



UNIVERSITY OF LATVIA
Institute of Mathematics and Computer Science

NORMUNDS GRŪZĪTIS

**FORMAL GRAMMAR AND SEMANTICS
OF CONTROLLED LATVIAN LANGUAGE**

Summary of Doctoral Thesis
in Computer Science

Riga 2010

This thesis has been developed with support from the European Social Fund.
Project “Support for the Doctoral Studies at the University of Latvia”.



Scientific Advisor:

*Senior Researcher, Dr.phys. Andrejs Spektors
Institute of Mathematics and Computer Science, University of Latvia*

Reviewers:

*Professor, Dr.math. Kārlis Podnieks
University of Latvia*

*Senior Researcher, Dr.sc.ing. Alla Anohina-Naumeca
Riga Technical University*

*Professor, Ph.D. Aarne Ranta
University of Gothenburg (Sweden)*

The defence of this thesis will take place in an open session of the Council for Promotion in Computer Science at the University of Latvia, on February 2, 2011, at 4:00 PM, at the Institute of Mathematics and Computer Science, University of Latvia (Room 413, Raina bulv. 29, Riga, Latvia).

The thesis (collection of publications) and its summary are available at the Library of the University of Latvia (Kalpaka bulv. 4, Riga, Latvia).

Head of the Council:

Jānis Bārzdiņš

Abstract

The research subject of this doctoral thesis is the formal, automatic grammatical and semantic analysis of the highly inflective, synthetic Latvian language. A novel hybrid grammar model is proposed, which is especially suited for languages with relatively free word order. The model has been tested on a syntactically restricted subset of Latvian, covering various constructions that can be found in simple extended sentences. The problem is then restricted also from the semantic perspective by developing a deterministic, yet natural subset of Latvian (accompanied with its parser and generator), whose semantics is defined in description logic. The author shows that the analysis of the information structure of a sentence is a reliable way to unambiguously identify the implicit quantifiers and coreferences in OWL terminological axioms, SWRL inference rules and SPARQL integrity queries that are given in a form of a controlled synthetic language. A two-level translation approach is proposed and implemented in a prototype that demonstrates the semantically precise machine translation from controlled Latvian to OWL (and vice versa) by using an existing controlled English as an interlingua and by reusing its readily available tools. In addition, a semi-automatic method is proposed to enable controlled, systematic polysemy and word sense disambiguation in controlled language texts, simultaneously dealing with the OWL ontology merging problem.

Acknowledgments

First of all I would like to thank my supervisor Andrejs Spektors who involved me in the exciting field of computer linguistics and also encouraged me to continue studies at the doctoral level, and has supported me throughout the study period and while writing this thesis. I am also grateful to Guntis Bārzdiņš from whom I have learned a lot and who was the leader of the State Research Programme's project SemTi-Kamols, in the scope of which a large part of this thesis has been developed, and that initiated the further investigations.

Many thanks to my colleagues for the friendly and productive working environment, especially to Gunta Nešpore and Baiba Saulīte for the constant consultations and discussions on linguistic issues, Inguna Skadiņa for the valuable comments, Kārlis Čerāns for the challenging and inspiring test cases, Ilmārs Poikāns and Jānis Džeriņš for their help on various technical issues, and Irēna Šuķe for proofreading the English translation (all the remaining mistakes are my own).

Thanks to Krasimir Angelov and Aarne Ranta for explaining essential aspects of Grammatical Framework, and for their comments on the initial results that encouraged me to actively continue the development of controlled Latvian. Also thanks to all the anonymous reviewers of the included publications for the constructive criticism and valuable suggestions that helped me grow.

I am very thankful for the opportunity to study in the Nordic Graduate School of Language Technology that was funded by NordForsk. Thanks also to the Faculty of Computing at the University of Latvia (UL) and to the Institute of Mathematics and Computer Science (UL) for the creative and stimulating environment.

This research has been mainly funded by the State Research Programme in Information Technology and by the European Social Fund.

Finally, the greatest thanks goes to my family and to my relatives for their support, especially to my wife for the understanding and for the tremendous support, my sons for the inspiration, strength and teaching how to use time more efficiently, and my parents for education and for all care. This thesis is dedicated to them.

Table of Contents

Topicality of the research and the achieved results	6
General characterization of the doctoral thesis	8
Introduction	10
1 What is the Sense of a Word?	13
2 From Language to Ontology	16
2.1 Dependency-based Syntactic Analysis	16
2.2 Multidimensional Ontology	22
3 Micro-ontologies as a Sense Inventory	25
3.1 Logic-based Controlled Natural Language	25
3.2 Micro-ontologies	26
3.3 Integration of Micro-ontologies and Word Sense Disambiguation	27
4 Controlled Latvian for Ontology Authoring and Verbalization	30
4.1 Analysis of the Information Structure of a Sentence	30
4.2 Extensions of the SVO Sentence Pattern	32
4.3 Syntactic Paraphrasing	33
4.4 Two-level Translation to/from OWL	37
Conclusion	42
List of included publications	43
References	44
Author's contribution to the included publications	50

Topicality of the research and the achieved results

Nowadays, when vast amounts of information are available in a machine-readable form, means for automatic analysis of such information is of great importance. This thesis, being a part of a larger research project, investigates facilities for formal analysis of information that is given in a form of natural language (text). The thesis mainly considers Latvian as the source and target language in the tasks of text meaning representation and language generation, largely relying on existing sources of formalized background knowledge — ontologies.

Latvian is a highly inflective, synthetic language: its rich morphology allows for syntactically free word order. In contrast to synthetic languages there are analytic languages (e.g., English): their simple morphology typically is compensated by the rigid word order. Computational linguistics has been mainly focusing on the formal analysis of English and other relatively analytic languages; models and methods that have been widely accepted for analysis of analytical languages have been often directly applied also to the analysis of synthetic languages, thus ignoring some important characteristics of this type of languages.

In the scope of this research, novel methods for syntactic and semantic analysis of Latvian have been developed, taking into account the characteristics of this highly synthetic language. The proposed methods can be applied (adapted) for other synthetic languages as well; moreover — some of the methods are language-independent and provide an alternative perspective in the context of the generally accepted approaches.

However, analysis of unrestricted Latvian is currently limited — natural language is highly ambiguous at all the levels of analysis (from morphology up to discourse), and, in general, it is hardly possible to ensure the choice of the most appropriate variant of analysis by exploiting rule-based methods alone. To ensure this, rule-based methods should be combined with statistical ones, but, in advance, the statistical language model has to be acquired (induced) from a large, manually annotated text corpus that contains both grammatical and semantic annotations; such a resource is currently not available for Latvian.

Therefore the main subject and the main result of this doctoral thesis is a syntactically and semantically controlled Latvian (together with its parser and generator). The restricted expressivity ensures that the meaning of controlled Latvian sentences is precisely and deterministically conveyed to the chosen formalism — the web ontology language OWL [15].

One of the primary aims of the semantic web technology is to make the on-line, decentralized and mainly unstructured information not only machine-readable, but also “machine-understandable”, thus facilitating the extensive automation and integration of information processes in a wide range of economic sectors and for the society in general. This technology is based on the development of domain ontologies (formal conceptual models), providing knowledge representation and reasoning capabilities.

In the development of domain-specific ontologies, not only knowledge engineers, but also domain experts play an important role, therefore controlled natural language (CNL) can be seen as an alternative (complementary) and a highly intuitive ontology language in contrast to the formal notations. An important “side effect” of the CNL interface is the ability for semantically precise (controlled) machine translation, facilitating the multilingual development and reuse of ontologies.

Main results

- An original dependency-based grammar model has been developed by implementing and extending the basic concepts of Tesnière's dependency grammar theory [16]. The proposed model is primarily aimed at inflective, synthetic languages. In contrast to the "classical" approaches to dependency grammar (in computational linguistics), it simultaneously describes both the vertical (dependency) relations between parts of sentence, and the horizontal (non-dependency) relations among the constituents of analytical word forms etc. This model has been tested in practice by formalizing a relatively wide subset of Latvian grammar, i.e., by covering various syntactic constructions that frequently appear in simple extended sentences¹ (e.g., analytical verb forms, expressing tense, mood and modality, as well as other complex parts of speech and their possible dependencies).
- A concept of micro-ontology has been defined, and a methodology for the consistent development and integration of such micro-domain ontologies has been proposed. A semi-automatic, language-independent method has been developed for the systematic partitioning of polysemous (inconsistent) classes (i.e., to introduce a controlled sense inventory), and for word sense disambiguation in factual (assertional) texts.
- A controlled, yet an utmost natural Latvian language has been developed², providing alternative, intuitive means of expression for the authoring and representation (verbalization) of OWL ontologies, in contrast to the explicitly formal notations. It has been shown that the analysis of the information structure of a sentence is sufficient for the unambiguous (deterministic) resolution of quantifiers (incl. the given and new information) in terminological axioms (OWL), inference rules (SWRL [17]) and data integrity queries (SPARQL [18]) that are given in a form of controlled synthetic language. The controlled Latvian (its parser/generator) has been efficiently implemented in Grammatical Framework (GF) [19]. To ensure the possibility of using different sentence patterns and syntactic constructions (incl. reductions) for expressing the same meaning (i.e., the same axiom or rule), as well as to allow for certain types of non-SVO (subject-verb-object) sentences or clauses (at the level of controlled language), two parallel grammars of controlled Latvian have been developed: one for a robust analysis, and the other — for the most natural and precise paraphrasing. Finally, a prototype has been developed that demonstrates a novel two-level approach (in the case of controlled languages) for translating from controlled Latvian and controlled English to OWL (and vice versa) by using the OWL subset of Attempto Controlled English [20] as an interlingua (together with its freely available parser and verbalizer).

The results of this research have been achieved in collaboration with co-authors. The author of this thesis has contributed significantly to all the stages of the underlying research. Author's contribution has been clarified in more details in the next sections of this summary and in the appendix.

¹ In other words, sentences that contain no subordinate or coordinate clauses.

² A subset of natural Latvian with few marginal exceptions.

General characterization of the doctoral thesis

The doctoral thesis “Formal Grammar and Semantics of Controlled Latvian Language” was developed from 2005 to 2010 at the Faculty of Computing, University of Latvia (UL) and at the Institute of Mathematics and Computer Science (IMCS UL) under the supervision of Andrejs Spektors. The research results are described in 11 publications [1–11], of which most have been prepared in the scope of the State Research Program in Information Technology project SemTi-Kamols, led by Prof. Guntis Bārzdīņš.

This thesis is a **thematically-related collection of publications**, which reflect the research that has been carried out by the author and its results, covering various aspects of formal syntactic and semantic analysis of natural language.

Research subject matter

The formal, automatic grammatical and semantic analysis of the highly inflective, synthetic Latvian language. A well-defined subset of Latvian as a knowledge representation language. Ontology-based word sense disambiguation.

Research aim

The wider aim of the undertaken research is to develop a novel method (model) of the formal grammatical analysis (representation) that would take into account the specific character of inflective, synthetic languages (by considering a commonly used subset of Latvian as an example), and to develop the generally known ontology-based text meaning representation approaches.

The narrower aim is to develop a syntactically restricted, deterministically parsable, while utmost natural subset of Latvian, whose semantics is defined in description logic — a decidable subset of first-order logic on which the Semantic Web ontology language OWL [15] and the inference rule language SWRL [17] is based.

Research stages

On the one hand, the results of the thesis make a rather complete research cycle, beginning with investigation on in-depth grammatical analysis of highly inflective, synthetic languages like Latvian, and on language-neutral modelling of deep lexical and syntactic semantics, and ending with investigation of logic-based controlled natural languages. As a result, a controlled Latvian together with the means for its analysis/generation has been developed for the deterministic authoring and verbalization of OWL ontologies. On the other hand, the achieved results should be treated as a prototype that has laid the base and encouraged continuing the research and development in this direction.

The underlying research can be split into four stages; the experience that was gained in the first stages provided a substantial base for successful completion of the last ones. These stages are organized in a thematic order that mostly corresponds also to the chronological order:

- Results of the **first stage** are described in Section 1 of this summary and in the corresponding publications [1–3]. The author has actively participated in the underlying research, and his contribution is more than 50%. The conclusions of this stage initiated further research, which led to the main results of the doctoral thesis.
- Results of the **second stage** are described in Section 2 of this summary and in the corresponding publications [4–6]. The author has actively participated in the underlying research, and his contribution is more than 40%. The main result of this

stage is a novel **hybrid dependency-based grammar** that is especially suited for languages with the relatively free word order, ensuring both flexible and detailed analysis (representation). The research and development of this stage provided a significant source of ideas for the further development of the controlled (logic-based) Latvian and the means for its deterministic analysis.

- Results of the **third stage** are described in Section 3 of this summary and in the corresponding publications [7, 8]. The author has actively participated in the underlying research, and his contribution is more than 60%. The main result of this stage is the proposed **concept of micro-ontology** and the **methodology for development and integration of such ontologies**. This course of research led to important conclusions and directly initiated the last and, in the scope of this thesis, the most important stage.
- Results of the **fourth stage** are described in Section 4 of this summary and in the corresponding publications [9–11]. The author has led the underlying research, and his contribution is more than 90%. The main result of this stage is the developed **controlled, deterministic Latvian and its parser and generator (paraphraser)**, which ensure that complex axioms and inference rules that are naturally and intuitively expressed in this language can be precisely interpreted in terms of the ontology language OWL, and vice versa — formal axioms and rules can be verbalized in the form of an utmost natural (subset of) Latvian.

Theoretical and practical significance

Under the SemTi-Kamols project, a partial parser has been developed for the proposed dependency-based hybrid grammar³. It is being used for semi-automatic and even automatic morphological annotation of various Latvian text corpora [21]. These experimental corpora are further used, for instance, to acquire factored language models for a statistical English-Latvian machine translation system [14, 22].

In the beginning of 2011, IMCS UL is planning to start an EU funded project “Semantic Database Framework for Domain Experts”, where the controlled Latvian (in parallel with controlled English and a UML-style graphical notation) will be further developed as a significant component (natural language interface). Natural language as an intuitive knowledge representation language can be seen as a supplement to other high-level notations for OWL, for instance, the various UML profiles [23], for which the target audience (similarly to a controlled natural language) to a large extent are domain experts, who might not be directly related to the IT field. Graphical languages help to unveil the interconnections among concepts and the overall structure of an ontology, but they are not well suited for defining and visualizing complex restrictions. This, however, can be conveniently and comprehensibly done by using a controlled natural language.

The rest of the summary is organized as follows. The main issues in natural language processing are briefly described in the introduction. Sections 1 to 4 concisely outline the underlying research and the achieved results, which are described in more detail in the corresponding publications. Some general conclusions and future tasks are outlined at the end. The list of the included publications and the author’s contribution to each of them is provided in the appendix.

³ The reference implementation of the parser is not a part of the author’s contribution. Author’s contribution is the grammar on which the parser is based on. Both components are freely available: www.semti-kamols.lv

Introduction

This thesis deals with issues in an interdisciplinary research field — computational linguistics. The central aim of computational linguistics is natural language understanding: automatic construction of text meaning representation and vice versa — language generation for the given model. Machine translation is a typical example that aims to ensure this roundtrip (in principle, by translating thought for thought). The main results of this thesis are also in a sense related to machine translation, however, we will address this concept not only for translation among natural languages (their subsets), but also between natural and formal languages. Although in practice, in most natural language processing (NLP) applications, the levels of deep semantic analysis are currently not reached, the grammatical (morphological and syntactic) analysis of the surface structure of a sentence already provides significant possibilities in information extraction, analysis and translation.

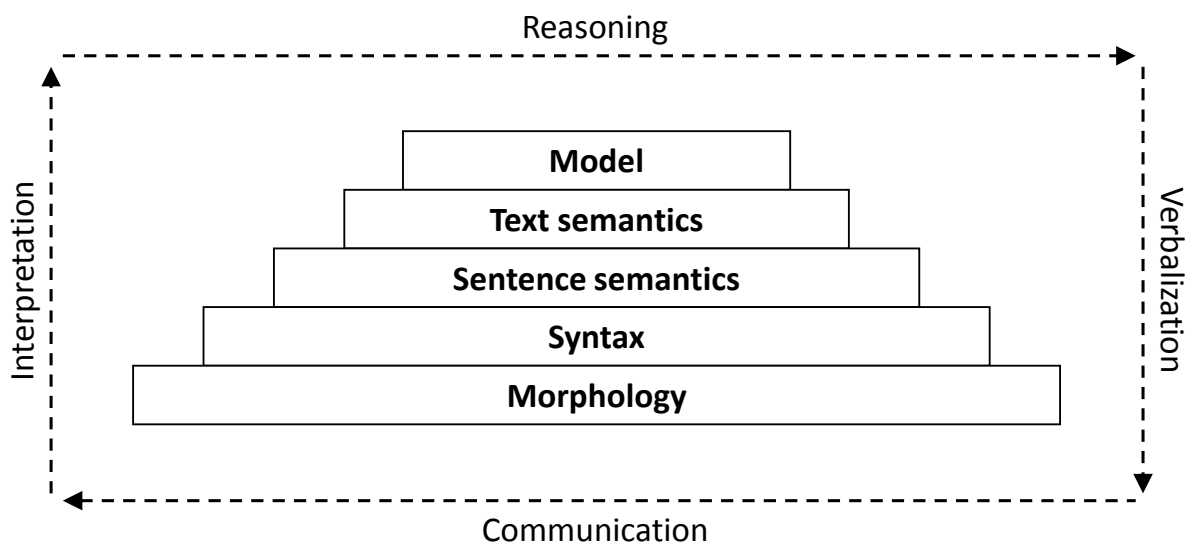


Fig.1. The levels of natural language processing and their role in the communication process.

The main problem in NLP is the excessive ambiguity of natural language (NL) in all the levels of analysis (see Figure 1): from the part-of-speech tagging⁴ up to detection of discourse referents. Although in general this makes the deterministic analysis of unrestricted NL virtually impossible, ambiguity is a phenomenal means that makes the use of NL possible for humans. The number of words, word forms, and syntactic constructions is finite, however, their combination and interpretation possibilities are theoretically infinite. It can be illustrated by the distribution of words in a text corpus, like the BNC corpus (see Figure 2).

According to Zipf's law [25] the frequency of any word is inversely proportional to its rank in the frequency table (against a text corpus). Thus the most frequent word occurs approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. In other words, relatively few words are used frequently — most words are used rarely. There are two consequences. First, words are intensively reused in different contexts, causing the polysemy, and, second, a huge⁵ corpus is necessary to cover the rarely used words/senses, so that the coverage would be statistically significant.

⁴ Especially for highly inflective languages like Latvian.

⁵ To compare, the size of the BNC corpus is 100 mill. running words, however, nowadays even billion word corpora are being constructed from the Web [26].

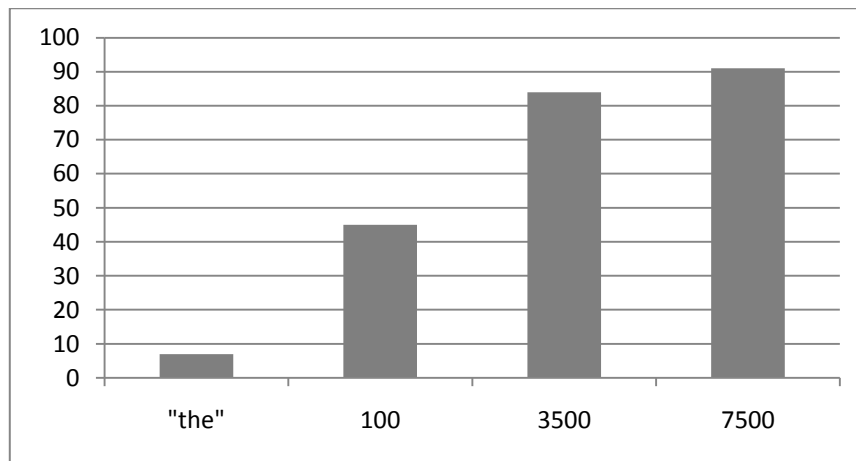


Fig.2. Distribution of words in the BNC corpus: 7500 most frequent words make up 90% of all tokens [24].

