How to Build a Controlled Vocabulary of Concepts for Deep Semantic Search

Hermann Bense

bense.com GmbH, Schwarze-Brüder-Str. 1, 44137 Dortmund, Germany
hb@bense.com, https://orcid.org/0000-0002-2562-224X

Abstract. *Semantic Search* (SS) is well understood and widely applied in search engines, WordNets (PWN) and ontologies. It is based on Semantic Relations (SR) between different concepts. SS in general does not take account of Intra-Concept Relations (ICR). This article describes how concepts can be formally composed as *Semantic Compounds* (SC) of sub concepts and *Semantic Primes* (SP) using ICRs. SMs are stored in a *Controlled Vocabulary of Concepts* (CVC) which is the fundamental for the implementation of a novel *Deep Semantic Search* (DSS) algorithm also called *Search by Meaning* (SbM). The evaluation of SbM has been conducted with a large energy ontology (EnArgus) and the mobile app rObby for the automated notification on scientific publications. SbM finds more relevant concepts and publications compared to traditional SS. The results of SbM are also evaluated against those of Google Search (GS) and *Princeton WordNet* (PWN).

Keywords: Concept Composition (CC), Concept Binary Tree (CBT), Search by Meaning (SbM), Controlled Vocabulary of Concepts (CVC), Word Sense (WS), SuPerSense (SPS), Semantic Compound (SC), Semantic Prime (SP), Semantic Relation (SR), PrincetonWordNet (PWN), Word Sense Definition (WSD)

1 Introduction

Prominent figures such as Gottfried Wilhelm Leibniz [13], Charles Sanders Peirce [17], Rudolph Wittgenstein [22], Rudolph Carnap [7], Rosch et al. [18], [19], [16], and Willem Levelt [14] have made significant contributions to the understanding and processing of language. Leibniz laid the foundation for formal language processing and tried to create a *lingua universalis* with the aim of creating a world catalog of ideas. Peirce focused on *semiotics* and meaning making. Wittgenstein's ideas on language and meaning influenced contemporary debates. Carnap emphasized clarity and logic in language use. Rosch's research on categorization shaped linguistic studies, and Levelt contributed to *language production models*. Their work has enriched various disciplines, including computational linguistics, semiotics, psychology, and philosophy of language. In this paper, we take on the challenge of creating a *Controlled Vocabulary of Concepts* (CVC) that allows the composition of term meanings through arbitrary semantic relations to sub terms.

So far, mainly relations such as *hypernym*, *hyponym*, *synonym*, *antonym*, and *meronym* have been available. To the best of our knowledge, there is currently no other general methodology for defining composite terms such as *functional nouns*, *conceptual noun phrases*, *prepositional concepts*, etc. We acknowledge this gap and aim to explore and present a novel approach to address this challenge, with a particular focus on demonstrating the benefits of such a methodology in improving *semantic search* (SS) capabilities by a method called *Search by Meaning* (SbM). By showcasing how the proposed method can enhance the accuracy and relevance of semantic search results, we aim to illustrate its practical value and potential impact on semantic search.

The central research question to be answered is: "How can entries of a *Controlled Vocabulary of Concepts* (CVC) be structured in such a way that they accurately represent one or more meanings by *Word Sense Definitions* (WSD)?" This leads to further questions: What is a minimal set of words to construct a *Foundational Core Glossary* (FCG) to support the automated translation between any pair of languages? How can the definition of complex *Semantic Compounds* (SC) be constructed from *Semantic Primes* (SP)? How to extend existing methods for *Semantic Search* (SS) to a *Deep Semantic Search* (DSS) that provides more relevant results.

The rest of the paper is organized as follows. The Related Works Section 2 discusses related work from the fields of linguistics, *Semantography*, and *Princeton WordNet* (PWN). In the Preliminaries Section 3, we introduce *Concept Binary Trees* (CBTs) along with other basic definitions. They are the basis for defining, representing, and querying complex linguistic data structures such as *Word Sense Definitions* (WSD). The central Methodology Section 4 describes which concepts are defined as *Semantic Primes* (SP) and how they are used in the composition of *Semantic Compounds* (SC). The new *Search by Meaning* (SbM) technology is described in Section 4.3, and the Discussion Section 5 explains the application of Search by Meaning (SbM) in several large-scale ontology applications and compares its efficiency in random examples with Google search and *Princeton WordNet* (PWN). The analysis in the Evaluation Section 6 takes up and shows how to solve the deficits described in the Related Work Section 2, and the Conclusion Section 7 summarizes the results.

2 Related Work

This section gives an overview on those approaches which are most relevant to this article and which limitations they entail. Most of them discuss *Word Senses* (WS) and provide methods for *WS Definitions* (WSD) and / or *WS Disambiguation*. We will discuss how much of their theory can be applied for constructing a *Controlled Vocabulary of Concepts* (CVC) which will be based on a *Foundation Core Glossary* (FCG) of lexical items.

2.1 Natural Semantic Metalanguage (NSM)

NSM as discussed by Wierzbicka and Goddard in [9] is a linguistic approach for semantic analysis or, in other words, for the detection of semantics in spoken communication. A substantial difference to other semantic analysis methods lies in the fact, that the description of a complex concept is done in natural language without neologisms, formal operators, abbreviations etc.

2.2 Longman´s Dictionary of Contemporary English

The Longman´s Dictionary of Contemporary English (LDCE) [1] contains 230,000 words, phrases and meanings and 165,000 corpus-based examples. In 1983 there were only 55,000 words in the LDCE. All definitions in the dictionary are based on the *Longman Defining Vocabulary* (LDV) of around 2,000 common words. We will use the words of the LDV also as a set of candidates for the *Foundation Core Glossary* (FCG).

2.3 BlissSymbolics

With *Semantography* Charles Bliss as the author of [4] and [5] has created a practical tool for general language independent communication, especially in science, industry, commerce, traffic, etc. It is a *non-alphabetical symbol writing system, readable in all languages for semantical education and based on the principles of ideographic writing and chemical symbolism.* I.e., one can write complete sentences using *Bliss Words* (BW), and the reader can understand them without knowing the language of the writer. Charles Bliss has been motivated by the principle of Chinese characters being constructed as logograms. Bliss symbols have been developed over decades. Therefore, the set of symbols can be regarded as quite robust. Meanwhile, the Bliss community has gathered a set of approximately 6,500 Bliss words, which are composed from 20 simple lines and shapes. Among them, there are about 650 basic nouns, 960 verbs, 566 adjectives and 14 indicators like *action, description, conditional, continuous form, definite form, future action, first person* etc. We will show that Bliss symbols are very helpful for the decomposition of concepts, since the combination of Bliss Words into Bliss symbols allows to identify *Semantic Compounds* (SC) more easily. One of the problems with Bliss symbols is that they do not have brackets and therefore their semantics are not always clear.

