



SEMPRIA-Whitepaper:
**Einführung in die bedeutungsorientierte
Suchmaschine SEMPRIA® Search**

SEMPRIA GmbH
Grafenberger Allee 277–287
40237 Düsseldorf
www.sempria.de

2011-05-11

Zusammenfassung

SEMPRIA® Search ist eine Suchmaschine neuen Typs, die erstmals eine **bedeutungsorientierte Suche** für das Deutsche anbietet. Im Gegensatz zu traditionellen Suchmaschinen kann SEMPRIA® Search die Mehrdeutigkeit von Wörtern auflösen und die Beziehungen zwischen den Begriffen einer Anfrage in die Suche einbeziehen. Damit wird ein entscheidender Schritt weg von der einfachen Stichwortsuche hin zum vollen Sprachverstehen getan. Durch die logisch-linguistische Fundierung der in SEMPRIA® Search eingesetzten Sprachtechnologie werden sowohl die **Vollständigkeit** als auch die **Genauigkeit** der Suche erhöht. Dabei kann SEMPRIA® Search logische Zusammenhänge zwischen Suchbegriffen und den Begriffen in den Zieldokumenten herstellen, z.B. zwischen *importieren* und *exportieren* oder zwischen *kaufen* und *Käufer*, und auch sprachliche Beziehungen zwischen Ausdrücken in unterschiedlichen Textteilen richtig deuten.

Inhaltsverzeichnis

1	Einleitung	2
2	Wissenschaftlicher Hintergrund	3
3	SEMPRIA® Search als wissensbasiertes System	5
4	Die Architektur von SEMPRIA® Search	9
5	Aufbau des Dokumentenarchivs als Datenbasis	11
6	Leistungen und Anwendernutzen	11

1 Einleitung

Traditionelle Suchmaschinen sind vorwiegend auf Stichwortsuche orientiert. Sie können weder die Mehrdeutigkeit von Wörtern auflösen (d.h. sie unterscheiden nicht zwischen einem *Pferd* als Tier und einem *Pferd* als Turngerät), noch können sie Beziehungen zwischen Begriffen in die Suche einbeziehen (d.h. *Gabriel kritisiert Merkel* und *Merkel kritisiert Gabriel* sind für sie das Gleiche, wodurch aber wichtige Unterschiede eingeebnet werden). Deshalb erhält man normalerweise bei einer reinen Stichwortsuche viel zu viele unzutreffende Antworten. Auf Grund dieses Mankos sind in letzter Zeit Suchmaschinen unter der werbewirksamen Bezeichnung *semantische Suchmaschine* auf dem Markt erschienen. Mit diesem Werbespruch verbinden sich zum einen elementare linguistische Funktionen (wie Reduzierung von deklinierten bzw. konjugierten Wortformen auf ihre Grundwörter - sogenannte Lemmatisierung) und zum anderen der Einsatz von Ontologien als Hintergrundwissen. Unter einer Ontologie versteht man dabei ein Begriffssystem, das mit Hilfe bestimmter Relationen strukturiert ist (hierzu gehören z.B. Unterordnungsbeziehungen, Synonymien und eventuell auch Teil-Ganzes-Beziehungen). Das ist grundsätzlich ein Schritt in die richtige Richtung, aber bei weitem noch nicht ausreichend zur wesentlichen Verbesserung der Suchqualität. SEMPRIA ist diesen Weg seit vielen Jahren konsequent weiter gegangen. SEMPRIA® Search setzt auf ein volles linguistisch und logisch fundiertes Sprachverstehen sowohl für Texte als auch für die auf die Texte zielenden Nutzeranfragen. Im Gegensatz zu den zuletzt genannten Suchsystemen, die nur erste Schritte in Richtung semantische Verarbeitung gehen, beruht SEMPRIA® Search auf einer voll ausgebauten Theorie der Bedeutungsdarstellung, den sogenannten **mehrschichtigen semantischen Netzen** (kurz: MultiNet), auf der als einheitliches Bindeglied alle Sprachverstehensprozesse von SEMPRIA® Search und alle logischen Suchprozesse basieren. Damit lassen sich Vollständigkeit und Genauigkeit der Suche deutlich steigern. Zum besseren Verständnis ist ein kurzer Einblick in die wissenschaftlichen Grundlagen erforderlich, den der nächste Abschnitt bieten will.



2 Wissenschaftlicher Hintergrund

Der Entwicklung von SEMPRIA® Search liegen über zwanzig Jahre Forschung auf den Gebieten der automatischen Wissensverarbeitung, der Computerlinguistik und der Computerlogik zugrunde (s. <http://pi7.fernuni-hagen.de/forschung/>). Um die Bedeutung von natürlichsprachlichen Ausdrücken (Fragen, Aussagesätzen, kurzen Phrasen oder ganzen Texten) auf dem Rechner überhaupt darstellen zu können, benötigt man ein geeignetes Wissensrepräsentationssystem. Dieses muss es gleichzeitig gestatten, Wortbedeutungen (zusammengefasst in einem Computerlexikon), Satzbedeutungen und logische Zusammenhänge zwischen Begriffen darzustellen. Nur dann gelingt es, die Bedeutungsanalyse natürlichsprachlicher Ausdrücke mit der logischen Antwortfindung in Suchsystemen kohärent zu verbinden. Mit dem oben erwähnten MultiNet-Formalismus, der vollständig in Helbig (2006) beschrieben ist, verfügt die SEMPRIA GmbH über ein solches Wissensrepräsentations-Paradigma. In seiner inneren Geschlossenheit, seiner logischen Fundierung, der Anbindung an sprachtechnologische und logische Komponenten sowie in der Unterstützung durch entsprechende Softwarewerkzeuge dürfte MultiNet auch im internationalen Vergleich führend sein.

