J. (1998). *The generative lexicon.* Cambridge etc.: The MIT Press.

Saeed, J. I. (2000). *Semantics.* Oxford & Massachusetts: Blackwell publishers.

Skadiņa, I., Veisbergs, A., Vasiļjevs, A., Gornostaja, T., Keiša, I. & Rudzīte, A. (2012). *The Latvian language in the digital age.* Springer.

Spektors, A., Auziņa, I., Darģis, R., Gruzitis, N., Paikens, P., Pretkalniņa, L., ... & Saulīte, B. (2016). Tēzaurs. lv: the Largest Open Lexical Database for Latvian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), pp. 2568-2571.

Taylor, J. R. (2009). *Linguistic categorization.* Oxford: Oxford UP.

Wierzbicka, A. (1996). *Semantics. Primes and universals.* Oxford & New York: Oxford UP.

Zuicena, I. (2010). Vārda nozīmes skaidrojums "Mūsdienu latviešu valodas vārdnīcā". (Explanation of word senses in the "Dictionary of Modern Latvian"). *Vārds un tā pētīšanas aspekti.* 14(1). Liepāja: LiePA, pp. 369–374.