2.4 Princeton WordNet (PWN)

Jurafsky and Martin in [12] provide some statistical information about WordNet: English WordNet 3.0 consists of three separate databases, one each for nouns ($\sim$ 188k) and verbs ($\sim$ 11.5k) and a third for adjectives ($\sim$ 22.5k) and adverbs ($\sim$ 4.5k). Each database contains a set of lemmas, each one annotated with a set of *Word Senses* (WS). The set of near synonyms for a WordNet sense is called a *SynSet* as abbreviation of *Synonym Set* (SS). 26 *Lexicographic Categories* (LC) are defined for nouns and 15 for verbs. These categories are often called *SuPerSenses* (SPS), as they act as coarse semantic categories or groupings of senses. SPSs have also been defined for adjectives and prepositions by Schneider et al. [21]. *Noun Categories* (NC) are the set $\mathcal{NC} = \{$*act, animal, artifact, attribute, body, cognition, communication, feeling, food, group, location, motive, natural event, natural object, other, person, phenomenon, plant, possession, process, quantity, relation, shape, state, substance, time*$\}$. PWN has as two kinds of *Taxonomic Entities* (TE) classes and instances. An instance is an individual or a proper noun that represents a unique entity.

2.5 Collaborative InterLingual Index (CILI)

The CILI project discussed by Bond et al. [6] identifies a number of problems with Princeton WordNet (PWN). Firstly, there are discrepancies in SynSet representation between different WordNets, where large SynSets in one WordNet are linked to small SynSets in another language through PWN, indicating underrepresentation. Additionally, different interpretations of relation names by editors and algorithms further contribute to inconsistencies. There are also noticeable differences in vocabulary coverage and polysemy across WordNets. Moreover, the absence of certain concepts in Princeton WordNet poses a disadvantage, as these concepts cannot be expressed. To investigate meaning across languages, cross-lingual linking of SynSets is necessary. However, the lack of coordination between projects leads to the introduction of similar or even identical concepts in multiple places, causing duplication and validation challenges. To overcome these issues, it is important to establish consistent meanings for semantic and lexical relations across languages. These challenges highlight the need for improved representation, coordination, and consistency to facilitate effective cross-lingual semantic analysis in order to ensure accurate and comprehensive interlingual communication. We will describe below, how many of the problems described can be addressed by creating a Controlled Vocabulary of Concepts as a formal ontology of linguistic concepts. Schneider et al. [21] introduce a new annotation scheme, a new corpus, and a new task for the *disambiguation of prepositions and possessives* in English. Starting from the insight that there is a semantic overlap between prepositions and possessives they produced a new hierarchical inventory (SNACS) of 50 super sense classes and a gold-standard corpus where all types and tokens of prepositions and possessives are disambiguated. In the Concept Composition section 4.1 and in the Discussion section 5 we will show how multi-word prepositions and possessives can be modeled as semantic compounds.

3 Preliminaries

Fig. 1 illustrates the basic principle of our approach. Terms such as *air*, *pressure*, *aid*, *thing*, and *measurement* are *Semantic Primes* (SP) that are not further

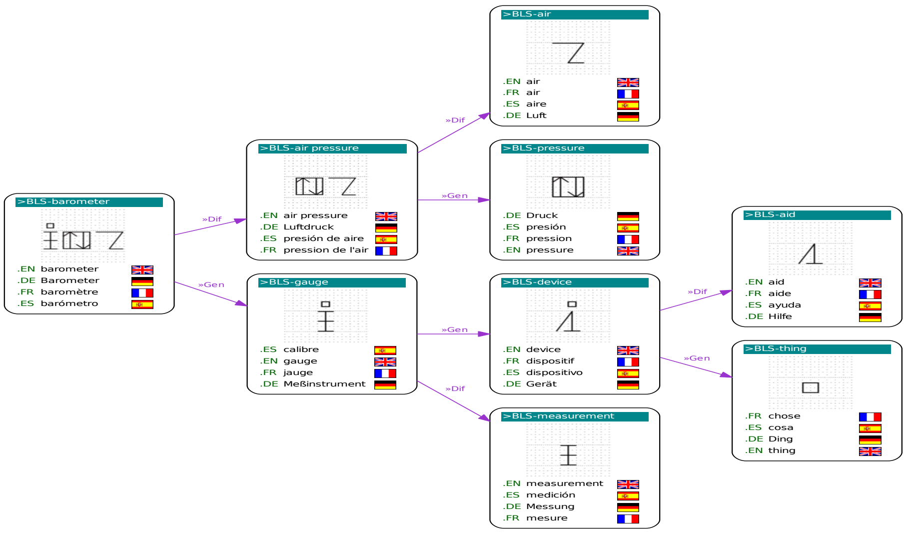

Fig. 1. Concept binary tree definition of barometer

decomposed. They are then used to construct compound terms, called *Semantic Compounds* (SC), such as *barometer*, *gauge*, *air pressure*, and *device*. When composing nouns, the *Genus-Differentiae Pattern* (GDP) is used, where the first term contains the more general concept (»Gen) and the second term specifies the first term (»Dif).A *Concept Binary Tree* (CBT) is a binary tree where each node has zero or two sub nodes. E.g., each node in the CBT of Fig. 1 represents exactly one language concept which we call *Basic Linguistic Symbol* (BLS). English labels are used for the names of BLSs, and the words for a BLS in any language is stored in the BLS and marked with a flag for that language. The pictograms in the BLSs are language independent BlissSymbols which can also be combined into BlissWords [4]. These make it possible to grasp the meaning of a term more quickly and also form a helpful basis for the construction of semantic compounds from sub terms. The following examples anticipate how nested strings for different languages can be generated from the representation of the terms in the tree structure, which we will use below for the implementation of the *Search by Meaning* (SbM) method.

