SEMPRIA White Paper:

Introduction to SEMPRIA Search
The Meaning-oriented Search Engine

Dr. Sven Hartrumpf, Prof. Dr. Hermann Helbig
SEMPRIA GmbH
Grafenberger Allee 277–287
40237 Düsseldorf, Germany
https://www.sempria.de/

2012-09-06, last major revision: 2020-10-29

Abstract

SEMPRIA Search is a search engine of a new type, which for the first time offers **meaning-oriented search** for German. In contrast to traditional search engines, SEMPRIA Search can resolve word ambiguities and exploit the relations between concepts of the query. In this way, a decisive step is made from simple keyword search to full language understanding. **Recall** and **precision** of search are increased by the **logical and linguistic foundation** of the language technology employed in SEMPRIA Search. To this end, SEMPRIA Search logically relates different concepts from queries and target documents with each other, e.g. *(to) import* and *(to) export* or *(to) buy* and *buyer*, and understands the linguistic relations between expressions in different document parts.

Contents

1 Introduction

Traditional search engines are mainly oriented towards keyword search. They cannot resolve *word sense ambiguities* (i.e., they do not distinguish *horse* as an animal and *horse* as a gymnastic equipment), nor can they exploit relations between concepts (i.e., *Obama criticizes Merkel* and *Merkel criticizes Obama* are seen as identical—a drastic and dangerous simplification). Hence, keyword search engines flood their users by many irrelevant answers. Due to this shortcoming, several search engines have appeared under the marketing name *semantic search engine* in recent years. Related to this slogan are basic linguistic functions (like reducing inflected forms to their base forms, so-called lemmatization) and the use of ontologies as background knowledge. An ontology is a conceptual system that is structured by means of certain semantic relations (for example, subsumption, synonymy, and maybe also part-whole relations). This step points into the right direction, but it is insufficient for a significant improvement of search quality.

SEMPRIA[1] has been pursuing this deep semantic (or meaning-oriented) approach for many years. SEMPRIA Search hence leverages full, linguistically and logically founded language understanding for texts and for users' queries aiming at these texts. In contrast to search engines that only achieved first steps of semantic processing (see above), SEMPRIA Search builds on a fully developed theory of meaning representation, the so-called *multi-layered extended semantic networks* (MultiNet). All language understanding processes and logical search processes in SEMPRIA Search are based on this solid theory. In this way, recall and precision of search results can be increased significantly. For a deeper understanding, a short overview of the scientific foundations is useful; it is presented in the following section.

[1] SEMPRIA is a registered trademark.

2 Scientific Background

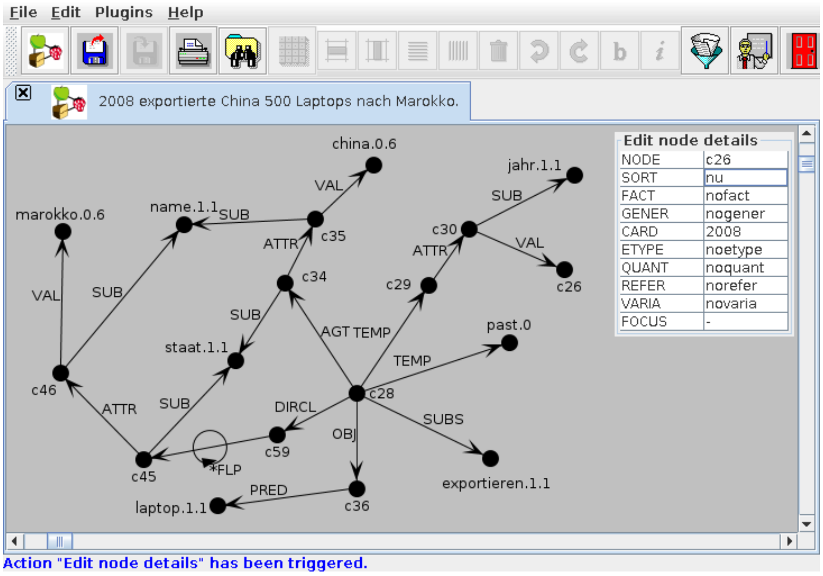
The development of SEMPRIA Search rests on 20 years of academic research in the areas of automatic knowledge processing, computational linguistics, and computational logics (see `http://pi7.fernuni-hagen.de/research/`). To represent the meaning of natural language expressions (questions, statements, short phrases, or whole texts) on a machine, one needs a suitable knowledge representation system. This system must allow to represent word meanings (collected in a **computational lexicon**), sentence meanings, and logical relationships between concepts. Only then, one can successfully link the meaning analysis of natural language expressions and the logical answer finding in a search system. With the MultiNet formalism mentioned above and fully described by Helbig (2006), the SEMPRIA GmbH owns such a knowledge representation paradigm. Its comprehensiveness, its logical foundation, the close linkage to components of language technology and logic processes, and the support by corresponding software tools turn MultiNet into a leading approach, even in international comparison. The language processing technology of SEMPRIA has been ranked among the ten most important AI technologies in Germany.[2]

