# Natural Language Understanding Applied to Research (an introduction and demo)

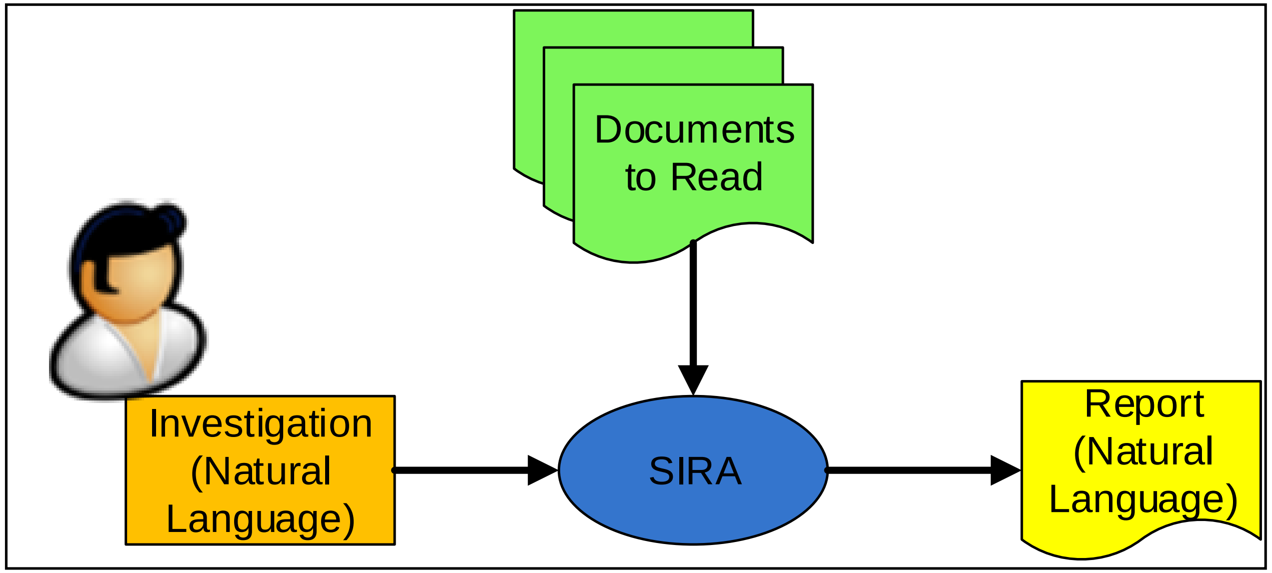
Chuck Rehberg

Chief Scientist, Semantic Insights

# Semantic Insights Research Assistant (SIRA) Technology

- The Semantic Insights Research Assistant (SIRA) Technology was developed to address the need to:
  - Capture and apply enough human knowledge to automate reading Natural Language prose to:
    - Read a potentially a vast number of documents,
    - Look for specific information of interest and
    - Render the findings in useful ways
- Along the way, we found ourselves in the thick of Natural Language Understanding, Reasoning, Machine Learning, Discovery, Natural Language generation, Big Data, Semantic Web, and Graph Databases (e.g. Triplestores).
- For more information visit: http://www.semanticinsights.com

# View of SIRA by General Users



1. An *Investigation* (i.e. query) is stated in Natural Language
2. The corpus of *documents* to “machine read” are identified
3. The kind of *report* is selected and generated

# About “The Investigation”

- *The Investigation* (actually the “query”) is a natural language text (English for now) consisting of one or more sentences.
- These sentences can also contain *open variables*
  - For example, the term “#?#” stands for an “unknown thing”
  - Sample usage: “#?# is traveling to Tunisia on Tuesday.”
- Sentences can also contain *closed variables*
  - For example; let #my-equities# refer to the list of stocks I am interested in.
  - Sample usage: “the price of #my-equities# fell.”