- ~translate(Δ>BLS-barometer, .EN) = (((thing, aid), measurement), (pressure, air))
- ~translate(Δ>BLS-barometer, .DE) = (((Ding, Hilfe), Messung), (Druck, Luft))
- ~translate(Δ>BLS-barometer, .FR) = (((chose, aide), measure), (pression, air))

First, some initial conventions, notations and definitions are introduced. The conventions for the naming of ontological *Knowledge Entities* and for the *Graph Visualization* are based on [3]. A *Data Property* (DP) is semantically synonym to an attribute and an *Object Property* (OP) is synonym to a relationship type. To avoid extra quoting of characters in words with blanks the underscore character

_ is used to mark the place of the blank: 'air pressure' = air_pressure and (gauge, air pressure) ≡ (gauge, air_pressure). A *Name of Pair* (NoP) for a concept pair (x, y) can be easily synthesized by concatenation: NoP (x, y) = x_y ≡ 'x y', e.g., NoP(gauge, air pressure) = gauge_air_pressure ≡ 'gauge air pressure'. Names of DPs are prefixed with a dot, e.g., .Name, .URI. Names of OPs are prefixed with the symbol ◊, e.g., ◊iof, ◊Def, ◊Gen, ◊Dif, ◊subClassOf, ◊Subject, ◊Relation and ◊Object. Be 'char' a single character, then the set of all literals is defined by $\mathcal{L} = \{\text{char}\}^*$. A *Basic Linguistic Symbol* (BLS) $\in \mathcal{BLS}$ represents a prime or compound concept and contains literals for the different languages represented by the set $\mathcal{LNG}$ = {.EN, .DE, .FR, .ES, .IT, .CN, .RU, ...} of top level domain names. An ID of a bls $\in \mathcal{BLS}$ is prefixed by '>BLS-' and has the form SID(concept) = '>BLS-concept', e.g., SID(air_pressure) = '>BLS-air_pressure'.

Concept Binary Trees (CBT) are binary trees which are composed from *Semantic Primes* (SP) and *Semantic Compounds* (SC) of the set of Basis Linguistic Symbols $\mathcal{BLS}$. For CBTs we define two functions isSC() and Δ_z to compute the complete structure of a CBT as nested string expression. Please note that Δ as generic function can take one or two parameters as an argument. Be x, y, z $\in \mathcal{BLS}$, »sub ∈ {»Subject, »Gen} and »obj ∈ {»Object, »Dif}. Using the pair (»Gen »Dif) for creating compounds from nouns forms a *Genus-Differentiae Pattern* (GDP). So, Δ_z = (x, y) is a GDP ⇔ ∃xyz : (z, »Gen, x) ∧(z, »Dif, y) → x, y are nouns. The function isSC(z): $\mathcal{BLS} \to \{T, F\}$ assigns an element of the set of Basic Linguistic Symbols $\mathcal{BLS}$ the value T|TRUE|1, if it is compound, else F|FALSE|0:

Df 31 (isSC = is Semantic Compound) *isSC(z) ← T ⇔ (∃x :(z, »subj, x) ∧ ∃y :(z, »obj, y)) or ∃x, y: (z = x_y), otherwise, isSC(z) ← F.*

Df 32 (Concept Binary Tree Expression) *If isSC(z) then ∃x, y such that* $\Delta_z \leftarrow (\Delta_x, \Delta_y)$*, otherwise* $\Delta_z \leftarrow z$*.*

If the function is called with a parameter z and the concept z is a semantic prime, then z is returned as the result. Or, if z is not a semantic prime, the symbol Δ indicates a semantic compound. The function preposition 'of' is absorbed, for example, $\Delta_{\text{tooth\_of\_dog} = (\text{tooth}, } \Delta_{\text{of\_dog}) = } (_{\text{tooth, }} \Delta_{\text{dog}}) = (_{\text{tooth, dog}})$, i.e., in general for a noun x we have $\Delta_{\text{of\_x}} = \Delta_{\text{x}}$. For handling recursive definitions such as $\Delta_{\text{holonym}} = \Delta_{\text{inv\_meronym}} = _{\text{(inv, meronym)}}$, $\Delta_{\text{meronym}} = \Delta_{\text{inv\_holonym}} = _{\text{(inverse, holonym)}}$ the function Δ uses a loop-detection mechanism and terminates when an argument appears a second time while rendering the tree structure. For compounds of nouns x and y where both represent semantic primes have the special condition $\Delta_{\text{>BLS-x\_y}} = (\Delta_{\text{>BLS-y}}, \Delta_{\text{>BLS-x}}) = (_{\text{y, x}}) = \Delta_{\text{y\_x}}$ for example, $\Delta_{\text{>BLS-air\_pressure}} = (\Delta_{\text{>BLS-pressure}}, \Delta_{\text{>BLS-air}}) = (_{\text{pressure, air}}) = \Delta_{\text{air\_pressure}}$. If y is not a noun it holds that $\Delta_{\text{>BLS-x\_y}} = (\Delta_{\text{>BLS-x}}, \Delta_{\text{>BLS-y}}) = (x, y) = \Delta_{\text{x\_y}}$. This means that in all compounds in which x and y are semantic primes but x is not a noun, the order (x, y) is retained, while in the case of noun pairs they are swapped to (y, x) to model a Genus Differentiate Pattern.

Concept Binary Trees (CBT) are not comparable to *Derivation Trees* that are otherwise used in Natural Language Processing (NLP). The latter are used to represent the grammatical structure of sentences and to determine the dependencies between the sentence components using LTAG Dependency Parsing. The main aspect of CBTs, on the other hand, is to represent lexical terms and phrases in such a way that semantically relevant sub-terms are arranged as high as possible in the CBT and less relevant terms as far down as possible.

4 Methodology

In the paper of Sebastian Löbner [15], page 1, a systematic determination of concept types has been elaborated. Concepts are used as the representation of an abstract or concrete thing of the real world in the mind of a human. In linguistics and ontologies, the naming and use of concepts often becomes cumbersome due to the different meanings words can have, e.g. *class*, *bank*, *ball*, *mouse* etc. This may result in *polysemy* as discussed by Habibi et al [10]. The approach presented here combines different methods to overcome these shortfalls. We show that concepts as *multi word units* can be composed from sub concepts, using *Intra Concept Relationships* (ICR) from the set $\mathcal{ICR} = \{\Diamond$is, $\Diamond$Def, $\Diamond$Gen, $\Diamond$Dif, $\Diamond$Subject, $\Diamond$Relation, $\Diamond$Object$\}$. *Textual Definitions of a Word* (TDW) or a concept in natural language – also called *glosses* - are replaced by precise formal *Concept Binary Tree* (CBT) definitions. *Semantic Relations* (SRs) between concepts are modeled with ICRs and with *Extra-Concept Relationships* (ECR): With $\Delta_{\text{hypernym}} = \Delta_{\text{inv\_hyponym}} =$ (inv, hyponym), $\Delta_{\text{hyponym}} = \Delta_{\text{inv\_hypernym}}$ = (inv, hypernym), $\Delta_{\text{holonym}} = \Delta_{\text{inv\_meronym}} =$ (inv, meronym), $\Delta_{\text{meronym}} = \Delta_{\text{inv\_holonym}} =$ (inverse, holonym), $\Delta_{\text{troponym}} = \Delta_{\text{hyponym\_verb}} =$ (hyponym, verb), we have $\mathcal{ECR} = \{$equal, inverse, antonym, hypernym, hyponym, holonym, synonym, troponym, homonym$\}$. *Concept Relationships* (CR) are defined as the union of intra and extra SRs: $\mathcal{CR} = \mathcal{ICR} \cup \mathcal{ECR}$. Our concept construction patterns only use CBTs to compose *Semantic Compounds* (SC) as super concepts from sub concepts or *Semantic Primes* (SP).