To illustrate the semantic representation of natural language expressions and to ease the understanding of the following sections, the meaning structures of an example sentence and of a corresponding question are provided in Figure 1 and Figure 2, respectively. Please note that the intelligent language understanding processes in SEMPRIA Search also work properly if a user enters only *short search phrases* like *export to Morocco*. SEMPRIA Search can offer search suggestions from one-word queries or two-word queries with unlinked concepts (e.g. *morocco export*, as preferred by Google users) by extending the query in a meaningful way. Thus, it supports users in better formulating their information needs.

Let us consider the sentence *2008 exportierte China 500 Laptops nach Marokko.* (*'In 2008, China exported 500 laptops to Morocco.'*), which could be part of a text about import and export business. The meaning of this sentence is represented in Figure 1 as a semantic network, which is automatically computed by the SEMPRIA-NetParser, our syntactico-semantic parser. A semantic network is a graph whose nodes represent concepts and whose edges are relations between concepts. For example, the node c45 stands for a country, which has a name attribute (ATTR relation) with value (VAL) Morocco.[3] Similarly, node c34 represents China. The whole sentence describes an export action (node c28), which took place in 2008, represented by two time edges (TEMP). Numbers are typically encoded in the inner structure of nodes, as shown by the popup menu of the SEMPRIA-NetLab tool for node c26 that can be seen in the screenshot. Node c26 represents the year 2008 (layer attribute CARD); the other inner attributes of c26 are irrelevant for this discussion. Finally, the edges AGT, OBJ, and DIRCL (local direction) represent in this order the so-called **roles** of the export action: the agent, the object (a collection of laptops, attached by the set predicate PRED), and the direction of the export. The relations SUB and SUBS denote the conceptual subordination for objects and actions/events, respectively. They serve to

[2] `https://ki50.de/die-zehn-bedeutenden-technologien-der-deutschen-ki-geschichte/`

[3] The indices of the concept names can be ignored; they are related to word ambiguities (for this reason, one cannot use words as node identifiers, see Section 3).



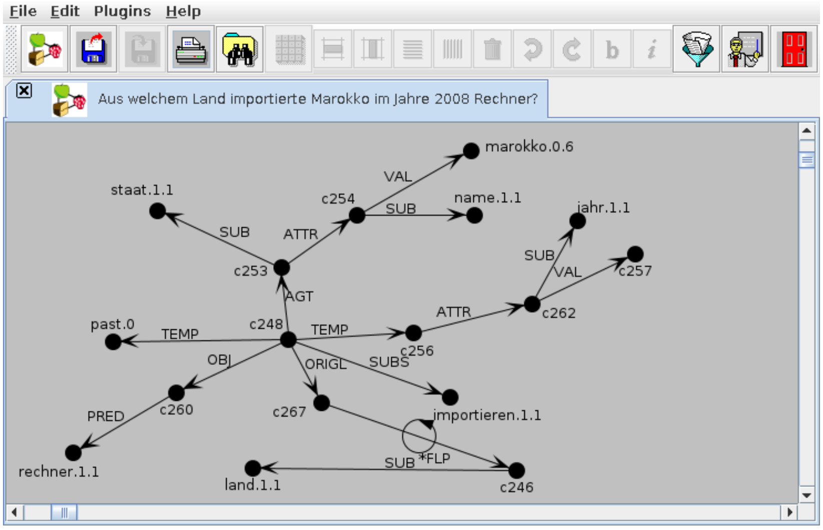
2008 exportierte China 500 Laptops nach Marokko.