Zur Veranschaulichung der semantischen Repräsentation natürlichsprachlicher Ausdrücke und zum besseren Verständnis der weiteren Ausführungen sind nachstehend die Bedeutung eines Satzes (s. Abbildung 1) und einer ausformulierten Frage (s. Abbildung 2) dargestellt. Es ist hervorzuheben, dass die intelligenten Sprachverstehensprozesse in SEMPRIA® Search auch dann funktionieren, wenn ein Nutzer nur **knappes Suchphrasen** wie *Export nach Marokko* eingibt. SEMPRIA® Search kann durch Vorschläge einem Nutzer helfen, Einwort-Anfragen und Zweiwort-Anfragen ohne inhaltliche Verbindung der Begriffe (*marokko export*, wie von Google-Nutzern bevorzugt) sinnvoll so zu erweitern, dass es seinem Informationsbedürfnis besser entspricht.

Betrachten wir einen Satz wie *2008 exportierte China 500 Laptops nach Marokko.*, den man sich als Bestandteil eines größeren Textes über Import-/Exportgeschäfte denken kann. Seine Bedeutung ist in Abbildung 1 als semantisches Netz dargestellt, wie man es automatisch vom SEMPRIA-NetParser als Ergebnis der syntaktisch-semantischen Analyse erhält. Ein semantisches Netz ist mathematisch gesehen ein Graph, dessen Knoten Begriffe darstellen und dessen Kanten die Beziehungen (Relationen) zwischen den Begriffen repräsentieren. So steht der Knoten c45 für einen Staat, der ein Namens-Attribut hat (Kante ATTR), dessen Wert (VAL) Marokko ist.¹ Analoges gilt für den Knoten c34, der China repräsentiert. Der gesamte Satz beschreibt eine Export-Handlung (repräsentiert durch den Knoten c28), die im Jahr 2008 stattfand (letzteres wird durch die zwei Zeitkanten (TEMP) dargestellt). Zahlen werden typischerweise in der inneren Struktur von Knoten codiert, die beispielhaft für den Knoten c26 mit dem Werkzeug SEMPRIA-NetLab in einem Pop-up-Menü sichtbar gemacht wurde. c26 repräsentiert demnach die Jahreszahl 2008 (Layer-Attribut CARD); die übrigen inneren Merkmale des Knotens c26 sind hier nicht weiter relevant.

¹Die Indizes an den Begriffsbezeichnungen vernachlässigen wir vorerst, sie hängen mit der Mehrdeutigkeit von Wörtern zusammen, weshalb man auch nicht einfach Wörter als Bezeichner von Knoten verwenden kann, s. Abschnitt 3.



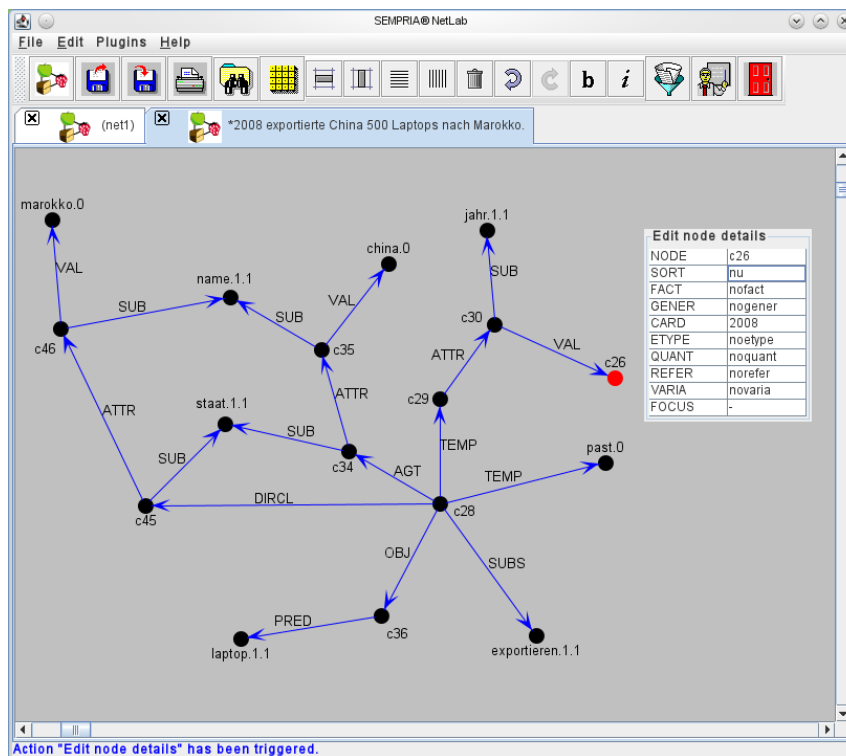


Abbildung 1: Vereinfachte Bedeutungsdarstellung des Satzes *2008 exportierte China 500 Laptops nach Marokko.*

