Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 36 (2003) 414–432
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Exploring semantic groups through visual approaches

Olivier Bodenreider* and Alexa T. McCray

Department of Health and Human Services, National Institutes of Health, National Library of Medicine, Lister Hill National Center for Biomedical Communications, MS 43, Bldg 38A Rm B1N28U, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received 30 October 2003

Abstract

Objectives. We investigate several visual approaches for exploring semantic groups, a grouping of semantic types from the Unified Medical Language System (UMLS) semantic network. We are particularly interested in the semantic coherence of the groups, and we use the semantic relationships as important indicators of that coherence.

Methods. First, we create a radial representation of the number of relationships among the groups, generating a profile for each semantic group. Second, we show that, in our partition, the relationships are organized around a limited number of pivot groups and that partitions created at random do not exhibit this property. Finally, we use correspondence analysis to visualize groupings resulting from the association between semantic types and the relationships.

Results. The three approaches provide different views on the semantic groups and help detect potential inconsistencies. They make outliers immediately apparent, and, thus, serve as a tool for auditing and validating both the semantic network and the semantic groups.

Keywords: Unified Medical Language System; Semantic network; Semantic relationships; Information visualization; Information exploration; Graph; Correspondence analysis

1. Introduction

Early in the Unified Medical Language System (UMLS) project,[1] we developed the UMLS semantic network in an effort to provide a semantic framework for the UMLS and its constituent vocabularies [1]. The current semantic network[2] consists of 134 semantic types[3] and 54 relationships, and it is expressed through two single-inheritance hierarchies, one for entities and another for events. The *isa* link allows nodes (i.e., semantic types) to inherit properties from higher-level nodes. In addition, there are five categories of associative relationships that interrelate the semantic types. A particular associative relationship may be physical (e.g., *connected_to*), functional (e.g., *causes*), spatial (e.g., *traverses*), temporal (e.g., *co-occurs_with*) or conceptual (e.g., *degree_of*). In the UMLS, semantic types are used to categorize the currently more than 800,000 concepts in the Metathesaurus, which interrelates some 60 families of vocabularies in the biomedical domain. While inter-concept relationships in the Metathesaurus generally instantiate specific knowledge, such as "kidney *location_of* nephroblastoma," semantic network relations represent general, high-level knowledge, such as "`Body Part`, `Organ`, or `Organ Component` *location_of* `Neoplastic Process`."

For some purposes, it is useful to classify the semantic types into a smaller number of semantic groups. In earlier work, we established fifteen high-level semantic groups that help reduce the conceptual complexity of the large domain covered by the UMLS [2] (see also [3] for a different attempt to partition the UMLS semantic network). Groupings of semantic types—the semantic groups—may prove to be useful in a number of applications including improved visualization

* Corresponding author. Fax: 1-301-480-3035.
E-mail address: olivier@nlm.nih.gov (O. Bodenreider).

[1] Information on the UMLS is available at this web site: umlsinfo.nlm.nih.gov. Unified Medical Language System (UMLS) and Metathesaurus are registered trademarks of the National Library of Medicine.

[2] Version 2002AC of the UMLS.

[3] A 135th semantic type, Drug Delivery Device, was added to the UMLS semantic network shortly after this study was performed.

1532-0464/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2003.11.002

and display of the knowledge in a particular domain [4]; natural language processing, where higher level categories are sometimes sufficient for semantic processing [5]; and auditing a domain for the valid representation of concepts and their interrelationships [6]. For example, if a particular concept in the UMLS has been assigned multiple semantic types and this assignment leads to the concept appearing in two different high-level groups, then it is possible that at least one of the semantic type assignments is incorrect. In our earlier work, we subjected the entire set of concepts in the 2000 version of the UMLS to this test, and we found a number of semantic type assignment errors through this method.

1.1. Grouping the semantic types

The groupings we established were subject to a set of general principles including, **semantic validity** (the groups must be semantically coherent); **parsimony** (the number of groups should be as small as possible[4]); **completeness** (the groups must cover the full domain); **exclusivity** (each concept in the domain must belong to only one group); **naturalness** (the groups must characterize the domain in a way that is acceptable to a domain expert); and **utility** (the groups must be useful for some purpose). Table 1 shows the groups that resulted from applying these principles.

The first column of Table 1 gives the name of the group, the second gives its abbreviation, the third lists the number of semantic types in that group, and the fourth lists the names of all of the semantic types that are members of that group. There is a variable number of semantic types in each group.[5] For example, both Chemicals & Drugs and Living Beings have a relatively large number of members, 26 and 23, respectively, while some groups like Devices and Occupations have only two members. Fig. 1 shows the distribution of semantic groups across the entire semantic network.

The left-hand side of each column in Fig. 1 lists the semantic types as well as displaying the hierarchical structure of the network. The right-hand side of the column shows the group to which the particular semantic type belongs. For example, the semantic type `Plant` belongs to the group Living Beings. An inspection of Fig. 1 shows that, in many cases, semantic types that are hierarchically related are also placed in the same group. For example, all of the chemicals are hierarchically related to each other, and they are also all in the group called Chemicals & Drugs. In other cases, a particular sub-tree in the semantic network has semantic types that are usefully placed in different groups. For example, the semantic types under `Phenomenon or Process` participate in three different groups, Phenomena (e.g., `Natural Phenomenon or Process`), Physiology (e.g., `Cell Function`), and Disorders (e.g., `Disease or Syndrome`). The group Disorders is interesting because it takes its members from several different trees in the semantic network. Anatomical abnormalities, for example, while they are, strictly speaking, anatomical structures, also share many of the same characteristics as disease processes. For example, an abscess is a physical entity that can be removed, and at the same it is a treatable disease. Likewise, injuries such as a leg fracture and poisonings such as carbon monoxide poisoning, while not pathologic functions, also share some of the characteristics of other disorders.

Among all of the principles we used to establish the groups, semantic validity is perhaps the most important one. In fact, without semantic coherence, it is hard to see how useful such groupings would be for any purpose. Assessing semantic coherence and validity, however, is not straightforward. One possible measure of coherence, to which we alluded in our previous work, is to analyze the relationships in which the semantic groups participate. These include not only the hierarchical relationship (*isa*), but also the many associative relationships observed in the biomedical domain (e.g., *treats*, *location_of*, *measures*). We would expect that many of the same relationships would be relevant for each of the members in a group, and also that there would be some consistency in the relationships that obtain across groups. For example, it would seem reasonable that all living beings would exhibit behaviors. Thus, if we find that a member of the group Living Beings does not share in the relationship *exhibits* with some member of the group Activities & Behaviors, we would find that surprising, and we would want to know what the reason for the anomaly was. In the following we inspect the full set of relationships between the semantic groups and explore these relationships through visual approaches.

Semantic Network relations can be represented as ordered triplets (ST_1, *rel*, ST_2), where *rel* is the relationship of semantic type ST_1 to semantic type ST_2. Examples of Semantic Network relations include (`Fully Formed Anatomical Structure`, *location_of*, `Biologic Function`), (`Pathologic Function`, *isa*, `Biologic Function`), and (`Pharmacologic Substance`, *treats*, `Pathologic Function`). The UMLS file SRSTR represents a total of 558 (ST_1, *rel*, ST_2) relations. By convention, inverse relations such as (ST_2, *inv_rel*, ST_1) where *inv_rel* is the inverse of *rel* are omitted from the file. Inverse

[4] Although many biomedical knowledge representation systems use 10–20 top-level categories, there is of course no absolute numerical bound on parsimony, the "ideal" number of groups being dependent on the purpose. For example, the medical subject headings (MeSH), developed for information indexing and retrieval, has 15 top-level classes, while SNOMED-CT, used for representing clinical information, has 19.

[5] The equal size of the groups is a criterion used in many partitioning and clustering algorithms. In the context of our semantic groups, however, we favored semantic coherence over equal size.