Edit node details

NODE	c26
SORT	nu
FACT	nofact
GENER	nogener
CARD	2008
ETYPE	noetype
QUANT	noquant
REFER	norefer
VARIA	novaria
FOCUS	-

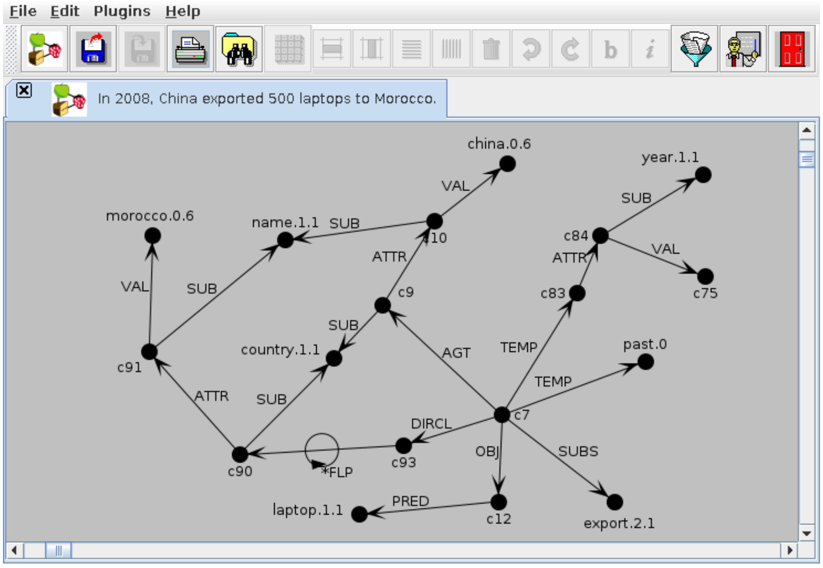
Action "Edit node details" has been triggered.

Figure 1: Simplified meaning representation of the German sentence *2008 exportierte China 500 Laptops nach Marokko.* (*'In 2008, China exported 500 laptops to Morocco.'*)



Aus welchem Land importierte Marokko im Jahre 2008 Rechner?

Figure 2: Simplified meaning representation of the German question *Aus welchem Land importierte Marokko im Jahre 2008 Rechner?* (*'From which country did Morocco import computers in 2008?'*)



In 2008, China exported 500 laptops to Morocco.

Figure 3: Simplified meaning representation of the sentence *In 2008, China exported 500 laptops to Morocco.*

distinguish individual concepts like c45 and c28, which stand for Morocco and a specific export event, from generic concepts like *Staat*/*'country'* or *exportieren*/*'export'*, respectively.[4]

Similar to the example sentence that describes an export event (see Figure 1), a question posed by a user with respect to the knowledge base (*Aus welchem Land importierte Marokko im Jahre 2008 Rechner?*, *'From which country did Morocco import computers in 2008?'*) is automatically analyzed; the resulting MultiNet graph is shown in Figure 2. Some nodes like c256 and c253 have a structure quite similar to certain nodes of the sentence from Figure 1 (in our case, the nodes c29 and c45, respectively). The remaining parts of the semantic network for the question are different: c248 describes an import event (not an export), the agent c253, connected by an AGT relation, is Morocco (not China), the object of the import are computers (not laptops). Finally, the question contains a relation of local origin (ORIGL) (not a local direction, DIRCL). Note that the SEMPRIA-NetParser determined automatically where the user's center of interest (the question focus) lies; it is represented as node c246. The sought object must be a country. In the meaning representation of the text (which comprises up to now only the semantic network from Figure 1), there is no mention of a country.

The representations are well suited for cross-lingual searching and for translation. This can be seen when comparing Figure 3 (the representation of the corresponding English sentence) to Figure 1 (the representation of the original German sentence). So far, so-called syntactico-semantic parsers for German, English, and Mandarin have been implemented that automatically analyze sentences and represent the results as semantic networks.