# About “The Source Documents”

- Must contain some form of natural language text
- Can be encoded in .html, .docx, .rtf, .txt, .pptx., .pdf (non-image), others
- Text can come from databases, email, web pages, documents, and social media feeds.
- Source Documents are machine read line-by-line/ sentence-by-sentence
- Source Documents are not preprocessed

# About "The Report"

- The Kind of Report is selected
  - The format of the report is taken from a template that has been previously created via a SIRA Report Editor.
- SIRA can currently generate reports in these renderings (.txt, .csv, .html, .pdf); others can be easily created
- The most common general report contains verbatim Natural Language text with Bibliography.

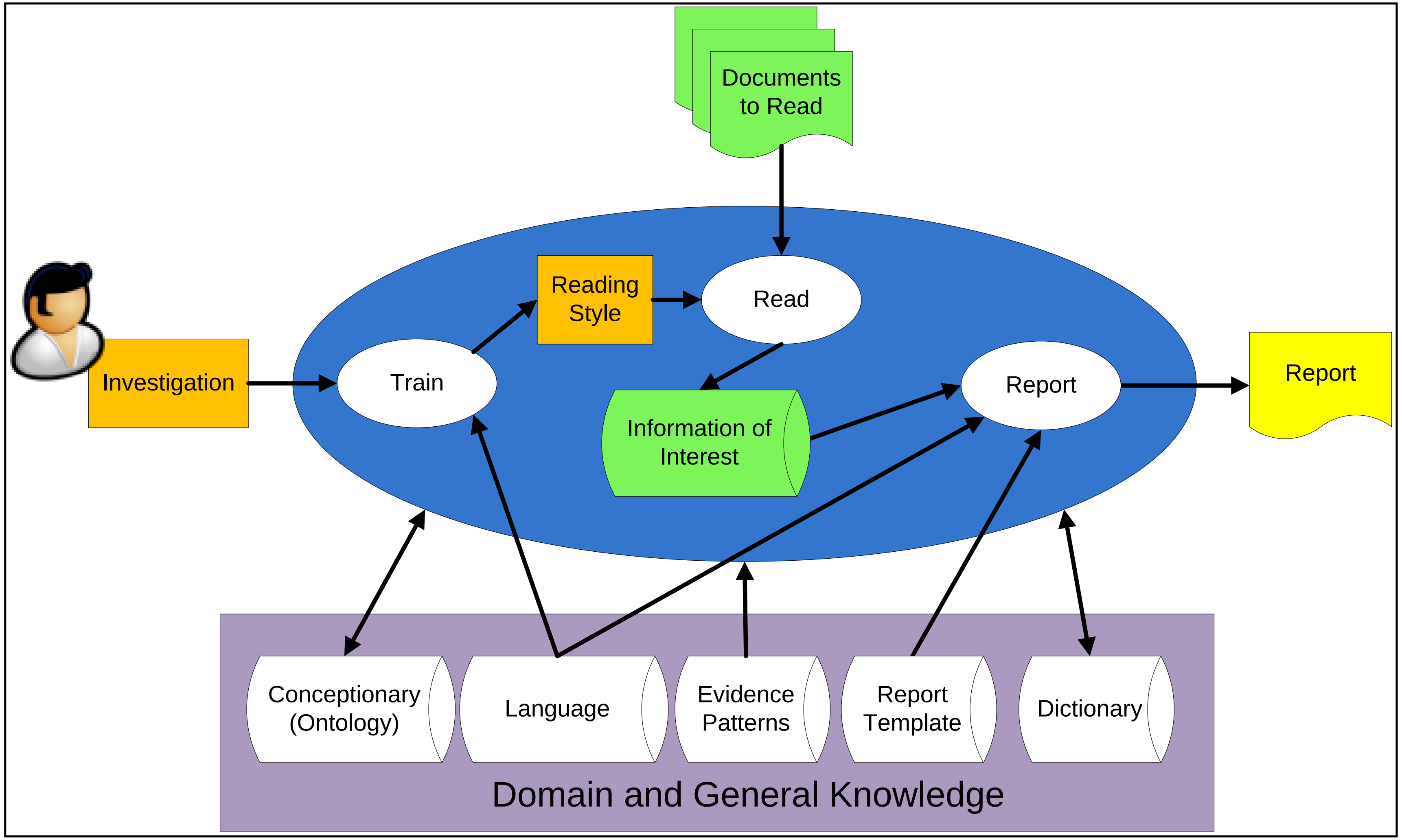
# Demos (two basic on-line tools)