4.1 Concept Composition

To tackle the problem of untranslatability of concepts identified in [9], our approach is based on the *Minimal English Lexicon* (MEL) and *Longmans Defining Vocabular* (LDV). In this section, methods and guidelines are introduced for the composition of the different word types nouns, verbs, adjectives, and propositions. Willem Levelt describes in [14], page 288 as important restriction that many of the word types can be combined, but some not, such as (noun, verb) or (verb, adjective). Many of the following patterns are derived from the work of Löbner [15].

Compound Nouns. *Compound NouNs* (CNN) are constructed with the Genus Differentiae Pattern (GDP) and define a hypernym relationship, for example

>BLS-air_pressure = $(\Delta_{\text{>BLS-pressure}}, \Delta_{\text{>BLS-air}})$ = (pressure, air). Many *Sortal NouNs* (SNN) take naturally adjectives like Δ_{tall} = (adj, tallness). *Individual NouNs* (INN) include proper names and personal pronouns along with nouns for unique institutions: Pope, Paris, weather(Munic), temperature(Sun); *Relational NouNs* (RNN) are binary predicate terms where the object is usually specified by a possessive construction and therefore call *possessor*. Examples of RNNs include most kinship terms and a broad variety of *Deverbal NouNs* (DNN). Here are some examples for kinship and for functional axiom / rule definitions:

- $\omega_{\text{grandfather\_\_maternal}}$= (>BLS-grandfather_(maternal), »Def, (father, mother));
- $\omega_{\text{grandfather\_\_paternal}}$=(>BLS-grandfather_(paternal), »Def, (father, father)).

This is extended by the hypernym definitions $\Delta_{\text{grandfather\_\_paternal}}$ = (is, grandfather); $\Delta_{\text{grandfather\_\_maternal}}$ = (is, grandfather) = $\Delta_{\text{is\_grandfather}}$. But please note, the definition $\Delta_{\text{grandfather}} = \Delta_{\text{father\_father}}$ has to be regarded as an anti-pattern, because it would conflict with $\Delta_{\text{grandfather}} = \Delta_{\text{father\_mother}}$ and be the source of errors in reasoning.

Functional Nouns. *Functional Nouns* (FNN) are unary function terms. Their meanings are functional concepts, involving the possessor as argument. Examples include role terms like Δ_{father}(Rosi) = $_{\text{(father, Rosi)}}$, and $\Delta_{\text{president}}$(France) = $_{\text{(president, France)}}$, as well as terms for unique parts such as Δ_{cover}(computer) = $_{\text{(cover, computer)}}$ and also terms for abstract aspects like Δ_{name}(Pope)= $_{\text{(name, Pope)}}$, dimensions like Δ_{age}(Picasso) = $_{\text{(age, Picasso)}}$, or Δ_{meaning}(meaning) = $_{\text{(meaning, meaning)}}$.

Conceptual Nouns Phrases (CNP). CNP = (determiner, noun): CNPs are characterized by modes of determination, e.g., *simply indefinite*: (a, table), *unspecific indefinite*: (some, people); *free choice*: (any, knife); *negative indefinite*: (no, employee), *contrastive demonstrative*: (this, car), (that, painting); *plural*: women = (plural, woman); *numerical*: $\Delta_{\text{century}} = \Delta_{\text{hundred\_years}}$ = (hundred, year); *quantitative*: (many, people), (several, products), (some, kilogram); *quantificational*: (every, hour), (each, person), (all, country), (both, side). CNPs go naturally with *singular definite determination*, e.g., definite articles indicate a conceptual type of reference like in (the, car), (the, plan).

Prepositional Concepts (PPC). These are composed from a preposition and a noun with PNP = (prep, noun) such as in $\Delta_{\text{under\_water}}$ = (under, water), $\Delta_{\text{behind\_Moon}}$ = (behind, Moon), $\Delta_{\text{through\_air}}$ = (through, air), $\Delta_{\text{inside\_house}}$ = (inside, house), and $\Delta_{\text{into\_nothing}}$ = (into, nothing). According to [15], *uniqueness* is a semantic property of individual and functional nouns while *relationality* is a semantic property of relational and functional nouns as in $\Delta_{\text{tooth\_of\_a\_dog}}$ = (tooth, (a, dog)).

Possessive / Property Chains. The example from [15] „*my father's wife's mother's car*“ can be represented as CBT car(mother(wife(my, father))). CBTs can be used for inferencing as follows: With Rosi ≡ $\Delta_{\text{mother\_Mary}}$ = (mother, Mary) and Tom ≡ $\Delta_{\text{father\_Rosi}}$ = (father, Rosi) → Tom = (father, $\Delta_{\text{mother\_Mary}}$) = $\Delta_{\text{father\_}\Delta\text{mother\_Mary}}$ = father(mother(Mary)). With grandfather_(maternal) = (father, mother) ≡ father(mother), the property chain matches and leads to the inference Tom ≡ (grandfather(paternal), Mary) and with grandfather_(paternal) = (is, grandfather) follows Tom ≡ (instance, grandfather). This inferencing / entailment result could then be stored as an assertion.

Activities: The pattern for the construction of compound concepts from activities is ({verb, adjective, result, cause}, *activity*): Adjectives of activities are constructed with the Genus Differentiae Pattern (GDP) and usually **do not** have comparatives and superlatives: $\Delta_{\text{print\_(to)}}$ = $\Delta_{\text{verb\_printing}}$ = (verb, printing), Δ_{printed} = $\Delta_{\text{adj\_printing}}$ = (adj, printing), Δ_{print} = $\Delta_{\text{product\_printing}}$ = (product, printing).

Latter corresponds to the definitions (>BLS_printing_product, »Gen, >BLS_product) and (>BLS_printing_product, »Dif, >BLS_printing).

Opposites. When defining concepts with the semantic relations inverse (inv), opposite_of (opp), and perpendicular (perp / orthogonal), as a general guideline we propose to take that concept as a base concept which is most *large*, *high*, *on_top* etc. In the case where an orientation or direction is involved, like in *north, east, south* and *west*, the suggestion is to go clockwise to choose the orthogonal/vertical concept like Δ_{south} = $\Delta_{\text{opp\_north}}$ = (opp, north); Δ_{east} = $\Delta_{\text{perp\_south}}$ = (perp, south) etc.