[4] The careful reader might have noticed the concept *staat.1.1* (superordinated to nodes c45 and c34), which is not contained in the sentence. This concept was found during the analysis in the background knowledge (see Section 3) and added to the semantic network.

To bridge the distance between question and the data provided by texts in a logical way and to analyze the semantics of natural language expressions, additional *background knowledge* is needed. This knowledge is described in the following section.

3 SEMPRIA Search as a Knowledge-based System

The Computational Lexicon SEMPRIA-NetLex Like a person learning a foreign language, a computer needs certain types of **background knowledge** for natural language *understanding*. A central part of this knowledge are the descriptions of syntactic and semantic properties of words (coming from computational linguistics).These are collected in a computational lexicon. Besides other, even more subtle difficulties, two main phenomena must be considered: **polysemy** and **homography**. A word (e.g. *star*) can have several meanings or readings; for instance, *star.1* – a celestial object, *star.2* – a public idol, and others. This phenomenon is called polysemy. In our meaning representation system, we use a numerical index to distinguish the word readings. But there are also words that – although spelled the same – differ both syntactically and semantically. This phenomenon is called homography; in the above example, a homograph to *star* in the mentioned readings would be the verbal reading *(to) star*. To distinguish such readings, we use a second numerical index. This explains why all identifiers of semantic network nodes bear two indices (with the exception of artificial concepts of the formalism, having only one index, namely *0*): The first distinguishes homographs, the second the readings of a polysemous word. (As demonstrated by our example, both phenomena can occur in connection with one word.)

The automatic distinction of readings, i.e., the reduction of ambiguous words or word forms to unambiguous concepts, is called disambiguation. Automatic disambiguation is a feature of the SEMPRIA search technology and one of the unique selling propositions (USPs) of the SEMPRIA search engine. This disambiguation is not performed by traditional search engines. For illustration purposes, some lexically relevant bits of information for the verbal concept *exportieren.1.1*/*'export.1.1'* are shown in Figure 4. (As an entry in the computational lexicon, this is called a *lexeme*.)

From the lexical entry of *exportieren.1.1*, one can see[5] that this concept (this lexeme) is characterized by the attribute SELECT comprising two obligatory roles (attribute OBLIG with value +) and two optional roles (attribute OBLIG with value -):

- an active agent (AGT – *Who is exporting?*),
- an object (OBJ – *What is exported?*),
- a local origin (ORIGL – *Where does the export come from?*), and
- a direction (DIRCL – *Where does the export go?*).

For every role, there is an entry specifying how it has to be syntactically embedded in the surface structure of a sentence (attribute SYN), and which entities are al-

[5] The morphological properties (attribute MORPH) and the syntactic properties (attribute SYN) of the verb *exportieren* itself, denoting the concept or lexeme with the name *exportieren.1.1* (see C-ID), will be ignored in this short explanation.