Table 1
List of semantic groups with semantic type members

Semantic Groups		Semantic Types	
Activities & Behaviors	ACTI	9	■ Activity ■ Behavior ■ Daily or Recreational Activity ■ Event ■ Governmental or Regulatory Activity ■ Individual Behavior ■ Machine Activity ■ Occupational Activity ■ Social Behavior
Anatomy	ANAT	11	■ Anatomical Structure ■ Body Location or Region ■ Body Part, Organ, or Organ Component ■ Body Space or Junction ■ Body Substance ■ Body System ■ Cell ■ Cell Component ■ Embryonic Structure ■ Fully Formed Anatomical Structure ■ Tissue
Chemicals & Drugs	CHEM	26	■ Amino Acid, Peptide, or Protein ■ Antibiotic ■ Biologically Active Substance ■ Biomedical or Dental Material ■ Carbohydrate ■ Chemical ■ Chemical Viewed Functionally ■ Chemical Viewed Structurally ■ Clinical Drug ■ Eicosanoid ■ Element, Ion, or Isotope ■ Enzyme ■ Hazardous or Poisonous Substance ■ Hormone ■ Immunologic Factor ■ Indicator, Reagent, or Diagnostic Aid ■ Inorganic Chemical ■ Lipid ■ Neuroreactive Substance or Biogenic Amine ■ Nucleic Acid, Nucleoside, or Nucleotide ■ Organic Chemical ■ Organophosphorus Compound ■ Pharmacologic Substance ■ Receptor ■ Steroid ■ Vitamin
Concepts & Ideas	CONC	12	■ Classification ■ Conceptual Entity ■ Functional Concept ■ Group Attribute ■ Idea or Concept ■ Intellectual Product ■ Language ■ Qualitative Concept ■ Quantitative Concept ■ Regulation or Law ■ Spatial Concept ■ Temporal Concept
Devices	DEVI	2	■ Medical Device ■ Research Device
Disorders	DISO	12	■ Acquired Abnormality ■ Anatomical Abnormality ■ Cell or Molecular Dysfunction ■ Congenital Abnormality ■ Disease or Syndrome ■ Experimental Model of Disease ■ Finding ■ Injury or Poisoning ■ Mental or Behavioral Dysfunction ■ Neoplastic Process ■ Pathologic Function ■ Sign or Symptom
Genes & Molecular Sequences	GENE	5	■ Amino Acid Sequence ■ Carbohydrate Sequence ■ Gene or Genome ■ Molecular Sequence ■ Nucleotide Sequence
Geographic Areas	GEOG	1	■ Geographic Area
Living Beings	LIVB	23	■ Age Group ■ Alga ■ Amphibian ■ Animal ■ Archaeon ■ Bacterium ■ Bird ■ Family Group ■ Fish ■ Fungus ■ Group ■ Human ■ Invertebrate ■ Mammal ■ Organism ■ Patient or Disabled Group ■ Plant ■ Population Group ■ Professional or Occupational Group ■ Reptile ■ Rickettsia or Chlamydia ■ Vertebrate ■ Virus
Objects	OBJC	5	■ Entity ■ Food ■ Manufactured Object ■ Physical Object ■ Substance
Occupations	OCCU	2	■ Biomedical Occupation or Discipline ■ Occupation or Discipline
Organizations	ORGA	4	■ Health Care Related Organization ■ Organization ■ Professional Society ■ Self-help or Relief Organization
Phenomena	PHEN	6	■ Biologic Function ■ Environmental Effect of Humans ■ Human-caused Phenomenon or Process ■ Laboratory or Test Result ■ Natural Phenomenon or Process ■ Phenomenon or Process
Physiology	PHYS	9	■ Cell Function ■ Clinical Attribute ■ Genetic Function ■ Mental Process ■ Molecular Function ■ Organ or Tissue Function ■ Organism Attribute ■ Organism Function ■ Physiologic Function
Procedures	PROC	7	■ Diagnostic Procedure ■ Educational Activity ■ Health Care Activity ■ Laboratory Procedure ■ Molecular Biology Research Technique ■ Research Activity ■ Therapeutic or Preventive Procedure

Entity ..... OBJC
- Physical Object ..... OBJC
  - Organism ..... LIVB
    - Plant ..... LIVB
      - Alga ..... LIVB
    - Fungus ..... LIVB
    - Virus ..... LIVB
    - Rickettsia or Chlamydia ..... LIVB
    - Bacterium ..... LIVB
    - Archaeon ..... LIVB
    - Animal ..... LIVB
      - Invertebrate ..... LIVB
      - Vertebrate ..... LIVB
        - Amphibian ..... LIVB
        - Bird ..... LIVB
        - Fish ..... LIVB
        - Reptile ..... LIVB
        - Mammal ..... LIVB
          - Human ..... LIVB
  - Anatomical Structure ..... ANAT
    - Embryonic Structure ..... ANAT
    - Anatomical Abnormality ..... DISO
      - Congenital Abnormality ..... DISO
      - Acquired Abnormality ..... DISO
    - Fully Formed Anatomical Structure ..... ANAT
      - Body Part, Organ, or Organ Component ..... ANAT
      - Tissue ..... ANAT
      - Cell ..... ANAT
      - Cell Component ..... ANAT
      - Gene or Genome ..... GENE
  - Manufactured Object ..... OBJC
    - Medical Device ..... DEVI
    - Research Device ..... DEVI
    - Clinical Drug ..... CHEM
  - Substance ..... OBJC
    - Chemical ..... CHEM
      - Chemical Viewed Functionally ..... CHEM
        - Pharmacologic Substance ..... CHEM
          - Antibiotic ..... CHEM
        - Biomedical or Dental Material ..... CHEM
        - Biologically Active Substance ..... CHEM
          - Neuroreactive Substance or Biogenic Amine ..... CHEM
          - Hormone ..... CHEM
          - Enzyme ..... CHEM
          - Vitamin ..... CHEM
          - Immunologic Factor ..... CHEM
          - Receptor ..... CHEM
        - Indicator, Reagent, or Diagnostic Aid ..... CHEM
        - Hazardous or Poisonous Substance ..... CHEM
      - Chemical Viewed Structurally ..... CHEM
        - Organic Chemical ..... CHEM
          - Nucleic Acid, Nucleoside, or Nucleotide ..... CHEM
          - Organophosphorus Compound ..... CHEM
          - Amino Acid, Peptide, or Protein ..... CHEM
          - Carbohydrate ..... CHEM
          - Lipid ..... CHEM
            - Steroid ..... CHEM
            - Eicosanoid ..... CHEM
        - Inorganic Chemical ..... CHEM
        - Element, Ion, or Isotope ..... CHEM
    - Body Substance ..... ANAT
    - Food ..... OBJC

[Entity] (continued)
- Conceptual Entity ..... CONC
  - Idea or Concept ..... CONC
    - Temporal Concept ..... CONC
    - Qualitative Concept ..... CONC
    - Quantitative Concept ..... CONC
    - Functional Concept ..... CONC
      - Body System ..... ANAT
    - Spatial Concept ..... CONC
      - Body Space or Junction ..... ANAT
      - Body Location or Region ..... ANAT
      - Molecular Sequence ..... GENE
        - Nucleotide Sequence ..... GENE
        - Amino Acid Sequence ..... GENE
        - Carbohydrate Sequence ..... GENE
      - Geographic Area ..... GEOG
  - Finding ..... DISO
    - Laboratory or Test Result ..... PHEN
    - Sign or Symptom ..... DISO
  - Organism Attribute ..... PHYS
    - Clinical Attribute ..... PHYS
  - Intellectual Product ..... CONC
    - Classification ..... CONC
    - Regulation or Law ..... CONC
  - Language ..... CONC
  - Occupation or Discipline ..... OCCU
    - Biomedical Occupation or Discipline ..... OCCU
  - Organization ..... ORGA
    - Health Care Related Organization ..... ORGA
    - Professional Society ..... ORGA
    - Self-help or Relief Organization ..... ORGA
  - Group Attribute ..... CONC
  - Group ..... LIVB
    - Professional or Occupational Group ..... LIVB
    - Population Group ..... LIVB
    - Family Group ..... LIVB
    - Age Group ..... LIVB
    - Patient or Disabled Group ..... LIVB

Event ..... ACTI
- Activity ..... ACTI
  - Behavior ..... ACTI
    - Social Behavior ..... ACTI
    - Individual Behavior ..... ACTI
  - Daily or Recreational Activity ..... ACTI
  - Occupational Activity ..... ACTI
    - Health Care Activity ..... PROC
      - Laboratory Procedure ..... PROC
      - Diagnostic Procedure ..... PROC
      - Therapeutic or Preventive Procedure ..... PROC
      - Research Activity ..... PROC
        - Molecular Biology Research Technique ..... PROC
      - Governmental or Regulatory Activity ..... ACTI
      - Educational Activity ..... PROC
  - Machine Activity ..... ACTI
- Phenomenon or Process ..... PHEN
  - Human-caused Phenomenon or Process ..... PHEN
    - Environmental Effect of Humans ..... PHEN
  - Natural Phenomenon or Process ..... PHEN
    - Biologic Function ..... PHEN
      - Physiologic Function ..... PHYS
        - Organism Function ..... PHYS
          - Mental Process ..... PHYS
        - Organ or Tissue Function ..... PHYS
        - Cell Function ..... PHYS
        - Molecular Function ..... PHYS
          - Genetic Function ..... PHYS
      - Pathologic Function ..... DISO
        - Disease or Syndrome ..... DISO
          - Mental or Behavioral Dysfunction ..... DISO
          - Neoplastic Process ..... DISO
        - Cell or Molecular Dysfunction ..... DISO
        - Experimental Model of Disease ..... DISO
  - Injury or Poisoning ..... DISO

Fig. 1. Semantic types trees for `Activity` and `Event`, with semantic groups (the tree structure is represented by the indentations).

relationships are provided as part of the definition of the Semantic Network (e.g., *has_location* for *location_of*). Some relationships are their own inverse (e.g., *associated_with*). In SRSTR, relations are represented at the highest level possible and, unless otherwise specified, associative relationships are meant to be inherited along the *isa* hierarchy. For example, (Fully Formed Anatomical Structure, *location_of*, Pathologic Function) is not present in SRSTR, but can easily be inferred from (Fully Formed Anatomical Structure, *location_of*, Biologic Function) and (Pathologic Function, *isa*, Biologic Function). The fully developed Semantic Network, including inherited relationships, is found in the SRSTRE* files. There are 6703 (ST_1, *rel*, ST_2) relations represented in the SRSTRE* files.[6]

In a Semantic Network relation (ST_1, *rel*, ST_2), each relationship *rel* is related to two semantic types ST_1 and ST_2. And, since each semantic type belongs to a unique semantic group, a relationship can be seen as connecting two semantic groups through a relation (SG_1, *rel*, SG_2). For example, at the level of the semantic groups, the relation (Pharmacologic Substance, *treats*, Pathologic Function) becomes (Chemicals & Drugs, *treats*, Disorders).

1.2. Related work

Several techniques have been developed for exploratory data analysis. The most relevant technique is correspondence analysis, developed for studying the association among the categories of two variables [7]. Because correspondence analysis is essentially a geometric method, its results can be expressed in a two-dimensional graphical representation, useful for visual exploration. Thus, correspondence analysis appears as a method of choice for studying the association between semantic types and relationships in the composition of the semantic groups.

Also a logical fit for this study are various kinds of graph visualization techniques. The importance of underlying graph theoretical methods in visualization has been studied [8,9], and our goal here is essentially to apply existing techniques rather than to develop new ones. (All graphs were created using GraphViz,[7] a freely available drawing package.)

