# An Universal Formal Ontology of Semantic Atoms and Semantic Compound Concepts

Hermann Bense , hb@bense.com, July 2021
bense.com GmbH, Schwarze-Brüder-Str. 1, 44137 Dortmund, Germany

**Keywords:** Family of Words (FoW), Family of Concepts (FoW), Concept Composition (CC), Search by Meaning (SbM), Cantor's Pairing Function (CPF), unique concept number (UCN), unique concept identifier (UCI), Formal Ontology Of Concepts (FOOC)

In [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* etc. The approach presented here combines different methods to overcome these shortfalls. The article shows that concepts (multi word units) can be composed from sub concepts and assigned a Unique Concept Number (UCN). Also a novel Search by Meaning (SbM) method is provided, which enables to search concepts by any combination of sub concepts contained in super concepts. The concept construction patterns use binary trees to compose Semantic Molecules (SM) as super concepts from sub concepts. E.g. with a *manometer* defined as a *gauge* of *air_pressure* and *air_pressure* composed from *air* and *pressure* the resulting binary tree is *manometer* = (*gauge*, *air_pressure*) = (*gauge*, (*pressure*, *air*)). FOOC is an implementation of the described methodology.

**Composition of Concepts**: In 1679 Gottfried Wilhelm Leibniz developed in nine manuscripts three different models for the representation of Aristotelian logic with numbers [2]. Corresponding to the idea of Leibniz concepts are containing or excluding higher concepts. In his method each Semantic Atom (SA) is a concept without a decomposition and is associated with a prime number. The problem with Leibniz´ numbering method lies in the commutativity of multiplication. E.g., with UCN(boat) = 3 and UCN(house) = 5 then the semantical different SMs **houseboat** and **boathouse** get the same numbers: UCN(3,5) = UCN(5,3) = 3 * 5 = 15. This problem is avoided in the approach presented here by applying Cantor's pairing function (CPF) [3]:
$\pi(x, y) = y + (x + y)(x + y + 1) / 2 \rightarrow \pi(3, 5) = 41 \neq \pi(5, 3) = 39$.
**Definitions**: The set of Semantic Concepts (SC) is SCs = SAs ∪ SMs, with SAs ∪ SMs = ∅. A SA is a concept, which is **not** composed from other concepts; SAs = {sa | sa is a SA}. A SM is composed from other SCs, with SMs = $\{sm_{xy} = sm(x, y) \mid x \in SCs \wedge y \in SCs\}$ where: $x = sc_{xy}[x]$ and $y = sc_{xy}[y]$. Be i a positive integer. SA[i] is the $i^{th}$ SA. CPF is used to assign a UCN to a SA[i] with UCN(SA[i],**0**) = $\pi$ (i,**0**); if $i \neq j \rightarrow$ SA[i] $\neq$ SA[j]. The UCNs of a SM are computed with UCN(sc[x], sc[y]) = $\pi(x, y)$. Then due to the properties of CPF all concepts of SCs have unique and different UCNs; Be w ∈

SCs ; Definition of Word **DoW**{w} = { $sc_{wd}$ ∈ SCs | sc[w] ∈ SCs ∧ sm[d] ='definition'}; Genus of Word **GoW** = {noun, adjective, adverb, …, }; Species of Word **SoW** (w) = {w} ∪ {a, an, all, any, both, some, no, the, plural, more, most} ∪ { $sm_{ws}$ ∈ SMs | sm[w] ∈ SCs ∧ sm[s] ∈ **GoW**} [1]; SynSet of Word **SSoW** = {equal , synonym, similar}; Relative of Word = **RoW** = **GoW** ∪ {is, hyperonym, hyponym, antonym, equal, synonym, similar, reason, cause, oppositeOf, perpendicularOf, partOf, measurement, measure}; Family of a Word **FoW**(w) = {w} ∪ { $sm_{wr}$ ∈ SMs | sm[w] ∈ SCs ∧ sm[r] ∈ RoW}; Family of a Concept **FoC**(c) = {c} ∪ **DoW** {w}∪ **FoW**(c) ∪ **FoW**(**FoW**(c)) ∪ **SoW**(**FoW**(c)). The Family of a Concept **FoC**(c) represents the set of all concepts including all of its word species, related concepts and definitions. Compared to [6] any number of definitions are allowed for a concept. Candidates for SAs have been identified by Wierzbicka and Goddard [4,5] in a decades lasting research on NSM (Natural Semantic Language). Their Minimal English (ME) Lexicon (MEL) in contrast to what they define as *Global English* provides around 400 words on the four layers ($ME_i$): ($ME_0$) 65 *semantic primes* plus 100 variant forms thereof (allolexes) , ($ME_1$) 70 *universal or near-universal semantic molecules*, ($ME_2$) 100 *semantic molecules found in many languages* and ($ME_3$) 60 *useful words for minimal English (not semantic molecules)*. Please note that Wierzbicka and Goddard use *semantic molecule* differently compared to the use of *SM* here. In the approach here, as an initial SAs the MEL is used. According to NSM research [4,5] 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 versions of other languages like Minimal German, Minimal Finnish, Minimal Russian or even Minimal Chinese. According to [6] in a SM the first argument (genus = GoW) is an existing definition that serves as a portion of the new definition, while the second argument (species = SoW) is the portion of the definition which is not provided by the genus, e.g., product = (result, production); boathouse (house, boat); produce_(to) = (verb, production); tall = (adjective, tallness); tallest = (superlative, tallness); smallness (oppositeOf, tallness); UCN(oppositeOf)=3; UCN(bigness) = 2415; UCN(smallness) = π (UCN(oppositeOf), UCN(bigness)) = π (3, 2415) = 5834640. Applying the definitions for SCs we obtain, e.g., following sets of word species and relatives: **SoW**(tallness) = {tallness, tall, taller, tallest}; **FoW**(tallness) = {bigness, quantity, smallness, growing, longness, altometer, odometer …}; **SoW** (west) = {west, western}; **FoW**(west) = {direction, east, south, north, … }; **SoW** (production) = {produce_(to), produced}; **FoW**(production) = {activity, product …}.