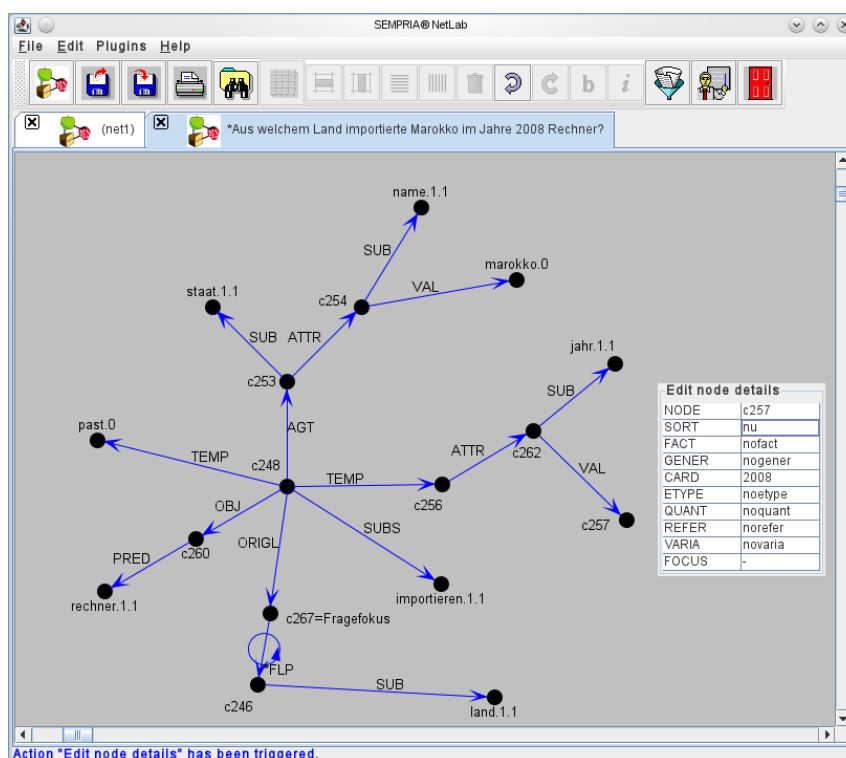


Abbildung 2: Vereinfachte Bedeutungsdarstellung der Frage *Aus welchem Land importierte Marokko im Jahre 2008 Rechner?*



Schließlich repräsentieren die Kanten AGT, OBJ und DIRCL (Direktion, räumliche Richtung) in dieser Reihenfolge die sogenannten Rollen der Export-Handlung: den Handelnden (Handlungsträger), das Objekt (eine Gesamtheit von Laptops, angeschlossen über das Mengenprädikat PRED) und die Richtung, wohin der Export erfolgt. Als wichtig zu erwähnen sind noch die Relationen SUB bzw. SUBS, die eine begriffliche Unterordnung für Objektbegriffe bzw. für Vorgänge und Handlungen bezeichnen. Sie dienen u.a. dazu, individuelle Begriffe, wie c45 und c28, die für Marokko bzw. einen speziellen Exportvorgang stehen, von generischen Begriffen, wie *Staat* oder *exportieren*, zu unterscheiden.²

Analog zum Beispielsatz, der einen Exportvorgang beschreibt (s. Abbildung 1), wird auch eine vom Nutzer an die Wissensbasis gestellte Frage *Aus welchem Land importierte Marokko im Jahre 2008 Rechner?* semantisch analysiert und als MultiNet-Struktur dargestellt (s. Abbildung 2). Einige Knoten wie c256 und c253 aus der Bedeutungsstruktur der Beispielfrage haben eine ganz ähnliche Struktur wie bestimmte Knoten des Aussagesatzes aus Abbildung 1 (in unserem Fall sind das c29 bzw. c45). Der übrige Teil des semantischen Netzes der Frage unterscheidet sich aber recht stark: c248 beschreibt einen speziellen Import-Vorgang (keinen Export), der Handlungsträger c253, angeschlossen über die AGT-Relation, ist Marokko (nicht China), das Objekt des Importierens sind Rechner (also zunächst erst einmal etwas anderes als Laptops) und schließlich enthält die Frage eine Relation der räumlichen bzw. lokalen Herkunft (ORIGL) und keine räumliche Richtung (DIRCL). Ganz wichtig ist die Tatsache, dass der SEMPRIA-NetParser automatisch festgestellt hat, wo das Zentrum des Interesses des Fragenden oder der sogenannte Fragefokus liegt (repräsentiert durch den Knoten c246). Dabei muss das gesuchte Objekt ein Land sein. In der Bedeutungsdarstellung des Textes (die uns bisher nur in Form des Netzes aus Abbildung 1 vorliegt) ist aber von keinem Land die Rede.

Um die Differenz zwischen Frage und den textuell gegebenen Daten logisch zu überbrücken, ja um natürlichsprachliche Ausdrücke überhaupt semantisch analysieren zu können, bedarf es zusätzlichen Hintergrundwissens, dem wir uns im folgenden Abschnitt zuwenden wollen.