Visualization of knowledge structures such as hierarchies has been explored by several research groups, often using cone trees and hyperbolic trees. It is beyond the scope of this paper to provide an overview of the field, but we refer interested readers to a recent review [10], and, in the medical domain, to [11]. Some of these techniques are used to visualize hierarchical structures in the medical subject headings (MeSH) [12]. Many knowledge exploration tools, however, use levels of indentation to represent items in hierarchical relationship, the top-level items usually being represented on the left. Well adapted to limited hierarchical structures such as file systems, this layout is also used for displaying biomedical hierarchies in environments such as Protégé-2000 [13] and the Gene Ontology browser AmiGo.[8] Also frequently studied are networks of items related by associative relationships such as proteins or documents. In this case, the nature of the associative relationship is either constant (e.g., protein interaction from the yeast *Saccharomyces cerevisiae* [14], synonymy in a lexical database [15]) or not known with precision (e.g., relatedness of documents based on the frequency of co-occurrence of words or descriptors [16], link between initial and final visit diagnosis [17]). The resulting graphical representations may be complex because of the sheer number of nodes in the graph. However, limiting the display to a single kind of relationship makes the representation simpler.

Our proposed work differs from existing work in several ways. First, we do not restrict our study to one particular kind of relationship; we use various kinds of associative relationships as well as the taxonomic relationship. And second, the kind of relationship that obtain among the semantic groups does matter in this study. In fact, an important part of this study actually relies on the semantics of the relationships.

1.3. Presentation of the three experiments

We analyzed the semantic types, groups, and relationships from a variety of perspectives. First, we exhaustively examined each pair of semantic groups, determining the nature and number of relationships that obtained between each pair. This would give us a perspective on the contribution of the relationships to the semantic coherence of the groups. Next, we investigated the groups from the point of view of the relationships themselves. Our hypothesis is that, in most cases, a given relationship applies to only a limited number of groups. Finally, we looked at the interaction of the semantic types and relationships, addressing the question of whether semantic types that share relationships also cluster naturally into the same or similar groups.

In all cases we used visualization techniques to help us express and also evaluate our results. For the perspective of the pairs, we generated matrices of semantic groups as

[6] For the relationships that are their own inverse (e.g., *associated_with*), the SRSTRE* files contain two copies of the relation (Occupational Activity, *associated_with*, Injury or Poisoning) and (Injury or Poisoning, *associated_with*, Occupational Activity), one of which is ignored in this count.

[7] http://www.graphviz.org/.

[8] http://www.godatabase.org/.

well as visually compelling diagrams, based on a radial layout. For the perspective of the relationships, in addition to generating an overall matrix of relationships and semantic groups, we created graphical representations of the data for each relationship. Finally, to illustrate the interaction of the semantic types and relationships, we created a two-dimensional graphical display to show how semantic types cluster when viewed from the perspective of the relationships in which they participate.

2. Experiment 1: perspective of the pairs

2.1. Methods

Once semantic groups have been formed, it is interesting to examine each group with regards to its interaction with other groups. First, for each of our fifteen semantic groups we looked to see which and how many relationships connected that group to each of the other groups. In practice, for each pair of semantic groups (SG_1, SG_2), we examine the triplets (ST_1, *rel*, ST_2) where the semantic type ST_1 belongs to the semantic group SG_1 and ST_2 to SG_2. For each pair of semantic groups, we consider on the one hand the number of types of relationships *rel* that obtain between the two groups, and, on the other, the number of triplets, providing an indication of the variety and strength of the relationships between the groups. Second, the connections that a given semantic group has with all of the other groups might give an interesting profile of that group, particularly when compared with the profiles of other groups. The strongest connections, in some cases, might be found within a group if the members of that group were linked by semantically related relationships. Finally, this method might help us discover outliers in the semantic relationships themselves. It could be the case that no relationships exist between a pair of groups, and this may be completely appropriate given the semantics of the groups. However, if we find that there are no relationships where some would be expected, then this is an indication that a change needs to be made to the semantic network itself. Similarly, the specific relationships that connect a pair of groups should be the expected ones, given the semantics of the two groups. If we find a relationship that looks unusual, this might be indication of an error in the semantic network.

As the first step in this investigation, we created two matrices of the semantic groups. The rows and columns are semantic groups and the values of each cell are the number of relationships that obtain between each of the groups. One matrix shows the number of triplets for each pair of groups. The other one shows the number of unique relationships for each pair of groups. Next, we derived a graphical representation from the matrix, showing a profile of each of the semantic groups with respect to all of the other groups. For these graphs, we used a radial layout, constraining the nodes (i.e., the fifteen semantic groups) to lie on a circle, with one semantic group at the center.

2.2. Results

2.2.1. Matrices of semantic groups

The matrices shown in Table 2 relate semantic groups to each other with respect to the number of relationships that obtain between the members of each pair of groups. Table 2A shows the total number of relationships (i.e., the number of triplets), while Table 2B shows the unique number (i.e., the number of types of relationships). Consider, for example, the last row of Table 2A. This shows that the group Procedures is related by 24 relationships to the group Activities & Behaviors, by 18 relationships to the group Anatomy, by 206 relationships to the group Chemicals & Drugs, and so on. Table 2B, on the other hand, represents the unique number of relationships between each pair of semantic groups. We see that the group Procedures shares two types of relationships with Activities & Behaviors, three with Anatomy, and four with Chemicals & Drugs. We also note that Procedures shares no relationships with the group Genes & Molecular Sequences.

2.2.2. Radial representation of semantic groups

Figs. 2 and 3 are radial diagrams that display all semantic groups in a constant circular arrangement. Each specific diagram then represents a different semantic group as the center of attention. For example, the diagram at the top of Fig. 2 has as its center the group Anatomy and represents the count of all the relationships that that group has with all other groups by the lines that radiate from the center. The top number is the number of unique relationships, and the number in parentheses is the total number.

The right-hand side of Fig. 2 shows the specific relationships between each pair of groups. Thus, for Anatomy there are 16 types of relationships between and among the semantic types that participate in the group Anatomy, i.e., relationships within the group Anatomy, listed under the heading ANAT–ANAT. The total number of relationships within Anatomy is 115, and the contribution that each relationship type makes to this total is also listed, e.g., there are 13 triplets involving the relationship *adjacent_to* within the group Anatomy. Analogously, there are 4 types of relationships between the groups Anatomy and Chemicals & Drugs, (*consists_of*, *disrupted_by*, *ingredient_ of*, and *produces*) with a total of 144 triplets. In this case the largest number of triplets involve the relationships *disrupts* and *produces*. For ease of understanding, we have listed the relationship name with the appropriate directionality. Thus, for

Table 2
Matrix of SG by SG (2A: all relationships, 2B: unique relationships)

A

	ACTI	ANAT	CHEM	CONC	DEVI	DISO	GENE	GEOG	LIVB	OBJC	OCCU	ORGA	PHEN	PHYS	PROC	Total
ACTI	29			13		77		3	111		28	16	5	28	24	334
ANAT		115	144	11		104	8		147	20	22		10	74	18	673
CHEM		144	413			604	27		64	80	52		92	364	206	2046
CONC	13	11		21		2	16	4	54	12	26	20	1	27	17	224
DEVI						38			24	6	4				6	78
DISO	77	104	604	2	38	575	11	10	418	44	24		213	465	207	2792
GENE		8	27	16		11	7		21	6	10		1	20		127
GEOG	3			4		10				1	2			1		21
LIVB	111	147	64	54	24	418	21		212	52	48	6	34	270	44	1505
OBJC		20	80	12	6	44	6	1	52	8	12	4	4	9	3	261
OCCU	28	22	52	26	4	24	10	2	48	12	6	8	12	18	28	300
ORGA	16			20					6	4	8	3			56	113
PHEN	5	10	92	1		213	1		34	4	12		44	187	17	620
PHYS	28	74	364	27		465	20	1	270	9	18		187	303	80	1846
PROC	24	18	206	17	6	207			44	3	28	56	17	80	12	718
Total	334	673	2046	224	78	2792	127	21	1505	261	300	113	620	1846	718	

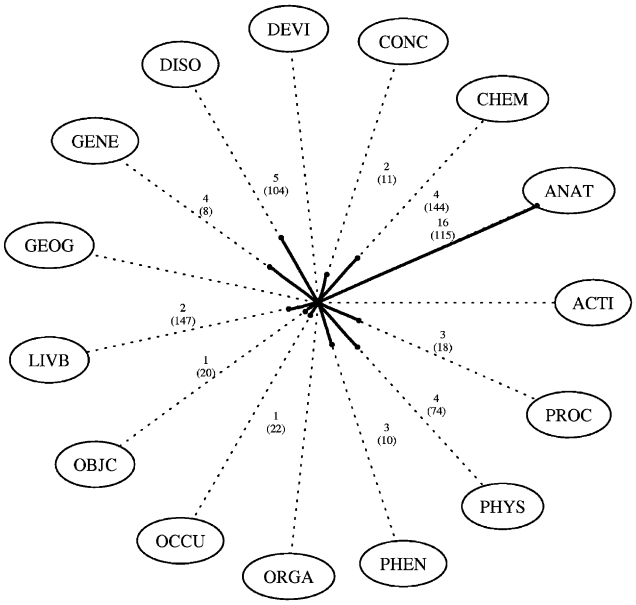
B

	ACTI	ANAT	CHEM	CONC	DEVI	DISO	GENE	GEOG	LIVB	OBJC	OCCU	ORGA	PHEN	PHYS	PROC	Total
ACTI	4			4		5		1	3		3	2	1	7	2	32
ANAT		16	4	2		5	4		2	1	1		3	4	3	45
CHEM		4	4			8	3		5	2	1		4	4	4	39
CONC	4	2		3		1	2	2	5	1	2	3	1	6	3	35
DEVI						3			2	1	1				1	8
DISO	5	5	8	1	3	15	2	1	8	3	1		15	10	9	86
GENE		4	3	2		2	4		2	1	1		1	4		24
GEOG	1			2		1				1	1			1		7
LIVB	3	2	5	5	2	8	2		3	3	2	2	2	4	2	45
OBJC		1	2	1	1	3	1	1	3	1	2	1	3	2	1	23
OCCU	3	1	1	2	1	1	1	1	2	2	3	1	1	1	2	23
ORGA	2			3					2	1	1	1			2	12
PHEN	1	3	4	1		15	1		2	3	1		9	11	5	56
PHYS	7	4	4	6		10	4	1	4	2	1		11	11	4	69
PROC	2	3	4	3	1	9			2	1	2	2	5	4	3	41
Total	32	45	39	35	8	86	24	7	45	23	23	12	56	69	41	

example, under the heading ANAT–DISO, we list *causes*, and *disrupted_by*, which is read as Anatomy *causes* Disorders, and Anatomy *disrupted_by* Disorders.