**SbM - Search by Meaning** [7]: The idea of search by meaning (SbM) is to enable the semantic search with sub concepts. In the project [8] more than 12.000 concepts have been modeled e.g. *electric_capacitor* with *electrical* = (*adjective, electricity*), *device* (*thing*, *aid*), *electrical_device* (*electrical, device*), *energy_storage* = (*storage*, *energy*), *for_energy_storage* = (*for*, *energy_storage*), *electric_capacitor* = (*electrical_device*, *for_energy_storage*). The complexity of a SM is characterized by the **N**umber **o**f its **C**oncepts **NoC** and by the depth of the defining binary tree **maxLevel**: ∀ sa ∈ SA: maxLevel(sa) = 1, NoC(sa) = 1; maxLevel (*electric_capacitor*) = 4, NoC(*electric_capacitor*) = 13. The search example [8] for *electricity storage* demonstrates how the result set can be enlarged by changing MaxLevel from 2 to 3 and then obtain *capacitor* also. Names of concepts are on creation automatically translated from English to all other languages currently supported by the deepl-API [9], namely German, French, Spanish, Italian, Dutch, Polish and Russian. Therefore it is possible to apply SbM with combinations of words in all of these languages, e.g., with “höchster mountain”, in case you forgot a word in a foreign language.

**Building a Concept Naming & Numbering System (CN$_2$S):** Finally, the methods presented are the foundations to build a **F**ormal **O**ntology **Of** **C**oncepts (**FOOC**) with unique normative naming and numbering of concepts. From the MEL [4,5] and Longman´s Defining Vocabulary [10] (around 2.000 words) SAs have been carefully selected. It has been shown how families of words (FoW) and families of concepts (FoC) are constructed as SMs. The experience from modeling has provided the insight that even for rather complex SMs the maxLevel of the defining binary tree tends not to be larger than 4 to 6. The leaves of the SM binary trees are always SAs. Having a UCN and a UCI for every SC in any language allows to identify concepts in analogy to the URIs/URLs of websites language independent. Since the UCN comprises the complete decomposable binary tree definition of a SC, documents and concepts can be indexed with all contained sub concepts with only one positive integer, e.g. UCN(smallness) = 5834640. This can be used for language independent SbM as well as for the automated tagging of documents. The UCI is generated from the UCN as unique semi-mnemotecnic ID, e.g., UCI(**ta**llness) = ‘**TA**7DB0P‘ and could also like the UCN be used as a persistent ILI ID [6]. For unique individuals like *Pope* and named entities like *Paris* [1,6] the same numbering mechanisms as for SAs is applied.

**Acknowledgement for support to**: Bernhard Humm, Sebastian Lothary, Ulrich Schade, Melanie Siegel.

## References

1. Sebastian Löbner, Concept Types and Determination, Journal of Semantics, Volume 28, Issue 3 , 2011, DOI: 10.1093/jos/ffq022, pp. 279–333
2. Klaus Grashoff, On Leibniz's Characteristic Numbers; Studia Leibnitiana 34 (2), pp. 161-1184, 2002, https://philpapers.org/rec/GLAOLC; last visited: 20.07.2021
3. Meri Lisi, Some Remarks on the Cantor Pairing Function, Le Mathematiche, Vol. LXII, Fasc. I , 2007, pp. 55-65
4. Cliff Goddard, Anna Wierzbicka, Global English, Minimal English: Symposium: Towards better intercultural communication, Australian Nat. University, Canberra, 2-3 July 2015
5. Cliff Goddard, Anna Wierzbicka, What is minimal English (and how to use it), Source: [4]
6. Francis Bond, Piek Vossen, John McCrae, Christiane Fellbaum, CILI: the Collaborative Interlingual Index, Proc. of the 8th Global WordNet Conference (GWC) , 2016, pp. 50-57
7. SbM: https://www.taoke.de/ke/CNS/SbM/2,electricity,storage.html, last visited: 20.07.2021
8. Schade, Bense, Dembach, Sikorski, Semantische Suche im Bereich der Energieforschungsförderung, in Ege, Humm, Reibold (eds.), Corporate Semantic Web – Wie semantische Anwendungen in Unternehmen Nutzen stiften , 2015, Springer / Vieweg
9. Deepl https://www.deepl.com/translator, last visited: 20.07.2021
10. Chris Fox, Rosalind Combley (eds.), LONGMAN Dictionary of Contemporary English, 2014, Pearson Education Limited