Summary of Binary Tree Compound Patterns. Compound and Functional noun definitions take the form NNP = (noun, noun), *Conceptual Noun Phrases* CNP = (determiner, noun), Adjective and Adverbial Object AAC = ({adj, adverb}, concept), *Prepositional Concepts* PNP = (prep, noun), *Adjectives of Quantities and Qualities* AQQ = (adj, noun), Comparatives CMP = (more, adj), Superlatives SUP = (most, adj), Opposites OPP = (opp, concept), and a recursively defined *Possessive / Property Chains* PPP = (noun, PPP). A PPP can be chain of n nouns, for example, Name(Wife(Grandfather_(maternal)) ≡ (Name, (Wife, Grandfather_(maternal))) and in general for n=3 we have x_of_y_of_z = x_y_z = (x, (y, z)).

4.2 Word Sense Definitions

Word Sense Definitions (WSD) as depicted in Fig. 1 define one or more meanings of Basic Linguistic Symbol (BLS). Be Δ_{X} the string definition of a linguistic concept. Then, WSDs are represented as triples of the form ω_{W} = (>BLS-W, »Def, Δ_{X}). We abbreviate this by $\omega_{\text{W}} \equiv \Delta_{\text{X}}$. Thus, with X ≠ Y and with $\Delta_{\text{X}} \neq$

Δ_{Y} we have two different definitions of the same concept W. The following examples except $\omega_{\mathrm{goodWeatherRule}}$ are taken from [9]. Please note that the definition for 'sad' shows how to define an equation with an variable X.

– $\Delta_{\text{>BLS-living\_thing}} = (\Delta_{\text{>BLS-living}}, \Delta_{\text{>BLS-thing}}) = (\text{living}, \text{thing})$
– $\omega_{\text{plant_(creature)}} = $ (>BLS-plants_(creature), »Def, $\Delta_{\text{>BLS-living\_thing}}) =$
$\omega_{\text{plant_(creature)}} = $ (>BLS-plants_(creature), »Def, (living, thing));
– $\omega_{\text{plant\_(creature)}} = ($>BLS-plants_(creature), »Def, $\Delta_{\text{>BLS-these\_things\_can't\_do\_something}})$,
where $\Delta_{\text{>BLS-these_things_can't_do_something}} = $ ((these, things), ((can, not), (do, something))
– $\omega_{\text{sad}} = $ (>BLS-sad, »Def, ((X, (feels, sad)), equal(X, (feels, something)))
– $\omega_{\text{goodWeatherRule}} = $ (>BLS-goodWeatherRule, »Def,
– ((((the, (pressure, air)), (equal, high)), then((the, weather), (equal, good)))

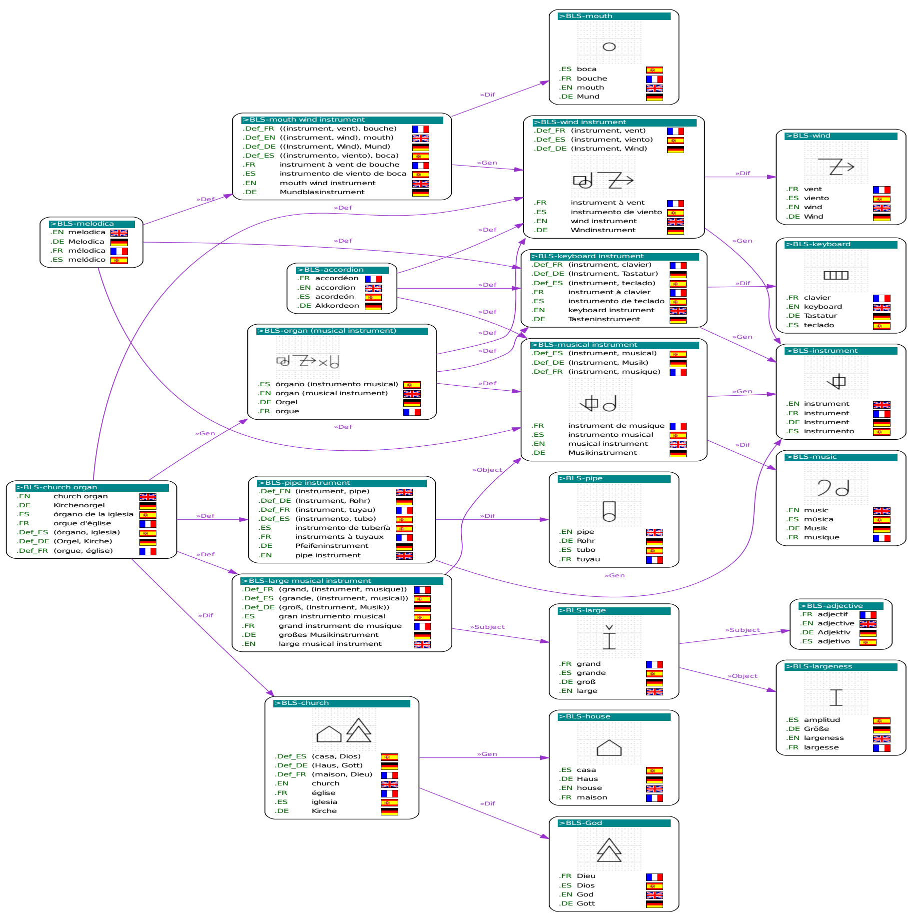

Fig. 2. Word sense definition of musical instruments

4.3 Search by Meaning

In this section, we explain the implementation of *Search by Meaning* (SbM). Later, in the Evaluation section 6 we give examples for the application of SbM in the fields of energy research and scientific publications and compare it with random examples to Google search, as well as to WordNet search. The definitions of the Basic Linguistics Symbols (BLS) can now be used to implement a performant SbM algorithm. By specifying a *Semantic Distance* (SD), it can be determined up to which definition level the terms should be included. The terms for the respective levels # are stored in the BLS for each language *lng* in the attribute .SoC_#_lng, for example, (>BLS-accordion, .SoC_2_EN, {instrument, music, keyboard, wind}), where SoC stands for *Set of Concepts*.