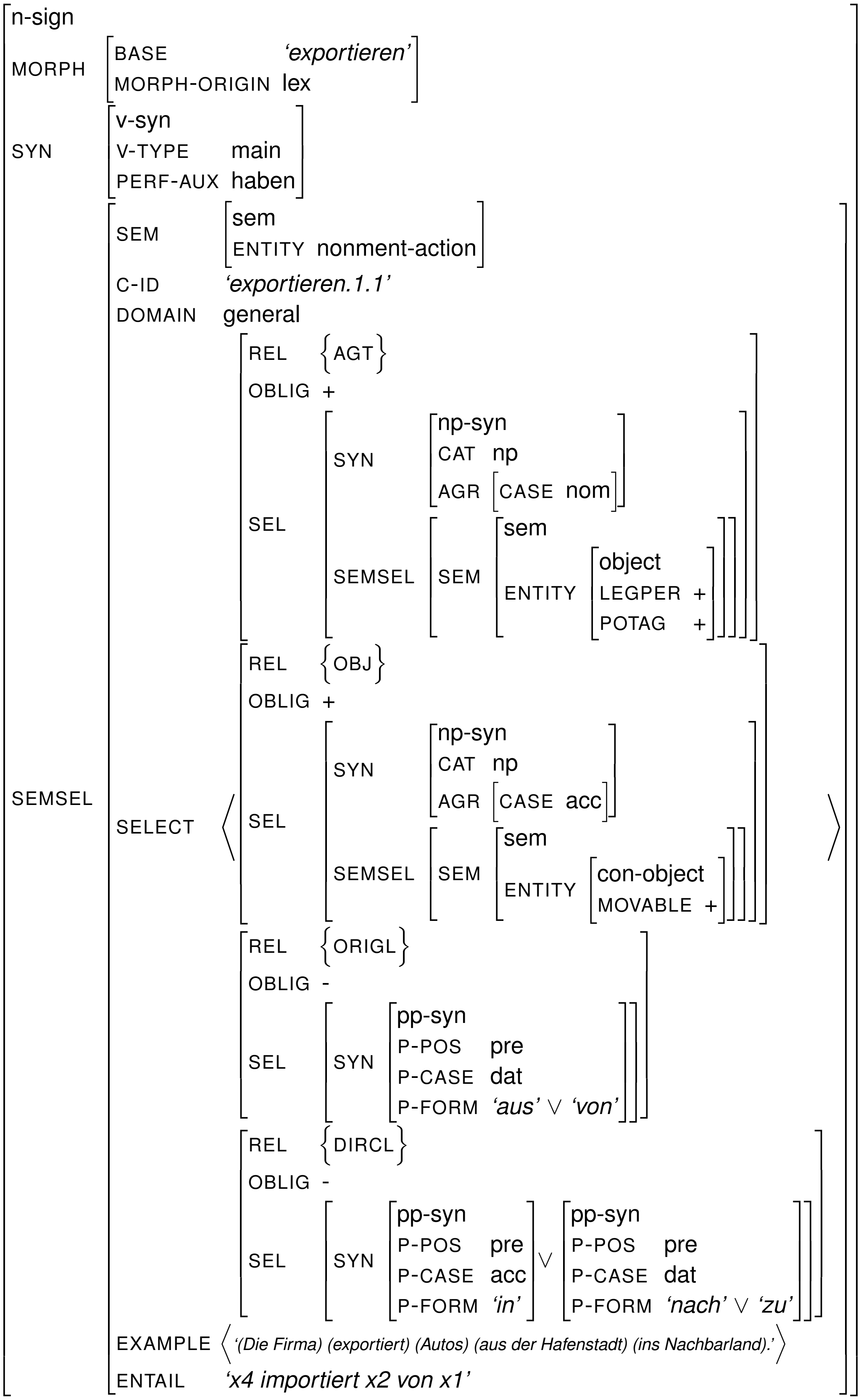


Figure 4: Simplified description of the lexeme *exportieren.1.1*/*'export.1.1'* in the computational lexicon

lowed to semantically fill this role (attribute SEM). For *exportieren.1.1* (*export.1.1*), the agent AGT has to be expressed by a noun group (noun phrase *np*) describing an object being able to act (attribute POTAG with value +) or a legal person (attribute LEGPER with value +). The object OBJ must be concrete (attribute value *con-object*) and movable (attribute MOVABLE with value +). The object has to be expressed by a noun group in the accusative case (attribute value *acc*). Semantically, the lexeme *exportieren.1.1*/*'export.1.1'* must denote a non-mental activity (*nonment-action*), as specified by the attribute SEM. The entailment attribute (ENTAIL) expresses the following relationship [6]: from the fact *x1 exports an x2 to x4* follows another fact *x4 imports x2 from x1*. Semantic relationships of exactly this kind allow SEMPRIA Search to establish connections between import and export activities, and many other semantically related concepts. The big advantage of a semantically oriented computational lexicon is that such logical relationships as shown in the example can often be transferred one-to-one into arbitrary other languages. Only the description of the syntactic properties (attribute SYN) and the morphological properties (attribute MORPH) of the words denoting a concept or a lexeme will differ to some degree from one language to another.

The lexical knowledge also comprises so-called *idiomatic phrases* (like *das Handtuch werfen*/*'(to) throw in the towel'* for *aufgeben*/*'(to) give up'*) and *light verb constructions* (like *in Verwahrung nehmen*/*'(to) take into safekeeping'* for *verwahren*/*'(to) keep'*). The knowledge about such semantic connections enables achievements of SEMPRIA Search that far surpass those of traditional search engines. With the computational lexicon SEMPRIA-NetLex and its tens of thousands of lexical entries, SEMPRIA Search also possesses a unique selling proposition (USP).