2.3. Interpretation

One of the difficulties of interpreting the matrices in Table 2 and the radial diagrams is that the number of relationships between two semantic groups is, in part, a function of the number of semantic types in these two groups. In some cases, a relationship obtains between all semantic types in a group and all semantic types in another group. For example, the 11 semantic types in the group Anatomy have a relationship *issue_in* to the two semantic types in the group Occupations, yielding 22 (2×11) relationships between the two groups. The number of types of relationships between two groups also influences the total number of relationships that obtain between the groups. For example, the group Disorders, although having fewer semantic types than Chemicals & Drugs, is connected to other groups by 2792 relationships, while Chemicals & Drugs only has 2046 relationships. On the other hand, Disorders is involved in more types of relationships (86) than Chemicals & Drugs (39). Finally, some relationships are specific to semantic types and are not expected to be widely shared. For example, the



SG: ANAT - Unique relationships
Number of relationships in graph: unique: 45 (total: 673)

ANAT-ANAT: 16 (115)
- adjacent_to (13)
- branch_of (1)
- conceptual_part_of (12)
- connected_to (6)
- consists_of (5)
- contains (9)
- derivative_of (1)
- developmental_form_of (8)
- interconnects (3)
- isa (10)
- location_of (11)
- part_of (15)
- produces (5)
- surrounds (13)
- traverses (2)
- tributary_of (1)

ANAT-CHEM: 4 (144)
- consists_of (8)
- disrupted_by (60)
- ingredient_of (1)
- produces (75)

ANAT-CONC: 2 (11)
- has_measurement (2)
- isa (9)

ANAT-DISO: 5 (104)
- causes (10)
- disrupted_by (6)
- inverse_isa (3)
- location_of (76)
- produced_by (9)

ANAT-GENE: 4 (8)
- contained_in (1)
- has_part (4)
- inverse_isa (2)
- produced_by (1)

ANAT-LIVB: 2 (147)
- location_of (28)
- part_of (119)

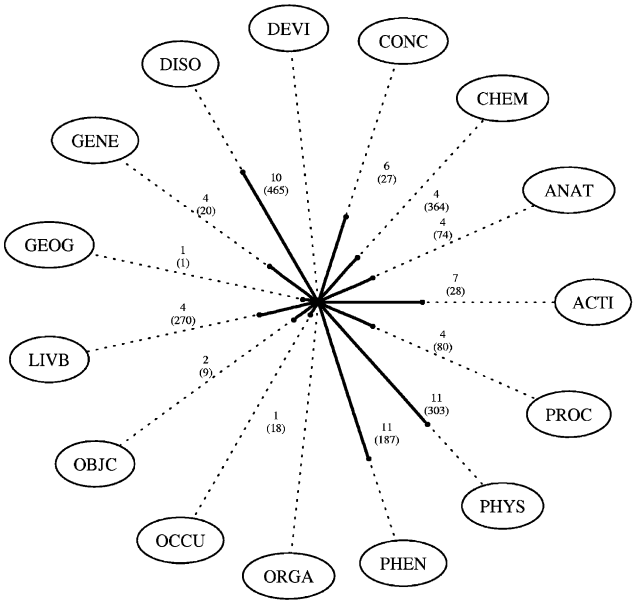
ANAT-OBJC: 1 (20)
- isa (20)

ANAT-OCCU: 1 (22)
- issue_in (22)

ANAT-PHEN: 3 (10)
- has_measurement (1)
- location_of (8)
- produced_by (1)

ANAT-PHYS: 4 (74)
- affects (7)
- location_of (56)
- produced_by (8)
- result_of (3)

ANAT-PROC: 3 (18)
- analyzed_by (2)
- location_of (12)
- measured_by (4)



SG: PHYS - Unique relationships
Number of relationships in graph: unique: 69 (total: 1846)

PHYS-ACTI: 7 (28)
- affected_by (3)
- affects (3)
- associated_with (6)
- has_manifestation (3)
- has_result (3)
- isa (7)
- result_of (3)

PHYS-ANAT: 4 (74)
- affected_by (7)
- has_location (56)
- has_result (3)
- produces (8)

PHYS-CHEM: 4 (364)
- affected_by (175)
- complicated_by (70)
- disrupted_by (70)
- produces (49)

PHYS-CONC: 6 (27)
- conceptual_part_of (2)
- has_conceptual_part (3)
- has_measurement (7)
- has_result (6)
- isa (2)
- occurs_in (7)

PHYS-DISO: 10 (465)
- affected_by (63)
- affects (42)
- associated_with (20)
- disrupted_by (7)
- has_evaluation (18)
- has_manifestation (77)
- has_process (42)
- has_result (70)
- process_of (42)
- result_of (84)

PHYS-GENE: 4 (20)
- affected_by (7)
- carried_out_by (2)
- has_location (7)
- has_result (4)

PHYS-GEOG: 1 (1)
- has_result (1)

PHYS-LIVB: 4 (270)
- affects (119)
- process_of (113)
- produces (4)
- property_of (34)

PHYS-OBJC: 2 (9)
- affected_by (7)
- isa (2)

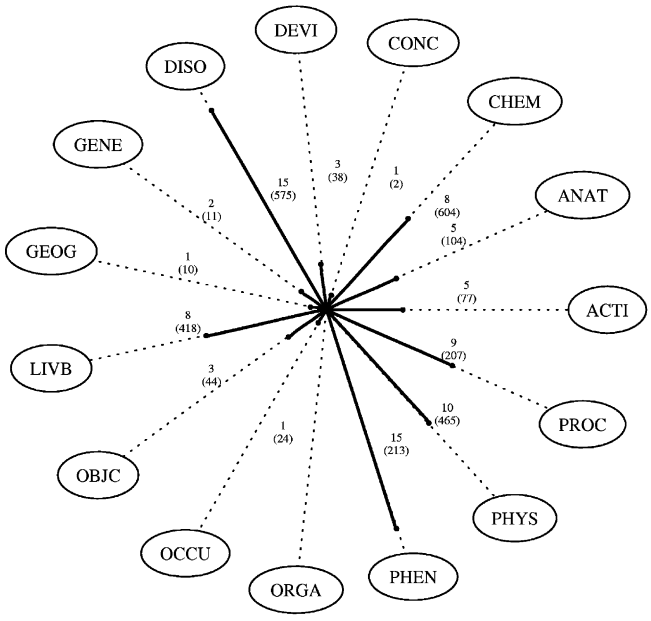
PHYS-OCCU: 1 (18)
- issue_in (18)

PHYS-PHEN: 11 (187)
- affected_by (14)
- affects (14)
- has_evaluation (9)
- has_manifestation (7)
- has_measurement (7)
- has_process (14)
- has_result (35)
- indicated_by (7)
- isa (21)
- process_of (14)
- result_of (45)

PHYS-PHYS: 11 (303)
- affects (63)
- associated_with (3)
- co-occurs_with (28)
- degree_of (9)
- isa (9)
- manifestation_of (14)
- measurement_of (14)
- occurs_in (2)
- precedes (49)
- process_of (49)
- result_of (63)

PHYS-PROC: 4 (80)
- affected_by (30)
- assessed_for_effect_by (7)
- complicated_by (7)
- measured_by (36)

Fig. 2. Radial diagrams for semantic groups Anatomy and Physiology.



DISO-ACTI: 5 (77)
- affects (3)
- associated_with (40)
- has_manifestation (3)
- isa (7)
- result_of (24)

DISO-ANAT: 5 (104)
- caused_by (10)
- disrupts (6)
- has_location (76)
- isa (3)
- produces (9)

DISO-CHEM: 8 (604)
- affected_by (150)
- caused_by (260)
- complicated_by (100)
- diagnosed_by (12)
- indicated_by (6)
- prevented_by (12)
- produces (42)
- treated_by (22)

DISO-CONC: 1 (2)
- isa (2)

DISO-DEVI: 3 (38)
- caused_by (20)
- prevented_by (7)
- treated_by (11)

DISO-DISO: 15 (575)
- affects (36)
- associated_with (41)
- co-occurs_with (56)
- complicates (78)
- conceptually_related_to (1)
- degree_of (37)
- diagnoses (10)
- evaluation_of (12)
- isa (10)
- location_of (18)
- manifestation_of (80)
- occurs_in (24)
- precedes (36)
- process_of (36)
- result_of (100)

DISO-GENE: 2 (11)
- disrupts (1)
- has_location (10)

DISO-GEOG: 1 (10)
- associated_with (10)

DISO-LIVB: 8 (418)
- affects (153)
- associated_with (10)
- caused_by (30)
- diagnosed_by (6)
- location_of (12)
- occurs_in (60)
- part_of (51)
- process_of (96)

DISO-OBJC: 3 (44)
- affected_by (6)
- caused_by (30)
- isa (8)

DISO-OCCU: 1 (24)
- issue_in (24)

DISO-PHEN: 15 (213)
- affected_by (12)
- affects (12)
- associated_with (10)
- co-occurs_with (2)
- evaluation_of (2)
- has_evaluation (6)
- has_manifestation (10)
- has_process (12)
- has_result (50)
- indicated_by (10)
- inverse_isa (1)
- isa (19)
- manifestation_of (5)
- process_of (12)
- result_of (50)

DISO-PHYS: 10 (465)
- affected_by (42)
- affects (63)
- associated_with (20)
- disrupts (7)
- evaluation_of (18)
- has_process (42)
- has_result (84)
- manifestation_of (77)
- process_of (42)
- result_of (70)

DISO-PROC: 9 (207)
- affected_by (24)
- assessed_for_effect_by (6)
- associated_with (70)
- complicated_by (6)
- diagnosed_by (20)
- measured_by (24)
- prevented_by (6)
- result_of (40)
- treated_by (11)

Fig. 3. Radial diagram for semantic group Disorders.

relationship *tributary_of* applies only to vascular structures and, therefore, only to the semantic type Body Part, Organ, or Organ Component. This contributes to the diversity of types of relationships observed within the semantic group Anatomy (16) and helps us understand why there are only 115 triplets in the group Anatomy overall.