3 SEMPRIA® Search als wissensbasiertes System

Das Computerlexikon SEMPRIA-NetLex Wie ein Mensch, der eine Fremdsprache erlernt, benötigt auch ein Computer bestimmtes Hintergrundwissen, ohne das er natürliche Sprache nicht *verstehen* kann. Hierzu gehört in erster Linie die computerlinguistische Beschreibung der syntaktischen und semantischen Eigenschaften von Wörtern, die in einem sogenannten Computerlexikon zusammengestellt werden müssen. Dabei sind (neben anderen, noch subtileren Schwierigkeiten) vor allem zwei Phänomene zu berücksichtigen: die Polysemie und die Homographie. Zum einen kann

²Dem aufmerksamen Betrachter wird nicht entgangen sein, dass in dem Netz der Begriff *Staat* enthalten ist (er ist jeweils den Knoten c45 und c34 übergeordnet), der in dem Ausgangssatz gar nicht vorkommt. Dieser Begriff, bezeichnet mit 'Staat.1.1' im Netz, wurde bei der Analyse aus dem Hintergrundwissen, s. Abschnitt 3, entnommen und automatisch hinzugefügt.



ein Wort (z.B. *Star*) mehrere Bedeutungen haben (*Star.1* – als Vogel, *Star.2* – als Augenkrankheit, *Star.3* – als Publikumsidol). Diese Erscheinung nennt man Polysemie, und um die verschiedenen Bedeutungen eines Wortes zu unterscheiden, benötigt man zum Wort schon einen ersten Index. Es gibt aber auch Wörter, die nur zufällig gleich geschrieben werden, sich aber syntaktisch und semantisch in ihrem Sprachgebrauch unterscheiden. Diese Erscheinung, die man Homographie nennt, findet man z.B. bei *sein*, das sowohl ein Pronomen als auch ein Hilfsverb bezeichnet. Um diese verschiedenen Lesarten zu unterscheiden, benötigt man einen zweiten Index, und das erklärt auch, warum alle Bezeichner von Knoten des semantischen Netzes (mit Ausnahme von Eigennamen, die nur einen Index, nämlich '0' tragen) zwei Indizes haben. Der erste steht für die Unterscheidung von Homographen und der zweite für die Unterscheidung von verschiedenen Lesarten eines polysemen Wortes (man beachte, dass semantisch beide Erscheinungen bei ein und derselben Wortform im Text vorkommen können). Die automatische Ausführung dieser Unterscheidung, d.h. die Zurückführung mehrdeutiger Wörter oder Wortformen auf eindeutige Begriffe, nennt man Disambiguierung – eine Leistung, die die SEMPRIA®-Sprachtechnologie auszeichnet und ein Alleinstellungsmerkmal für die SEMPRIA®-Suchmaschine darstellt. Gerade dies wird von traditionellen Suchmaschinen nicht geleistet. Zur Illustration sind in Abbildung 3 einige lexikalisch relevante Informationen zum Begriff *exportieren.1.1* gezeigt (als Eintrag im Computerlexikon bezeichnet man diesen als Lexem).

Aus dem Lexikoneintrag für *exportieren.1.1* kann man entnehmen,³ dass dieser Begriff (dieses Lexem) unter dem Merkmal SELECT durch zwei obligatorische Rollen (Merkmal OBLIG mit Wert +) und zwei optionale Rollen (Merkmal OBLIG mit Wert -) gekennzeichnet ist, nämlich:

- einen Handlungsträger (AGT – *Wer exportiert?*),
- ein Objekt (OBJ – *Was wird exportiert?*),
- eine Herkunft (ORIGL – *Woher wird etwas exportiert?*) und
- eine Richtung (DIRCL – *Wohin wird etwas exportiert?*).

Für jede Rolle ist wiederum angegeben, wie sie in der Oberflächenstruktur des Satzes syntaktisch eingebettet werden muss (Merkmal SYN) und durch welche Entitäten sie semantisch ausgefüllt werden kann (Merkmal SEM). Das bedeutet: der Handlungsträger AGT ist im Satz durch eine Substantivgruppe (Nominalphrase *np*) im Nominativ (*nom*) auszudrücken, die semantisch ein handlungsfähiges Objekt (Merkmal POTAG mit Wert +) oder eine juristische Person (Merkmal LEGPER mit Wert +) beschreibt. Das Objekt OBJ muss ein Konkretum sein (Merkmalswert *con-object*) und es muss beweglich sein (Merkmal MOVABLE mit Wert +). Syntaktisch wird es durch eine Substantivgruppe im Akkusativ (Merkmalswert *acc*) ausgedrückt. Die beiden fakultativen Rollen ORIGL und DIRCL werden analog charakterisiert. Hier besteht die Besonderheit, dass die Herkunft (ORIGL) durch eine Präpositionalgruppe (mit den Präpositionen *aus* bzw. *von*) und die Zielrichtung (DIRCL) durch eine andere Präpositionalgruppe (mit den Präpositionen *nach* bzw. *zu*) beschrieben werden muss.

³Auf die morphologischen Eigenschaften (Merkmal MORPH) bzw. die syntaktischen Eigenschaften (Merkmal SYN) des Verbs *exportieren* selbst, das den Begriff bzw. das Lexem mit dem Konzept-Namen (der C-ID) *exportieren.1.1* beschreibt, soll hier nicht weiter eingegangen werden.