Theoretically, every subsequent level of NL analysis helps to eliminate or at least to reduce the ambiguity of the previous level: morphological ambiguities are reduced by syntactic analysis, syntactic ambiguities — by lexical and syntactic semantics, and, finally, semantic ambiguities — by the context and the extra-linguistic (world) knowledge.

Let us consider a simple sentence: *“es ceļu māju”* (“*I am building a house*”). If we analyse the morphological features of each individual word form, we get:

- “*es*” — the personal pronoun “*I*” (first person, singular);
- “*ceļu*” — the verb “*to build*” (indicative mood, present tense, first person), or the noun “*path*” (singular, accusative OR plural, genitive), or the noun “*knee*” (plural, genitive);
- “*māju*” — the noun “*house*” (singular, accusative OR plural, genitive), or the verb “*to wave*” (indicative mood, present tense, first person).

The syntactic analysis decreases the morphological ambiguity, but, unfortunately, it is ambiguous itself:

- “*es*”(“*I*”)SUBJECT + “*celt*”(“*to build*”)VERB + “*māja*”(“*house*”)OBJECT
- “*es*”(“*I*”)SUBJECT + “*ceļš*”(“*path*”)OBJECT + “*māt*”(“*to wave*”)VERB

In order to choose the correct syntactic parse, one needs knowledge of:

- lexical semantics, for instance, “*māja*” (“*house*”) is a hyponym of “*celtne*” (“*building*”), but “*celt*” (“*build*”) is a synonym of „*būvēt*” (“*construct*”);
- syntactic valence of a verb — its typical complements, for instance, the verb “*māt*” (“*to wave*”) is intransitive (in Latvian), thus is used together with a prepositional construction (e.g., answers to the question *with what?*) instead of a direct object (*what?*);
- syntactic semantics (semantic valence) — what are the typical semantic categories (semantic roles) of verb complements, for instance, “*māt*”(“*to wave*”) attracts INSTRUMENT, but not PATIENT;
- world (ontological) knowledge, for instance, *everything that has been built by someone is a building*.

In most cases, humans deal with such disambiguation tasks very well⁶, however, for the automated grammatical and semantic analysis it presents an enormous obstacle — due to the difficulty of formalizing linguistic and world knowledge.

⁶ In practice, when syntactically [27] and semantically [28] annotated corpora are being developed, there is often no consensus even among linguists on how to interpret a sentence or a word. There are various reasons for this: insufficient context, a too fine-grained sense inventory, open questions in syntax as such.

There are two general approaches in computational linguistics: the rule-based approach, which aims at deep, systemic description of NL, according to the linguistic theories, and the statistical approach, which aims at a shallow approximation of the language models, exploiting statistical and machine learning methods. Again, the industry of the (statistical) machine translation (SMT) can be mentioned as a well-known example, where the language and translation models are acquired by calculating n-gram probabilities in parallel corpora [29]. To some extent, statistics allow us to avoid the need for in-depth formalization of (linguistic) knowledge, and help to ensure a comparatively wide (unrestricted) coverage of language. However, the higher the level of analysis involved, the deeper (more structural) knowledge is needed⁷. The idea is depicted in Figure 3.

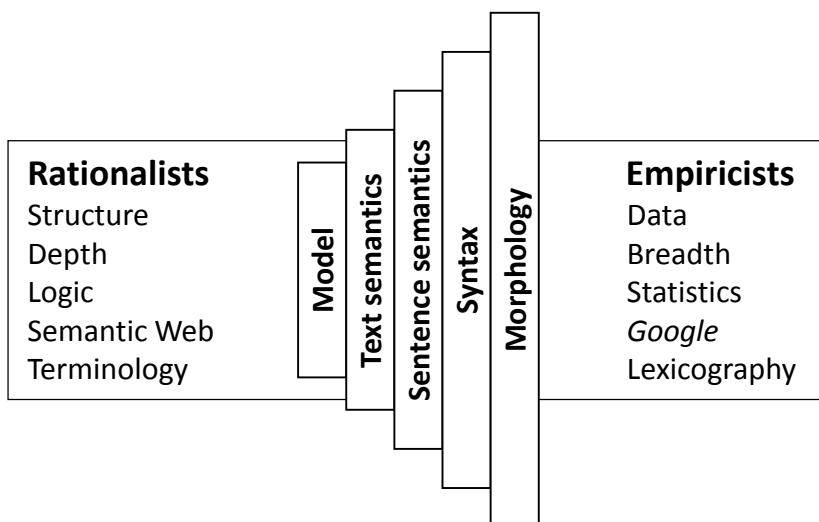


Fig.3. Two general approaches to natural language processing (the differentiation between rationalists and empiricists is taken from A. Kilgarriff [31]).

In practice, the most promising is the hybrid approach by merging rule-based and statistical means. For instance, a high-precision and wide coverage parser can be acquired by learning grammar rules and their weights from a treebank [32], whereas better results in MT can be achieved by combining statistical and rule-based models [33]. However, sufficiently large syntactically and semantically annotated corpora have to be prepared in advance. For instance, one of the largest corpora of that type is the German SALSA/TIGER corpus that covers ca. 50 000⁸ syntactically annotated sentences and semantic role labelling has been done for ca. 20 000 verb instances [34]. On the basis of this corpus the relatively high-precision semantic parser Shalmaneser has been trained for semantic role labelling in unrestricted texts [35]. It should be noted that the higher (deeper) level of analysis is considered, the more laborious and expensive is the development, and therefore a comparatively smaller corpus can be prepared. In addition, domain-specific knowledge takes a more and more important role.

For Latvian there is no such corpus available, and the aim of this research is the in-depth analysis of Latvian, therefore the (shallow) statistical methods are not considered and used. Moreover, only a subset of Latvian is being considered: initially, by imposing syntactic restrictions⁹, and then — also semantic restrictions.

⁷ The current state of the art in SMT implicitly confirms it [30].

⁸ <http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/>

⁹ Manual formalization of a wide coverage grammar is a complex, time-consuming process, and there are several grammatical phenomena in Latvian that have to be more systematically categorized by philologists.

1 What is the Sense of a Word?

In this section the initial research on the modelling of lexical and ontological semantics, and the lessons learned are briefly outlined. On this basis, the further investigations were initiated that are described in Sections 2–4 and where the main results are achieved.

The initial course of the research [1–3] reflects the expectations that the semantic analysis of a text is possible in a top-down manner — that the primary disambiguation task is to differentiate word senses (according to a predefined dictionary, i.e., ontology). The most popular model for the development of such lexical ontologies (taxonomies) is the Princeton WordNet [36]; similar taxonomies have been developed afterwards for many languages all over the world¹⁰. WordNet is based on the concept of synonym set (synset): every synset represents a concept, containing words (lexical signs of the concept) that are mutual synonyms in some particular senses¹¹. There are different types of predefined semantic relations possible among the synsets: hyponymy (generalization), meronymy (part-whole), entailment, causativity and others.

A number of algorithms for the word sense disambiguation (WSD) task have been developed that exploit the large and detailed WordNet taxonomy by calculating the semantic distance¹² between words (their potential senses) that appear in some context “window”, by measuring the degree of direct and indirect semantic interconnections in a set of potential word senses etc. [37]. However, as the evaluation shows, the performance¹³ of the state of the art WSD systems of that time was only slightly above 65% [38]. Note that the naive approach by always selecting the most frequent sense of each word¹⁴ would have ensured a score of 60% for the given test corpus [38].

Despite this, a similar lexicon for Latvian was considered necessary, however, taking into account the amount of labour in the case of its manual construction¹⁵, an investigation on semi-automatic WordNet construction methods was carried out. One of the most appropriate sources of lexical knowledge in such case are (machine-readable) dictionaries that actually contain most of the semantic relations that are necessary for a WordNet kind of a taxonomy. The “only” problem is that they are given in an implicit and informal way — in the definitions of word senses and in their usage examples. The results that have been achieved by analysing definitions from dictionaries of English (by applying lexico-syntactic patterns as illustrated in Figure 4) [40, 41] inspired the author to carry out a similar experiment for Latvian. As a result of the fully-automated quick-and-dirty experiment, rough verb synsets and hyponymy relations among them were acquired [2].

<p>such NP₀ as {NP}* {(and or)} NP “.. works by such authors as Herrick, Goldsmith, and Shakespeare ..” ⇒ HYPONYM(Herrick, author) HYPONYM(Goldsmith, author) HYPONYM(Shakespeare, author)</p>
--

Fig.4. An example of a lexico-syntactic pattern for extracting hyponymy relations in a text corpus [42].
Hypothesis: $\forall NP : \text{hyponym}(NP, NP_0)$, where NP is a noun phrase.

¹⁰ http://www.globalwordnet.org/gwa/wordnet_table.htm

¹¹ Note that in natural language there are few absolute synonyms.

¹² WordNet is an oriented, acyclic graph.

¹³ A weighted average of the precision and recall using the same test corpus.

¹⁴ Zipf's rule can be applied to word senses and their frequency as well [39].

¹⁵ The original (and largest) Princeton WordNet was developed for about 20 years.

During the experiment, word senses were not differentiated (partitioned into different synsets) — if two words appeared to be synonymous in any one of their senses, they were treated as absolute synonyms. As it turned out, if the defined patterns are fine-grained enough, they can to some extent compensate for the rough assumption: the average size of the acquired synsets was actually the same as in the case of Princeton WordNet¹⁶. It gave confidence that it is possible to acquire a practicable result by including more detailed constraints in the set of heuristic patterns¹⁷, and by using bilingual or even multilingual resources (translational dictionaries or parallel text corpora) for automatic word sense partitioning [44], as well as by performing some manual pre- and post-processing [1, 45].

The original WordNet approach, however, has several well-known shortcomings: word sense partitioning is too fine-grained¹⁸, there is a lack of domain-specific knowledge, the formalization is shallow and, thus, the possibilities for the automatic reasoning are limited. For practical applications, domain-specific ontologies are often used, in which the necessary background-knowledge is formalized more precisely, but in order to link such ontologies with natural language, lexical knowledge is needed. The border, where the lexical knowledge ends and the ontological knowledge starts, is vague and hard to define [46].

A reasonable trade-off is to separate and, meanwhile, to align the two knowledge sources, leaving the shades of meaning at the level of the lexicon, but at the ontological level — the conceptual disambiguation (i.e., WSD) and the text meaning representation (see Figure 5). This scenario was adopted in the first stage of the research [1], following the state of the art approach in ontological semantics — OntoSem [47]. The development of this approach had been carried out for more than a decade, claiming to cover rich common-sense knowledge.

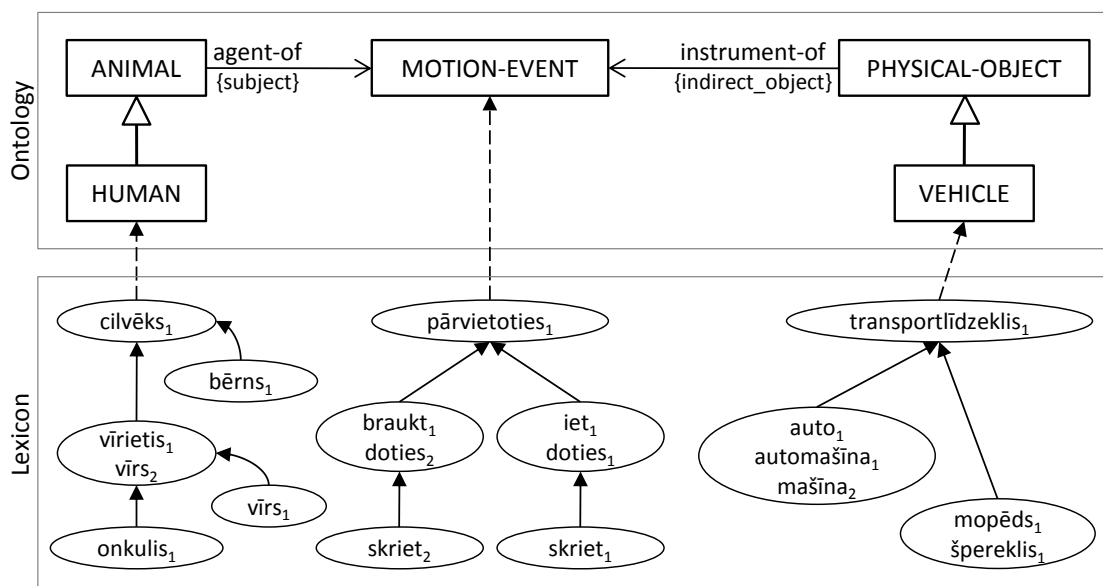


Fig.5. A simplified illustration of separated, but aligned lexical and world knowledge.

Besides OntoSem, there are other well-known common-sense (“upper-level”) ontologies, for instance, SUMO, which has been aligned with WordNet [48], but, in addition to the lexical mapping, the OntoSem approach provides also a syntactic mapping. However, several

¹⁶ The average size of verb synsets in the Princeton WordNet is about 1.82 (<http://goo.gl/jcFgA>); the same results were achieved in the experiment [2]. However, the dictionary of Latvian that was used covers 30% less verbs — equal sets would slightly affect the results due to the lack of word sense partitioning.

¹⁷ By performing an actual syntactic analysis, adjusted to the sentence patterns that are used in definitions [43].

¹⁸ As practice shows, even in the case of manual WordNet-based sense tagging in a corpus, the inter-annotator agreement is only about 70% [38].

aspects turned out in the OntoSem approach that makes it difficult (non-flexible) to reuse and develop further. Although OntoSem claims to be able to construct a meaning representation of an unrestricted text, its relatively wide coverage is shallow — for in-depth analysis each domain has to be described separately [49] by defining the relevant concepts and by specifying the lexicon¹⁹. This is a complex process: besides the concepts that represent objects and events, specific scripts for complex events have to be provided, and specific (ad-hoc) meaning procedures for calculation of context-sensitive parameters should be defined in the lexicon as well [50] — all this in addition to the syntactic patterns. Thus, the approach is highly dependent on the particular OntoSem implementation. Last but not least, the formalism that is used by OntoSem is not compatible with the widely used standards, OWL²⁰ in particular, which means that the efficient, general-purpose OWL reasoners cannot be used and the integration with existing domain-ontologies is difficult.

OntoSem ontology is defined in a proprietary frame-based formalism that, on the one hand, is more expressive than OWL, but, on the other hand, does not support the use of logical restrictions, therefore the possibility for automatic reasoning is limited. Moreover, its semantics is partially encoded into the text analysis applications that are driven by OntoSem. In order to make the ontology compliant with the Semantic Web standards and tools, it was converted into OWL DL [3].

Description logic uses the open world assumption: all classes overlap by default — it has to be explicitly specified, which classes are mutually disjoint (cannot have any instances in common). Such restrictions are needed both for word sense partitioning and for automatic reasoning (to ensure WSD), therefore they were gradually introduced by manually analyzing the class taxonomy (in the top-down direction). However, the ontology soon became inconsistent, unveiling a fundamental problem — unavoidable ambiguity also at the ontological level²¹ (see Figure 6). How to systematically differentiate the ambiguous concepts? What (formally) is a word sense? Are there predefined word senses at all [51]? This presents a problem not only for OntoSem, but for common-sense ontologies as such. Thus, is it possible to create a consistent common-sense (multi-domain) ontology at all?

To better understand these essential questions on how word senses arise, it was decided to search for the answers from the bottom, by analysing NL as a **united system**: from the grammar up to the ontology, formalizing regularities and connections in depth and in detail²².

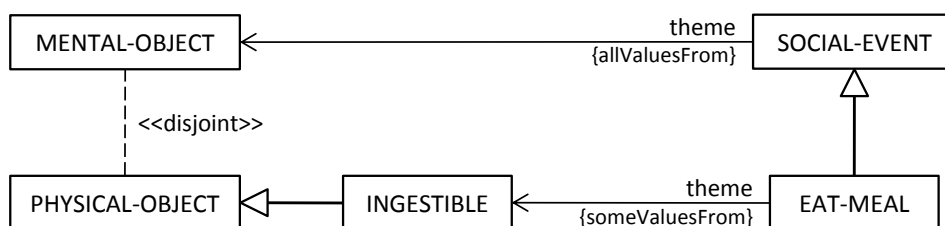


Fig. 6. An inconsistent fragment from the OWL compliant OntoSem ontology. It was pointed out by the automatic reasoner (no instances are possible for the `EAT-MEAL` class).

¹⁹ Thus, the OntoSem ontology is by no mean balanced — in the result of individual research projects only several domains are described in-depth (e.g., politics, economics and biomedicine).

²⁰ In this thesis, by OWL is meant OWL DL (incl. the latest OWL 2), which is based on description logic (DL) — a subset of first-order logic (FOL) that ensures effective reasoning capabilities.

²¹ It concerns not only the word senses, but also concepts as such: different schools and beliefs exist in parallel (e.g., whether the human is an animal — see Figure 5).

²² Note that the shallow WSD systems are showing promising results, achieving the precision over 80% — if a coarse-grained sense inventory is used instead of the fine-grained WordNet [52]. However, it is not sufficient to guess only the word senses; relations among tokens (incl. co-references) have to be detected as well.

2 From Language to Ontology

In this section, the principles and results of the second research stage are outlined [4–6]. In contrast to the previous stage, the roots of text meaning and ontology have been investigated in the bottom-up direction, i.e., from the morpho-syntactic analysis up to frame semantics and discourse, trying not to miss any formal detail. Although for practical reasons the considered subset of Latvian has been syntactically restricted, no particular semantic restrictions were imposed.

2.1 Dependency-based Syntactic Analysis

Latvian is a member of the Baltic language group for which synthetic word forms are dominant: due to the fact that grammatical senses are expressed mainly via flections (cases), the word order is syntactically free — parts of sentence can be recognized by flections and the syntactic agreement between them. But even in highly synthetic languages the word order is rather free: it is restricted by both analytical forms and the semantic bounding.

Two general approaches can be distinguished in the syntactic analysis: phrase structure grammar (PSG), also known as context-free grammar [53], and dependency grammar [16] (see Figures 7 and 8 respectively). At the first glance, assuming that only projective parse trees are produced, i.e., trees have no crossing edges (while preserving the linear order of words in a sentence), both formalisms are weakly equivalent²³, and their grammars can be mutually transformed [54]. However, both approaches provide essentially different perspectives and methodology, with respective pros and cons.

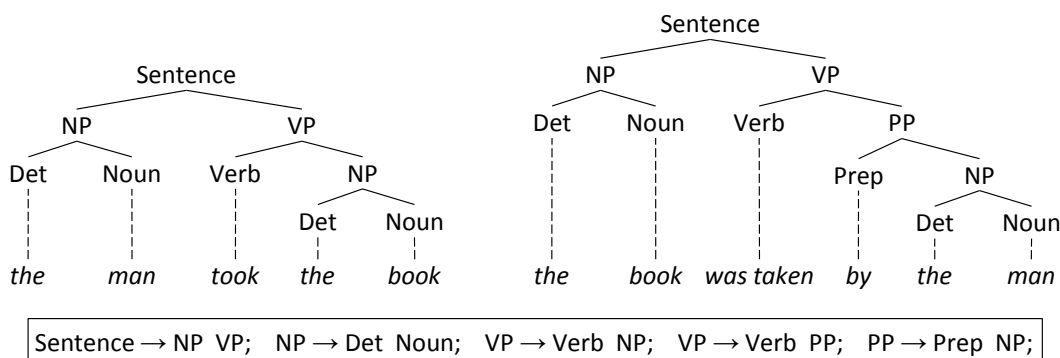


Fig.7. A simple example, showing how English sentences can be represented in a PSG grammar (parse trees for the active/passive voice differ due to the lack of differences at the morphological level).