The radial diagrams proved helpful for comparing the profiles of various groups. In Fig. 2, we can see that there are strikingly different profiles for each of the two groups, Anatomy and Physiology. It is clear at a glance that the group Anatomy shares the largest number of relationships with its own members. The right-hand side of the diagram shows the specific relationships that are involved, with many of them being physical relationships, such as *branch_of*, *connected_to*, and *part_of*. One exception is *conceptual_part_of*. This can be accounted for by the fact that the semantic types `Body System`, `Body Location or Region`, and `Body Space or`

Junction have been grouped with other anatomical terms, even though they are conceptual entities, rather than physical entities. For some purposes it may be useful to group them in this way, but their location as conceptual entities in the semantic network itself is necessary for appropriate reasoning. The profile for Physiology shown in the bottom half of Fig. 2 shows that this group shares almost equivalent numbers of relationships with Disorders (10), Phenomena (11), and with itself (11). This makes sense, given that all three groups are closely related in meaning. Each group consists of either natural or human-caused processes and functions, and, therefore, it is not surprising that they participate in some of the same functional relationships, such as *affects*, *causes*, and *process_of*. The profile for Disorders shown in Fig. 3 confirms, at a glance, that this group and the group Physiology have similar profiles, with, however, some notable exceptions. The link to Chemicals & Drugs is stronger and more diverse for Disorders than it is for Physiology. Relationships like *treats*, *prevents*, and *causes* are relevant for these two groups, and are seen again in the relationships that bind Disorders to Devices. No such relationships exist between the group Physiology and Disorders or Devices. In fact, no relationships at all are stated between the group Physiology and Devices. This latter may represent an omission in the semantic network, since there are undoubtedly devices that, for example, monitor normal function. On a similar note, the lack of relationships between the groups Genes & Molecular Sequences and Procedures is unexpected, since the semantic type Molecular Biology Research Technique is a member of the group Procedures. This is, therefore, also a case where an omission in the semantic network becomes readily apparent and needs to be rectified.

These matrices may be helpful as the semantic network is developed further. As new relationships are added to a particular pair of semantic types, it would make sense to check if they apply to other members of the semantic groups to which these semantic types belong. For example, if a relationship is added between the semantic types Disease or Syndrome and Organism, then each of the members of the group Disorders and each of the members of the group Living Beings should be inspected for the possible applicability of that relationship.

3. Experiment 2: perspective of the relationships

3.1. Methods

3.1.1. Association between relationships and semantic groups

The simplest representation of associations between relationships and semantic groups is a matrix with 49 rows for the relationships used in the Semantic Network[9] and one column for each of the 15 semantic groups. The column (rel_i, SG_j) in the matrix contains the number of semantic relations (ST_1, *rel*, ST_2) in which *rel* equals rel_i and at least one of the semantic types ST_1 or ST_2 belongs to the semantic group SG_j.[10]

Although there is no definitive method for analyzing such a matrix, our assumption is that the matrix should reflect some of the principles on which the semantic network and semantic groups were built. Here are some of the indicators we propose:

- The row margin for the relationship *rel* contains the number of semantic relations (ST_1, *rel*, ST_2) in which *rel* is involved. Knowing that the semantic network relationships are generally coarse (compared, for example, to relationships in GALEN[11]), small counts could be indicative of unusually specific relationships, omissions, or possible errors. The same reasoning applies to counts for a specific group.
- The column margin for the semantic group SG contains the number of semantic relations (ST_1, *rel*, ST_2) in which ST_1 or ST_2 belongs to SG. Since semantic network relationships can be inherited along the *isa* hierarchy, the number of semantic relations involving a semantic group is expected to be somewhat proportional to the number of semantic types in the group. Therefore, extreme values for the ratio number of semantic relations/number of semantic types for a group could indicate issues with this group.

The goal of this method is to provide a bird's eye view on the relationships in order to assist humans in the analysis of the semantic groups.

3.1.2. Subsets of related semantic groups

We hypothesized that, in most cases, a given relationship applies to only a limited number of groups. What this means practically is that the constitution of the groups takes into account not only the semantics of the types, but also that of the relationships. For example, since what can be treated generally belongs to the realm of disorders, it is expected that the semantic types involved with the relationship *treats* will be clustered mostly in the semantic group Disorders. Moreover, the limited number of relationships across groups is generally concentrated around a few groups which play a central role in the relationship. For example, *treats* applies only to Disorders and Living Beings. When it applies to Disorders, the semantic groups involved can only be

[9] Out of the 54 relationships in the Semantic Network, five relationships (*brings_about*, *functionally_related_to*, *physically_related_to*, *spatially_related_to*, and *temporally_related_to*) do not appear in actual semantic relations.

[10] What is represented in this matrix is the association between groups and relationships, not directionality. What concerns us for this purpose is the existence of a relationship.

[11] www.opengalen.org.

Chemicals & Drugs, (e.g., Antibiotic *treats* Disease or Syndrome), Devices (e.g., Medical Device *treats* Injury or Poisoning), and Procedures (e.g., Therapeutic or Preventive Procedure *treats* Congenital Abnormality). When it applies to Living Beings, the only semantic group involved is Living Beings (e.g., Professional or Occupational Group *treats* Patient or Disabled Group).

From the perspective of graph theory, a partition of the semantic network can be represented as a directed graph where semantic groups are the nodes and relationships the edges. The number of types of relationships with which a semantic group is involved constitutes the degree of a node. More precisely, the degree of each node can be divided into the in-degree (for "incoming" relationships) and the out-degree (for "outgoing" relationships). We hypothesize that semantic coherence should translate, for a given relationship, into a small number of nodes (called pivot nodes) with high in- or out-degree, while most nodes are of degree 1 or 0. In other words, the set of edges for a given relationship is easily decomposed into subsets organized around pivot nodes and the number of such subsets is generally small. In the example above, the two pivot nodes for the relationship *treats* are the semantic groups Disorders (degree = 3) and Living Beings (degree = 1). The set of four edges involving the relationship *treats* is thus decomposed into two subsets organized around these two nodes: {Chemicals & Drugs–Disorders, Devices–Disorders, Procedures–Disorders} and {Living Beings–Living Beings}. The procedure used to find the smaller number of subsets for a given relationship is as follows. The first subset of edges corresponds to the node of highest degree. All edges involved with this node are removed from further processing and the degree of each node is recomputed after excluding these edges. The procedure is applied iteratively until no edges remain. Applied to the example above, this procedure first identifies Disorders as the node of highest degree (3), creating a first subset from the three corresponding edges. Then, the only remaining node is Living Beings, whose self-edge becomes the only member of the second subset. This procedure was applied to the 49 relationships used in the Semantic Network—including *isa*. The total number of subsets of edges in the Semantic Network is computed as the sum for all relationships of the number of subsets of edges for each relationship.

3.1.3. Creating random partitions

In order to validate our hypothesis that semantically coherent groups should result in a small number of such subsets of edges in the whole Semantic Network, we demonstrate that the number of subsets of edges (NSE) should be higher when the semantic groups are not designed to be semantically coherent, e.g., in randomly created semantic groups. We generated random partitions by assigning the semantic types to random groups, keeping the number of groups and the number of members in each group similar to that in our original semantic groups, so that the only factor influencing NSE is the semantic group assignment. This procedure is usually referred to as permutation test. Since the number of possible rearrangements is close to 134!, we used a Monte Carlo approach to examine only a random sample [18, p. 45]. What we want to show is that it is extremely unlikely that, by chance only, the small NSE observed in the original semantic groups is also observed in partitions resulting from the random assignment of the semantic group labels. Not examining all possible rearrangements, it is not possible to calculate an exact p value. It is, however, possible to get an estimate of this probability by calculating the upper bound for p.

3.2. Results

3.2.1. Association between relationships and semantic groups

The matrix containing the number of semantic relations by relationships and by semantic groups is shown in Table 3. The matrix can be analyzed from two perspectives, relationships, and groups. From the perspective of relationships, the total number of semantic relations (ST_1, *rel*, ST_2) in which the relationship *rel* equals rel_i, shown in the rightmost column of Table 3, ranges from 1 (for *branch_of*, *derivative_of*, and *tributary_of*) to 1968 (for *affects*), with a median of 89. From the perspective of the semantic groups, the total number of semantic relations, shown in the last row of Table 3, ranges from 21 (for Geographic Areas) to 2792 (for Chemicals & Drugs), with a median of 334.