In addition to the lexical knowledge stored in the computational lexicon, there is another type of *background knowledge* applied by SEMPRIA Search. It comprises the following aspects:

Ontological knowledge: The most prominent part is constituted by the subordination relations SUB and SUBS between concepts (the first relation is connected to objects and the second to states or events). These relations introduce a hierarchical order into the world of concepts, e.g. (*laptop* SUB *computer*) or (*workstation* SUB *computer*). The synonymy of concepts does also belong to this area (e.g. the concepts *country* and *state* are often synonymously used). Information of this kind (especially, that laptops are computers) is crucial to answering the question of Figure 2 on the basis of the sentence given in Figure 1.

Logical properties of relations: Names of relations attached to the edges of the semantic networks are not isolated labels, they all bear links to logical relationships. Thus, the subordination relation SUB between concepts is transitive. This means, from (a SUB b) and (b SUB c) it follows that (a SUB c); or in a concrete example: If a rose (a) is a flower (b), and a flower (b) is a plant (c), then a rose is also a plant. Knowing these interconnections, one may also successfully ask for plants instead for roses (*Which plants are growing in your garden?*). Another useful property of some relations is their symmetry. For example, the synonymy

[6] x1 to x4, by convention, stand in this order for the roles (arguments) in the lexeme specification (see Figure 4), in this case AGT, OBJ, ORIGL, and DIRCL, respectively.

relation SYNO itself is symmetric: from (a SYNO b) follows (b SYNO a). Also the relationships between different semantic relations are important. For example, there is a connection between the causal relation CAUS and the relation of temporal successorship ANTE: generally speaking, from (a CAUS b) one can deduce (a ANTE b) because the effect can never take place before the cause. This has the consequence for search systems that they can correctly answer questions about the temporal succession of events if causal relationships are known between these events.

Logical entailments between concepts: Many concepts are logically connected by entailments, as in the example *export* and *import* above.

World knowledge: During language understanding, human beings use a large stock of information far surpassing the knowledge about the language itself and comprising knowledge about the world in the broadest sense. For instance, people typically know that Morocco is a state in North Africa (a part-whole relationship) or that a hammer is a tool (a subordination relation). This knowledge can be used to connect questions containing the second concepts of these pairs with texts containing the first concepts, respectively. Knowledge of this kind can mostly automatically be acquired by means of the SEMPRIA language technology. To this end, several publicly available sources like Wikipedia can be used.

Altogether, MultiNet provides around 140 relations (each connected with its logical apparatus). These can be used to represent, to structure, and to store the whole knowledge pertinent to all applications. Every piece of knowledge integrated into the SEMPRIA knowledge base (*background knowledge*) automatically enhances the power of every SEMPRIA application, especially that of SEMPRIA Search. This means that every user of SEMPRIA products profits from each extension of the knowledge base – be it the computational lexicon or the background knowledge – without any update of its application software.

To support the processes of automatic natural language understanding (NLU), a whole repository of logically and linguistically founded technological means has been developed on the basis of the knowledge representation paradigm MultiNet.[7] The following tools are relevant to SEMPRIA Search:

- a semantic parser (SEMPRIA-NetParser), which automatically translates natural language expressions (be it short phrases, sentences, or whole texts) into their meaning representations.
- Logical proof mechanisms and validation techniques, intelligently establishing a semantically precise connection between the meaning representation of a question and the semantic networks produced by the parser from the text archives or given by the background knowledge.
- Workbenches for the computational lexicographer (SEMPRIA-LexLab) and the knowledge engineer (SEMPRIA-NetLab), supporting the acquisition and maintenance of the background knowledge required by the parser and the logical answer finding (see Section 3).

The next section explains how these techniques cooperate during meaning-oriented

[7] Figures 1, 2, and 4 have been produced by using these tools.