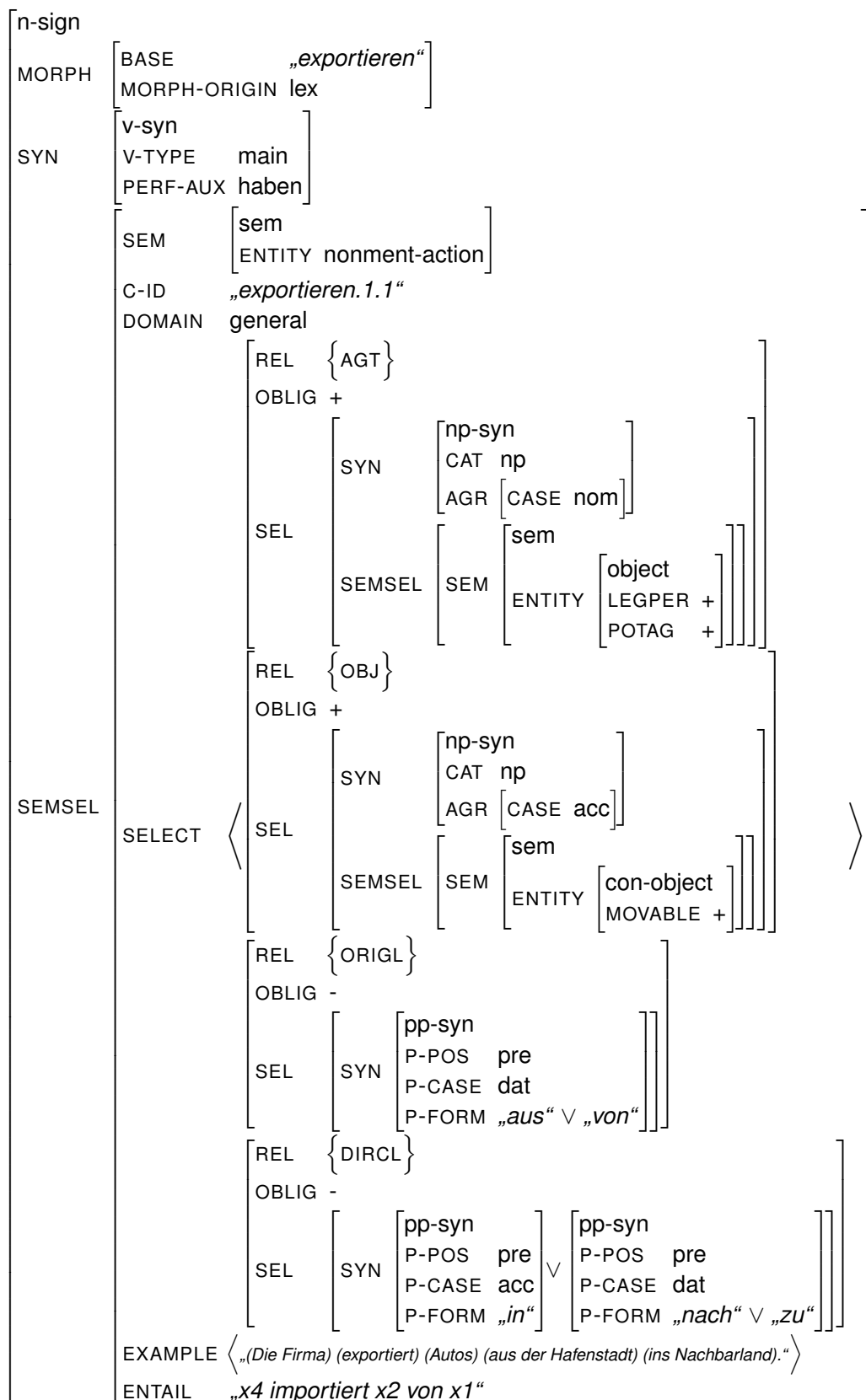


Abbildung 3: Vereinfachte Beschreibung des Lexems *exportieren.1.1* im Computerlexikon



Das Lexem *exportieren.1.1* selbst muss semantisch (Merkmal SEM) eine nichtmentale Handlung (*nonment-action*) sein. Durch ein sogenanntes Entailment (Merkmal ENTAIL) wird Folgendes ausgedrückt⁴: aus dem Sachverhalt *x1 exportiert ein x2 aus dem x3 nach x4* folgt der Sachverhalt *x4 importiert x2 von x1*. Es sind semantische Beziehungen genau dieser Art, die es SEMPRIA® Search ermöglichen, logische Zusammenhänge zwischen Import- und Export-Handlungen (und umgekehrt) herzustellen. Der große Vorteil eines semantisch orientierten Lexikons besteht darin, dass die hier dargestellten inhaltlichen Zusammenhänge praktisch eins zu eins auf beliebige andere Sprachen übertragen werden können. Lediglich die Beschreibungen der Syntax und der Wortformen (Merkmal MORPH) der Wörter, die einen Begriff bzw. ein Lexem beschreiben, werden sich mehr oder weniger stark unterscheiden.

Zum lexikalischen Wissen gehören auch sogenannte idiomatische Wendungen (wie *das Handtuch werfen* gleich *aufgeben*) oder Funktionsverbgefüge (wie *in Verwahrung nehmen* gleich *verwahren*). Das Wissen über solche Sinnzusammenhänge führt zu Leistungen von SEMPRIA® Search, die weit über diejenigen traditioneller Suchmaschinen hinausgehen. Mit SEMPRIA-NetLex und seinen Zehntausenden von semantischen Lexikoneinträgen verfügt SEMPRIA® Search über ein Alleinstellungsmerkmal, das vergleichbaren Systemen fehlt.

Über das lexikalische Wissen aus dem Computerlexikon hinaus wird noch weiteres **Hintergrundwissen** eingesetzt, das u.a. folgende Aspekte umfasst:

Ontologisches Wissen: Ganz wichtige Beziehungen sind die Unterordnungsbeziehungen zwischen Begriffen (SUB, SUBS, etc.). Diese strukturieren die Begriffswelt hierarchisch, z.B. (*Laptop* SUB *Rechner*) oder (*Workstation* SUB *Rechner*). Auch die Synonymie von Begriffen gehört in diesen Bereich (z.B. werden die Begriffe *Land* und *Staat* synonym gebraucht). Informationen dieser Art (insbesondere, dass Laptops Rechner sind) sind entscheidend für die Beantwortung der Frage aus Abbildung 2 vor dem Wissenshintergrund der Aussage aus Abbildung 1.