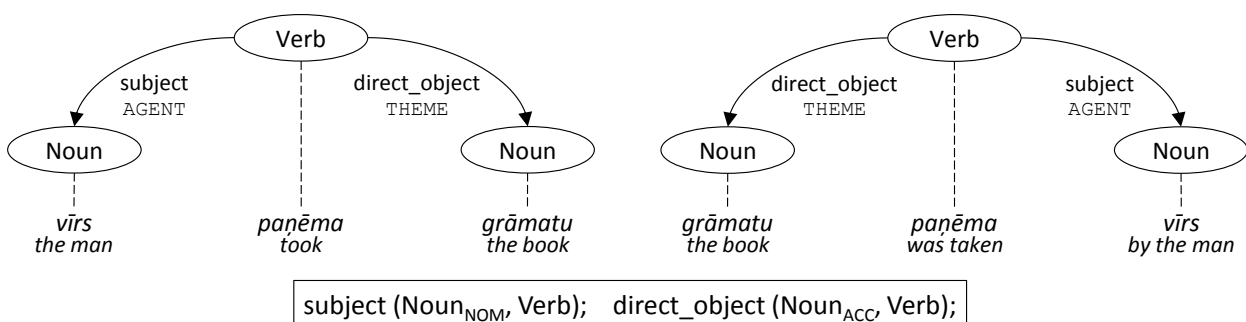


Fig.8. An example showing how the Latvian counterparts (see Figure 7) are represented in a dependency grammar. The voice does not have to be changed in order to change the word order — both parse trees are equal (the same functions are used); only the linear order of verb dependents is different.

²³ Their grammars generate equivalent sets of sentences, but they do not assign the same sentence structure.

In the case of analytical languages (like English), where the word order is rigid, traditional is the top-down approach: by analysing the linear order of words, sentences are recursively split into phrases and subphrases (constituents) according to a set of production rules²⁴. In the case of synthetic languages, traditional is the bottom-up approach: by recognizing parts of sentence and dependency relations between them, according to a set of syntactic functions. Besides, dependency grammar, in contrast to PSG, allows for an adequate modelling of non-projective structures, i.e., discontinuous constituents that is a rather frequent phenomenon in languages with relatively free word order²⁵. In practice, however, the type of language is not considered much: to formalize a synthetic language and/or to develop a treebank for that language, the PSG approach is often chosen, but during the last five years or so, together with the emerging efficient data-driven parsers²⁶, the popularity of the dependency approach has significantly increased — also in the case of analytical languages, for which a number of automatic converters have been developed to acquire the dependency representation of existing PSG treebanks [59]. The benefits of dependency grammar have been recognized, for instance, in information extraction [60], automatic semantic role labelling in unrestricted texts [61, 62], and in other tasks of semantic analysis.

The type of language is not the only criterion that should be considered. On the one hand, dependency grammar provides more robust means for analysis: the same subset of NL can be covered by a smaller set of simpler rules (functions), without specifying the word order. On the other hand, it is not appropriate for language generation²⁷ that is the strength of PSG. However, in the scope of this thesis, a more important criterion is the subsequent possibility for mapping of parse trees into their meaning representations. From this perspective, the dependency approach is beneficial, since it places the verb (predicate) at the centre, together with its argument structure that is directly encoded by dependency links. The verb is central not only syntactically, but also semantically — as an event. Thus, the dependency relations can be straightforwardly mapped into the corresponding semantic relations (see Figure 8). It should be noted that such a direct bidirectional mapping between the syntactic and semantic layers facilitates also the possibility for using lexical (incl. frame) semantics in the syntactic disambiguation [63].

There are various theories, representations and implementations of the dependency approach [58]. There are also various hybrid formalisms that extend the PSG approach by using elements of dependency grammar. For instance, the well-known HPSG (head-driven PSG) approach [64] that, apart from other, allows to specify, which is the head constituent of a phrase and what is its valence. A hybrid annotation schema has been used also for the already mentioned TIGER corpus, where edges are labelled by syntactic functions, and secondary functional links are used in addition to the structural edges, allowing for representation of discontinuous constituents [65].

Dependency grammars are usually treated and implemented in a very simplified manner, if compared to Tesnière’s original approach [16], by sacrificing the linguistic nuances for the benefit of efficient parsing algorithms [66]. In the result, each word (either a function word or a content word) actually is treated as a separate part of sentence, which is involved in a

²⁴ The parsing algorithm, of course, can be implemented (and usually is implemented) as a bottom-up search.

²⁵ For instance, in the Prague Dependency Treebank, 23% of all sentences contain non-projective links [55].

²⁶ The complexity of the classical PSG algorithms (e.g., the Earley parser [56]) is $O(n^3)$. The complexity of projective dependency parsing is the same [57]. However, the recognition of non-projective structures, and the satisfaction of various global constraints over a parse tree, in general, is a NP complete problem [55, 58].

²⁷ Although in inflective languages the order of parts of sentence is syntactically free, pragmatically it is bound.

separate dependency relation²⁸ [57]. Also it is not taken into account that the word order is not syntactically absolutely free even in highly synthetic languages²⁹. By taking into account these aspects, a hybrid dependency-based model (a.k.a. SemTi-Kamols grammar) for syntactic analysis has been developed by implementing and extending Tesnière’s original approach [4, 5].

In Tesnière’s structural syntax, the basic concepts are the following [16]: syntactic relations (*connexions structurales*), junctions (*jonctions*) and transference (*translation*). For the notion of a node (in a parse tree), Tesnière has introduced also the concept of a syntactic nucleus (*nucléus*). Nuclei are inseparable units that apart from the node itself (a content word) may contain additional constituents (usually, function words). Dependency links appear only at the level of nuclei, and not at the very surface level. Nuclei are acquired via the morpho-syntactic transference that is an operation that changes the original category (function) of the content word (the head constituent of a nucleus), depending on the function word(s). Finally, junctions combine parts of sentence that are in a coordination (horizontal) relationship instead of a subordination (vertical) relationship.

The concept of syntactic nucleus is implemented in the proposed SemTi-Kamols grammar, by using a mechanism that is very similar to the transference operation. Moreover, the notion of nucleus has been extended by covering also junctions. For this purpose the concept of complex words or “x-words” has been introduced (see Figure 9). X-words are devices that cancel off substrings during the parsing. Due to their dual nature, they act as glue between the PSG approach and the simplified dependency grammar approach:

- from the PSG point of view, an x-word is a non-terminal symbol that substitutes all constituents forming the respective node (i.e., x-words define the horizontal relations);
- from the dependency grammar point of view, x-words are regular words that can act as head and/or dependent nodes in the dependency (i.e., vertical) relations.

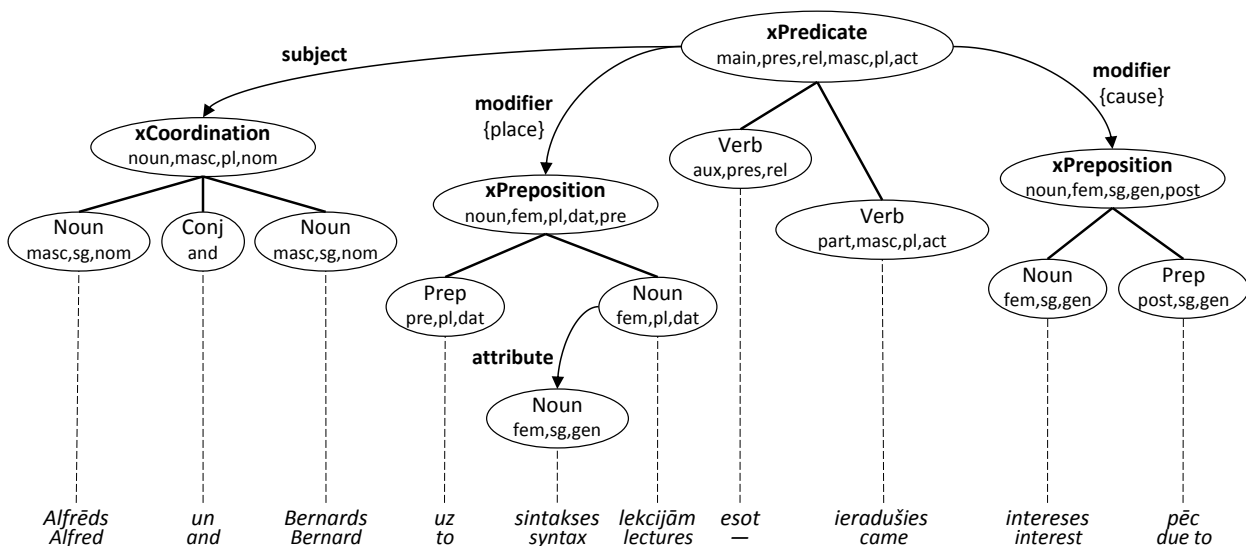


Fig.9. A simplified parse tree according to the hybrid, dependency-based grammar.

As a result of transference, new artificial nodes representing x-words are created in the parse tree. X-words have their own morpho-syntactic features that are mostly inherited from their constituents, but additional information that specifies the x-word as a whole can be

²⁸ For instance, the analytical verb form “was taken” is artificially split into two parts of speech (two nodes of the parse tree), so that the main verb depends on the function verb or vice versa.

²⁹ Consider the prepositional phrases, for instance.

included as well, allowing to check for additional agreement restrictions while applying the dependency functions. By recursively substituting all analytical word forms and coordinated parts of sentence with x-words, we end up with a simple sentence structure that can be analysed by the well-known methods that are used in the case of the simplified dependency approach. For instance, the classical Covington’s algorithm [57] could be applied, which accepts word by word and tries to link it as a head or dependent to one of the nearest previous words (taking into account that each node, except the root, has exactly one parent). Covington’s cubic-time algorithm can be modified to parse x-words as well [67].

The primary constituents of the complex words have a fixed order, however dependents of these constituents may interleave in between in a free order. For instance, in Figure 9 the noun “*lekcijām*” that is the head of the prepositional phrase “*uz lekcijām*” has the attribute “*sintakses*”. Therefore x-words in the SemTi-Kamols grammar are defined by specifying only the primary constituents (and their order); the secondary (indirect) constituents are included by the application of dependency functions (see Figure 10). Note that constituents of an x-word and their potential dependents may be both regular words and x-words.

Word forms (the result of the morphological analysis):
w('Alfrēds', [noun, masc, sg, nom]).
w('un', [conj, and]).
w('Bernards', [noun, masc, sg, nom]).
w('uz', [prep, pre, sg, gen]).
w('uz', [prep, pre, pl, dat]).
w('sintakses', [noun, fem, sg, gen]).
w('sintakses', [noun, fem, pl, nom]).
w('sintakses', [noun, fem, pl, acc]).
w('lekcijām', [noun, fem, pl, dat]).
Patterns of x-words (production rules):
x([[_, [verb, aux, TENSE, MOOD]], [_, [verb, part, GEND, NUM, VOICE]]], ['xPredicate', [verb, main, TENSE, MOOD, GEND, NUM, VOICE]]) :- member(TENSE, [pres, fut]).
x([[_, [noun, _, _, CASE]], ['un', [conj, and]], [_, [noun, _, _, CASE]]], ['xCoordination', [noun, _, pl, CASE]]).
x([[_, [prep, pre, NUM, CASE]], [_, [noun, GEND, NUM, CASE]]], ['xPreposition', [noun, GEND, NUM, CASE, pre]]).
x([[_, [noun, GEND, NUM, CASE]], [_, [prep, post, NUM, CASE]]], ['xPreposition', [noun, GEND, NUM, CASE, post]]).
Syntactic functions (dependency rules):
f(subject, _, [noun, GEND, NUM, nom], [verb, main, _, MOOD, GEND, NUM, _]) :- member(MOOD, [ind, cond, rel]).
f(modifier_place, _, [noun, _, _, pre], [verb, main, _, _, _, _]).
f(modifier_cause, _, [noun, _, _, post], [verb, main, _, _, _, _]).
f(attribute, left, [noun, _, _, gen], [noun, _, _, _]).

Fig.10. A highly simplified fragment from the SemTi-Kamols Latvian grammar, specifying the sentence that is given in Figure 9. The morphological tagset³⁰ contains some features/values that are not directly related to morphology, but are useful for elimination of some potential ambiguities or incorrect parse trees (e.g., by indicating the permissible position of each preposition against the head constituent).

The SemTi-Kamols model can adequately represent various syntactic constructions, including analytical verb forms (e.g., the relative mood as in Figure 9 — “*esot ieradušies*”), prepositional constructions (“*uz lekcijām*”, „*intereses pēc*”) and coordination (“*Alfrēds un Bernards*”). Theoretically, the concept of x-word can be extended to cover also coordinated

³⁰ http://www.semti-kamols.lv/doc_upl/TagSet.pdf

and subordinated clauses: coordinated clauses could be treated similarly as coordinated parts of sentence, by specifying their roots (predicates) as the primary constituents of the x-word, whereas an x-word representing a subclause could be linked to the head node in the main clause by a dependency function. In parsing, however, such x-words would cause substantial ambiguity³¹. Therefore, in the formalization of Latvian grammar, only a syntactic subset has been considered — simple extended sentences — making a simplified assumption that semantically each clause can be considered separately and linked to its context at the discourse analysis level³². Moreover, as the development of a formally precise grammar that would extensively cover even simple extended sentences is a very complex and laborious task³³, in the scope of the underlying project, a partial parser (chunker) has been developed: if a full parse tree is not found, the longest possible subtrees are returned. Thus the grammar already can be used in partial parsing of unrestricted texts³⁴.

In order to avoid the construction of incorrect or hardly probable parse trees and, thus, to reduce the possibility for unjustified ambiguities (that usually cause a combinatorial explosion), the application of rules in the SemTi-Kamols grammar and parser is restricted by various local and global constraints (in addition to the unification of morphological features). It is specified, which x-word constituents can potentially act as heads (make a subtree) and which cannot (usually function words). For certain parts of sentence (in certain cases) a fixed position relatively to their heads can be specified (e.g., an attribute in Latvian is usually used to the left from the noun). And last but not least, it is restricted (in the parser's configuration), which parts of sentence are unique (e.g., the subject) and which are not (e.g., verb modifiers). However, the satisfaction of the latter constraint is a NP complete problem [67] that makes the current prototype implementation of the SemTi-Kamols parser impractical — an exhaustive search is performed to return all the possible parse trees³⁵. A practical solution is possible, for instance, by applying the dynamic programming technique (similarly to the Early algorithm) to acquire all the possible variants (without any uniqueness constraints), and by reducing the initial problem to the problem of efficient searching in the acquired chart [67]. Nevertheless, development of an efficient parsing algorithm is not the aim of this thesis; the aim is the model of analysis as such (a corpus annotation schema), avoiding any premature optimisation. Moreover, the grammar-based parser eventually will be replaced by a data-driven (corpus-based) statistical parser. In the case of the simplified dependency approach, statistical algorithms work in linear time [32], even by performing heuristic analysis of non-projective dependencies [55].

It should be noted that the current implementation of the SemTi-Kamols parser, similarly to most dependency parsers [58], does not support analysis of non-projective dependencies (the model itself does not impose such a restriction), although in Latvian, similarly to other inflective languages, this is not a rare phenomenon. However, the modern Latvian philology allows reducing most of such cases to the construction of a projective parse tree — by

³¹ For instance, it would be very hard to differentiate between coordinated predicates and coordinated clauses.

³² Theoretically, every coordinated or subordinated sentence can be transformed into one or (usually) more simple extended sentences by using, for instance, anaphoric pronouns instead of subordinating conjunctions (see Section 2.2). Regarding the coordinated sentences, Tesnière expresses a similar opinion [16].

³³ By periodically (for two years) extending the grammar with typical and occasional rules, about 300 x-word patterns and more than 100 dependency rules are defined and tested. Due to various technical optimisations, the initial set of rules (450/200) has been substantially reduced, meanwhile substantially extending the coverage.

³⁴ The SemTi-Kamols grammar has been used in the automatic morphological tagging of Latvian text corpora [21], keeping all ambiguous variants (that are less than if a morphological analyzer alone would be used).

³⁵ The automatic morphological tagging of the Latvian Web Corpus was performed in the BalticGRID [21].

relying on the concept of semi-predicative component³⁶ (SPK) [68]. It is believed that the non-projective dependent (SPK) addresses the whole sentence (or clause), i.e., both the direct head and the predicate. This is justified from the semantic point of view: as it has been emphasized several times, the syntactically free word order is pragmatically bound, i.e., it reflects a different meaning (or a shade of the meaning) of the sentence. However, by allowing for two edges, the principle that each dependent has a unique head would be violated (see Figure 11). Therefore, in the SemTi-Kamols grammar, only the relationship between the SPK and the predicate is specified, avoiding both non-projective trees and multiple heads. The fact that also the subject (or some other part of sentence — depending on the case) is indirectly modified can be inferred heuristically.

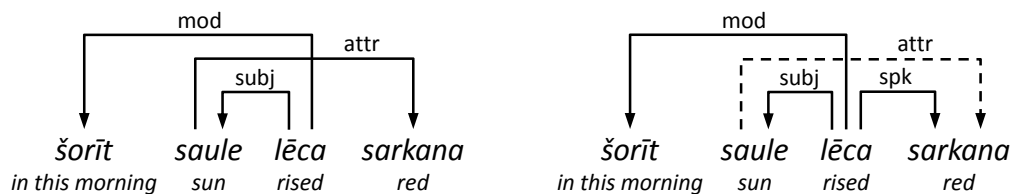


Fig.11. A sentence whose structure in a typical dependency grammar would be represented as a non-projective tree (on the left), in the SemTi-Kamols grammar is treated as a projective tree (on the right) by relying on the concept of semi-predicative component (its simplified interpretation).

The proposed model bears some similarity to the HPSG approach where features of a phrase are passed over via its head, i.e., the phrase is represented by its head. This is not the first time when the shortcomings of the simplified dependency approach have been pointed out, arguing that it does not allow for adequate representation of a number of essential syntactic phenomena that are present in both analytical and synthetic languages. One of the closest implementations is the constraint grammar by Järvinen and Tapanainen that is to some extent implemented in the Connexor parser. Their grammar is in between the simplified approach and Tesnière’s original theory: a descriptively adequate dependency model that is formally explicit and parsable (by using only local heuristic constraints) [66]. The difference of the SemTi-Kamols model is that it separates between simple and complex words (nuclei) and extends the notion of nucleus by including coordinated parts of sentence and other phenomena³⁷. Besides, judging by the current implementation of the Connexor parser³⁸, technically it is still based on a simplified type of dependency representation where not only junctions, but also elements of nuclei are encoded by predefined dependency links.

Another formalism that very closely reflects the basic concepts of Tesnière’s theory is Tesnière’s Dependency Structure (TDS) [69] that has been verified by converting the Penn Treebank [70] — one of the major English treebanks that is annotated according to a pure PSG grammar. But the TDS model can be developed further: in the original Tesnière’s approach, dependency types are simplified, and dependency links as such are actually anonymous — instead of syntactic functions Tesnière uses generalized parts of speech that are assigned to nuclei during the transference operations. Although it allows for a more robust analysis, the representation of a sentence structure in the TDS model is comparatively shallow. Therefore the SemTi-Kamols model can be seen as an extension also to the TDS

³⁶ On the condition that we are considering only simple extended sentences. In the case of a subordinate sentence, edges could cross the clause boundaries. Such a situation can occur, for instance, due to a relative clause (e.g., literally: “[...] for such language analysis that has free word order.”). Note that relative clauses often can be systematically converted as attributes, obtaining simple extended sentences (see Section 4.3).

³⁷ Järvinen and Tapanainen propose to link the elements of a junction by the predefined `cc`: relation [66].