- Research Assistant™ is a Google Chrome™ plugin that
  - Starting from the current web page, and a research description you provide,
  - Research Assistant “reads” the given web page + links + links of links and
  - generates a research report with bibliography
  - Demo
    - “What causes Autism?”
    - http://en.wikipedia.org/wiki/Autism
- Research Librarian™ is a website that
  - Starting from a selected set of information sources, and a research description you provide,
  - Research Librarian “reads” each document in the selected source + links and
  - Generates a research report with bibliography
  - Demo
    - Example: “What inhibits CYP2D6?”
    - http://52.205.34.110:8082/ResearchAssistant/loginResearchLibrarian

# Another on-line tool with more information

- PriArt is an on-line Research tool that
  - Gather's information based on your research statements and
  - Generates research reports with bibliography and hyperlinks containing information relevant to your research
  - PriArt provides more control of language and presentation than Research Assistant or Research Librarian
- Demo
  - http://www.autism-society.org/what-is/

# A little more detail



# Unsupervised Learning of a given Domain

- *Domain Knowledge Bootstrap*

  1. Load pre-existing domain-dictionaries
  2. Run the *Unrecognized Term Identifier* on a set of domain-specific documents to identify terms missing from the dictionary and looks them up terms in common and domain-specific web dictionaries.
  3. Use the *Hypernym Recognizer* to generate hierarchy relationships in the Ontology and add the terms to the dictionary.
  4. Use a *set of domain-specific documents as the investigation* to bootstrap the domain dictionary (terms) and Ontology (relationships)

- *To learn what else is known about a given topic*

  1. Start with a set of statements of interest (i.e. the initial Investigation)
  2. Read a corpus of documents
  3. Add the relevant sentences in the results to the current investigation and re-read the corpus
  4. Continue this process until condition is reached (e.g. no new relevant sentences)