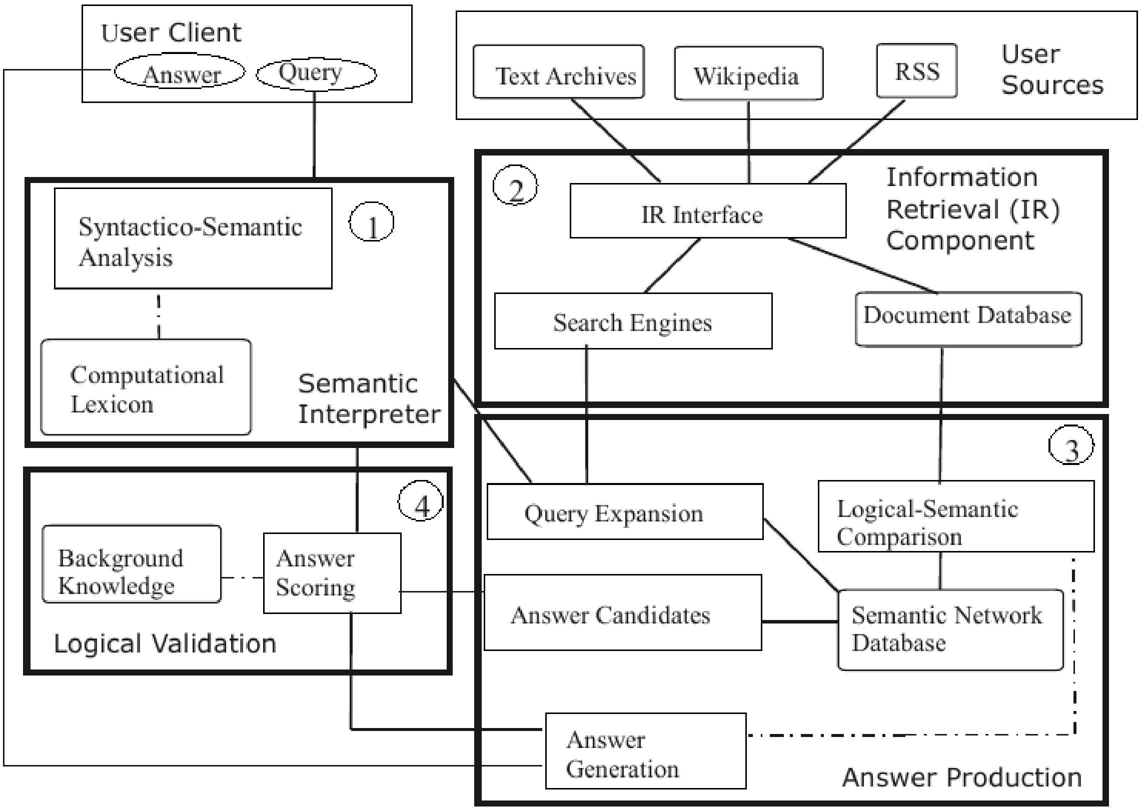


Figure 5: Structure and functionality of SEMPRIA Search

search.

4 The Architecture of SEMPRIA Search

The SEMPRIA Search engine consists of four main components, each with their respective inner structure (see Figure 5):

1. A semantic interpreter: It translates the user queries, which are input via the user interface (user client), by means of information from the computational lexicon into MultiNet meaning structures.
2. An information retrieval component searching for relevant answer candidates in the textual sources of the user by means of different special search engines; this process is based on a semantically enhanced query including synonyms, superordinated concepts, etc.
3. A component of answer production: Here, a logical-semantic comparison of answer candidates with the query is carried out. This step is necessary because the search engines from component (2) also comprise so-called flat methods as a fallback strategy, which possibly deliver imprecise or entirely inappropriate answer candidates (comparable to the results of standard search engines).
4. A logical validation component: Answer candidates produced by component (3) are evaluated according to their logical quality (i.e., it is investigated to which degree query and answer are logically consistent or to which degree they differ).

This process is supported by background knowledge to finally deliver a *logically founded ranking* of the answers.

The following section shows how the data of the user is integrated into the system.