Logische Eigenschaften von Relationen: Die an den Kanten vermerkten Relationsnamen sind nicht einfach tote, isolierte Namen, sondern sie verweisen auf logische Zusammenhänge. So ist die Subordinationsbeziehung SUB zwischen Begriffen transitiv. Das bedeutet, aus (a SUB b) und (b SUB c) folgt (a SUB c); oder in einem konkreten Beispiel: Wenn eine Rose (a) eine Blume (b) ist, und eine Blume (b) eine Pflanze (c) ist, dann ist auch eine Rose eine Pflanze. Wenn man diese Zusammenhänge kennt, dann kann man statt nach Rosen erfolgreich auch nach Pflanzen fragen (*Welche Pflanzen wachsen in Ihrem Garten?*). Eine andere nützliche Eigenschaft mancher Relationen ist ihre Symmetrie (z.B. ist die Synonymie-Beziehung SYNO symmetrisch; d.h. aus (a SYNO b) folgt (b SYNO a)). Aber auch die Beziehungen zwischen verschiedenen Relationen sind von Bedeutung. Hierzu gehört z.B. die Beziehung zwischen der Kausalrelation CAUS und der zeitlichen Nachfolgerrelation ANTE: aus (a CAUS b) folgt etwas vereinfacht (a ANTE b), weil die Wirkung nicht vor der Ursache stattfinden kann. Das hat für Suchsysteme die Konsequenz, dass man Fragen nach der zeitlichen Abfolge logisch korrekt beantworten kann, wenn man Ursache-Wirkungs-

⁴x1 bis x4 stehen nach Lexikonkonvention in dieser Reihenfolge für die Rollen (Argumente) AGT, OBJ, ORIGL und DIRCL in Abbildung 3.



Zusammenhänge kennt.

Logische Entailments zwischen Begriffen: Viele Begriffe sind logisch miteinander verbunden durch Entailments, wie im Beispiel oben für *exportieren* und *importieren*.

Weltwissen: Der Mensch setzt beim Sprachverstehen eine große Menge von Informationen ein, die weit über das Sprachwissen hinausgehen und Wissen über die Welt im weitesten Sinne umfassen, z.B. weiß er, dass Marokko ein Staat in Nordafrika ist (eine Teil-Ganzes-Beziehung) oder dass ein Hammer ein Werkzeug ist (eine Unterordnungsbeziehung). Dieses Wissen kann mit Hilfe der SEMPRIA®-Sprachtechnologie zu einem großen Teil automatisch gewonnen werden, wofür eine Reihe von allgemein zugänglichen Quellen (wie z.B. die Wikipedia) genutzt werden können.

Insgesamt stehen in MultiNet weit über hundert Relationen (verknüpft mit einem logischen Apparat) zur Verfügung, mit denen dieses Wissen dargestellt, strukturiert und für alle Anwendungen gespeichert werden kann. Mit jedem Stück Wissen, das so in die SEMPRIA®-Wissensbasis aufgenommen wird (dem sogenannten **Hintergrundwissen**), wächst die Leistungsfähigkeit jeder SEMPRIA®-Anwendung, also auch die von SEMPRIA® Search. Das bedeutet, dass jeder Nutzer von jedem Wissenszuwachs – sei es im Computerlexikon oder im Hintergrundwissen – profitiert, ohne irgend ein Update seiner Anwendungssoftware durchführen zu müssen.

Zur Unterstützung der Sprachverstehensprozesse wurde auf der Grundlage des Wissensrepräsentationsparadigmas MultiNet ein ganzes Repertoire von sprachtechnologischen Werkzeugen⁵ entwickelt, von denen die folgenden für die Suchmaschine SEMPRIA® Search relevant sind:

- Ein semantischer Parser (SEMPRIA-NetParser), der natürlichsprachliche Ausdrücke (seien es kurze Phrasen, Sätze oder ganze Texte) in ihre Bedeutungsdarstellung übersetzt.
- Logische Beweisverfahren und Validierungstechniken, die es gestatten, die Bedeutungsdarstellung der Frage mit den semantischen Netzen, die aus Textarchiven und dem Hintergrundwissen gewonnen wurden, inhaltlich präzise und intelligent zu verbinden.
- Werkbänke für den Computerlexikographen (SEMPRIA-LexLab) und den Wissensingenieur (SEMPRIA-NetLab), die es gestatten, das für den Parser und die logische Antwortsuche erforderliche Hintergrundwissen (s. Abschnitt 3) zu erstellen und zu pflegen.

Die Frage, wie diese Techniken bei der bedeutungsorientierten Suche zusammenwirken, wird im nachstehenden Abschnitt behandelt.

4 Die Architektur von SEMPRIA® Search

Die SEMPRIA®-Suchmaschine besteht im wesentlichen aus vier Komponenten mit einer entsprechenden inneren Struktur (s. Abbildung 4):

⁵Die Abbildungen 1, 2 und 3 sind mit diesen Werkzeugen erstellt worden.