³⁸ <http://www.connexor.eu/technology/machine/demo/syntax/>

approach. Finally, it should be noted that in the Prague Dependency Treebank (PDT), which is the leading dependency treebank, a multi-layer annotation scheme is used by adapting the Functional Generative Description theory [71]: at the deep or tectogrammatical layer that represents the linguistic meaning of a sentence (exploiting no background knowledge about the event), analytical word forms, appositions and other similar constructions are collapsed under common (single) nodes, whereas at the surface or analytical layer a simplified dependency grammar is used, introducing artificial dependencies between function and content words.

The SemTi-Kamols model that has been tested on a subset of Latvian can be adapted also for other inflective, synthetic languages; the grammatical structure of Baltic languages is very close to Slavic languages, for instance. It should be mentioned that currently (at IMCS UL) there is work in progress to approve the adequacy of the proposed approach in the case of unrestricted texts, by developing a pilot Latvian treebank³⁹.

2.2 Multidimensional Ontology

Assuming that the proposed hybrid grammar model ensures precise and adequate means to formalize Latvian grammar, the question is how to proceed with formalization of lexical and sentence (frame) semantics, and the discourse as a whole. Even if we had a large treebank, and on its basis — a high precision data-driven parser, the syntactic structure of a sentence follows from the semantic structure [16]. Therefore, by taking into account the initially stated principles, possible solutions on how to formally represent the various aspects of semantics and pragmatics were investigated in the rest of this research stage, so that the layers of text analysis would not be separated apart, but on the contrary — merged into a single model⁴⁰.

One of the most advanced approaches in frame semantics is FrameNet [74] that provides a deep insight into lexical semantics and its relationship to the syntactic structure of a sentence. FrameNet formalizes lexical semantics by linking words (their senses) to frames or idealized situations. Unlike OntoSem (see Section 1) that defines thousands and thousands of concepts, FrameNet describes less than 1000 frames. Although the set of frames is not complete and balanced (to a large extent, it has been acquired by analyzing only a small collection of domain-specific texts⁴¹), FrameNet's approach in its nature is flexible and universal. It has been approved, for instance, by its successful exploitation in the semantic annotation of the wide coverage SALSA/TIGER corpus (see Section 1). Promising results have been achieved also in FrameNet-based automatic (statistical) semantic role labelling in unrestricted texts [75]. The role of WordNet should be mentioned here as well — it has been exploited to extend the FrameNet's coverage, because the original lexicon that is manually aligned with FrameNet frames is very limited⁴² [76].

However, the development of FrameNet has been mainly focused on the linguistic aspects, thus the level of its formalization is insufficient. More rigorous ontological formalization would improve the usability of this significant resource in the tasks of automatic semantic analysis. Several investigations have been carried out to achieve this aim: FrameNet has been partially converted to OWL [77], and it has been aligned with the common-sense SUMO ontology [78]. This has been mainly done to enforce the strict ontological control over about 40 predefined semantic types that are intended for the restriction

³⁹ A SemTi-Kamols profile for the Prague Mark-up Language (PML) has been developed [72], allowing to exploit the existing PML-compatible tools like TrEd [73], which has been verified in the development of PDT.

⁴⁰ Note that at the level of semantic analysis, the choice of the model does not depend on the language type.

⁴¹ <http://goo.gl/S8Two>

⁴² This means that the development of Latvian WordNet for certain use-cases is still topical (see Section 1).

of the fillers of frame elements⁴³ (FE), and for the formalization of the frame inheritance structure. But the issue here is that these semantic types have been used inconsistently and occasionally — due to their insufficient and ambiguous specification. On the other hand, in the FrameNet formalization attempts, no special attention has been devoted to the formal part of FrameNet — the binding of lexical units (LU) to frames and frame elements (valences).

To preserve all the dimensions of analysis that are provided by FrameNet, as well as to allow for the possibility of introducing new dimensions (e.g., for discourse annotations), a non-standard approach has been proposed to combine all the levels and aspects of analysis in a single, universal model — multidimensional ontology — a finite set of points that are arranged in a multidimensional space⁴⁴ [6]. Such a model allows for capturing of all linguistic and world knowledge that can be encoded during the manual annotation of a text. Assuming that general and domain-specific texts will be annotated in parallel, the accumulated facts could eventually form the basis for extracting and learning correlations: both for training a statistical semantic parser, and for the development of a justified common-sense ontology.

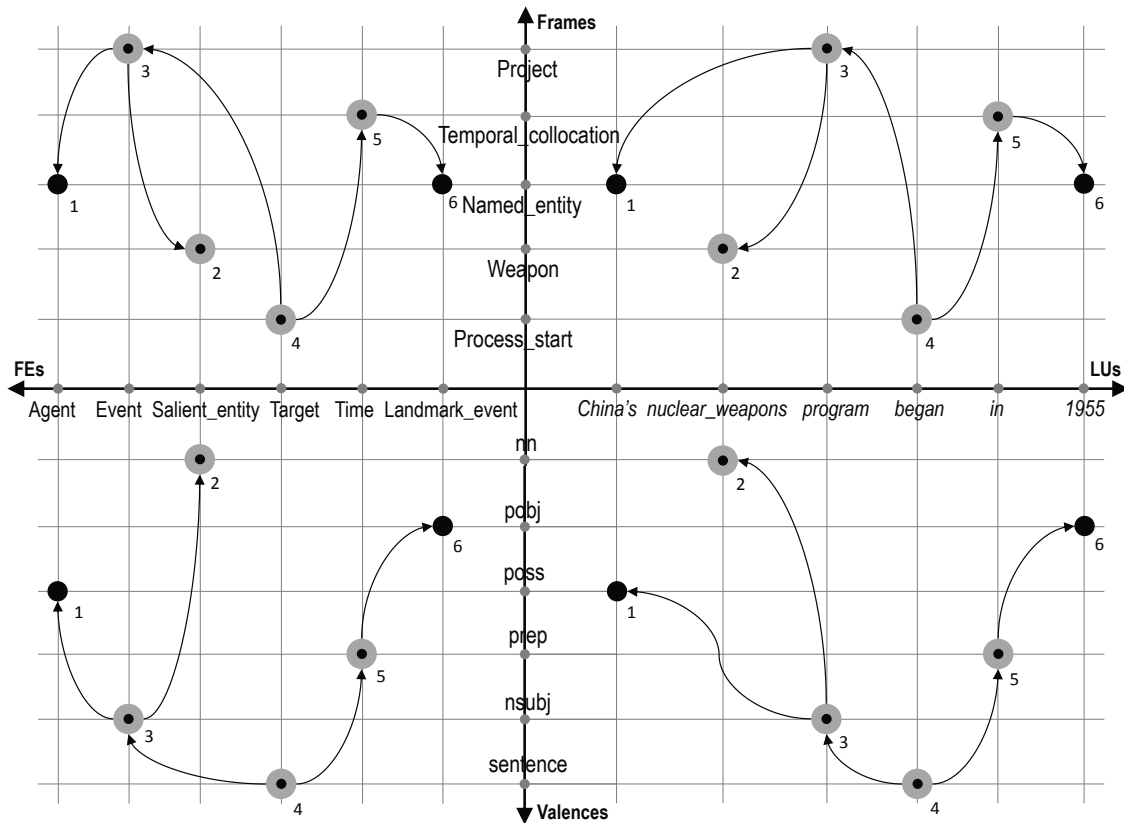


Fig.12. The four panes visualize various 2D aspects of the 4D annotation graph of the given sentence. The non-highlighted vertices represent named entities that are used for the filling of FEs, but do not invoke frames. The oriented edges are dependency relations.

The proposed approach is based on the idea to geometrically arrange a FrameNet-annotated text in the multidimensional space in such a way that annotations are located on the multidimensional coordinate axis, whereas tokens of a given sentence are located as anonymous points (instances) that are linked by anonymous dependency relations (see Figure 12). If we imagine that in the FrameNet’s four dimensional (4D) space there are

⁴³ The term “frame element” is also known as “slot” or “semantic role”.

⁴⁴ By introducing the concept of multidimensional ontology, we temporarily derogate from OWL and FOL in general as the knowledge representation language.

points and links placed that represent not only a single sentence, but a whole text corpus, the resulting filling would be complex and seemingly messy, but still forming a structure: the arrangement of points and links most likely would outline certain subspaces⁴⁵. Namely, the patterns of linguistic and world knowledge, obeyed and validated by a human-annotator, would be explicitly and implicitly exhibited geometrically: co-locations of the anonymous instances and their relations, and the “density” of such co-locations, from the ontological point of view, can be interpreted as classes and their properties.

However, 4D annotations, in general, are not sufficient to fully disambiguate the meaning of a sentence: if the discourse is not taken into account, several equally correct interpretations can be possible for an arbitrary sentence. Anaphora resolution is one of the primary tasks in discourse analysis. To ensure this, first, the current 4D space has to be extended by one more dimension that would allow for sequential enumeration of sentences and clauses that form a text. Second, to link the anaphors and their antecedents, a new type of relation has to be introduced (in addition to the dependency relation). As a result, sentences (and their clauses) would be linked by discourse referents and by co-references among them. Moreover, it would be pointed out, which discourse referents (points) have the same identity.

The arrangement (clusters) of instances in the multidimensional space, together with the co-reference annotations and the consequent statements like “*the same as*” (`owl:sameAs`), “*similar to*” (`similarTo`⁴⁶) and “*kind of*” (`rdf:type`) actually reflect the ontological knowledge that has been captured in a text corpus, and theoretically can be used as the basis to systematize and restrict the least formal part of FrameNet — semantic types of FEs. It should be mentioned that the proposed model is conceptually compatible also with the Corpus Pattern Analysis (CPA) approach that can be seen as a complement to FrameNet [80]. In CPA, valences are defined for each sense of a verb separately, instead of generalized frames; the permissible fillers are restricted by a more consistent ontology [81] (if compared to FrameNet’s semantic types).

The result of this part of the research is theoretical and has been tested in practice by manually annotating only some occasional sentences and paragraphs (both, in Latvian and in English). Even though in the scope of the SemTi-Kamols project, a prototype tool (a GUI) has been developed, considerably facilitating the process of the manual, multidimensional annotation, and the annotation model as such is conceptually simple, the experience shows that this process is highly laborious. Although it has been shown that, by exploiting the existing statistical syntactic and semantic parsers for English, a draft multidimensional annotation of a text can be acquired [82] that “only” has to be manually edited via the GUI tool, in the case of Latvian, we face the already mentioned chicken-and-egg problem — the lack of a data-driven, wide coverage statistical parser.

⁴⁵ Such a tendency was observed by annotating small fragments of occasional texts.

⁴⁶ The author agrees with Pat Hayes et.al. [79] that `owl:sameAs` semantics is often too strict (precise) to use it appropriately in practice.

3 Micro-ontologies as a Sense Inventory

Due to the lack of a syntactically and semantically annotated Latvian text corpus, the next two stages of this research [7–11] are devoted to the investigation of the capability of deterministic NL analysis, by narrowing the considered subset of Latvian not only syntactically, but also semantically, while preserving the practical usability in specific applications.

3.1 Logic-based Controlled Natural Language

There are several sophisticated controlled natural languages (CNL) that cover relatively large subsets of NL, for instance, ACE (Attempto Controlled English) [20], providing both expressivity and precise (automatic) interpretation [83]. CNLs can be divided into two general types according to the formalist or the naturalist approach⁴⁷ [84]. The formalist approach supports the deterministic, bidirectional mapping of CNL to some formal language like first-order logic (FOL) or its decidable description logic (DL) subset [85], allowing the integration with existing tools for reasoning, consistency checking and model building [7]. Although logic-based CNL is a seemingly informal, high-level means for knowledge representation, essentially it is still a formal language that is no more and no less expressive than the corresponding formalism, and whose interpretation is (intuitively) predictive. In the naturalist approach, possible ambiguities are decreased, but not excluded, thus allowing for a wider coverage of NL and more informal use of CNL, while preserving highly precise, still probabilistic interpretation: shallow parsing, context-sensitive heuristics and WordNet-based methods are applied for WSD, i.e., searching for the “best” interpretation is done. In this thesis the formalist approach has been chosen, and a subset of NL whose semantics is defined in the ontology language OWL.

To enable the unambiguous interpretation of logic-based CNL texts (i.e., to enable the construction of discourse representation structures (DRS) [86] and the translation of DRS into the chosen formal notation), basically two kinds of restrictions are imposed: a small set of interpretation rules for possibly ambiguous syntactic constructions, and a monosemous lexicon — content words are not interpreted, but are treated as predicate identifiers whose meaning is defined only by FOL formulas derived from the text being analyzed. Thus the sense of a word is implicitly defined in the CNL text itself. Additionally, to resolve the discourse referents, it is assumed that the antecedent of an anaphor is the most recent and the most specific noun phrase (NP) that is not in some closed scope [87]. Moreover, if the anaphor itself is a NP (in English: marked by the definite article), both NPs must be lexically equivalent; such constraint is often violated in NL [7].

While the syntactic restrictions “only” enforce the user to be precise (explicit) and consistent, and always (in all contexts) select the same “correct” parse tree, the lexical restrictions make a fundamental limitation, as the natural language lexicon is inherently polysemous. The consequence of the latter restriction is that logic-based CNLs are mainly used in domain-specific applications, e.g., in the development and verbalization of domain ontologies, assuming that in the scope of a domain the terminology is monosemous.

However, polysemy actually is the only natural means that would allow the integration of the rich multi-domain background knowledge in a CNL text. The root cause of polysemy is that there is a finite set of words in NL, but the set of concepts and contexts (relatively, domains) is potentially unlimited. Although new words (terms) are invented over time as well, this happens comparatively rarely and slowly — the new words have to be accepted

⁴⁷ Here we can draw a parallel with the division in rationalists and empiricists (see the introduction).

and used by the community, therefore reuse of existing words in different contexts is a common “workaround”. There are two main ways how words are reused: metaphorically⁴⁸ by relying on similarity (e.g., “*mouse*” for “*computer pointing device*”) and metonymically by relying on relationship (e.g. “*library*” for “*building of library*”) [89]. Fortunately, various senses of the same word typically fall into different domains [90]; explicit, possibly natural indication of the appropriate domains in a CNL text would enable support for polysemy, preserving the characteristics of deterministic CNL.

Continuing the investigation on the formal features of a word sense, a method for partitioning and disambiguation of polysemous class names (nouns and noun collocations) has been proposed in the thesis — with minimal and systematic use of lexical domain identifiers, while integrating and referring multi-domain ontologies [8]. In this way, controlled polysemy is enabled in the formalist approach, meanwhile preserving the possibility of deterministic analysis. The proposed solution can be extended to cover also polysemy of properties (verbs) by combining declarative (static) and procedural (dynamic) background knowledge [8], but the dynamic aspects are outside the scope of this thesis.

3.2 Micro-ontologies

A typical solution to differentiate polysemous (inconsistent) same-named classes is to introduce complex (often ad-hoc) lexemes by explicitly pointing out the specific meaning (e.g., “*library-building*” versus “*programming-library*”). However, consistent principles usually are not followed, and the dependency on such an ad-hoc lexicon makes the language non-systematic and, thus, non-user-friendly [91]. The method proposed in this thesis aims at eliminating (or at least reducing) this deficiency by introducing such multi-word units (MWU) systematically and largely automatically (while merging monosemous domain-specific ontologies), and only where it is necessary.

Internally monosemous and consistent domain ontologies that follow a linguistically motivated naming convention [96] will be called micro-ontologies. Instead of trying to make the lexicon globally monosemous, it is suggested to split the background knowledge into multiple, relatively small though lexically (terminologically) unambiguous micro-domains (see Figure 13). The benefit of this approach is its flexibility and scalability⁴⁹.

By substituting a common-sense or multi-domain ontology with numerous domain-specific micro-ontologies, the ontology merging problem is inevitably introduced. In this context, the problems of ontology merging and WSD are tightly intertwined, and the lack of definitive success in solving of any of them is largely due to addressing these issues separately⁵⁰. The author proposes to address them simultaneously by applying a semi-automatic method for word sense partitioning during the merging of OWL micro-ontologies. Namely, it is proposed to use the same semantically precise micro-ontologies as the sense inventory (dictionary), instead of external lexicons. It is expected that a logically reasonable sense partitioning will facilitate the (semi-)automatic WSD capabilities in CNL texts that describe facts (i.e., refer to the merged ontology).

Sentences that are accepted and generated by OWL-based CNLs can be divided into two groups: terminological (ontological) and assertional (factual) ones, corresponding to the

⁴⁸ “Language is a graveyard of dead metaphors” [88].

⁴⁹ The proposed concept of micro-ontology bears some similarity to the concepts of environment or viewpoint in [92] and to the Cyc micro-theories [93]. The difference of the micro-ontology approach is that it utilises the standard OWL DL reasoning abilities and is compatible with existing CNLs.

⁵⁰ Although there is a vast literature on these issues [94], majority of methods rely on shallow similarity measurements based on an external lexical taxonomy like WordNet [95].

description logic TBox and ABox axioms respectively [96]. Ontological statements define categories (classes) and the possible relationship (properties) between them (e.g., “*Every mouse_[zoology] is an animal.*”), whereas factual statements talk about instances that belong to specific categories (e.g., “*This mouse_[computer] is connected to the workstation.*”).

	Domain	Terminology
T-Box	General	<i>Every building is a physical_entity.</i> <i>Every collection is an abstract_entity.</i> <i>No physical_entity is an abstract_entity.</i>
	Building	<i>Everything that has a roof is a building.</i> <i>Every library is a building.</i> <i>Every green_roof is a roof.</i>
	Programming	<i>Everything that contains something is a collection.</i> <i>Every library is a collection.</i> <i>Every function is something.</i>
A-Box	Facts	
	<i>There is a library_[building] that has a green_roof.</i>	
	<i>AbsoluteValue is a function. [...] The library_[programming] contains AbsoluteValue.</i>	

Fig.13. The three micro-ontologies (verbalized in ACE) and the separately given facts illustrate the emergence of polysemy for the lexeme “*library*” (an automatic reasoner would discover a contradiction). To create a consistent merger, the appropriate sense (namespace) should be explicitly indicated in each utterance of “*library*”.

In contrast to the current practice, the two kinds of sentences in the proposed approach are strictly separated into ontological and factual texts. Ontological texts are, by definition, monosemous within the scope of a micro-ontology, thus the WSD problem becomes limited to the factual texts. The rationale for this separation is that an “average” CNL user is relieved from providing ontological knowledge, allowing him/her to concentrate on the factual content and querying, whereas the development and merging of micro-ontologies is left to domain experts and knowledge engineers — as is the common practice already.

3.3 Integration of Micro-ontologies and Word Sense Disambiguation

Note that in the proposed approach the WSD process is separated into two steps: first, during the merging of micro-ontologies, a minimal sense partitioning is acquired and recorded in the merger⁵¹, and, second, on the basis of the acquired partitioning, word senses are disambiguated during the parsing (in a wider sense) of factual texts that correspond to the merged micro-ontology. The same instrument is used to carry out both steps — an OWL reasoner that can automatically check the consistency of both the acquired sense partitioning (TBox) and the interpretation of a factual text (ABox).

The first approximation of the general algorithm for micro-ontology merging is very simple. First, if there are same-named classes among several micro-ontologies, we put forward the hypothesis that the concepts that are represented by these classes overlap and, thus, these classes are equivalent (`owl:equivalentClass`). Second, we verify the hypothesis: if the merged ontology is still consistent, we accept the hypothesis and the equivalence axiom; otherwise we reject the hypothesis and the equivalence axiom, and rename the inconsistent classes by including domain name of origin⁵² (see Figure 14).

⁵¹ Remind that in this thesis, in the case of micro-ontologies, the problem of polysemy is limited to class names (nouns); property names (verbs) are assumed to be globally monosemous among all micro-ontologies [8].

⁵² Making of MWUs instead of creating new words is a common technique in terminology.