3.2.2. Subsets of related semantic groups

We computed the NSE for each of the 49 relationships used in the Semantic Network—including *isa*. The NSE per relationship ranges from 1 to 13 with a median of 2. Not surprisingly, the highest count is for the relationship *isa*. Since the members of a semantic group often come from a subtree of the semantic network, the relationship *isa* logically appears within most groups. The maximum NSE for the other relationships is 6. Examples of subsets of edges are presented in Fig. 4 (relationship *treats*) and Fig. 5 (relationship *location_of*). The 321 triplets involving the relationship *location_of* can be reduced to 15 pairs of semantic groups. In turn, in the graph, the corresponding 15 edges are organized around five semantic groups, playing the role of pivot nodes. Nodes are represented with an oval shape when they receive no edge, i.e., when their in-degree is 0 (e.g., Organizations). Nodes represented with an octagon both emit and receive edges (e.g., Anatomy). The other nodes have a rectangular shape when they only receive edges (e.g., Procedures) or are not involved

Table 3
Matrix of relationships by semantic groups

	ACTI	ANAT	CHEM	CONC	DEVI	DISO	GENE	GEOG	LIVB	OBJC	OCCU	ORGA	PHEN	PHYS	PROC	Total
adjacent_to		13														13
affects	18	7	375	10		501	7		297	14		4	128	547	60	1968
analyzes		2	50												52	104
assesses_effect_of			50			6							2	7	65	130
associated_with	76		7	3		201		13	35		6		10	29	70	450
branch_of		1														1
carries_out	8						2					36		2	28	76
causes		10	260		20	350			30	30						700
complicates			180			184							11	77	14	466
conceptual_part_of	2	12		10							1		1	5	2	33
conceptually_related_to	1			1		1										3
connected_to		6														6
consists_of		13	9				1									23
contains		10					1									11
co-occurs_with						58							3	28		89
degree_of						37								9		46
derivative_of		1														1
developmental_form_of		8														8
diagnoses			12			48			6						20	86
disrupts		66	140			14	11							77		308
evaluation_of	8			15		38							18	27	7	113
exhibits	45								45							90
indicates			9			16			3				18	7		53
ingredient_of		1	28							2						31
interacts_with			325						174							499
interconnects		3														3
isa	58	44	140	71	6	50	23	4	83	188	5	11	54	41	25	803
issue_in	18	22	52	24	4	24	10	2	46	10	268	8	12	18	14	532
location_of	8	191	42			116	22		86			36	9	63	40	613
manages									6			6				12
manifestation_of	6					175							23	101		305
measurement_of		3	25	15			4	1		2			35	28		113
measures		4	100	8		24							8	36	180	360
method_of	7										18				24	49
occurs_in				7		84			60					9		160
part_of		138				51	22		187							398
performs	48								90						42	180
practices									2		2					4
precedes						36								49	1	86
prevents			12		7	25									6	50
process_of	1					240			226				73	274		814
produces		99	194	30	12	51	16		46	6		12	8	61		535
property_of			2	6			4		40					34		86
result_of	30	3		6		418	4	1					207	317	42	1028
surrounds		13														13
traverses		2														2
treats			22		11	44			1						11	89
tributary_of		1														1
uses			12	18	18				42	9					15	114
Total	334	673	2046	224	78	2792	127	21	1505	261	300	113	620	1846	718	
Sem. Types (ST)	*9*	*11*	*26*	*12*	*2*	*12*	*5*	*1*	*23*	*5*	*2*	*4*	*6*	*9*	*7*	*134*
Relationships per ST	*37.1*	*61.2*	*78.7*	*18.7*	*39.0*	*232.7*	*25.4*	*21.0*	*65.4*	*52.2*	*150.0*	*28.3*	*103.3*	*205.1*	*102.6*	*87.0*

and have their name displayed only for illustrative purposes (e.g., Devices). The 15 edges can be grouped into five subsets, centered on the five pivot nodes (Anatomy, Disorders, Genes & Molecular Sequences, Living Beings, and Organizations). For example, the subset centered on Genes & Molecular Sequences comprises the edges of this node to Disorders, Living Beings, Phenomena, and Physiology. The legend on the right side of the graph provides details about the number of semantic relations represented by each edge. For example, 76 triplets participate in the relationship of Anatomy to Disorders.

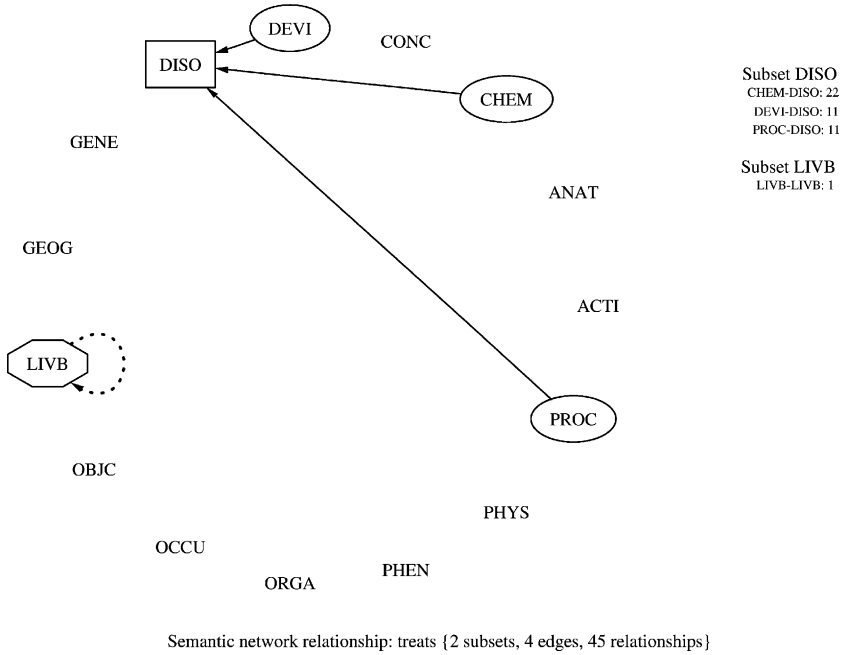


Fig. 4. Subsets of edges for the relationship *treats* (each style of line corresponds to a subset of edges: plain for DISO, dotted for LIVB).

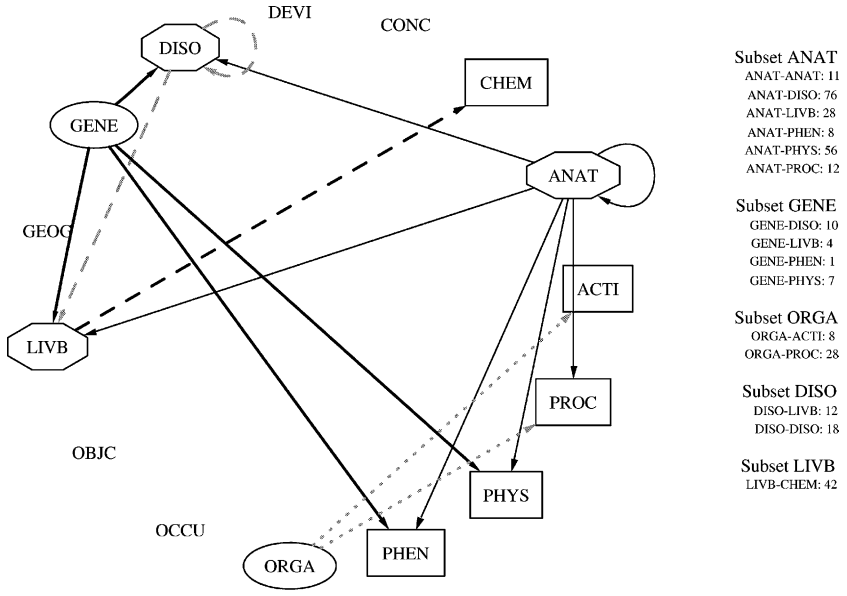


Fig. 5. Subsets of edges for the relationship *location_of* (each style of line corresponds to a subset of edges, e.g., grey and dotted for ORGA).

3.2.3. Random partitions

The total NSE in the Semantic Network, computed as the sum for all relationships of the NSE for each relationship, is 116. We generated 20,000 random partitions and computed the total NSE for all relationships. Counts range from 219 to 301, with a median of 261. From this experiment, we can conclude that the probability p of obtaining a total NSE of 116 by random is at most 0.001 ($p < 0.001$). Although this experiment does not prove that a small value for NSE is indicative of semantic coherence, it shows that groups generated randomly, i.e., without regard to semantic coherence, never exhibit this property.

3.3. Interpretation

Interesting observations can be made by studying the margins in Table 3, i.e., the total number of relationships for each relationship (rightmost column) and for each semantic group (last row).

Most relationships with a count lower than 25 are associated with the semantic group Anatomy, some of them

being specific to subdomains such as blood vessels (*branch_of*, *tributary_of*) or embryologic development (*developmental_form_of*). A majority of them are spatial (*surrounds*, *adjacent_to*, *traverses*) or physical relationships (*connected_to*, *interconnects*) and are, therefore, not necessarily applicable to other subdomains of the semantic network. With the exception of *interacts_with*, all the relationships with a count greater than 300 are associated with the semantic group Disorders. Examples of these relationships include *complicates*, *causes*, *process_of*, *result_of*, and *affects*. High-level semantic network relationships (e.g., *associated_with*) and broadly applicable relationships (*affects*, *interacts_with*) are also involved in a large number of semantic network relations.

From the perspective of the semantic groups, we found that, as expected, the groups that have the larger number of members also tend to have a larger number of semantic network relations, shown in the last row of Table 3. Examples of such groups include Chemical & Drugs, Living Beings, Disorders, and Anatomy. However, the group Concepts & Ideas, although having as many members as Disorders only has a fraction of its semantic relations. So, proportionality with the number of semantic types does not strictly explain the number of semantic relations in the groups. The seven semantic groups representing clinical medicine and physiopathology account for 70% of the semantic types, but 88% of the semantic relations, confirming the rich representation of this subdomain in the semantic network.