The set of words of a bls $\in \mathcal{BLS}$ for level n in language *lng* is $\mathrm{soc_n}$(bls, lng) = .SoC_lng_n(bls). The function *contains(soc_n(bls, lng), kw)* : $\mathcal{L} \rightarrow$ {T, F} returns T (true) if the keyword *kw* is contained in $\mathrm{soc_n}$(bls, lng), otherwise F (false). Then, the function SbM takes the form *SbM(keywords, lng, sd)* : $2^{\mathcal{L}} \rightarrow \mathcal{BLS}$ and returns for the set of *keywords* of all result terms in the language *lng* which are contained within the semantic distance *sd* in any $\mathrm{soc_i}$(bls,lng), where i $\leq$ sd. Now we can define the *Search by Meaning* function (SbM) where all keywords of KW have to appear in at least one $\mathrm{soc_i}$ of concept bls:

Df 41 (SbM = Search by Meaning) *SbM(KW,lng,sd)={bls|∀i ≤sd ∧ ∀kw∈KW: contains(soc_i(bls,lng),kw).}*

The example depicted in Fig. 2 shows the decomposition of musical instruments where each level n+1 is a decomposition of terms in level n. For the concepts $\Delta_{\text{>BLS-church organ}}$, $\Delta_{\text{>BLS-melodica}}$, and $\Delta_{\text{>BLS-accordion}}$ we get:

$\Delta_{\text{>BLS-church organ}} = \{\Delta_{\text{>BLS-pipe instrument}}, \Delta_{\text{>BLS-large musical instrument}}, \Delta_{\text{>BLS-organ (musical instrument)}}\} \cup \{(\Delta_{\text{>BLS-house}}, \Delta_{\text{>BLS-God}})\}$

- SoC0: {church organ}
- SoC1: {pipe instrument, large musical instrument, organ (musical instrument), church}
- SoC2: {instrument, pipe, large, wind instrument, keyboard instrument, musical instrument, house, god}
- SoC3: {instrument, music, wind, keyboard}, SoC4: ∅

$\Delta_{\text{>BLS-melodica}} = \{\Delta_{\text{>BLS-keyboard instrument}}, \Delta_{\text{>BLS-mouth wind instrument}}, \Delta_{\text{>BLS-musical instrument}}\}$

- SoC0: {melodica}
- SoC1: {keyboard instrument, mouth wind instrument, musical instrument}
- SoC2: {instrument, keyboard, mouth, wind instrument, music}
- SoC3: {instrument, wind}, SoC4: ∅

$\Delta_{\text{>BLS-accordion}} = \{\Delta_{\text{>BLS-keyboard instrument}}, \Delta_{\text{>BLS-wind instrument}}, \Delta_{\text{>BLS-musical instrument}}\}$

- SoC0: {accordion}
- SoC1: {keyboard instrument, wind instrument, musical instrument}
- SoC2: {instrument, keyboard, wind, music}, SoC3: ∅

SbM Examples. The following examples demonstrate the functionality of Search by Meaning. For the Basic Linguistic Symbols in the result set, the ~translate function is applied. Sample 6 demonstrates the use of the mixed languages FR, ES, and DE.

1. SbM ('church organ', 'DE', 0) = {>BLS-church_organ}
2. ~translate(SbM ('church organ', 'DE', 0)) = {Kirchenorgel}
3. ~translate(SbM ('music, instrument', 'EN', 1)) = {melodica, accordion}
4. ~translate(SbM ('keyboard, instrument', 'EN', 2)) = {church organ, accordion, melodica}
5. ~translate(SbM ('wind, instrument', 'FR', 3)) = {orgue d'église, accordéon, mélodica}
6. ~translate(SbM ('grand, tubo, instrument', 'EN', 3)) = {church organ}

5 Discussion

5.1 Word Embeddings

A *word embedding* is an embedding in which words or other symbols are each assigned to a vector v with $v \in \mathbb{R}^n$. This is primarily used in machine learning. The goal here is to obtain an abstract representation of the meaning of the words or symbols while simultaneously reducing dimensions. Historically, one of the main limitations of static word embeddings or word vector space models is that words with multiple meanings are conflated into a single representation (a single vector in the semantic space). In other words, polysemy and homonymy are not handled properly. E.g., there is no possibility of distinguishing between homonyms, and it is important to note that synonyms must have the same value. So, our method offers a couple of features which we cannot get from word embeddings: We represent words with multiple meanings with different Word Sense Definitions. For homonyms we introduce different Basic Linguistic Symbols. The same holds for synonyms where each synonym can be related to another synonym with the Extra Concept Relation »SynonymOf. But the main difference is that word embeddings are not intended to compose the meaning of words from other words. That's why word embeddings cannot be used to implement Search by Meaning in the way we would need it.

Natural Semantic Metalanguage (NSM). Wierzbicka and Goddard in [9] state: *The only 65 declared NSM primes have stabilized as a list of irreducible meanings, coded as English words with specific senses. These primes are hypothesized to be language universals, with most of them having been tested across a wide variety of languages without encountering disconfirmation.* As we have shown in the Word Sense Definition (WSD) section 4.2 for a concept c any number of WSDs are possible. Candidates for *Semantic Primes* have been identified by the authors in a decades lasting research on *Minimal English* (ME). In contrast to what the authors describe as *Global English* (GE) their *Minimal English Lexicon* (MEL) provides around 400 words on four layers: ME_0 contains 65 semantic primes plus 100 variant forms thereof (*allolexes*). The layer ME_1 has

70 universal or near-universal semantic compounds. The third layer ME_2 adds 100 semantic compounds found in many languages, and ME_3 collects 60 useful words for Minimal English which are not semantic compounds. According to the authors, one can assume that all translations of concepts using concepts of the layers ME_0 and $ME_{1,}$ and partially even ME_2 are *universal or near-universal* translatable to minimal lexicon versions of other languages like Minimal German, Minimal Finnish, Minimal Russian or Asian language like Minimal Chinese. This means that all language concepts that we have defined with WSDs based on concepts from the layers ME_0 through ME_2 can be automatically translated into all target languages without the need for lexicographer support. Compared to the translations of SynSets in WordNet this saves vast amounts of work.

Longman´s Dictionary of Contemporary English. All definitions in the dictionary are based on the *Longman Defining Vocabulary* (LDV) of around 2,000 common words. *"The words in the LDV have been carefully chosen to ensure that the definitions are clear and easy to understand, and that the words used in explanations are easier than the words defined".* The following definitions for *captain* demonstrates, that a word can have several meanings (polysemy):

$TDW_{captain}$ = {the sailor in charge of a ship, the pilot in charge of an aircraft, a military officer with a fairly high rank, someone who leads a team or other group of people}. In our approach we have shown with the example of ω_{church_organ} how to use Word Sense Definitons (WSD) such as $\omega_{ship\_captain}$, $\omega_{aircraft\_pilot}$, $\omega_{military\_captain}$ and $\omega_{team\_captain}$ to avoid the disadvantages of polysemy. The words of the LDV have been used as the second layer of our *Foundation Core Glossary* (FCG).

Princeton WordNet (PWN). As we already discussed at the beginning, the Semantic Relations described in the methodology section 4 that we have designated as Extra-Concept Relations (ECR) are essentially available in PWN. This also makes it possible to specify generic terms, synonyms, antonyms etc. However, in PWN there is no way of specifying a restriction in the sense of a Genus-Differentiae Pattern (GDP) for a term. Also, it is not possible to model concepts with the extended possibilities of Intra-Concept Relations (ICR) as discussed in the Concept Composition section 4.1. Consequently, in contrast to PWN, our Search by Meaning (SbM) methodology enables a *Deep Semantic Search* (DSS) that takes advantage of the composition of terms from others. For example, deep searching using queries like ‘air pressure’, ‘air gauge’, and ‘pressure gauge’ answers with the concept ‘barometer’, while PWN only finds the basic terms involved[1]. On the other hand, DSS also returns the ’barometer’ concept for all search terms combinations. PWN uses ILIs as unique identifiers for synsets, which are terms that can broadly be considered to represent synonyms. We, on the other hand, use speaking identifiers with the Basic Linguistic Symbols to make editing easier for lexicographers.