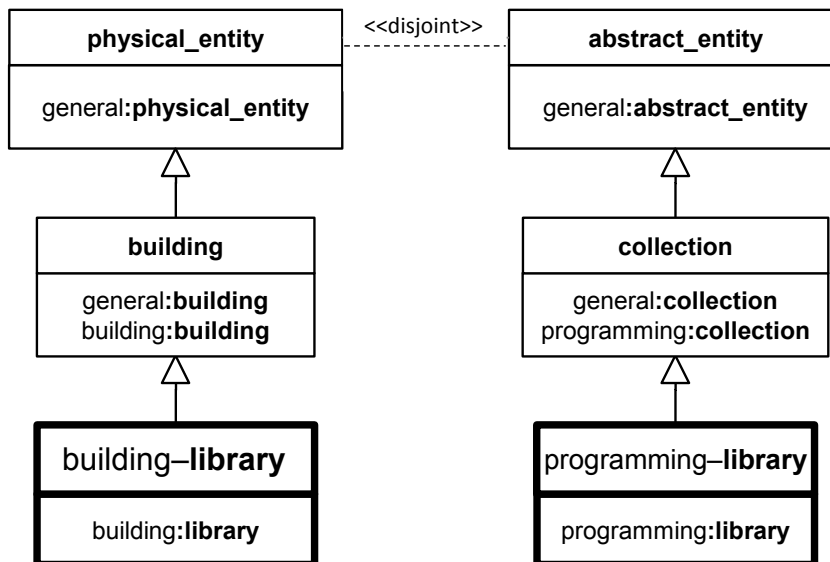


Fig.14. An ontology that has been acquired by merging micro-ontologies given in Figure 13. The sets of equivalent classes and their relationship are illustrated. Each set consists of one or more classes that belong to some micro-ontology (references to the original namespaces are preserved) and of a newly created class that belongs to the default namespace of the merged ontology. In the case of polysemy, the local prefix of the namespace of origin is included in the name of the newly created class.

In addition, two more aspects have to be taken into account:

- It is possible that each set of same-named classes can be split into subsets whose elements are (internally) mutually consistent, but the sets are mutually inconsistent.
- It is possible that the same name is used for classes, one of which actually is a subclass of the other (e.g., “moon” in the astronomy micro-ontology might denote a satellite of any planet, while in the agriculture micro-ontology it would more likely denote only the satellite of Earth).

Therefore the actual algorithm for the automatic merging of micro-ontologies, assuming that for the given set of micro-ontologies there is a unique “conceptually correct” sense partitioning, is as follows:

- Among all the given micro-ontologies:
 - (1) For each pair of same-named classes
 - in each of both directions
 - a) insert the `rdfs:subClassOf` relation and check the consistency of the current merger by exploiting an OWL reasoner;
 - b) if the current merger is consistent, keep the newly inserted relation; otherwise reject it.
 - (2) For each pair of same-named classes (X, Y)
 - insert `<X owl:equivalentClass Y>` if both `<X rdfs:subClassOf Y>` and `<Y rdfs:subClassOf X>` hold.

The proposed automatic procedure will return the single “correct” sense partitioning only if the following conditions are satisfied in advance:

- The given micro-ontologies contain sufficient restrictions to cause a contradiction, if the relation `rdfs:subClassOf` is inserted between any two same-named classes that actually do not overlap in the “conceptually correct” partitioning of senses.

- The given micro-ontologies contain sufficient restrictions to cause a contradiction, if for any two same-named classes, for which the axiom `<X rdfs:subClassOf Y>` holds in the “conceptually correct” partitioning of senses, but the opposite axiom does not hold, the axiom `<Y rdfs:subClassOf X>` is inserted.

Unfortunately, there is no automatic way to tell whether the produced merged ontology is conceptually correct or wrong — such verification is left to knowledge engineers (and domain experts), which means that the iterative fine-tuning of the selected micro-ontologies and the verification of their merger until the above mentioned conditions are met is, in general, a laborious task⁵³. Nevertheless, this semi-automated method is more systematic and flexible if compared to the manual merging of micro-ontologies or to the development of a universal (macro-) ontology from scratch. It is also formally more precise than the fully automatic but shallow methods for ontology matching and merging.

WSD, again, is possible by exploiting a reasoner, i.e., by merging the axioms derived from a text with the axioms of the merged ontology and by searching for such a set of senses (MWUs) for which the merger is still consistent. In general (due to the insufficient context), the proposed method ensures only partial WSD by excluding the contextually inconsistent variants. Difficulties might arise also in the choice of natural and intuitive namespace identifiers (prefixes), especially if subdomains of a wider domain are being merged [8]. Assuming that, on the average, the resulting sense partitioning will be rather coarse-grained, numeric sense identifiers could be used instead of the domain prefixes, similarly as it is done in WSD approaches that are based on traditional dictionaries or WordNet, but avoiding the problem of an overly fine-grained (unjustified) sense inventory (see Section 1). In any case, a user-friendly CNL editor is a predictive (look-ahead) editor [87, 97] that would help the users not only to create syntactically correct sentences, but also to choose the correct senses for the potentially ambiguous terms.

Although the proposed method has not been tested on real life examples, if compared to the “legacy” common-sense ontologies (like OntoSem), the micro-ontology approach theoretically offers several significant advantages: it is more scalable — reality does not have to be compressed into a restricted number of approximate categories; no single consistent model of the world is imposed, allowing distinct (mutually inconsistent), non-predefined points of view to co-exist. Micro-ontologies can be seen as snapshots, supporting the categorization of non-stable and temporal entities. Last but not least, the distinction between lexical and world knowledge is not needed anymore, only the distinction between ontological and factual texts.

The micro-ontology approach has been demonstrated on (controlled) English examples, however, the method itself is language-independent and can be directly applied also in the case of controlled Latvian.

⁵³ It might be necessary to introduce also a bridging micro-ontology that defines bridging axioms (restrictions) among concepts of different micro-ontologies [8].

4 Controlled Latvian for Ontology Authoring and Verbalization

In the last research stage [9–11], a subset of Latvian is considered for which the syntactic and semantic parsing is deterministic and, thus, can be done fully automatically. In contrast to the previous stage (see Section 3), the focus is on ontological (TBox) texts instead of factual ones, i.e., on the development (authoring) of micro-ontologies via a possibly natural, intuitive and unambiguous CNL, and on the representation (verbalization) of existing ontologies (both single and merged micro-ontologies) in a possibly natural and precise CNL (i.e., on the generation of ontological texts). As it was already mentioned, the development of (micro-) ontologies is not only a laborious, but also a knowledge-intensive process in which domain experts should be involved in addition to experienced knowledge engineers.

Several high-level notations are widely used to make the OWL ontologies more intelligible for both domain experts and knowledge engineers (if compared to the mathematical notation of DL or the various serialization formats for OWL). These notations can be divided in at least three groups: graphical languages, like UML and its profiles [23], controlled natural languages, like ACE [20], and “human-readable” formal syntaxes, like the Manchester OWL Syntax [98] that explicitly unveils the OWL semantics and therefore requires a substantial training to obtain acceptable reading and writing skills.

CNL provides the most informal and perhaps the most intuitive means for knowledge representation, i.e., no training is necessary for a human to be able to interpret (with high precision) an automatically generated CNL text. Controlled languages have been successfully used in practice — in domains where the involvement of domain experts is crucial, e.g., by facilitating the development of a large-scale ontology for Great Britain’s national mapping agency [99]. Whereas graphical notations provide an alternative view, unveiling the high-level structure of an ontology, and have been successfully used, for instance, in the reengineering of medical statistics registries in Latvia [100]. In fact, these perspectives are mutually complementary: a graphical language is not well suited for representation of restrictions (class expressions), property chains and inference rules that can be easily and comprehensibly done by using natural language.

In this thesis the focus is on untrained domain experts and ontology end-users, and, thus, on CNL as an interface to OWL that has to be as natural as possible. Moreover, the focus is on multilingual ontology authoring and verbalization, facilitating ontology localization and reuse (assuming that, at the conceptual level, knowledge as such is language-independent). This can be seen as a machine translation problem, but not in the traditional sense.

The well-known CNLs for OWL [85] are based on English — a highly analytic language with rigid word order, simple morphology and consistent use of articles that significantly facilitate the translation of CNL sentences into their semantic representation (DL axioms and rules). Besides, English is often used as a “meta-language” for naming the logical symbols (class and property names) at the ontology level.

4.1 Analysis of the Information Structure of a Sentence

It has been recently shown that the Grammatical Framework (GF), a formalism and a resource grammar library that provide means for developing parallel grammars (their parsers/generators), is a convenient framework for rapid and flexible implementation of multilingual CNLs [101]. This allows for easy reuse of existing tools that are developed for the English-based CNLs (e.g., for translation to/from OWL) also for those languages, which do not have all the necessary tools readily available. However, in the case of highly synthetic

languages⁵⁴ (like Slavic and Baltic), the bidirectional translation to English is not straightforward even for a subset of NL, especially if we are dealing not only with terminological (TBox) statements, but also with inference rules (SWRL [17]). In rules, anaphoric NPs are frequently used: in English they are marked by the definite article, while in Baltic and in most of the Slavic languages such markers are, in general, not explicitly used and are not encoded even in noun endings. Therefore one of the main problems during the semantically precise translation is how to distinguish between axioms and rules, and how to convey, which information is new (potential antecedents) and which is already given (anaphors).

Although in Latvian it could be theoretically possible to impose the mandatory use of artificial determiners, by using, for instance, indefinite and demonstrative pronouns, this would make the language unnatural. The lack of articles is even more apparent in Lithuanian (another Baltic language), which, in contrast to Latvian, has no historic influence from the comparatively analytical German. It does not mean that in Latvian (and in other synthetic languages) there are no formal features that help to grasp the information structure of a sentence; there are such features, but, in general, they appear in the surface structure (linearization or the linear order of word forms) implicitly.

It turns out that there is a connection between the given and new information and the word order that can be formulated as the topic and focus articulation (TFA) [102]. The topic (theme) is what we are talking about (the already known information), whereas the focus (rheme or comment) — what we are saying about it (the new information). Although the topic and focus parts of a sentence, in unrestricted Latvian, are not reflected by systematic (deterministic) changes in the word order [103], the author has put forward the hypothesis that, in the case of a controlled, logic-based Latvian, changes in the neutral word order can be exploited as reliable and intuitively satisfiable (by a native speaker) formal features [9] that compensate the “missing” articles. The basic principle of the simplified TFA method is as follows: all NPs that appear before the predicate belong to the topic part, whereas the predicate and all NPs that follow the predicate — to the focus part (see Figure 15). To illustrate the developed controlled Latvian, fragments from simplified university and wildlife ontologies will be alternately used (verbalized) in the remaining part of this section.

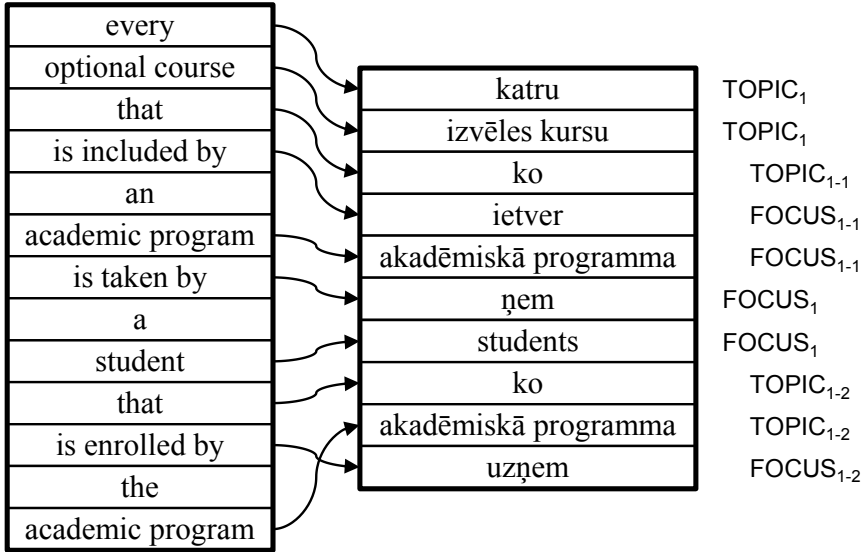


Fig.15. An inference rule, verbalized in English and in Latvian in parallel. The alignment of terms shows how the changes in word order affect the information structure of a sentence; in English, different syntactic constructions are used (e.g., the passive voice instead of the active voice).

⁵⁴ Rich morphology, syntactically free word order, definite/indefinite articles are not used.

To verify the hypothesis in practice, an initial GF grammar of the controlled Latvian was developed⁵⁵ by covering the main constructions in OWL (except cardinalities, literal values and other secondary constructions) and by supporting inference rules and basic data integrity queries⁵⁶. The parsing was fully based on the analysis of the information structure of a sentence, ensuring precise translation to ACE and partially also to OWL⁵⁷. Two additional GF grammars were developed for this purpose: one that covers the corresponding subset of ACE, and the other — the corresponding subset of OWL functional-style syntax [15]. By choosing a small set of representative samples (covering axioms and rules of various complexity), a survey was carried out, aiming at representatives of different domains (about 80 respondents participated) [10]. The results of the survey confirm⁵⁸ that the simple conditions on the word order are intuitively satisfied by a native speaker, and the TFA method is appropriate for the automatic resolution of given and new information (i.e., co-references) in controlled Latvian sentences. The survey confirmed other important aspects as well:

- due to the rich morphology, there are various alternatives possible in the syntactic realization of a sentence, while preserving the same meaning (e.g., a relative clause can be replaced by an attribute; such transformations can be applied to a limited extent also in English: “*animal that eats an animal*” → “*animal-eating animal*”);
- explicit determiners (“articles”) in certain cases are preferred, as they can improve the reading (interpretation) of a sentence: an indefinite pronoun (e.g., “*kāds*” ≈ “*a*”) usually is assigned to the object, if it is not restricted by a relative clause (or an attribute), whereas a demonstrative pronoun (e.g., “*šis*” ≈ “*the*”) helps in complex rule statements (in addition to the word order);
- sentences in the plural are often more concise and more natural;
- limitations of the OWL expressivity (SVO triples only, no time dimension etc.) to some extent can be seemingly lessened on the surface level of the CNL (while preserving the possibility for the deterministic bidirectional translation), for instance, by using (where appropriate) adverbial modifiers of place instead of direct objects, nouns (roles) instead of verbs (actions), and the present perfect tense instead of the simple tense.

Therefore, in addition to a grammar that generates the best possible (default) linearization (taking into account the information structure), a parallel grammar that accepts the various syntactic alternatives and the optional use of determiners is necessary (and is developed).

4.2 Extensions of the SVO Sentence Pattern

Theoretically, all predicates in OWL ontologies syntactically conform to the SVO pattern (generalization axioms are a special case). In practice, however, it can be very hard or even impossible to come up with an appropriate verb for a property, or to use a syntactic object (in the accusative case), so that the statement would remain (semantically) natural.

In the first case, individuals of two classes, in natural language, most likely can be associated at least by a role. The leading English-based CNLs (incl. ACE) support different

⁵⁵ The GF grammar of controlled Latvian (incl. its extensions that are described in the following subsections) follows the PSG approach and can be equally well used for both parsing and generation.

⁵⁶ The developed grammars and an interactive demonstration are available on-line: <http://valoda.ailab.lv/cnl/>

⁵⁷ GF does not support dealing with anaphoric references, although theoretically it could be possible [104]. Anaphors (incl. explicit variables) may appear not only in rules, but also in property axioms (see Figure 21). Therefore the translation to OWL in GF can be provided only for class axioms (regardless of the complexity of class expressions). An alternative framework that is specifically designed for the development of CNL grammars, supporting anaphora resolution, is Codeco [87], but it does not support multilingual grammars.

⁵⁸ <http://goo.gl/3CdDj> (Results of the survey on Lithuanian are available here: <http://goo.gl/2q8Gu>)

ways to define and refer to such properties (see the next subsection), but in controlled Latvian they can be expressed in a uniform way (incl. the inverse use of a property): by making a NP that consists of a class name (term) in the genitive (possessive) case followed by its role name whose case depends on the context. For instance, “*ikviens kurss ir kādas akadēmiskās programmas daļa*” (ACE: “*every course is a part of an academic program*”) versus “*ikvienas akadēmiskās programmas daļa ir kāds kurss*” (ACE: “*for every academic program its part is a course*”).

In the second case, typical pseudo-objects in ontological statements are adverbial modifiers of place (expressed by NPs), for instance, “*every course is included in an academic program*” versus “*every course is included by an academic program*”. It should be noted that, in the current implementation, only such modifiers are considered that in Latvian are expressed by the locative case. In the future, modifiers of place that are expressed by a prepositional construction should be supported as well, but it is unlikely that support for other types of verb modifiers would be needed in practice. On the other hand, the type of a modifier syntactically makes (almost) no difference, moreover, the use of certain types of indirect objects could be enabled as well (e.g., those answering the question *with what?*) — prepositions themselves are not interpreted; they are appended as suffixes to property names (see Section 4.4). However, it should be taken into account that, in an inflective language, the syntactic valence of a preposition can be ambiguous (e.g., “*uz_{on} galda_{GEN}*” versus “*uz_{to} lekciju_{ACC}*”). This issue has not been investigated in more detail in this thesis, but a potential solution would be to include the valence information (restriction) in the domain lexicon: in a sentence, valence of a preposition depends on the valence of a verb. In natural language, the verb valence, of course, can be ambiguous, depending on the verb sense, however, in the scope of a fixed domain, such a solution could be appropriate. In the direction from OWL to CNL, the syntactically appropriate translation equivalent (for a preposition in English) would facilitate the resolving of ambiguity (“*on a table*” → GEN, “*to a lecture*” → ACC).

4.3 Syntactic Paraphrasing

Taking into account the results of the survey and the explicitly and implicitly provided suggestions, the initial GF grammar of controlled Latvian has been significantly improved [10, 11]. Exploiting the GF support for parallel grammars, a mechanism for syntactic paraphrasing of input sentences has been provided, ensuring that the user-provided sentences (or the axioms of an existing ontology) are automatically rewritten in a possibly natural and, meanwhile, precise form, preserving the possibility for the deterministic analysis. It is important that both the original information structure and the abstract syntax tree are preserved during the paraphrasing, i.e., no semantic changes are introduced.

```

abstract Ontology ..
fun
  RestrictedClass : C → R → C_Bar ;
  Restriction : Lnk → Pred → E_Bar → R ;
  Link : Lnk ;
  ExistentialQuantification : EC_Bar → E_Bar ;
  ExistentialClass : Ex → C_Bar → EC_Bar ;
  Some : Ex ;
  Class : C → C_Bar ;

```

Fig.16. A fragment from the abstract grammar that defines functions for constructing the parse trees (independently from the language and the domain). The abstract grammar is implemented by the parallel grammars of controlled Latvian, as well as by the grammars that cover the corresponding subsets of ACE and OWL, specifying how the parse trees are linearized.

Let us consider four grammatically different, but semantically equivalent NPs:

- (1) „*kādā akadēmiskajā programmā iekļautais kurss*”;
- (2) „*kurss, kas ir iekļauts kādā akadēmiskajā programmā*”;
- (3) „*kurss, kas iekļauts akadēmiskajā programmā*”;
- (4) „*kurss, kas ietilpst akadēmiskajā programmā*”.