# About reading the Internet…

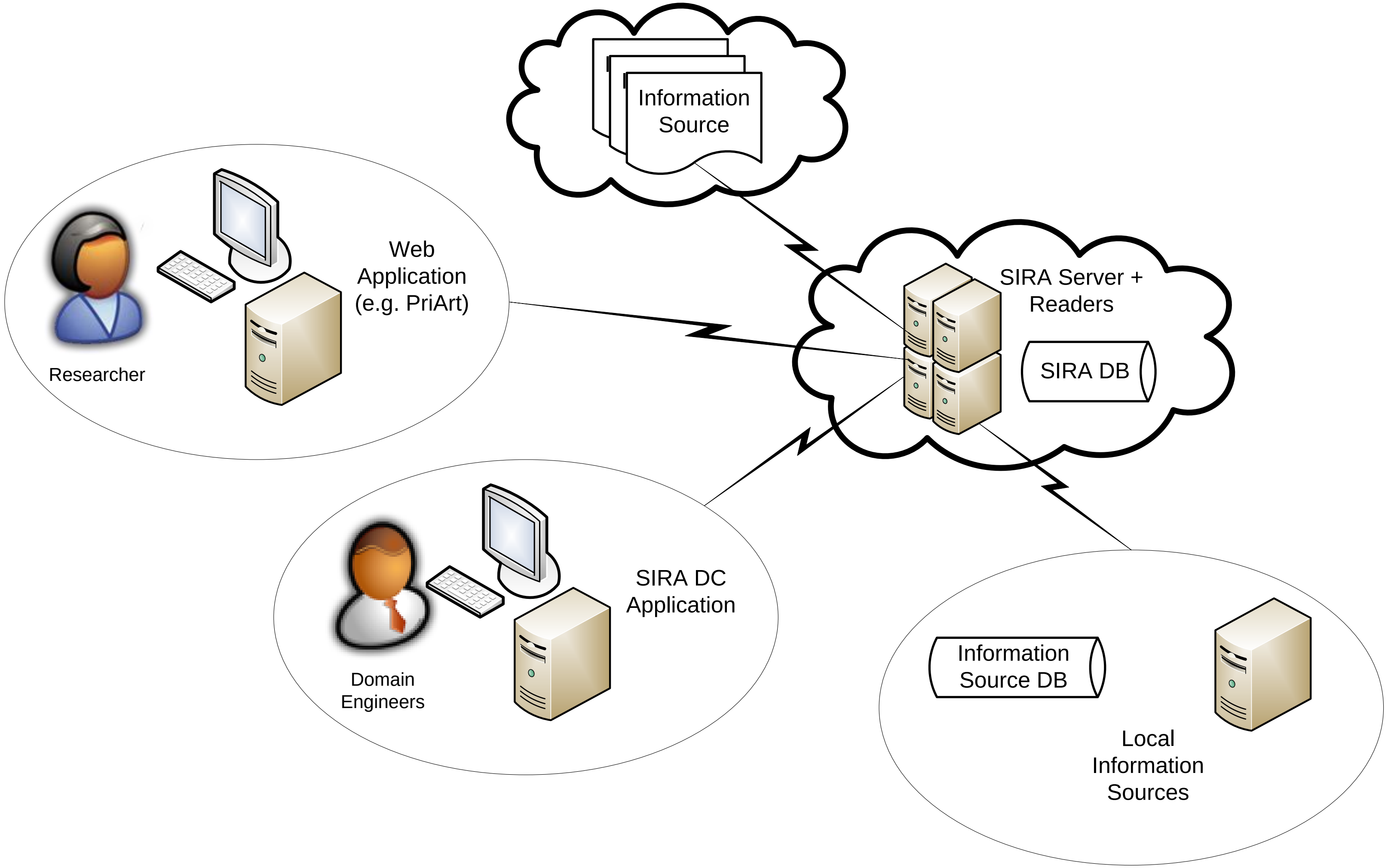
- PriArt doesn't really read the whole internet. Of course that would take a long time. The main limiting factor is bandwidth.
- However, if you want PriArt to effectively read a large document corpus (like the internet), you can direct PriArt to automatically gather relevant documents by using a keyword search engine.
- PriArt will create and execute a search strategy for you. This includes executing a number of keyword queries, each designed to find the most relevant documents based on examining your statement of investigation.
- You still may get millions of documents to read. So, you can direct PriArt to read a specific number of most relevant documents found according to the search engine ranking.

# Learning as you go

- Every time you provide SIRA with an investigation, the investigation is subjected to deep linguistic analysis
- This analysis can add to the Dictionary and create new relationships in the Ontology
- Other forms on experienced-based learnings are under development