[1] http://wordnetweb.princeton.edu/perl/webwn?s=barometer

Disambiguation of Prepositions and Possessives. Using concept binary trees, the examples of [21] can be modeled as follows: Using the absorption of the preposition 'of', Multiword prepositions, e.g., $\Delta_{\text{out\_of}}$ = out, $\Delta_{\text{in\_front\_of}}$ = (in, front) are covered as well as *idiomatic prepositional phrases* like $\Delta_{\text{at\_large}}$ = (at, large). The 50 super senses are organized in the three hierarchies *circumstance* (non-core properties of events), *participant* (entity playing a role in an event) and *configuration* (thing, usually an entity or property, involved in a static relationship to some other entity). Their *Universal Semantic Tagset* (UST) defines a cross-linguistic inventory of semantic classes for content and function words. Some more examples are $\Delta_{\text{after\_all}}$ = (after, all), $\Delta_{\text{as\_well\_as}}$ = ((as, well), as) and *Prepositional Multi Word Expressions* (PMWE) such as $\Delta_{\text{in\_town}}$ =(in, town) and also *non-prepositional* MWE like $\Delta_{\text{take\_care}}$ = take_care.

6 Evaluation

6.1 Levels of Knowledge Organization

Hoppe in [11], pp. 115-129 defines five *Levels Of Knowledge Organization* (LOKO), all of which are covered by our Controlled Vocabulary of Concepts (CVC): The lowest level 1 is the *controlled vocabulary.* Sets of concepts may have assignments to classes, to the set of concept definitions, and to the set of synonyms. The second LOKO level is the *taxonomy level.* It adds hypernym / hyponym relationships. Thesaurus level 3 also adds *hierarchical relationships* between concepts, which are not necessarily hypernym / hyponym relationships. The WordNet level 4 also includes *lexical types* such as nouns, verbs, adjectives, and adverbs of concepts, as well as SynSets, senses, and glosses. Finally, ontology level 5 adds *ontological concepts* such as instantiations of classes, axioms, and rules for inference. The CVC can also support automated indexing of documents [11], p. 161ff. By passing the *Semantic Distance* (SD) as a parameter to SbM, the user can decide in which depth of the Concept Binary Tree of a concept he wants to find the entries in the controlled vocabulary. The query SbM ('lion', 'EN',1) returns {leo (in the zodiac), lion, roar (lion)(to), sea lion} as a result set and demonstrates how disambiguation can be supported [11], p. 172ff. 'Lion' in German and 'lion' in English are polysemes. Based on a given context, it should be possible in the majority of cases to find out which of the concepts is the relevant one. For example, SbM ('animal lion', 2) returns exactly the one relevant concept 'lion'.

Using Search by Meaning (SbM) in Apps. The rObby App[2] is a hyper personalized notification and search app for iOS and Android devices that also offers multilingual group chat functionality. Based on user-defined rules, it sends intra-day push alerts for exactly those news stories from press, sports, finance, culture, technology, and science that are highly relevant to the user. By using previous query results and testing queries with SbM support, the user can refine

[2] https://rob.by/en/App/

the rules to get even more relevant results. In the sciences, researchers spend a significant amount of time researching literature. The following sample scenario is based on rObby's Publication Database (RPDB) and shows how the rObby app can help the researcher to save work and time. The RPDB currently contains about 6,5 million scientific publications and its knowledge graph has more than 550 million triples with a database size of more than 260 gigabytes. Crawlers add several thousand new publications from scientific publisher sources daily as Conceptual Binary Trees (CBT). Throughout the day, the rObby rule processor checks the rules of all users to see if they match the newly added documents. If a rule is triggered, the user is notified with a push alert in the rObby App. The push alert includes a link to the source of the document. This way, the user is automatically informed about new relevant searches in his field on the same day they are published online.

SbM Compared with Google Search. Google is by far the most popular search engine in the western hemisphere. There is almost no information accessible how large the *knowledge graph* (KG) behind Google's Search (GS) is and how is supports search. Though, one could expect, that the KG is also used to support semantic search. Therefore, we executed some sample queries with GS. GS delivers as *result snippets* (RS) links to websites (URIs) which contain all or some keywords of the query, and short excerpts of the target page, containing the text fragments which contain the keywords. The following examples discuss which of the results of GS can be regarded as results of semantic search. For that purpose, the around 10 results of the first result page are analyzed for relevant hits. RSs not containing semantically related hits, are called Non-semantic Search Results (NSSR): (1) GS('wind, keyboard, instrument') → 22,5 million RSs[3]; Page 1: only ca. 5 hits for *melodica* and the rest are NSSRs; (2) GS(', large, wind, keyboard, instrument') → 19,1 million RSs; Page 1 contains only NSSRs; (3) GS('large, pipe, wind, keyboard, instrument') → 66,4 million RSs; Page 1 lists correctly 'pipe organ' as first result, the rest contains only NSSRs. From the examples given, there is no indication that GS, to some extent supports *Deep Semantic Search* (DSS).

SbM Compared with ChatGPT. ChatGPT[4] is not a typical search engine but it can assist to retrieve semantical relevant concepts. When querying ChatGTP with "Give a list of wind keyboard instruments" it delivers as result "Organ, Harmonium, Accordion, Melodica, Claviola, Free-Reed Organ, Sho, Regal Harmonium and Harmoniflute" along with quite precise descriptions of the instruments. On the other hand, there is no indication that ChatGPT represents the concepts in a way comparable to Concept Binary Trees (CBT). So, it is not possible to control the search by the Semantic Distance (SD) of concepts to contained sub concepts for the purpose of narrowing or widening the search results.

[3] https://www.google.com/search?q=wind+keyboard+instrument
[4] https://chat.openai.com

Applying Deep Semantic Search in the EnArgus Project. The EnArgus project was funded by the German Federal Ministry of Economics and Energy. The goal of the project is to make government funding policies in the field of energy research more transparent and to facilitate the evaluation of technological developments. The *EnArgus Internet Portal* (EIP)[5], provides information on current and completed research projects in the field of energy research. The EnArgus2 ontology was originally developed in Protégé and then imported into our controlled vocabulary (CVC) and extended by concept binary trees to facilitate and support retrieval by SbM (Deep Semantic Search). The paper [20] shows how *Search by Meaning* (SbM) has been used to improve the semantic search of the EIP.