Each of the given NPs corresponds to the category *C_Bar* in the developed grammar (see Figure 16). By applying the grammatical rules (see Figure 17 and 18) and the lexical rules (see Figure 19), a single, common parse tree is acquired for all of them:

```
RestrictedClass Course (Restriction Link includedIn (Existential-
Quantification (ExistentialClass Some (Class AcademicProgram)))
```

```
concrete OntologyLavVar of Ontology = open OntologyLav ..
lin
  RestrictedClass c r = {
    str = case r.type.type of {
      RelCl ⇒ c.str ++ "," ++ r.str ;           -- Relative clause
      Attr ⇒ r.str ++ c.str                    -- Attribute
    } ;
    restr = {restr = Restr ; type = r.type.type}
  } ;
  Restriction lnk pred e_bar = case e_bar.restr.restr of {
    Restr ⇒ mkRelCl lnk pred e_bar ;
    NonRestr ⇒ mkRelCl lnk pred e_bar | mkAttr lnk pred e_bar ;
  } ;
  ExistentialQuantification ec_bar = {
    str = ec_bar.str ;
    restr = ec_bar.restr ;
  } ;
  ExistentialClass ex c_bar = {
    str = case c_bar.restr.restr of {
      Restr ⇒ mkExClass (Impl | Expl) ex c_bar ;
      NonRestr ⇒ mkExClass (Expl | Impl) ex c_bar
    } ;
    restr = nbar.restr
  } ;
  Class c = {
    str = c.str ;
    restr = {restr = NonRestr ; type = NoType}
  } ;
```

Fig.17. A highly simplified fragment from the domain-independent controlled Latvian grammar: excluding the inflection paradigms, syntactic agreement etc., but including the possible alternatives.

```
resource OntologyLav ..
oper
  mkRelCl : Pronoun → Verb → Noun → RelClause = λpron,verb,noun → {
    str = pron.str ! Expl ++ verb.str ++ noun.str ;
    type = {restr = noun.restr.restr ; type = RelCl}
  } ;
  mkAttr : Pronoun → Verb → Noun → RelClause = λpron,verb,noun → {
    str = pron.str ! Impl ++ noun.str ++ verb.adj ;
    type = {restr = noun.restr.restr ; type = Attr}
  } ;
  mkExClass : Usage → Determiner → Noun → Str = λuse,det,noun →
    det.str ! use ++ noun.str ;
```

Fig.18. A simplified fragment (without the syntactic agreement) from the controlled Latvian resource grammar: common functions that can be applied in any of the parallel controlled Latvian grammars.

Domain-specific, but language-independent concepts

C — classes, Pred — properties (predicates), NPred — nominalized properties (predicate nominatives)

abstract University = Ontology ** ..

fun

AcademicProgram, Course, MandatoryCourse, Student : C ;
enrolls, enrolledIn, includes, includedIn, takes : Pred ;
partOf : NPred ;

Domain-specific Latvian lexicon

concrete UniversityLavVar of University = OntologyLavVar ** **open** ResLav ..
lin

AcademicProgram = **mkMWU** (**mkAdjective** "akadēmisks") (**mkNoun** "programma") ;
Course = **mkNoun** "kurss" ;
MandatoryCourse = **mkMWU** (**mkAdjective** "obligāts") (**mkNoun** "kurss") ;
Student = **mkNoun** "students" ;
enrolls = (**mkParticiple** "uzņēmis") | (**mkVerb** "uzņem" Object) ;
enrolledIn = **mkParticiple** "uzņemts" ;
includes = **mkVerb** ("iekļauj" | "ietver") Object ;
includedIn = (**mkParticiple** "iekļauts") | (**mkVerb** "ietilpst" Place) ;
takes = **mkVerb** ("apgūst" | "ņem") Object ;
partOf = **mkNoun** "daļa" ;

Domain-specific English lexicon

concrete UniversityEngVar of University = OntologyEngDef ** **open** ResEng ..
lin

AcademicProgram = **mkMWU** (**mkAdjective** "academic") (**mkNoun** "program") ;
Course = **mkNoun** "course" ;
MandatoryCourse = **mkMWU** (**mkAdjective** "mandatory") (**mkNoun** "course") ;
Student = **mkNoun** "student" ;
enrolls = (**mkParticiple** "enrolled" Object) | (**mkVerb** "enroll" Object) ;
enrolledIn = (**mkParticiple** "enrolled" In Place) |
 (**mkVerb** "enroll" In Place) ;
includes = **mkVerb** "include" Object ;
includedIn = **mkParticiple** "included" In Place ;
takes = **mkVerb** "take" "taken" Object ;
partOf = **mkNoun** "part" ;

Fig.19. A fragment from a domain-specific, multilingual lexicon: content words and MWUs with the possible variants (both grammatical and lexical). The alignment of the translation equivalents is ensured by the abstract lexicon. Functions that define the inflectional paradigms (e.g., **mkNoun**) are defined in separate, application-independent grammar modules (**ResLav** and **ResEng**), similarly as in Figure 18.

The second of the four NPs provided above is the default linearization (according to the domain-independent grammar and the domain-specific lexicon) that most precisely represents the meaning⁵⁹ without losing the naturalness. This NP will be generated both while automatically paraphrasing any of the other NPs and while verbalizing an existing ontology. The default linearization can be partially changed by the user: in the grammar and the lexicon, given in Figure 17 and 19, all the possible alternatives are included, out of which only the first one is kept in the default grammar (and lexicon)⁶⁰. Thus the choice of the most appropriate synonym (e.g., for the property **takes** in the Latvian lexicon), as well as the use of the simple/perfect tense (e.g., in the case of the property **enrolls**) is in the control of the user (domain expert). Note that a verb that makes a non-SVO sentence (see the previous subsection) cannot be paraphrased by using another verb (or another form of the same verb) that attracts the direct object, so that the meaning would not be changed (e.g., the properties **includes** and **includedIn** are different — mutually inverse) or the sentence would not become ungrammatical (see the next subsection).

⁵⁹ .. Course(x) ∧ ∃y(AcademicProgram(y) ∧ includedIn(x,y))..

⁶⁰ Thus, the default grammar and lexicon can be derived automatically.

All kinds of paraphrasing that are provided by the parallel grammars (and domain lexicons) of controlled Latvian can be differentiated in the following groups:

- Lexical paraphrasing, allowing the arbitrary use of synonyms for denoting the same concept. This applies also to function words (e.g., by using the universal quantifier “*katrs*” instead of “*ikviens*”) that themselves can be ambiguous (e.g., “*ikviens*” can be interpreted either as “*every*” or as “*everything*”), but in a sentence they are syntactically disambiguated. Note that the lexical paraphrasing could be potentially applied also in the semi-automatic WSD process described in Section 3 by enumerating the ambiguous class names as secondary alternatives for the corresponding monosemous MWUs. For instance, by indicating the word “*library*” as an alternative for both concepts: `programming-library` (in addition to the MWU “*programming library*”) and `building-library` (in addition to “*library building*”). In this case, the use of the polysemous word “*library*” will result in two parse trees and, respectively, in two paraphrases from which the user himself can choose the correct one.
 - In cases when the universal class (`owl:Thing`) is (implicitly) referenced, it is syntactically impossible to differentiate whether the subject/object is an animate or inanimate thing⁶¹, and which grammatical gender (masculine or feminine) should be used. Such information could be partially encoded in the lexicon (by specifying the verb valence), however, both cases can be possible for some properties, and this information is not available in (existing) ontologies. Therefore the most neutral alternatives are preferred: inanimate thing and masculine gender. The latter one is, in fact, a grammatically correct choice, but the assumption of inanimate thing (like in ACE) is a trade-off. However, in Latvian it is often possible to use the demonstrative pronoun “*tas*” — if it is followed by a relative clause — making the sentence both grammatically and semantically correct (see sentences 1, 5 and 6 in Figure 22).
- The use of indefinite and demonstrative pronouns (“articles”) is completely optional; in paraphrases, the indefinite pronoun is used if the NP is not restricted by a relative clause. The demonstrative pronoun is always used in paraphrases.
- In certain cases, it is allowed that the word order does not correspond to the information structure (in terms of the simplified TFA assumption), and is corrected in the paraphrase (e.g., “[*..*], *ko apgūst šīs_{the} students*” → “[*..*], *ko šīs_{the} students apgūst*”). In certain other cases, the “correct” information structure is modified for the sake of naturalness (e.g., “[*..*] *ir iekļauts kaut kur_{somewhere}*” → “[*..*] *ir kaut kur_{somewhere} iekļauts*”).
- Both the present simple and the present perfect tense are accepted (to express a past event that has present consequences). These alternatives are defined by the user, and the paraphrase depends on the user-specified order in the domain lexicon.
 - In the case of the present perfect tense, the auxiliary verb “*ir*” may be omitted.
- In certain cases, syntactically substantially different, but semantically equivalent constructions may be used on the surface structure (e.g., by substituting an attribute with a relative clause, as it was shown in Section 4.1).
- Sentences in the plural are paraphrased to singular sentences, or vice versa (depending on the configuration). Moreover, in plural TBox statements, it is natural to avoid explicit quantifiers (e.g., “*lions_{FEM} hunt giraffes_{FEM}*”⁶² → “*every lion hunts a giraffe*”).

⁶¹ For instance, “*ikviens*” (“*everyone*”) versus “*jebkas*” (“*everything*”).

⁶² In the plural, if the subject and the object are in the feminine gender, only the neutral (SVO) interpretation is possible [105]. This is due to the fact that noun endings in the nominative and the accusative case are equal.

4.4 Two-level Translation to/from OWL

The possible steps that are performed while translating controlled Latvian sentences to OWL (and vice versa) are illustrated in Figure 20. *LavDefSg* is a grammar that defines the default verbalization patterns, using singular sentences, *LavDefPl* is its counterpart for plural sentences⁶³, and *LavVar* is an extended combination of both default grammars, extensively allowing for free variations (see the previous subsection). *LavVar* is used for robust, still deterministic parsing of the input sentences⁶⁴, while one of the default grammars (depending on the choice of the end-user) — for paraphrasing *LavVar* sentences and for verbalizing existing ontologies. Finally, *AceOwl* implements a subset of ACE (ACE-OWL) that is being used as an interlingua, ensuring that the translation between ACE-OWL and OWL (incl. SWRL) can be done by exploiting the existing ACE tools [20, 106]. In addition, a prototype of controlled English has been developed (*EngVar* and *EngDef*) that is based on full ACE⁶⁵ with some improvements:

- the support for the present perfect tense has been extended (e.g., by allowing sentences like *“every academic program has enrolled a student”*);
- to provide an alternative way for expressing inverse nominalised properties, a pattern from the Sydney OWL Syntax (SOS) [107] has been taken — it is used also in *EngDef* (in main clauses, e.g., by using *“every branch has a leaf as a part”* instead of *“for every branch its part is a leaf”* or *“every branch has-part a leaf”*);
- for compliance with the Latvian counterpart, the English grammar has to accept and generate constructions where a property whose “object” is a modifier of place is inversely referred (e.g., *“every savannah is a place where a lion lives in”*).

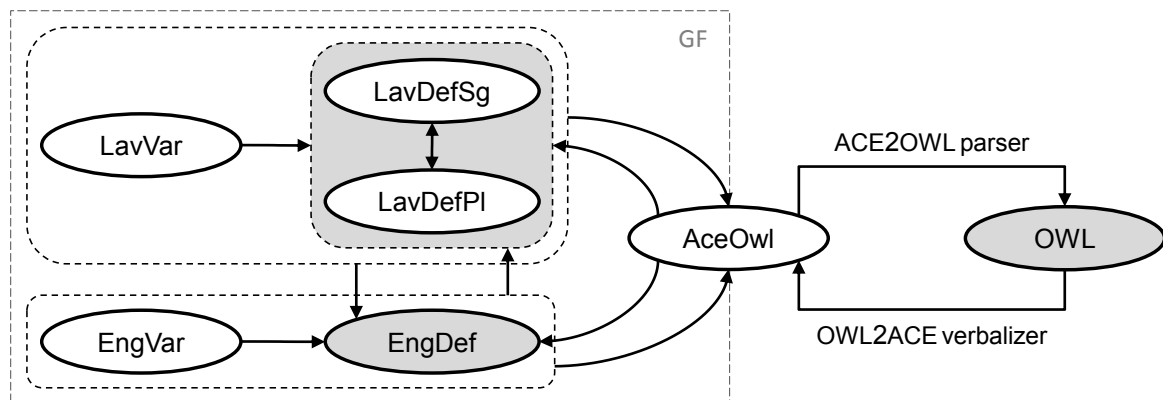


Fig.20. The overall data flow of the automatic translation process among controlled Latvian, English, and OWL, using ACE-OWL as an interlingua. Existing tools are exploited for the transition to/from OWL; all the other transitions are ensured by the parallel GF grammars.

Although the development of controlled English is not the aim of this thesis, the parallel translation to/from English, apart from the multilingual aspect, can be seen as a supplemental means for paraphrasing. Nevertheless, in the prototype, the English grammar is defined from scratch, however from the grammar engineering point of view, the universal GF resource grammar library should be reused [108].

⁶³ The grammar *LavDefPl* (and the corresponding parts of *LavVar*) is currently under development.

⁶⁴ In rare cases syntactic ambiguity is still possible (e.g., if there are three or more sequential, coordinated relative clauses). To resolve it, a small set of simple interpretation rules have to be applied (similarly as it is done by the ACE parser; see <http://goo.gl/6mmGL>).

⁶⁵ The full ACE supports prepositional phrases, adjectives a.o. constructions that are not allowed in ACE-OWL.

Any other CNL that is designed for the authoring and verbalization of OWL ontologies could be used as an interlingua instead of ACE. The well-known ACE has been chosen because of its easily available (open source) tools and web services. Moreover, the verbalization of existing ontologies could be done directly from OWL, involving no interlingua at all (using fixed patterns for verbalizing property axioms), however, by implementing the CNL to OWL direction, the reverse direction is provided by GF “for free”.

To illustrate the results and the intermediate results that are acquired in the proposed two-level translation approach, let us consider an ontology that is visualized in Figure 21. Note that this example also demonstrates that UML and CNL are mutually complementary notations: the complex restrictions on classes `herbivore` and `carnivore` are depicted in the formal Manchester syntax — CNL statements could be used instead.

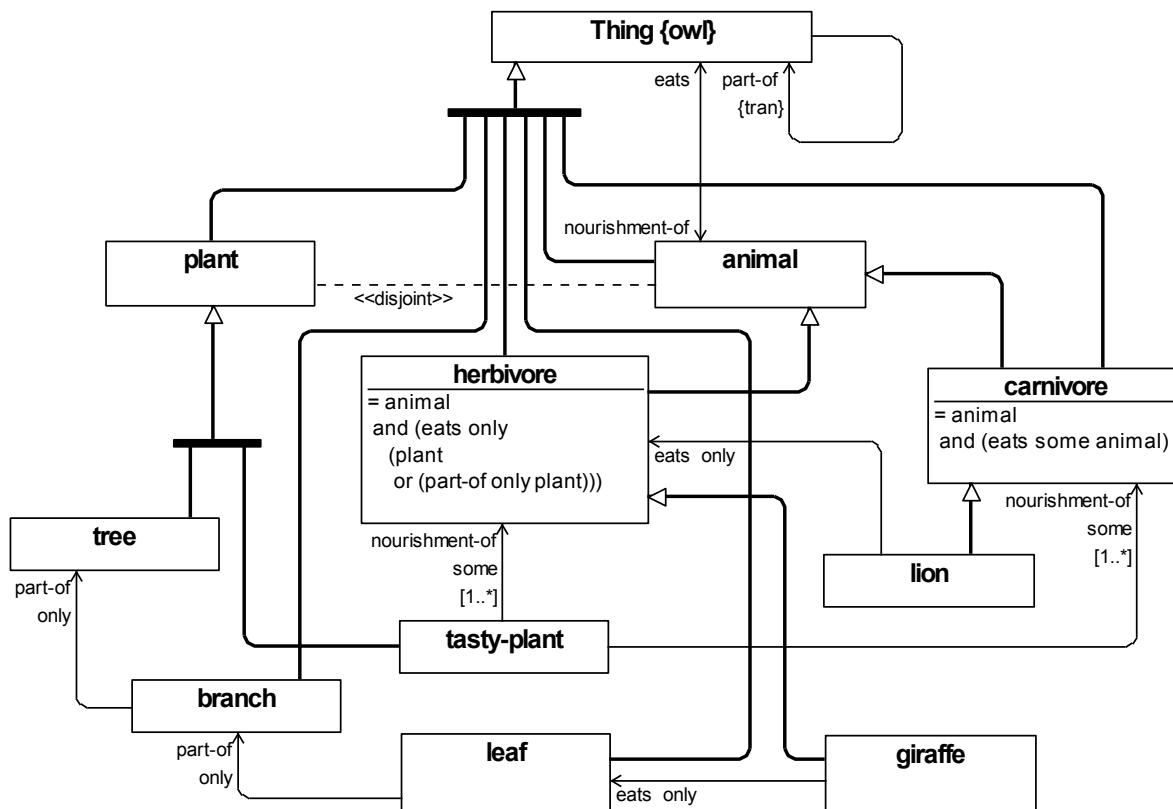


Fig.21. A simplified wildlife ontology [109], visualized in a UML-style graphical notation [23]. The original version has been slightly modified (in a linguistically motivated way [96]) by renaming the property `eaten-by` to `nourishment-of`, i.e., by using a role name instead of an inconsistent verb phrase.

The given ontology is (partially) verbalized in Figure 22. This has been ensured (in parallel) by the `LavDefSg` and `EngDef` grammars. Apart from other things, this example unveils also several aspects for which the verbalization in Latvian should be improved:

- for TBox axioms, plural sentences mostly would be more natural and more concise (e.g., “*tas, ko ēd žirafes, ir lapas*” instead of “*tas, ko ēd kāda žirafe, ir lapa*”);
- the indefinite pronoun “*kaut kas*” should be paraphrased by using the demonstrative pronoun “*tas*” (see the 3rd statement), if it is preceded by the restriction “*tikai*” (“*only*”, i.e., `owl:allValuesFrom`); in other cases it might cause ambiguity;
- in certain cases (if a relative clause follows), analytical forms should be used in the verbalization of genitive constructions (roles), e.g., “*daļa no kaut kā*” instead of “*kaut kā daļa*” (see the 10th statement).

1	Tas, kas kaut ko <i>ēd</i> , ir <i>dzīvnieks</i> .	Everything that <i>eats</i> something is an animal .
2	Ikviens <i>plēsējs</i> ir <i>dzīvnieks</i> , kas <i>ēd kādu dzīvnieku</i> . Ikviens <i>dzīvnieks</i> , kas <i>ēd kādu dzīvnieku</i> , ir <i>plēsējs</i> .	Every <i>carnivore</i> is an animal that <i>eats an animal</i> . Every <i>animal</i> that <i>eats an animal</i> is a <i>carnivore</i> .
3	Ikviens <i>zālēdājs</i> ir <i>dzīvnieks</i> , kas <i>ēd tikai kaut ko</i> , kas ir <i>augš</i> vai kas ir tikai <i>auga daļa</i> .	Every <i>herbivore</i> is an animal that <i>eats</i> nothing but things that are a plant or that are a part of nothing but <i>plants</i> .
4	Ikviens <i>žirafe</i> ir <i>zālēdājs</i> .	Every <i>giraffe</i> is a herbivore .
5	Tas, ko <i>ēd kāda žirafe</i> , ir <i>lapa</i> .	Everything that is <i>eaten</i> by a giraffe is a leaf .
6	Tas, kura <i>daļa</i> ir <i>kāda lapa</i> , ir <i>zars</i> .	Everything that has a leaf as a part is a branch .
7	Ikviens <i>garšīgs augš</i> ir <i>kāda plēsēja barība</i> .	Every <i>tasty plant</i> is a nourishment of a carnivore .
8	Neviens <i>dzīvnieks</i> nav <i>augš</i> .	No <i>animal</i> is a plant .
9	Ja X-s <i>ēd</i> Y-u, tad Y-s ir X-a <i>barība</i> .	If X <i>eats</i> Y then Y is a nourishment of X.
10	Ja X-s ir kaut kā, kas ir Y-a <i>daļa</i> , <i>daļa</i> , tad X-s ir Y-a <i>daļa</i> .	If X is a part of something that is a part of Y then X is a part of Y.