Intuitively, it makes sense that the semantic relationships be organized around a limited number of pivot semantic groups rather than equally distributed among the groups. With the exception of *isa*, which applies to all groups, we observed that most relationships tend to associate with some groups. In the examples we presented earlier, *treats* and *location_of*, it was easy to imagine a small number of pivot groups. More surprisingly, relationships such as *associated_with*, *issue_in*, and *result_of* exhibit a similar behavior. Interestingly enough, this behavior is not found in semantic groups resulting from random partitions of the semantic network. Although it would require more investigation, we believe that, for a given number of groups, a small number of subsets of edges in the Semantic Network may reflect that semantic types sharing a given relationship were appropriately grouped together.

4. Experiment 3: interaction between semantic types and relationships

4.1. Methods

In the previous sections, we used (SG_1, *rel*, SG_2) relations to explore the semantic groups and the relationships represented among them, first focusing on the semantic groups and then on the relationships. While these methods provide a useful summary of the 6703 semantic network relations, they provide less insight into the role played by relationships among semantic types on the composition of the semantic groups. Relationships among semantic types may influence the constitution of the semantic groups for two major reasons. First, relationships are inherited along the *isa* hierarchy, so that, except when a relationship is explicitly blocked, the descendants of a semantic type ST_i inherit the relationships of ST_i. And, because they are semantically close, the descendants of ST_i are likely to belong to the same semantic group as ST_i. Therefore, the semantic types in a semantic group are likely to share at least part of their relationships. For example, all the descendants of `Pathologic Function` (e.g., `Neoplastic Process`) inherit a relationship to `Chemical` (`Chemical` *causes* `Pathologic Function`). In other words, the property "caused by chemical" is shared by all the descendants of `Pathologic Function`. The second reason is that, even if they do not necessarily have common ancestors in this group, the semantic types in a semantic group often share properties with other semantic types in the group. These properties are usually represented as relationships to other semantic types. For example, disorders have in common the property of being treated by, say, drugs. Therefore, semantic types involved in a relationship *treats* with Pharmacologic Substance will likely belong to the semantic group Disorders. This is why the group Disorders includes not only `Pathologic Function` and its descendants, but also `Congenital Abnormality` and `Injury or Poisoning`, which are not hierarchically related to `Pathologic Function` (and should not be).

What we were interested in exploring is how the semantic groups reflect the properties of semantic types—expressed through the relationships in which they participate. The association between semantic types and relationships can be summarized in a matrix where the number of times a semantic type ST_i is involved in a relationship rel_j constitutes the intersection of row ST_i and column rel_j., i.e., the number of semantic network relations (ST_1, *rel*, ST_2) in which *rel* is equal to rel_j and either ST_1 or ST_2 is equal to ST_i. Such a matrix expresses the observed association between two categorical variables, semantic type and relationship and is also called a two-way contingency table. The method of choice for analyzing this kind of two-dimensional data is correspondence analysis. A succinct description of this method is given below and we refer interested readers to [7] for more details.

Correspondence analysis is an exploratory technique related to principal component analysis, which finds a multidimensional representation of the association between the row and column categories of a two-way contingency table. Correspondence analysis provides a method for representing both the row categories and the

column categories in the same space, so that the results can be visually examined for structure. To reduce dimensionality, only the first two or three axes of the new space are plotted. In the two-dimensional graphical display, the overall quality of representation of the points can be expressed as a proportion of the total variation (called inertia in correspondence analysis parlance). If a large percentage of the total inertia lies along the principal axes displayed, it means that most points are well represented with respect to these axes. Distance among points reflects similarity in the shape of their profiles. These two semantic types are therefore expected to appear very close to each other on the two-dimensional graphical display.

We created a matrix, described above, of 134 rows (categories of the variable semantic type) and 49 columns (categories of the variable relationship). The statistical package MVSP[12] was used to perform the correspondence analysis.

Correspondence analysis is generally used to display both the row categories and the column categories in the same graph, using, for example, the structure (groupings) of column categories to suggest explanations about the structure of row categories. In this study, however, we display only row categories, i.e., the semantic types, because we are mainly interested in comparing the groups resulting from the analysis to the groups we created manually, i.e., the semantic groups. Moreover, to facilitate the comparison with our original partition, the semantic types are represented with symbols reflecting the semantic group to which they were assigned. For the correspondence analysis to validate our original groupings, two conditions must be fulfilled. First, the symbols corresponding to a given semantic group must appear close to each other on the display. Second, and conversely, semantic types belonging to different groups should be apart on the display.

4.2. Results

A portion of the two-way contingency table used in the correspondence analysis is presented in Table 4. This matrix can be thought of as a series of profiles for each semantic type. The list of relationships to which a semantic type is associated, along with the frequency of each association constitutes the profile of this semantic type. By simply scanning the table, it is noticeable that, with the exception of `Finding` and `Sign` or `Symptom`, most semantic types from the semantic group Disorders have similar profiles. As we mentioned earlier, in correspondence analysis, the similarity of profiles translates to a small distance among the corresponding points.

[12] Multi-Variate Statistical Package, www.kovcomp.com.

The first two principal axes account only for about 19% of the total inertia, which means that some points may not be correctly represented with respect to these two axes. The two-dimensional graphical display using these two axes is presented in Fig. 6. For validation purposes, we compared this display to representations using additional principal axes.

The grouping of semantic types observed on the display are as follows:

- The groups Occupations and Organizations are both very cohesive and quite distinct from other groups.
- The groups Anatomy, Chemicals, and Genes & Molecular Sequences are essentially cohesive, with the exception of one member in each group.
- The groups Disorders, Physiology, and Phenomena exhibit a more complex pattern. As for the groups above, these groups are essentially cohesive, but at least one member of each group is isolated from the others. Moreover, the majority of the semantic types in these three groups are so close that they appear as one unique group and their isolated members also form one group.
- The groups Activities & Behaviors and Living Beings are organized around several distinct poles. To some extent, the group Chemicals could also be seen as having three poles.
- The groups Concepts & Ideas and Objects exhibit a large dispersion, often overlapping other groups. The group Procedures is also disperse.
- Finally, the groups Geography and Devices are more difficult to interpret because of their small number of members. However, if Geography seems distinct from other groups, it is not the case for Devices.

4.3. Interpretation

4.3.1. Cohesive groups, except for one member

The semantic types located away from the other members of their group include `Body System` (Anatomy), `Clinical Drug` (Chemicals & Drugs), and `Gene or Genome` (Genes & Molecular Sequences). In the three cases, the semantic type in question, although semantically related to them, does not belong to the same part of the semantic network hierarchy as most of the other members of the group. For example, although an anatomical type, `Body System` is a conceptual entity, most of the other types in the group Anatomy are physical entities. However, this difference does not provide a full explanation. In fact, in displays using different principal axes (e.g., axes 2 and 3, not displayed here), the outlier in the group Anatomy is the semantic type `Body Substance`, a physical entity, differing from other anatomical types by specific properties represented through specific relationships (e.g., *causes*) and by different frequencies of association with relationships common to the other members (e.g., only 1 for *location_of*). Also, in

Table 4
Matrix of semantic types by relationships (partial representation)

SG	Semantic Type \ Relationship	adjacent_to	affects	analyzes	assesses_effect_of	associated_with	branch_of	carries_out	causes	complicates	conceptual_part_of	conceptually_related_to	connected_to	consists_of	contains	co-occurs_with	degree_of	derivative_of	developmental_form_of	diagnoses	disrupts	evaluation_of	exhibits	indicates	ingredient_of	interacts_with	interconnects	isa	issue_in	location_of	manages	manifestation_of	measurement_of	measures	method_of	occurs_in	part_of	performs	practices	precedes	prevents	process_of	produces	property_of	result_of	surrounds	traverses	treats	tributary_of	uses	Grand Total
ANAT	Anatomical Structure																										12	2	4						17															35
	Body Location or Region	3									3		1														4	2	24			1											1		1				40
	Body Part, Organ, or Organ Component	5					1				1		4	2	2				3		11						1	4	2	27						24					16			2	1		1	107
	Body Space or Junction	5									2		2		3												2	4	2	27			1			1							1	3				53
	Body Substance			2					10		1		1	10	8			1						1				3	2	1			1	4		2					20			2					69
	Body System										9																	4	2															1					16
	Cell	2									1			1	1				2		11						1	4	2	25						23					16			3					92
	Cell Component	2	7								1				1						11							4	2	26						22					17			2					95
	Embryonic Structure														1				4		11						1	3	2	18						19								3					62
	Fully Formed Anatomical Structure										3				1						11							8	2	22						17					16								80
	Tissue	3									1		2	1	2			1	3		11							4	2	26						22					19			6					103
DISO	Acquired Abnormality		24			18			32	22						10				3				1				4	2	18		17				6	17						50			4			228
	Anatomical Abnormality		24			18			32	22						9				3				1				5	2	18		17				6	17						47			4			225
	Cell or Molecular Dysfunction		76		1	24			37	27						10	11			6		3		2				5	2	12		25		4		10				11	4	46	8		52			4			380
	Congenital Abnormality		24			18			32	22						9				3				1				4	2	18		17				6	17						47			4			224
	Disease or Syndrome		76		1	24			37	27		1				10	11			6		3		2				7	2	12		25		4		15				11	4	46	9		52			4			389
	Experimental Model of Disease		76		1	24			37	27		1				10	11			6		3		2				5	2	12		25		4		10				11	4	46	8		52			4			381
	Finding					10										3						16						4	2			18													53
	Injury or Poisoning					18			32	25						8				3	14			1				2	2	8		9				12					1			52			4			191
	Mental or Behavioral Dysfunction		79		1	24			37	27						10	11			6		3		2				6	2	12		28		4		15				11	4	40	9		52			4			387
	Neoplastic Process		76		1	24			37	27						10	11			6		3		2				6	2	12		25		4		15				11	4	46	9		52			4			387
	Pathologic Function		76		1	24			37	27						10	11			6		3		2				9	2	12		25		4		10				11	4	46	8		52			4			384
	Sign or Symptom					10										3	1			10		16						3	2			18															4			67
GENE	Amino Acid Sequence																										5	2				1											2	1					11
	Carbohydrate Sequence																										5	2				1												1					9
	Gene or Genome		7					2						1	1						11							4	2	22						21					16	2							89
	Molecular Sequence																										7	2				1												1					11
	Nucleotide Sequence																										5	2				1				1							2	1					12