ICIB Classification. Another use case for SbM is to synthesize a nomenclature of knowledge fields based on the ICIB-1 classification by Ingetraut Dahlberg [8]: SbM('knowledge, mathematics', 2) = {analysis (mathematics), applied mathematics, logical foundations of mathematics, mathematician, mathematics}. This again demonstrates how SbM finds more semantically related concepts with intra-concept search as opposed to SS, only using Extra Concept Relations (ECRs). It is at the same time an example for Multi-Language Support: When new concepts are entered into the CVC, names of concepts are automatically translated from English to German, French, Spanish, Italian, Dutch, Polish and Russian using the deepl-API[6]. Therefore, it is possible to apply SbM with combinations of words in all of these languages, e.g., a German-English combination such as "Statistik science".

7 Conclusion

The methodology presented in this paper are the basis for building a Controlled Vocabulary of Concepts (CVC) for the support of Deep Semantic Search. The elements of the *Foundation Core Glossary* (FCG) were carefully selected from the *Minimal English Lexicon* (MEL) [9], Longman's Defining Vocabulary [1] with about 2,000 words, and the noun categories of Wordnet. The experience gained from modeling has shown that even for rather complex *Semantic Compounds* (SC), the *maxLevel*of CBT definitions tends not to be larger than 4 to 6. The leafes of the Concept Binary Trees as depicted in Fig. 1 are always semantic primes, e.g. *air, pressure, aid, thing,* and *measurement*. In the approach presented here, each concept is bootstrapped from English and supplemented with translations from the other languages. Consequently, it is not necessary to link concepts between languages via SynSets. Instead, concepts are already linked at the deepest glossary level. This saves an enormous amount of data and work. Compared to WordNet Search, SbM allows to search for arbitrary ontological sub-concepts. Therefore, SbM can be a valuable tool for editors to

[5] https://www.enargus.de/
[6] https://www.deepl.com/translator

check if concepts are already included in the CVC before inserting a new concept. Other lexicons and thesauri enrich concepts and their meanings mainly by extra-concept relations to other concepts. With intra-concept relations (ICRs), we have added new modeling possibilities. To the best of our knowledge, there is no comparable approach. This paper shows that Search by Meaning (SbM) goes beyond traditional search by systematically considering the internal ontological and linguistic semantics of concepts.

8 Future Work

The CVC represents lexical concepts with the methodology of the Semantic Web. Therefore, CVC could be a basis to build a bridge between the world of linguistics and the world of formal ontologies. We are currently working on a method to represent each concept binary tree, which represent a basic linguistic symbol with an integer, analogous to the ideas of Leibniz. We will avoid the disadvantages of Leibniz' coding by using Cantor's pairing function instead of prime numbers for pair formation. All partial meanings of a term can be recovered from the integer number. This would also eliminate the need for randomly assigning IDs and satisfy another demand of the CILI project.

References

1. Longman Dictionary of Contemporary English. 2014.
2. Proceedings of the 11th GlobalWordnet Conference. 2021. ISBN: 978-9-464027-31-0.
3. Hermann Bense. The Unique Predication of Knowledge Elements and their Visualization and Factorization in Ontology Engineering. *Kutz O, Garbacz P (eds.), Proceedings of the Eighth International Conference (FOIS 2014), Rio de Janeiro, Brazil, Sept. 22-25, 2014*, pages 241–250, 2014.
4. Charles K. Bliss. Semantography, a non-alphabetical symbol writing, readable in all languages. *Sydney: Institute for Semantography. OCoLC: 26684585*, 1949.
5. Charles K. Bliss. Semantography (Blissymbolics). A simple system of 100 logical pictorial symbols. *Sydney: Semantography (Blissymbolics) Publications. OCoLC: 1014476.*, 1965.
6. Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. CILI: the Collaborative Interlingual Index. *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57, 2016.
7. Rudolph Carnap. Meaning and Necessity - A Study in Semanctics and Modal Logic. *The University of Chicago Press*, 1947.
8. Ingetraut Dahlberg. ICIB 1 - International Classification and Indexing Bibliography, Classification Systems and Thesauri, 1950 - 1982. *INDEKS Verlag, Frankfurt, Germany*, 1982.
9. Cliff Goddard and Anna Wierzbicka. Global English, Minimal English: Towards better intercultural communication. *Symposium: Towards better intercultural communication, Australian National University, Canberra, 2-3 July 2015*, 2015.
10. Amir Ahmad Habibi, Bradley Hauer, and Grzegorz Kondrak. Homonymy and Polysemy Detection with Multilingual Information. *[2]*, pages 26–35, 2021.

11. Thomas Hoppe and Bernhard Humm. Semantische Suche. *Springer Vieweg*, 2020. ISBN: 978-3-658-30427-0.
12. Daniel Jurafsky and James H. Martin. Word Senses and WordNet. 2020.
13. Gottfried Wilhelm Leibniz. Sämtliche Schriften und Briefe. *Sechste Reihe Philosophische Schriften, Vierter Band, Akademie Verlag*, 1999.
14. Willem J. M. Levelt. Speaking - From Intention to Articulation. *MIT Press*, page 537, 1989. ISBN: 13 978-0-262-12137-8.
15. Sebastian Löbner. Concept Types and Determination. *Journal of Semantics, Volume 28, Issue 3*, page 279–333, 2011.
16. Carolyn B. Mervis and Eleanor Rosch. Categorization of Natural Objects. *Annual Rev. Psychol.*, 1981.
17. Charles Sanders Peirce and Kenneth Laine Ketner. Reasoning and the Logic of Things. *The Cambridge Conferences Lectures of 1898, Harvard University Press, 1992*, page 297. ISBN: 0674749669, 9780674749665.
18. Eleanor Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General 104*, pages 192–233, 1975.
19. Eleanor Rosch. Prototype Classification and Logical Classifications: The Two System. *New Trends in Cognitive Representation: Challenges to Piaget´s Theory, ed. by E.Scholnik, Lawrence Erlbaum, Hilsdale*, pages 73–86, 1978.
20. Ulrich Schade, Hermann Bense, Michael Dembach, and Lukas Sikorski. Semantische Suche im Bereich der Energieforschungsförderung. *Corporate Semantic Web – Wie semantische Anwendungen in Unternehmen Nutzen stiften, Börteçin Ege, Bernhard Humm, Anatol Reibold (eds.), Springer / Vieweg*, 2015.
21. Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. Comprehensive supersense disambiguation of English prepositions and possessives. *ACL*, 2018.
22. Ludwig Wittgenstein. Tractatus logico-philosophicus, Logisch-philosophische Abhandlung. *edition suhrkamp 12, Erste Auflage 1963, (c) Basil Blackwell Oxford, 1959*.