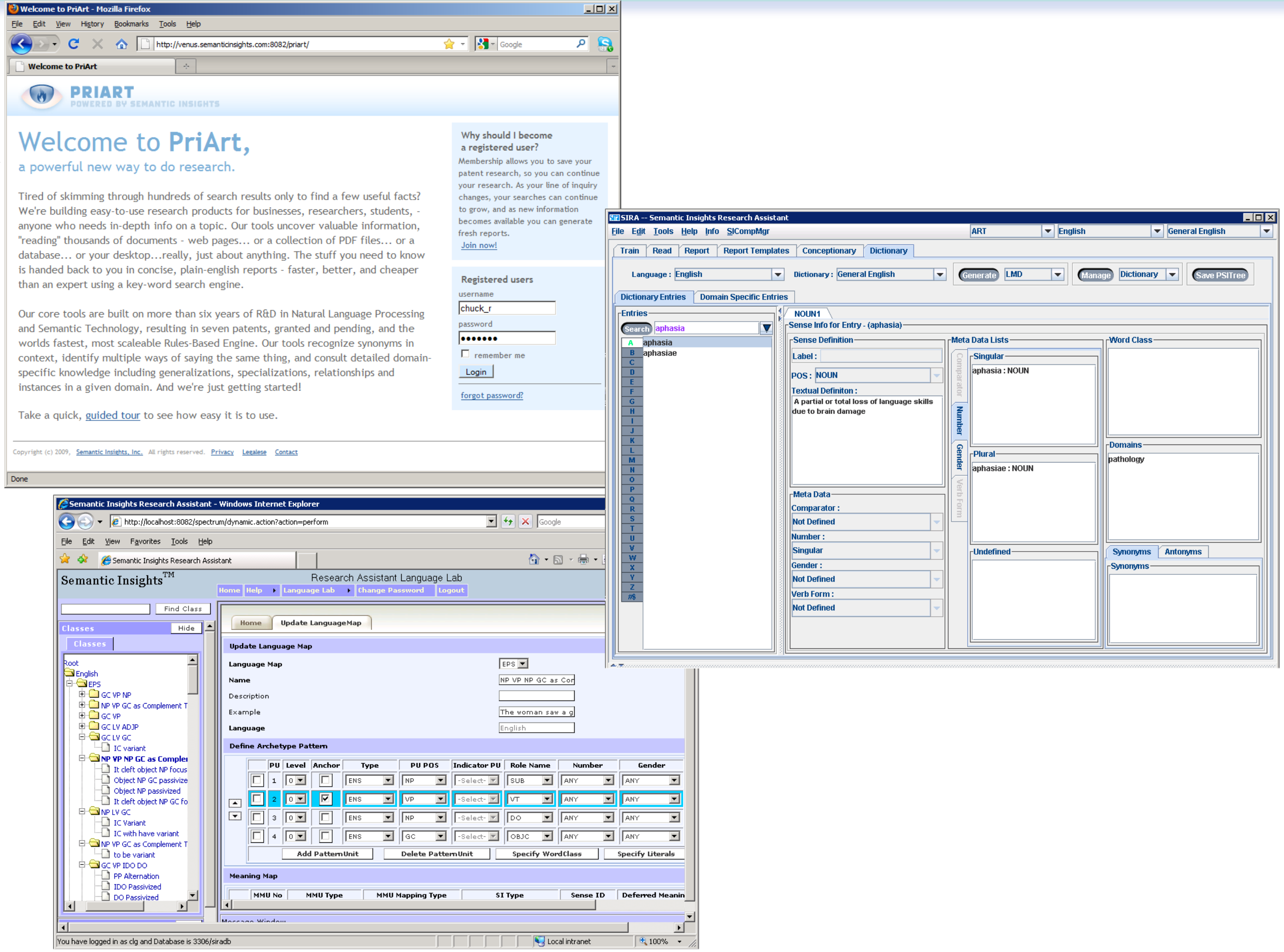
# Physical SIRA Configuration

- SIRA Server
- SIRA Readers
- Web Applications
- SIRA Development Center (thick Client)
- Multiple Information Sources



# Development Tools Provided with SIRA

1. **PriArt**: A Semantic Research Assistant
   - Browser-based UI
   - Used to demonstrate and test via Browser Interface
2. **SIRA Development Center**
   - Desktop Application
   - Used to develop and manage World Knowledge
     - Ontologies, Dictionaries, Testing and Training
   - Used to develop and manager Report Templates
3. **Language Lab**
   - Browser-based UI
   - Used to define Language and Genre
   - Syntax, Grammar and Meaning Maps



Welcome to PriArt - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://venus.semanticinsights.com:8082/priart/
Google
Welcome to PriArt
PRIART
POWERED BY SEMANTIC INSIGHTS

Welcome to PriArt,
a powerful new way to do research.