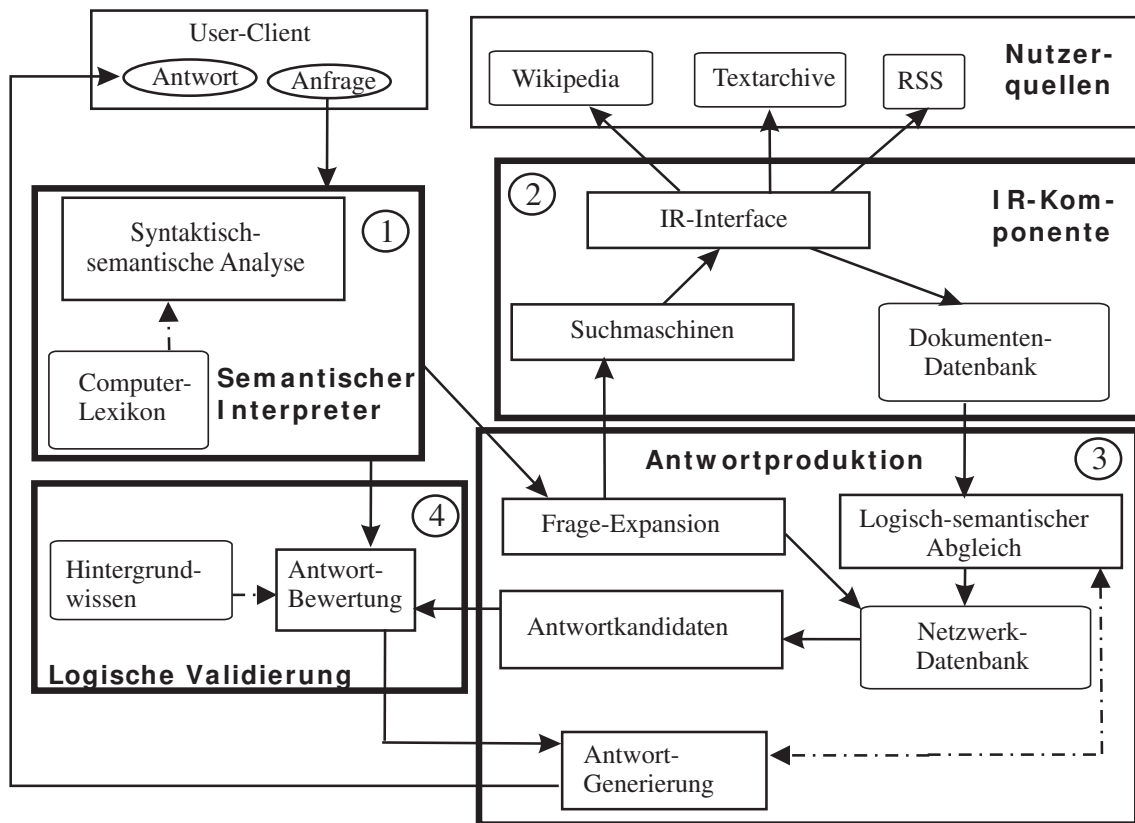


Abbildung 4: Aufbau und Wirkungsweise von SEMPRIA® Search

1. einem semantischen Interpreter: Er übersetzt die Nutzeranfragen, die über eine Nutzerschnittstelle (*user client*) eingegeben werden, mit Hilfe der Informationen aus dem Computerlexikon in MultiNet-Bedeutungsstrukturen.
2. einer Information-Retrieval-Komponente, die auf der Basis einer semantisch angereicherten Frage (Einbeziehung von Synonymen, Oberbegriffen, Entailments (s.u.) u.a.) mit Hilfe verschiedener Spezial-Suchmaschinen in den Nutzerquellen nach passenden Antwortkandidaten und zugehörigen Belegstellen sucht.
3. einer Komponente zur Antwortproduktion: Hier erfolgt ein logisch semantischer Abgleich mit der Frage. Dieser ist erforderlich, da unter den Suchmaschinen aus Komponente (2) auch sogenannte flache Verfahren als Rückfall-Systeme eingesetzt werden, die u.U. unpräzise oder gar nicht passende Antwortkandidaten liefern (wie das bei Standardsuchmaschinen auch der Fall ist).
4. einer Validierungskomponente: Die in Komponente (3) erzeugten Antwortkandidaten werden nach ihrer logischen Qualität bewertet (d.h. es wird untersucht, bis zu welchem Grad Frage und Antwort logisch zueinander passen oder voneinander abweichen). Auch bei diesem Prozess wird wieder Hintergrundwissen eingesetzt, um schließlich eine **logisch begründete** Rangordnung (*ranking*) zwischen den Antworten zu ermitteln. Dies ist ein weiterer Unterschied zu traditionellen Suchmaschinen, deren Ranking oft von der Zahl der Links bestimmt wird, die auf ein bestimmtes Dokument verweisen.

Wie die Daten des Nutzers in das System integriert werden, zeigt der folgende Abschnitt.