5 Construction of the Document Archive as a Data Base

Typically, the document archive of the user is transferred to SEMPRIA via Internet. The documents can be provided in diverse formats and they are semantically analyzed for inclusion into the search. To this end, a preprocessing is necessary, called indexing, which is normally carried out offline. In contrast to traditional search engines, labor-intensive and complex processes are initiated by SEMPRIA Search during this indexing, aiming at a possibly far-reaching language understanding, as explained in the foregoing sections. The documents are translated into semantic networks according to the MultiNet formalism and connected to the knowledge already gathered (data integration into the document base). During the retrieval phase, the search engine can finally work with these coherent semantic networks; that means it can logically compare and semantically deduce. On the technical level, customer-specific information and general background knowledge are strictly separated during the construction of the document archive, thus, always guaranteeing the protection of proprietary or confidential data.

Currently, SEMPRIA Search is able to integrate the following formats (this list can easily be extended according to user requirements):

- Pure text (encoded in ASCII, ISO Latin 8859, Unicode UTF-8, . . .)
- HTML
- DOC, RTF, LibreOffice, OpenOffice, etc.
- PDF, PostScript, DVI, etc.
- WordPress, Drupal, Joomla, Typo3, and other content management systems (CMS).

It goes without saying that a comprehensive and clean integration of metadata is strictly observed by SEMPRIA Search. Only in a few cases, manual steps are required during this process. These include the provision of additional information connected to certain formats (such as tables) or the correction of errors resulting from the application of text recognition systems (OCR). Indexing and updating of documents are carried out during times agreed upon and after predefined time intervals. They are initiated by transferring a simple list of URLs, using a standard protocol (like HTTP, HTTPS, FTP, or SFTP) for the transfer actions.

6 Achievements and User Benefits

Summarizing, we can state that the application of SEMPRIA Search warrants a considerable advantage to the user compared with traditional search engines. In this context, only the most important aspects shall be emphasized:

- The system allows for a natural language access on the basis of the latest achievements in the fields of knowledge processing, computational linguistics, and computational logics.
- By its logical and linguistic foundation and by achieving a truly deep semantic language processing, SEMPRIA Search is able to adequately deal with linguistic phenomena far beyond the range of traditional search engines. These phenomena include:
  - the treatment of multiword expressions (which are now properly recognized as a single semantic entity);
  - the automatic resolution of ambiguities (lexical ones – ambiguities of words — as well as structural ones – if there exist several possibilities how parts of sentences can be related)
  - the understanding of metonymies (*Washington protests . . .*), of idiomatic figures of speech (*das Handtuch werfen*/*'(to) throw in the towel'*) and light verb constructions (*zum Abschluss bringen*/*'(to) bring to a conclusion'* for *abschließen*/*'(to) finish'*);
  - the correct treatment of temporal constructs (these include absolute as well as relative constructs, like *on July 1st, 2012* or *yesterday*, respectively) as well as of numbers and measurements;
  - the resolution of references (e.g. the references of pronouns), which often reach over sentence borders;
  - the understanding of the relationships between objects and the role of participants in an event (e.g. that the actor of an action *(to) sing* is a *singer*) and their proper treatment in queries;
  - the construction of semantic descriptions for objects and situations/events from several sentences or documents; and
  - the generation of semantic search suggestions from documents.
- The system is knowledge-based, i.e., background knowledge (linguistic knowledge and so-called world knowledge) can be included into the search. This knowledge is mostly automatically accumulated by SEMPRIA GmbH. Therefore, all users benefit from every extension of the knowledge base or the computational lexicon without the need of a repeated updating on the user side.
- SEMPRIA Search can be made more robust against errors in queries or documents by employing different correction modules regarding, for example, orthography, compound building, and others.
- Altogether, precision and completeness (recall) of the search can be significantly improved and the satisfaction and efficiency of the user can be considerably increased by the application of the modern language technology provided by SEMPRIA.
- Last but not least, the use of a linguistically founded language processing also opens a way to better connect the search modules to methods of acoustic speech recognition (access to data archives via smart phones), since it is more

natural for users to speak entire questions rather than keywords. It opens also the possibility to attach further applications like sentiment analysis or opinion mining, topic spotting, semantic recognition of duplicates (or even plagiarism), readability tests, machine translation, and many others.

References

Helbig, Hermann (2006). Knowledge Representation and the Semantics of Natural Language. Berlin: Springer.