Tired of skimming through hundreds of search results only to find a few useful facts? We're building easy-to-use research products for businesses, researchers, students, - anyone who needs in-depth info on a topic. Our tools uncover valuable information, "reading" thousands of documents - web pages… or a collection of PDF files… or a database… or your desktop…really, just about anything. The stuff you need to know is handed back to you in concise, plain-english reports - faster, better, and cheaper than an expert using a key-word search engine.

Our core tools are built on more than six years of R&D in Natural Language Processing and Semantic Technology, resulting in seven patents, granted and pending, and the worlds fastest, most scaleable Rules-Based Engine. Our tools recognize synonyms in context, identify multiple ways of saying the same thing, and consult detailed domain-specific knowledge including generalizations, specializations, relationships and instances in a given domain. And we're just getting started!

Take a quick, guided tour to see how easy it is to use.

Why should I become a registered user?
Membership allows you to save your patent research, so you can continue your research. As your line of inquiry changes, your searches can continue to grow, and as new information becomes available you can generate fresh reports.
Join now!

Registered users
username: chuck_r
password: •••••••
□ remember me
Login
forgot password?

Copyright (c) 2009, Semantic Insights, Inc. All rights reserved. Privacy Legalese Contact

Done

SIRA -- Semantic Insights Research Assistant
File Edit Tools Help Info SICompMgr
ART
English
General English
Train Read Report Report Templates Conceptionary Dictionary
Generate
Manage
Save PSITree
Dictionary Entries
Domain Specific Entries
Entries
Search
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z #$
NOUN1
Comparator
Number
Gender
Verb Form
Synonyms
Antonyms

Language : English
Dictionary : General English
LMD
Dictionary

Search: aphasia
aphasia
aphasiae

Sense Info for Entry - (aphasia)

Sense Definition
Label :
POS : NOUN
Textual Definiton : A partial or total loss of language skills due to brain damage

Meta Data
Comparator : Not Defined
Number : Singular
Gender : Not Defined
Verb Form : Not Defined

Meta Data Lists
Singular: aphasia : NOUN
Plural: aphasiae : NOUN
Undefined:

Word Class:
Domains: pathology
Synonyms:

Semantic Insights Research Assistant - Windows Internet Explorer
http://localhost:8082/spectrum/dynamic.action?action=perform
Google
File Edit View Favorites Tools Help
Semantic Insights Research Assistant
Semantic Insights™
Research Assistant Language Lab
Home Help Language Lab Change Password Logout
Find Class
Classes Hide
Classes
Home
Update LanguageMap
Add PatternUnit
Delete PatternUnit
Specify WordClass
Specify Literals
Message Window
Local intranet
100%

Root
English
EPS
GC VP NP
NP VP GC as Complement T
GC VP
GC LV ADJP
GC LV GC
IC variant
NP VP NP GC as Compler
It cleft object NP focus
Object NP GC passivize
Object NP passivized
It cleft object NP GC fo
NP LV GC
IC Variant
IC with have variant
NP VP GC as Complement T
to be variant
GC VP IDO DO
PP Alternation
IDO Passivized
DO Passivized

Update Language Map

| | |
|---|---|
| Language Map | EPS |
| Name | NP VP NP GC as Cor |
| Description | |
| Example | The woman saw a g |
| Language | English |

Define Archetype Pattern

| | PU | Level | Anchor | Type | PU POS | Indicator PU | Role Name | Number | Gender |
|---|---|---|---|---|---|---|---|---|---|
| □ | 1 | 0 | □ | ENS | NP | -Select- | SUB | ANY | ANY |
| □ | 2 | 0 | ☑ | ENS | VP | -Select- | VT | ANY | ANY |
| □ | 3 | 0 | □ | ENS | NP | -Select- | DO | ANY | ANY |
| □ | 4 | 0 | □ | ENS | GC | -Select- | OBJC | ANY | ANY |

Meaning Map

| MMU No | MMU Type | MMU Mapping Type | SI Type | Sense ID | Deferred Meanin |
|---|---|---|---|---|---|

You have logged in as clg and Database is 3306/siradb

SEMANTIC INSIGHTS™

A Division of Trigent Software, Inc.