5 Aufbau des Dokumentenarchivs als Datenbasis

Typischerweise werden die Dokumentenarchive der Anwender über das Internet an SEMPRIA übertragen. Dabei können die Dokumente aus den Textarchiven der Nutzer in verschiedensten Formaten semantisch aufbereitet und in die Suche einbezogen werden. Dazu ist eine Vorverarbeitung nötig, die üblicherweise offline erfolgt und als Indexieren bezeichnet wird. Beim Indexieren mit SEMPRIA® Search werden – anders als bei traditionellen Suchmaschinen – aufwändige und komplexe Prozesse angestoßen, die auf ein möglichst weitgehendes Sprachverstehen abzielen, wie sie in den vorangehenden Abschnitten beschrieben wurden. Die Dokumente werden in Übereinstimmung mit dem MultiNet-Formalismus in semantische Netze übersetzt und mit dem bereits vorhandenen Wissen verbunden (Datenintegration in die Dokumentenbasis). Mit diesen in sich kohärenten Netzen kann die Suchmaschine bei der Recherche später rechnen, also logisch vergleichen und Inhalte semantisch erschließen. Beim Aufbau des Dokumentenarchivs werden kundenspezifische Informationen und allgemeines Hintergrundwissen technisch klar voneinander getrennt, so dass der Schutz der kundeneigenen Daten jederzeit gewährleistet ist.

SEMPRIA® Search kann zur Zeit folgende Formate problemlos integrieren (diese Liste ließe sich aber auf Kundenwunsch ohne weiteres erweitern):

- reiner Text (kodiert in Zeichensätzen wie ASCII, Latin-1, UTF-8 ...)
- HTML
- DOC, ODF, RTF u.a.
- PDF
- PostScript, DVI u.a.

Eine umfassende und saubere Metadaten-Integration ist in SEMPRIA® Search selbstverständlich. In wenigen Fällen sind noch manuelle Arbeitsschritte beim Integrieren der Metadaten oder der Dokumente nötig. Dazu gehört die Eingabe von Zusatzinformationen bei bestimmten Formaten (wie Tabellen) und die Korrektur von Fehlern beim Einsatz von Texterkennung (OCR).

Die Erstindexierung und Aktualisierung der Dokumente erfolgt zu vereinbarten Zeitpunkten und in definierten Abständen. Sie werden durch Übermittlung einer einfachen URL-Liste angestoßen, wobei zur Übertragung die Protokolle HTTP, HTTPS, FTP und SFTP verwendet werden.

6 Leistungen und Anwendernutzen

Zusammenfassend kann man feststellen, dass der Einsatz von SEMPRIA® Search für den Anwender einen deutlichen Fortschritt gegenüber traditionellen Suchmaschinen bringt. Hier sollen nur die wichtigsten Aspekte hervorgehoben werden:

- Das System gewährleistet einen natürlichsprachlichen Zugang, dem der neueste Stand der Forschung auf den Gebieten der Wissensverarbeitung, Computerlinguistik und Computerlogik zugrunde liegt.



- Durch die logisch-linguistische Fundierung und das Erreichen einer wirklich tiefen semantischen Sprachverarbeitung können sprachliche Phänomene berücksichtigt werden, die weit außerhalb des Leistungsspektrums traditioneller Suchmaschinen liegen. Hierzu gehören:
 - die Beherrschung von Mehrwortausdrücken (die als semantische Einheit erkannt werden);
 - die automatische Auflösung von Mehrdeutigkeiten (lexikalische – Mehrdeutigkeit von Wörtern; strukturelle – mehrere Möglichkeiten, Satzteile einander zuzuordnen)
 - das Verstehen von Metonymien (*Washington protestiert . . .*), von idiomatischen Wendungen (*das Handtuch werfen* für *aufgeben*) und von Funktionsverbgefügen (*zum Abschluss bringen* für *abschließen/beenden*);
 - die korrekte Verarbeitung von Zeitangaben (sowohl absolute als auch relative, wie *am 9.11.1989* bzw. *gestern*) sowie von Zahlen und Maßen;
 - die Auflösung von Referenzen (z.B. Bezüge von Pronomen), auch über Satzgrenzen hinweg;
 - das Verstehen von Beziehungen zwischen Objekten und der Rolle von Beteiligten in Ereignissen (z.B. dass der Handlungsträger einer Handlung *singen* ein Sänger ist) und ihre richtige Behandlung beim Beantworten von Fragen;
 - die Konstituierung von semantischen Beschreibungen für Objekte und Sachverhalte aus mehreren Sätzen und Dokumenten;
 - die Generierung semantischer Suchvorschläge aus den Dokumenten.
- Das System arbeitet wissensbasiert, d.h. in die Suche kann auch Hintergrundwissen (sprachliches Wissen oder sogenanntes Weltwissen) einbezogen werden. Dieses Wissen wird unabhängig vom konkreten Anwender von der SEMPRIA GmbH (zum großen Teil sogar automatisch) akkumuliert. Dadurch profitieren die Nutzer von jeder Erweiterung der Wissensbasis oder des Computerlexikons, ohne ein Update ihrer Anwendungssoftware durchführen zu müssen.
- SEMPRIA® Search kann durch den Einsatz verschiedener Korrekturmodule (Rechtschreibung, Zusammenschreibung, etc.) für den Nutzer fehlerrobuster gemacht werden.
- Insgesamt lassen sich durch den Einsatz modernster Sprachtechnologien Genauigkeit und Vollständigkeit der Suche bedeutend verbessern und die Zufriedenheit und Effizienz der Nutzer erheblich steigern.
- *Last but not least* eröffnet der Einsatz einer linguistisch fundierten Sprachverarbeitungstechnologie strategisch den Anschluss zur akustischen Sprachverarbeitung (Stichwort: Zugang zu Datenarchiven über Smartphone) und Möglichkeiten zum Anschluss weiterer Applikationen wie Stimmungsanalyse (*sentiment analysis, opinion mining*), Themen-Spotting, semantische Duplikatserkennung, Lesbarkeitstests und viele weitere.

Literatur

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