Fig.22. A (partial) verbalization of the ontology, given in Figure 21, in controlled Latvian (LavDefSg) and English (EngDef). The domain-specific terms (classes and properties) are given in italic; underlined are properties that are expressed by nouns (roles), but are interpreted as (binary) predicates.

It should be explained that the necessary and sufficient conditions for class equivalence, i.e., definitions in ACE are stated in two separate statements (see the 2nd paragraph; in the 3rd paragraph the reverse direction is omitted). In other CNLs (e.g., in SOS), definitions are verbalized explicitly (“..defined as..”), but then the user has to understand the meaning. Whereas, if variables are used, explicit endings have to be attached to them in Latvian in order to recognize (unambiguously) the subject and the object.

1	Everything that v:eats something is an n:animal.	ObjectProperty: eats Domain: animal
2	Every n:carnivore is an n:animal that v:eats an n:animal. Every n:animal that v:eats an n:animal is a n:carnivore.	Class: carnivore EquivalentTo: animal and (eats some animal)
3	Every n:herbivore is an n:animal that v:eats nothing but things that are a n:plant or that v:part-of nothing but n:plant.	Class: herbivore EquivalentTo: animal and (eats only (plant or (part-of only plant)))
4	Every n:giraffe is a n:herbivore.	Class: giraffe SubClassOf: herbivore
5	Everything that is v:eats by a n:giraffe is a n:leaf.	Class: inverse (eats) some giraffe SubClassOf leaf
6	Everything that is v:part-of by a n:leaf is a n:branch.	Class: inverse (part-of) some leaf SubClassOf branch
7	Every n:tasty-plant v:nourishment-of a n:carnivore.	Class: tasty-plant SubClassOf: nourishment-of some carnivore
8	No n:animal is a n:plant.	Class: animal DisjointWith: plant
9	If X v:eats Y then Y v:nourishment-of X.	ObjectProperty: eats InverseOf: nourishment-of
10	If X v:part-of something that v:part-of Y then X v:part-of Y.	ObjectProperty: part-of Characteristics: Transitive

Fig.23. An automatically generated ACE-OWL text, translated from Figure 22 (by the AceOwl grammar) or verbalized from the original ontology (by the ACE verbalizer). For the sake of clarity, terms are marked by POS prefixes that are accepted by the ACE parser, but are actually not used. The semantic interpretation is acquired by the ACE parser and is given in parallel (in the Manchester notation).

The translation (reduction) from controlled Latvian or English to ACE-OWL is an internal step of the translation process. During its execution, all non-SVO statements are converted to artificial SVO statements (e.g., “*part of something*” → “*part-of something*”), and all terms are normalized into fixed forms that are conveyed as is to the ontology⁶⁶ (see Figure 23). The intermediate result is purely technical and, from the natural language point of view, ungrammatical⁶⁷, but it is a good illustration, which explicitly unveils the nature and limitations of OWL. Note that some of the sentences in the interlingua are still grammatical (see the 1st, 2nd, 4th and 8th sentence), and some others could be made grammatical (e.g., by using the past participle form in the 5th sentence), but it is not necessary to do so — the end-user normally is not aware of this intermediate step.

The same steps are performed for interpreting inference rules⁶⁸ and can be potentially performed for interpreting data integrity queries [110] (see Figure 24). It should be mentioned that the ACE parser (currently) does not provide the translation to SPARQL, however, it constructs a DRS for an interrogative sentence, thus the conversion from DRS to SPARQL can be implemented independently (similarly as it has been done in [111]).

	SWRL	SPARQL
LavDef	Ikvienu obligāto kursu, kas <u>ir iekļauts kādā</u> akadēmiskajā programmā, <u>apgūst</u> ikviens students, ko <u>šī</u> akadēmiskā programma <u>ir uzņēmusi</u> .	Vai ir <u>kāds</u> students, kas <u>apgūst</u> kursu, kas <u>nav iekļauts</u> akadēmiskajā programmā, kas <u>šo</u> studentu <u>ir uzņēmusi</u> ?
EngDef	Every mandatory course that <u>is included in an</u> academic program <u>is taken</u> by every student that <u>is enrolled</u> by <u>the</u> academic program.	Is there <u>a</u> student that <u>takes a</u> course that <u>is not included in an</u> academic program that <u>has enrolled the</u> student?
AceOwl	Every n:mandatory-course that <u>v:included-in</u> an n:academic-program is <u>v:takes</u> by every n:student that is <u>v:has-enrolled</u> by the n:academic-program.	Is there a n:student that <u>v:takes</u> a n:course that does not <u>v:included-in</u> an n:academic-program that <u>v:has-enrolled</u> the n:student?
Interpretation	mandatory-course(?x1), academic-program(?x2), student(?x3), included-in(?x1,?x2), has-enrolled(?x2,?x3) -> takes(?x3,?x1)	ASK WHERE { ?x1 rdf:type student. ?x1 takes ?x2. ?x2 rdf:type course. ?x3 rdf:type academic-program. ?x3 has-enrolled ?x1. NOT EXISTS {?x2 included-in ?x3}}

Fig.24. Translation of an inference rule and an integrity query in the corresponding formal notation. ACE parser provides translation of rules (and axioms) in OWL/FSS and OWL/XML format; in the example, Manchester syntax is used for the sake of clarity. Translation to SPARQL is not supported.

It should be emphasized once more that the task of logic-based CNL is to ensure the deterministic interpretation of its sentences, so that the user could easily and precisely predict and grasp the meaning of a specification (ontological text) that is being written or read, and

⁶⁶ The normalization of word forms is provided by a domain-specific AceOwl lexicon that can be automatically derived from the EngDef lexicon (see Figure 19).

⁶⁷ From the ACE parser point of view, these sentences are still grammatical. This is achieved by passing an auto-generated user lexicon (to the ACE parser), where all word forms of each term are equivalent, including the one that is used for the logical symbol (see <http://goo.gl/qQq7E>).

⁶⁸ The user himself does not have to differentiate between axioms and rules; the user may even not be aware of the fact that some sentences are interpreted as rules.

so that roundtrips between CNL and OWL would not introduce any semantic changes (unless the user himself makes some changes). Therefore, for instance, the well-known, sophisticated CPL (Computer Processable Language) [84] is not appropriate in this case, because its interpretation is not deterministic (although the context-sensitive lexical and syntactic disambiguation is ensured with a very high precision). Besides, the experience with CPL unveils an interesting fact: to inform the user of what is the interpretation of a given CPL text, it is being paraphrased into the simpler, deterministic CPL-Lite language; it turns out that users themselves gradually switch to the deterministic subset of CNL [84].

One more aspect should be mentioned here: before generating an ACE-OWL text, the ACE verbalizer performs some refactoring of the ontology. As a result, many axioms are rewritten in a more general form (preserving the same semantics), for instance, the subject of a predicate may become a complex class (class expression) instead of an atomic class as it was in the original version. This allows to verbalize an ontology in a more natural English, but the translation back to OWL leads to a structurally modified ontology. This is not convenient if, for instance, a UML-style notation or the Manchester syntax is used in parallel (e.g., in the Protégé editor). However, this should not be seen as a drawback of the CNL interface — similar issues are caused also by using the graphical languages, therefore each of the OWL notations (interfaces) should carry out an appropriate refactoring (normalization).

The proposed two-level translation approach has allowed to develop a rather complex Latvian grammar on top of the comparatively limited ACE-OWL (in terms of naturalness). The controlled Latvian has been tested on real examples by defining and verbalizing complex axioms, inference rules and integrity queries. In addition, it has been demonstrated that an existing controlled English grammar can be independently and flexibly modified and extended in this way as well. Of course, ACE-OWL itself could be developed to be equally natural, but the benefit of the proposed approach is that (i) it allows for more flexible and independent extensions and adjustments, (ii) constructions of different CNLs can be mixed or used in parallel, and (iii) the interlingua can be relatively easily changed.

It is interesting that a recent empirical research on the basis of an OWL ontology “corpus” shows that the full potential of OWL is not used in practice: by analyzing more than 600 000 axioms from 200 ontologies, it turned out that 90% of axioms are *SubClassOf*, *DisjointClasses* and *EquivalentClasses* axioms, but 8% are ABox axioms [112]. Whereas 69% of those TBox axioms have both the subject and the object as an atomic class, but in 25% of cases — an atomic subject and a simple restriction on the object: *ObjectSomeValuesFrom (ObjectProperty, Class)*. The authors of this research conclude that CNL, thus, can be used as an adequate notation for OWL. The author of this thesis, however, would like to extend this conclusion: the statistics not only confirm the adequacy of CNL for ontology verbalization, but perhaps also indicates that the involvement of domain experts in the ontology development is insufficient, and that the formal notations are difficult to use for a non-experienced knowledge engineer. This raises the hope that the (parallel) use of CNL as a natural and flexible, still formal knowledge representation language, involving more domain experts, will eventually make changes in the statistics of the OWL corpus.

Conclusion

In overall, the aims of this thesis are accomplished and, summarizing the main results, they are as follow:

- An original dependency-based grammar model has been developed, which is primarily aimed at languages with relatively free word order and provides both flexible and detailed analysis and representation of the syntactic structure of a sentence. The proposed model has been tested in practice by formalizing a relatively wide subset of Latvian grammar, covering various syntactic constructions that frequently appear in simple extended sentences.
- A concept of OWL micro-ontology has been defined, and a methodology for the consistent development and merging of such ontologies has been proposed. It has been shown that by solving the merging problem we simultaneously solve also the word sense partitioning problem, acquiring a formal and semantically precise (clear) sense inventory. It has been also shown that the merged micro-ontologies facilitate semi-automatic WSD in factual, logic-based CNL texts.
- A logic-based, deterministic Latvian (its parser/generator) has been implemented in GF, providing the syntactically and semantically precise parsing of complex axioms and rules that are naturally expressed in the controlled Latvian, and vice versa — the verbalization of existing OWL ontologies in a precise, intuitive subset of Latvian. For translating controlled Latvian sentences to/from OWL, a novel, two-level translation approach has been proposed and experimentally implemented, allowing to use an existing CNL as an interlingua (incl. its readily available infrastructure) and providing the end-user with a possibly natural, flexible interface. The results of this research stage are the major achievement of this thesis.

Meanwhile, there are several directions for future developments — both for the near future and in the long term. The tasks for the near future are related to the development of the controlled Latvian by supporting plural constructions, patterns for specifying cardinalities etc. It would be interesting to investigate the possibility for adapting the deterministic, TFA-based method also for analysis of ABox statements. To make the further development of controlled Latvian more flexible and more rapid (also for other use-cases), modules of the Latvian resource grammar have to be implemented and included in the GF resource grammar library. The long term tasks are related to the development of a Latvian treebank and a data-driven parser.

The results of this research can be adapted also for other inflective, synthetic languages, like Slavic (especially regarding the hybrid model of syntactic analysis and the information structure based method for the analysis of the given and new information in a logic-based controlled language). Some of the proposed methods are language-independent, namely, the micro-ontology approach for the formalization of lexical semantics. In fact, the hybrid grammar model itself is language-independent as well, and theoretically could be attractive also for analytic languages.

Although analysis of unrestricted natural language most likely is impossible without exploiting corpus-based (statistical) methods, it is important that the annotated language resources that are used for training the statistical systems represent the structure of the language as precisely and deeply as possible. The approach that has been followed in this thesis, by significantly restricting the considered subset of NL and by investigating it in-depth, has made some contribution to the study and formalization of the structure of Latvian.

Bibliography

List of included publications

1. Bārzdiņš G., Grūzītis N., Kudiņš R., Nešpore G., Spektors A. *Latviešu valoda semantiskajā tīmeklī* (*Latvian Language in the Semantic Web*). Proceedings of the Latvian Academy of Sciences, Section A, Vol. 60, No. 6, 2006, pp. 26–42
2. Grūzītis N., Nešpore G., Saulīte B. *Hierarhisku attieksmju izgūšana no latviešu valodas skaidrojošās vārdnīcas* (*Extracting Hierarchical Relations from the Dictionary of Latvian*). Proceedings of the 11th International Conference “The Word: Aspects of Research”, 2007, pp. 147–159
3. Bārzdiņš G., Grūzītis N., Kudiņš R. *Re-engineering OntoSem Ontology Towards OWL DL Compliance*. Proceedings of the 7th Joint Conference on Knowledge-Based Software Engineering, Frontiers in Artificial Intelligence and Applications, Vol. 140, IOS Press, 2006, pp. 157–166
4. Bārzdiņš G., Grūzītis N., Nešpore G., Saulīte B. *Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order*. Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA), 2007, pp. 13–20
5. Nešpore G., Saulīte B., Bārzdiņš G., Grūzītis N. *Comparison of the SemTi-Kamols and Tesnière’s Dependency Grammars*. Proceedings of the 4th International Conference on Human Language Technologies — the Baltic Perspective, Frontiers in Artificial Intelligence and Applications, Vol. 219, IOS Press, 2010, pp. 233–240
6. Bārzdiņš G., Grūzītis N., Nešpore G., Saulīte B., Auziņa I., Levāne-Petrova K. *Multi-dimensional Ontologies: Integration of Frame Semantics and Ontological Semantics*. Proceedings of the 13th EURALEX International Congress, 2008, pp. 277–283
7. Bārzdiņš G., Grūzītis N., Nešpore G., Saulīte B., Auziņa I., Levāne-Petrova K. *Ontological Word Sense Disambiguation for Discourse Representation*. Proceedings of the 3rd Baltic Conference on Human Language Technologies, 2007, pp. 33–40
8. Grūzītis N., Bārzdiņš G. *Polysemy in Controlled Natural Language Texts*. Revised papers of the Workshop on Controlled Natural Language (CNL 2009), LNCS/LNAI, Vol. 5972, Springer, 2010, pp. 102–120
9. Grūzītis N. *Word Order Based Analysis of Given and New Information in Controlled Synthetic Languages*. Proceedings of the 1st Workshop on the Multilingual Semantic Web (at WWW 2010), CEUR, Vol. 571, 2010, pp. 29–34
10. Grūzītis N., Nešpore G., Saulīte B. *Verbalizing Ontologies in Controlled Baltic Languages*. Proceedings of the 4th International Conference on Human Language Technologies — the Baltic Perspective, Frontiers in Artificial Intelligence and Applications, Vol. 219, IOS Press, 2010, pp. 187–194
11. Grūzītis N., Bārzdiņš G. *Towards a More Natural Multilingual Controlled Language Interface to OWL*. Proceedings of the 9th International Conference on Computational Semantics (IWCS), 2011 (accepted for publication)

List of supplemental publications (not included in the thesis)

12. Milčonoka E., Grūzītis N., Spektors A. *Natural Language Processing at the Institute of Mathematics and Computer Science: Ten Years Later*. Proceedings of the 1st Baltic Conference on Human Language Technologies, 2004, pp. 6–11

13. Viļķelis M., Grundspenķis J., Grūzītis N. *Natural Language Based Concept Map Building*. Proceedings of the 6th International Conference on e-Learning and the Knowledge Society, 2010, pp. 179–184
14. Skadiņa I., Auziņa I., Grūzītis N., Levāne-Petrova K., Nešpore G., Skadiņš R., Vasiļjevs A. *Language Resources and Technology for the Humanities in Latvia*. Proceedings of the 4th International Conference on Human Language Technologies — the Baltic Perspective, Frontiers in Artificial Intelligence and Applications, Vol. 219, IOS Press, 2010, pp. 15–22

References

15. *OWL 2 Web Ontology Language*. W3C Recommendation, 2009 (<http://www.w3.org/TR/owl2-primer/>)
16. Tesnière L. *Éléments de syntaxe structurale*. Klincksieck, Paris, 1959 (Tulkojums krievu valodā: Теһнер Л. *Основы структурного синтаксиса*. Ред. В.Г. Гак. Москва, Прогресс, 1988)
17. *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. W3C Member Submission, 2004 (<http://www.w3.org/Submission/SWRL/>)
18. *SPARQL Query Language for RDF*. W3C Recommendation, 2008 (<http://www.w3.org/TR/rdf-sparql-query/>)
19. Ranta A. *Grammatical Framework: A Type-Theoretical Grammar Formalism*. Journal of Functional Programming, Vol. 14, No. 2, 2004, pp. 145–189
20. Fuchs N.E., Kaljurand K., Kuhn T. *Attempto Controlled English for Knowledge Representation*. Proceedings of the 4th International Reasoning Web Summer School, LNCS, Vol. 5224, Springer, 2008, pp. 104–124
21. Džeriņš J., Džonsons K. *Harvesting National Language Text Corpora from the Web*. Proceedings of the 3rd Baltic Conference on Human Language Technologies, 2007, pp. 87–94
22. Skadiņa I., Brālītis E. *English-Latvian SMT: knowledge or data?* Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA), NEALT, Vol. 4, 2009, pp. 242–245
23. Bārzdīņš J., Bārzdīņš G., Čerāns K., Liepiņš R., Sproģis A. *OWLGrEd: a UML style graphical notation and editor for OWL 2*. Proceedings of the 7th International Workshop on OWL: Experiences and Directions (OWLED), CEUR, Vol. 614, 2010
24. Atkins B.T.S., Rundell M. *The Oxford Guide to Practical Lexicography*. Oxford University Press, 2008
25. Zipf G.K. *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley, 1949
26. Kehoe A., Gee M. *New corpora from the Web: making Web text more 'text-like'*. Pahta P., Taavitsainen I., Nevalainen T., Tyrkkö J. (Eds.) *Towards Multimedia in Corpus Studies, Studies in Variation, Contacts and Change in English*, Vol. 2, 2007 (http://www.helsinki.fi/varieng/journal/volumes/02/kehoe_gee/)
27. Brants T. *Inter-Annotator Agreement for a German Newspaper Corpus*. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC), 2000, pp. 107–112
28. Chklovski T., Mihalcea R. *Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation*. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), 2003

29. Brown P.F., Cocke J., Della Pietra S.A., Della Pietra V.J., Jelinek F., Lafferty J.D., Mercer R.L., Roossin P.S. *A statistical approach to machine translation*. Computational Linguistics, Vol. 16, No. 2, 1990, pp. 79–85
30. Callison-Burch C., Koehn P., Monz C., Schroeder J. *Findings of the 2009 Workshop on Statistical Machine Translation*. Proceedings of the 4th EACL Workshop on Statistical Machine Translation, 2009, pp. 1–28
31. Kilgarriff A. *Ontologies and terminology and how they relate to lexicography*. Course notes of the 7th Workshop in Lexicography and Lexical Computing (Lexicom), 2007
32. Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., Marsi E. *MaltParser: A language-independent system for data-driven dependency parsing*. Natural Language Engineering, Vol. 13, No. 2, 2007, pp. 95–135
33. Eisele A., Federmann C., Uszkoreit H., Saint-Amand H., Kay M., Jellinghaus M., Hunsicker S., Herrmann T., Chen Yu. *Hybrid Machine Translation Architectures within and beyond the EuroMatrix project*. Proceedings of the 12th Annual Conference of the European Association for Machine Translation, 2008, pp. 27–34
34. Burchardt A., Erk K., Frank A., Kowalski A., Padó S., Pinkal M. *The SALSA Corpus: a German Corpus Resource for Lexical Semantics*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006
35. Erk K., Padó S. *Shalmaneser — a flexible toolbox for semantic role assignment*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006
36. Fellbaum C. (Ed.) *WordNet: an Electronic Lexical Database*. MIT Press, 1998
37. Budanitsky A., Hirst G. *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*. Computational Linguistics, Vol. 32, 2006, pp. 13–47
38. Snyder B., Palmer M. *The English all-words task*. Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SensEval), 2004, pp. 41–43
39. Kilgarriff A. *How Dominant is the Commonest Sense of a Word?* Proceedings of the 7th International Conference on Text, Speech and Dialogue (TSD), LNCS, Vol. 3206, Springer, 2004, pp. 103–112
40. Chodorow M.S., Byrd R.J., Heidorn G.E. *Extracting Semantic Hierarchies from a Large On-Line Dictionary*. Proceedings of the 23rd Annual Conference of the Association for Computational Linguistics, 1985, pp. 299–304
41. Ide N., Véronis J. *Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation*. Proceedings of the 3rd International EAMT Workshop on Machine Translation and the Lexicon, 1993, pp. 19–34
42. Hearst M.A. *Automated Discovery of WordNet Relations*. Fellbaum C. (Ed.), *WordNet: an Electronic Lexical Database*. The MIT Press, 1998, pp. 131–152
43. Montemagni S. *Architecture and Functioning of a System for the Acquisition of Taxonomical Information from Dictionary Definitions*. Proceedings of the 4th International Conference on Computational Lexicography, 1996, pp. 173–182
44. Dyvik H. *Translations as a semantic knowledge source*. Proceedings of the 2nd Baltic Conference on Human Language Technologies, 2005, pp. 27–38
45. Piasecki M., Szpakowicz S., Broda B. *A Wordnet from the Ground Up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2009
46. Poesio M. *Domain modelling and NLP: Formal ontologies? Lexica? Or a bit of both?* Applied Ontology, Vol. 1, No. 1, 2005, pp. 27–33

47. Nirenburg S., Raskin V. *Ontological Semantics*. Cambridge: The MIT Press, 2004
48. Niles I., Pease A. *Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology*. Proceedings of the IEEE International Conference on Information and Knowledge Engineering, 2003, pp. 412-416
49. Nirenburg S., McShane M., Zabłudowski M., Beale S., Pfeifer C. *Ontological Semantic text processing in the biomedical domain*. Working Paper #03-05, Institute for Language and Information Technologies, UMBC, 2005
50. McShane M., Beale S., Nirenburg S. *Some Meaning Procedures of Ontological Semantics*. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), 2004
51. Kilgarriff A. *I don't believe in word senses*. Computers and the Humanities, Vol. 31, No. 2, 1997, pp. 91–113
52. Navigli R., Litkowski K.C., Hargraves O. *SemEval-2007 Task 07: Coarse-Grained English All-Words Task*. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval), 2007, pp. 30–35
53. Chomsky N. *Three models for the description of language*. IRE Transactions on Information Theory, Vol. 2, No. 3, 1956, pp. 113–124
54. Gaifman H. *Dependency Systems and Phrase-Structure Systems*. Information and Control, Vol. 8, No. 3, 1965, pp. 304–337
55. Nivre J. *Constraints on Non-Projective Dependency Parsing*. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2006, pp. 73–80
56. Earley J. *An Efficient Context-Free Parsing Algorithm*. Communications of the ACM, Vol. 13, 1970, pp. 94–102
57. Covington M.A. *A Fundamental Algorithm for Dependency Parsing*. Proceedings of the 39th Annual ACM Southeast Conference, 2001, pp. 95–102
58. Nivre J. *Dependency Grammar and Dependency Parsing*. MSI report 05133. Växjö University: School of Mathematics and Systems Engineering, 2005
59. Johansson R., Nugues P. *Extended Constituent-to-Dependency Conversion for English*. Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA), 2007, pp. 105–112
60. Stevenson M., Greenwood M. *Comparing Information Extraction Pattern Models*. Proceedings of the Workshop on Information Extraction Beyond the Document (at COLING/ACL), 2006, pp. 12–19
61. Johansson R., Nugues P. *LTH: Semantic Structure Extraction using Nonprojective Dependency Trees*. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval), 2007, pp. 227–230
62. Johansson R., Nugues P. *Comparing Dependency and Constituent Syntax for Frame-semantic Analysis*. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), 2008
63. Eisner J.M. *Bilexical grammars and their cubic-time parsing algorithms*. Bunt H., Nijholt A. (Eds.), *Advances in Probabilistic and Other Parsing Technologies*, Kluwer, 2000, pp. 29–62
64. Pollard C., Sag I.A. *Head-driven phrase structure grammar*. University of Chicago Press, 1994

65. Brants S., Hansen S. *Developments in the TIGER Annotation Scheme and their Realization in the Corpus*. Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC), 2002, pp. 1643–1649
66. Järvinen T., Tapanainen P. *Towards an implementable dependency grammar*. Proceedings of the Workshop on Processing of Dependency-Based Grammars, 1998, pp. 1–10
67. Podiņš K. *Ātrs atkarību gramatiku sintaktiskais analizators (An Efficient Dependency Parser)*. Maģistra darbs (Master thesis). Latvijas Universitātes Datorikas fakultāte, 2008
68. Lokmane I. *Sekundāri predikatīvs komponents kā teikuma loceklis (Semi-predicative Component as Part of Sentence)*. Baltu filoloģija (Baltic Philology), Vol. 11, No. 1, Riga: UL, 2002, pp. 47–64
69. Sangati F., Mazza C. *An English Dependency Treebank à la Tesnière*. Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories, 2009, pp. 173–184
70. Marcus M.P., Santorini B., Marcinkiewicz M.A. *Building a large annotated corpus of English: The Penn treebank*. Computational Linguistics, Vol. 19, No. 2, 1993, pp. 313–330
71. Sgall P., Hajičová E., Panevová J. *The Meaning of the Sentence in Its Pragmatic Aspects*. Reidel, 1986
72. Pretkalniņa L., *Integrēta sistēma sintaktiski anotēta latviešu valodas tekstu korpusa izveidei (An Integrated System for the Development of Latvian Treebank)*. Maģistra darbs (Master thesis). Latvijas Universitātes Datorikas fakultāte, 2011 (in progress)
73. Petr P., Stepanek J. *Recent Advances in a Feature-Rich Framework for Treebank Annotation*. Proceedings of the 22nd International Conference on Computational Linguistics (COLING), 2008, pp. 673–680
74. Fillmore C.J., Johnson C.R., Petruck M.R.L. *Background to FrameNet*. International Journal of Lexicography, Vol. 16, 2003, pp. 235–250
75. Baker C., Ellsworth M., Erk K. *SemEval'07 Task 19: Frame Semantic Structure Extraction*. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval), 2007, pp. 99–104
76. Tonelli S., Pianta E. *A novel approach to mapping FrameNet lexical units to WordNet synsets*. Proceedings of the 8th International Conference on Computational Semantics (IWCS), 2009, pp. 342–345
77. Scheffczyk J., Baker C.F., Narayanan S. *Ontology-based reasoning over lexical resources by means of ontologies*. Proceedings of the Workshop on Ontologies and Lexical Resources (OntoLex), 2006, pp. 1–8
78. Scheffczyk J., Pease A., Ellsworth M. *Linking FrameNet to the Suggested Upper Merged Ontology*. Proceedings of the Fourth International Conference on Formal Ontology in Information Systems (FOIS), Frontiers in Artificial Intelligence and Applications, Vol. 150, IOS Press, 2006, pp. 289–300
79. Halpin H., Hayes P.J. *When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web*. Proceedings of the WWW2010 Workshop on Linked Data on the Web (LDOW), CEUR, Vol. 628, 2010
80. Hanks P., Pustejovsky J. *A Pattern Dictionary for Natural Language Processing*. Revue Française de linguistique appliquée, Vol. 10, No. 2, 2005
81. Pustejovsky J., Havasi C., Littman J., Rumshisky A., Verhagen M. *Towards a Generative Lexical Resource: The Brandeis Semantic Ontology*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006

82. Meņšikovs K. *Teksta nozīmes vizualizēšana (Visualization of the Text Meaning)*. Maģistra darbs (Master thesis). Latvijas Universitātes Datorikas fakultāte, 2009
83. Wyner A., Angelov K., Barzdins G., Damljanovic D., Davis B., Fuchs N., Hoefler S., Jones K., Kaljurand K., Kuhn T., et al. *On Controlled Natural Languages: Properties and Prospects*. Revised papers of the Workshop on Controlled Natural Language (CNL 2009), LNAI, Vol. 5972, Springer, 2010, pp. 281–289
84. Clark P., Murray W.R., Harrison P., Thompson J. *Naturalness vs. Predictability: A Key Debate in Controlled Languages*. Revised papers of the Workshop on Controlled Natural Language (CNL 2009), LNAI, Vol. 5972, Springer, 2010, pp. 65–81
85. Schwitter R., Kaljurand K., Cregan A., Dolbear C., Hart G. *A Comparison of three Controlled Natural Languages for OWL 1.1*. Proceedings of the 4th International Workshop on OWL Experiences and Directions (OWLED), CEUR, Vol. 496, 2008
86. Blackburn P., Bos J. *Working with Discourse Representation Structures*. Representation and Inference for Natural Language: A First Course in Computational Linguistics, Vol. 2, 1999
87. Kuhn T. *Codeco: A Grammar Notation for Controlled Natural Language in Predictive Editors*. Pre-Proceedings of the 2nd Workshop on Controlled Natural Languages (CNL 2010), CEUR, Vol. 622, 2010
88. Leary D.E. (Ed.) *Metaphors in the history of psychology*. Cambridge: Cambridge University Press, 1994
89. Ravin Y., Leacock C. *Polysemy*. Oxford: Oxford University Press, 2000
90. Magnini B., Strapparava C., Pezzulo G., Gliozzo A. *The role of domain information in word sense disambiguation*. Natural Language Engineering, Vol. 8, No. 4, 2002, pp. 359–373
91. Rinaldi F., Dowdall J., Hess M., Molla D., Schwitter R., Kaljurand K. *Knowledge-Based Question Answering*. Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES), LNCS/LNAI, Vol. 2773, Springer, 2003, pp. 785–792
92. Wilks Y., Barnden J., Wang J. *Your metaphor or mine: belief ascription and metaphor interpretation*. Proceedings of the 12th International Joint Conference on Artificial Intelligence, 1991, pp. 945–950
93. Lenat D. *Cyc: A Large-Scale Investment in Knowledge Infrastructure*. Communications of the ACM, Vol. 38, No. 11, 1995, pp. 33–38
94. Euzenat J., Shvaiko P. *Ontology Matching*. New York: Springer, 2007
95. Banek M., Vrdoljak B., Tjoa A.M. *Word Sense Disambiguation as the Primary Step of Ontology Integration*. Proceedings of the 19th International Conference on Database and Expert Systems Applications, LNCS, Vol. 5181, Springer, 2008, pp. 65–72
96. Schwitter R. *Creating and Querying Linguistically Motivated Ontologies*. Proceedings of the Knowledge Representation Ontology Workshop, Conference in Research and Practice in Information Technology, Vol. 90, 2008, pp. 71–80
97. Kuhn T., Schwitter R. *Writing Support for Controlled Natural Languages*. Proceedings of the Australasian Language Technology Association Workshop, 2008, pp. 46–54
98. Horridge M., Drummond N., Goodwin J., Rector A., Stevens R., Wang H. *The Manchester OWL syntax*. Proceedings of the 2nd International Workshop on OWL: Experiences and Directions (OWLED), 2006

99. Dimitrova V., Denaux R., Hart G., Dolbear C., Holt I., Cohn A.G. *Involving domain experts in authoring OWL ontologies*. Proceedings of the 7th International Conference on the Semantic Web (ISWC), LNCS, Vol. 5318, Springer, 2008, pp. 1–16
100. Bārzdīņš G., Liepiņš E., Veilande M., Zviedris M. *Semantic Latvia Approach in the Medical Domain*. Proceedings of the 8th International Baltic Conference on Databases and Information Systems, 2008, pp. 89–102
101. Angelov K., Ranta A. *Implementing controlled languages in GF*. Revised papers of the Workshop on Controlled Natural Language (CNL 2009), LNAI, Vol. 5972, Springer, 2010, pp. 82–101
102. Hajičová E. *Issues of Sentence Structure and Discourse Patterns*. Prague: Charles University, 1993
103. Saulīte B. *Linguistic Markers of Information Structure in Latvian*. Proceedings of the 18th International Congress of Linguists, 2008, pp. 3067–3076
104. Ranta A. *Type Theoretical Grammar*. Oxford: Oxford University Press, 1994
105. Lokmane I. *Vārdu secības funkcijas latviešu valodā (Functions of Word Order in Latvian)*. Latvistika un somugristika Latvijas Universitātē (Latvian Studies and Finno-Ugristics at the University of Latvia), Riga: UL, 2010, pp. 59–68
106. Kaljurand K., Fuchs N.E. *Verbalizing OWL in Attempto Controlled English*. Proceedings of the 3rd International Workshop on OWL: Experiences and Directions (OWLED), CEUR, Vol. 258, 2007
107. Cregan A., Schwitter R., Meyer T. *Sydney OWL Syntax — towards a Controlled Natural Language Syntax for OWL 1.1*. Proceedings of the 3rd International Workshop on OWL: Experiences and Directions (OWLED), CEUR, Vol. 258, 2007
108. Ranta A. *Grammars as Software Libraries*. Bertot Y., Huet G., Lévy J.-J., Plotkin G. (Eds.), *From Semantics to Computer Science*, Cambridge University Press, 2009, pp. 281–308
109. Antoniou G., van Harmelen F. *A Semantic Web Primer*. Cambridge: MIT Press, 2003
110. Sirin E., Tao J. *Towards Integrity Constraints in OWL*. Proceedings of the 6th International Workshop on OWL: Experiences and Directions (OWLED), CEUR, Vol. 529, 2009
111. Bernstein A., Kaufmann E., Göhring A., Kiefer C. *Querying ontologies: A controlled English interface for end-users*. Proceedings of the 4th International Conference on the Semantic Web (ISWC), 2005, pp. 112–126
112. Power R., Third A. *Expressing OWL axioms by English sentences: dubious in theory, feasible in practice*. Posters of the 23rd International Conference on Computational Linguistics (COLING), 2010, pp. 1006–1013

Appendix

Author's reports on the results of this thesis

1. **Hierarhisku attieksmju izgūšana no latviešu valodas skaidrojošās vārdnīcas** (Extracting Hierarchical Relations from the Dictionary of Latvian) — *11th International Conference "The Word: Aspects of Research"*, Liepaja (Latvia), December, 2006
2. **Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order** — *16th Nordic Conference of Computational Linguistics (NODALIDA)*, Tartu (Estonia), May, 2007
3. **Ontological Word Sense Disambiguation for Discourse Representation** — *3rd Baltic Conference on Human Language Technologies*, Kaunas (Lithuania), October, 2007
4. **Polysemy in Controlled Natural Language Texts** — *1st Workshop on Controlled Natural Language*, Marettimo Island (Italy), June, 2009
5. **Word Order Based Analysis of Given and New Information in Controlled Synthetic Languages** — *1st Workshop on the Multilingual Semantic Web (at WWW 2010)*, Roli (NC, USA), April, 2010
6. **Verbalizing Ontologies in Controlled Baltic Languages** — *4th International Conference on Human Language Technologies — the Baltic Perspective*, Riga (Latvia), October, 2010

Author's contribution to the included publications

Authors	Publication	Contrib. (%)	Notes
G. Bārzdiņš N. Grūzītis R. Kudiņš G. Nešpore A. Spektors	<i>Latviešu valoda semantiskajā tīmeklī</i>	70	<ul style="list-style-type: none"> • Participation in the development of the ideas. • Development of the basic principles for using existing ontologies in the lexical disambiguation (in Latvian). • Development of the basic principles and their experimental verification for the semi-automatic alignment of the Latvian lexicon with the ontological concepts, preserving the fine-grained word senses.
N. Grūzītis G. Nešpore B. Saulīte	<i>Hierarhisku attieksmju izgūšana no latviešu valodas skaidrojošās vārdnīcas</i>	90	<ul style="list-style-type: none"> • Development of the idea. • Defining the lexico-syntactic patterns for the extraction of semantic relations. • Development of the experimental tool for the semi-automatic extraction of a lexical taxonomy.
G. Bārzdiņš N. Grūzītis R. Kudiņš	<i>Re-engineering OntoSem Ontology Towards OWL DL Compliance</i>	40	<ul style="list-style-type: none"> • Analysis of the OntoSem approach. • Participation in the development of the basic principles for converting and debugging of the ontology. • Development of the basic principles for debugging an ontology by extracting test cases from a dictionary.

G. Bārzdiņš N. Grūzītis G. Nešpore B. Saulīte	<i>Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order</i>	70	<ul style="list-style-type: none"> • Participation in the development of the idea. • Development of the initial model of the SemTi-Kamols dependency grammar for covering the various syntactic constructions. • Formalization of a subset of Latvian grammar by covering simple extended sentences (incl. complex analytic constructions).
G. Nešpore B. Saulīte G. Bārzdiņš N. Grūzītis	<i>Comparison of the SemTi-Kamols and Tesnière's Dependency Grammars</i>	30	<ul style="list-style-type: none"> • Comparison of the SemTi-Kamols dependency grammar to other approaches; formulation of its essential characteristics in the context of the original Tesnière's grammar.
G. Bārzdiņš N. Grūzītis G. Nešpore B. Saulīte I. Auziņa K. Levāne-Petrova	<i>Multidimensional Ontologies: Integration of Frame Semantics and Ontological Semantics</i>	50	<ul style="list-style-type: none"> • Participation in the development of the ideas, incl. in defining and development of the concept of multidimensional ontology. • Analysis of the FrameNet model and its mapping to the model of multidimensional ontology; investigation of the FrameNet conversion possibilities.
G. Bārzdiņš N. Grūzītis G. Nešpore B. Saulīte I. Auziņa K. Levāne-Petrova	<i>Ontological Word Sense Disambiguation for Discourse Representation</i>	60	<ul style="list-style-type: none"> • Participation in the development of the ideas. • Participation in the development of a Protégé plug-in for converting and extending OWL ontologies to FOL, and for building and visualizing the minimal model (by exploiting Mace4).
N. Grūzītis G. Bārzdiņš	<i>Polysemy in Controlled Natural Language Texts</i>	50	<ul style="list-style-type: none"> • Participation in the development of the ideas. • Formulation and development of the micro-ontology approach and the methodology for the systematic development and merging of domain ontologies, ensuring semi-automatic partitioning of polysemous classes.
N. Grūzītis	<i>Word Order Based Analysis of Given and New Information in Controlled Synthetic Languages</i>	100	<ul style="list-style-type: none"> • Development of the idea. • Designing the initial (basic) version of controlled, logic-based Latvian and its implementation if GF, showing that the analysis of the information structure of a sentence is a precise and sufficient means for the unambiguous resolution of quantifiers and co-references in OWL terminological axioms, SWRL inference rules and SPARQL data integrity queries that are given in a form of a controlled, highly synthetic language.

<p>N. Grūzītis G. Nešpore B. Saulīte</p>	<p><i>Verbalizing Ontologies in Controlled Baltic Languages</i></p>	<p>90</p>	<ul style="list-style-type: none"> • Development of the questionnaire, and analysis of the results of the survey. • Designing a substantially extended version of the controlled Latvian by <ul style="list-style-type: none"> - providing vast alternatives on how the same construction of the abstract grammar may be linearized, - allowing to express the predicate as a role (noun), and the object — as a modifier of place (noun). • Implementation of parallel grammars for the robust, still deterministic parsing, and for the possibly natural, still precise generation of controlled Latvian; efficient implementation in GF.
<p>N. Grūzītis G. Bārzdiņš</p>	<p><i>Towards a More Natural Multilingual Controlled Language Interface to OWL</i></p>	<p>90</p>	<ul style="list-style-type: none"> • Development of the idea. • Implementation of a prototype that demonstrates the proposed two-level translation approach for translating sentences in controlled Latvian to OWL (and vice versa) by using the Attempto Controlled English (ACE) subset for OWL as an interlingua (and by exploiting its readily available tools). • Implementation of an experimental parallel grammar, showing that the proposed approach allows also for the development of extended and adjusted controlled English on the basis of the restricted ACE-OWL — in a flexible and independent manner.