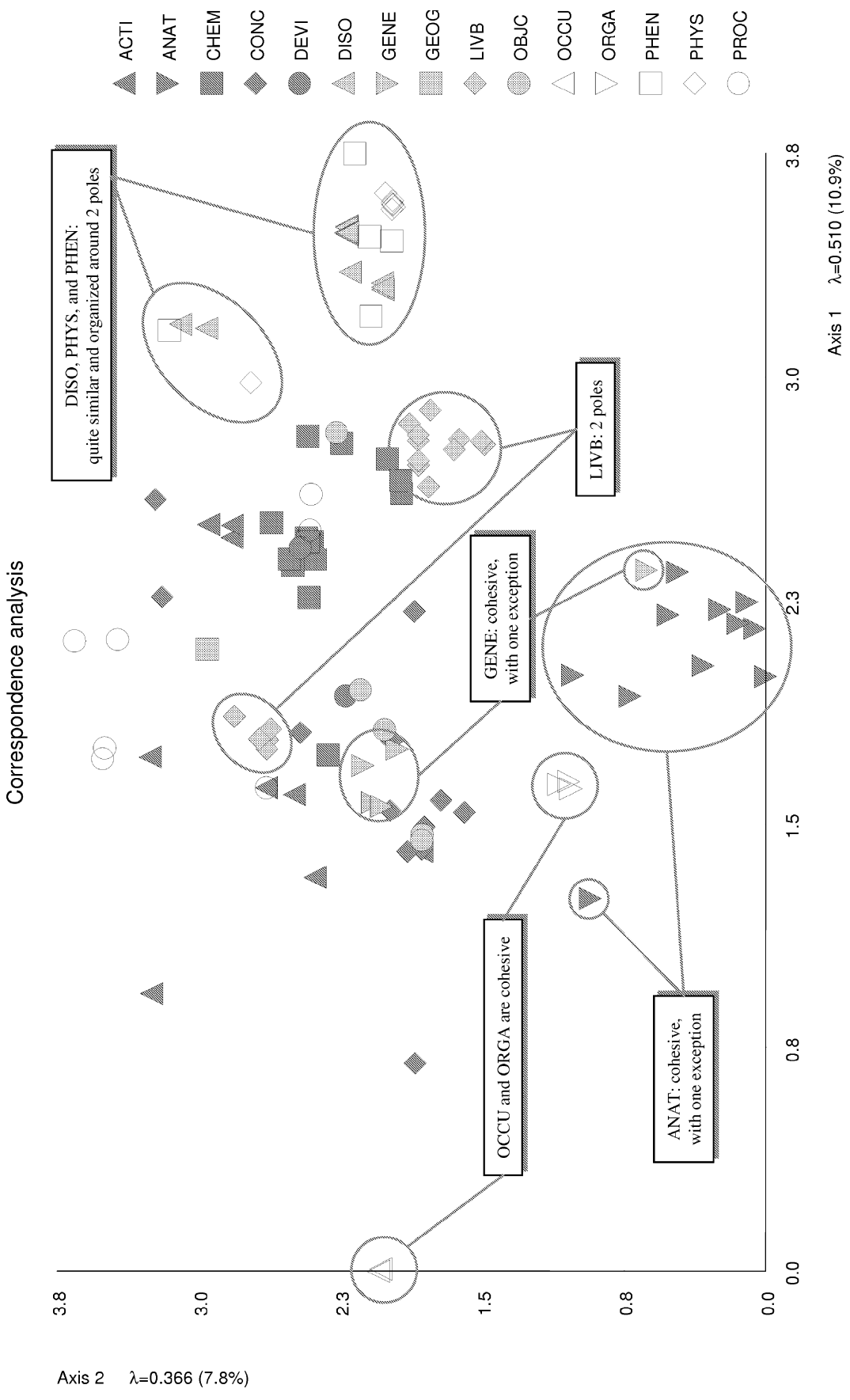


Fig. 6. Correspondence analysis diagram.

addition to Body Systems, two other semantic types in the group Anatomy—Body Space or Junction and Body Location or Region—are conceptual entities. Nevertheless, these two semantic types are consistently represented closer to the physical entities than to the other conceptual entity. In the other two groups, the outlier also belongs to a different part of the semantic network hierarchy than the other members of the group.

4.3.2. Several subgroups

Two semantic groups seem organized around several poles, namely Activities & Behaviors and Living Beings. The very name of the group Activities & Behaviors indicates that it is more a cluster than anything else. Not surprisingly, one of the poles is around the behaviors and the other one is around the activities. The group Living Beings consists of, on the one hand, the semantic type Organism and its descendants, which are physical entities, and, on the other hand, the semantic type Group and its descendants, which are conceptual entities. Unlike what happens in the group Anatomy, here, the distinction between the two subgroups is the opposition between physical and conceptual. Finally, although relatively close to each other, the majority of the semantic types in the group Chemical & Drugs form two distinct subsets. Not surprisingly, these subsets reflect the organization of the semantic network hierarchy starting with the semantic type Chemical, i.e., two separate subtrees for functional and structural views on chemicals.

4.3.3. Disorders, physiology, and phenomena

The three semantic groups Disorders, Physiology, and Phenomena exhibit an interesting pattern. As in the group Anatomy, one or two members of these groups differ from the others by their profile. These semantic types are Finding and Sign or Symptom for Disorders, Organism Attribute and Clinical Attribute for Physiology, and Laboratory or Test Result for Phenomena. With respect to correspondence analysis, the members of these three groups are very close to each other. Another characteristic of these groups is that the five semantic types Finding, Sign or Symptom, Organism Attribute, Clinical Attribute, and Laboratory or Test Result appear closer to each other than to the other members of their group. In other words, in Fig. 6, the members of these three groups form two subgroups, one for the majority of the members, and one for the exceptions. One characteristic common to the four exceptions is not their nature, but the role they play in the diagnostic process.

4.3.4. Less cohesive groups

Groups exhibiting less coherence on the two-dimensional graphical display include Concepts & Ideas, Objects, and Procedures. The groups Concepts & Ideas and Objects tend to include higher-level semantic types than other groups (e.g., Physical Object and Temporal Concept). These are not specific to the biomedical domain and actually belong to an upper-level ontology. Following the hierarchical organization of the semantic network, these high-level semantic types were grouped into two groups, one for entities, and one for events. Logically, the root of each hierarchy of the semantic network, i.e., the semantic types Entity and Event, is a member of the corresponding semantic group. It is therefore not surprising that these groups appear less consistent than other groups more specific to the domain.

For different reasons, the semantic group Procedures is not very consistent either. Actually, examining the contributions of individual relationships, it appears that the display is largely influenced by the relationship *measures*. The semantic types associated with the relationship *measures* with a relatively high frequency are Diagnostic Procedure, Laboratory Procedure, Research Activity, and Molecular Biology Research Technique.

5. Conclusions

Our goal in this paper has been to use visualization techniques to investigate the semantic type groupings we developed in earlier work. We were particularly interested in the semantic coherence of the groups, and we have used the semantic relationships as important indicators of that coherence. Our study has revealed some issues about the composition of the groups, and, interestingly, about the semantic network itself. In particular, in some cases, expected relationships between groups are missing, and this has revealed that additions need to be made to the semantic network. For example, we noted that there are no relationships expressed between Procedures and Genes & Molecular Sequences and, in fact, we would expect them, since the semantic type Molecular Biology Research Technique is a member of the group Procedures. One possible relationship that could be added would be *analyzes*, e.g., Molecular Biology Research Technique *analyzes* Amino Acid Sequence, etc. The methods described in this paper have made these and other outliers immediately apparent, and, thus, serve as a tool for auditing and validating both the semantic network and the semantic groups.

From the point of view of the relationships in which semantic types do or do not participate, some semantic types appear to be "loners" in the semantic group in which they have been placed. This might be addressed either by placing them in some already existing group, if this is appropriate and is borne out by further

investigation, or by establishing a new group, particularly if some other "loner" semantic types appear to cluster with this type. This might be the case, for example, for those semantic types that describe clinical attributes of various kinds, such as `Finding`, `Laboratory or Test Result`, and others.

In some cases, and for some purposes, a single group might be split into two groups. For example, we saw a clear division of the group Living Beings into two subgroups, when considering the relationships in which the constituent semantic types participate. One group of semantic types clustered around the semantic type `Organism` in the semantic network, and the other around the semantic type `Group`. This latter type is actually a conceptual entity that classifies individuals according to certain characteristics such as age, profession, etc. Another group that might be split for some purposes would be Chemicals & Drugs. There are two clusters here, the chemicals viewed from their structural perspective and those viewed functionally. Relationships such as *treats* and *prevents* apply to the functional perspective, but are not obviously relevant for, for example, inorganic chemicals. The tradeoff here is between parsimony on the one hand (create as few groups for your purposes as possible) and semantic coherence on the other. The methods described in this paper have allowed us to pose these types of questions using a variety of visual techniques.

References

[1] McCray AT. The scope and structure of the first version of the UMLS Semantic Network. Proc Annu Symp Comput Appl Med Care 1990:126–30.

[2] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001;10(Pt 1):216–20.

[3] Chen Z, Perl Y, Halper M, Geller J, Gu H. Partitioning the UMLS semantic network. IEEE Trans Inf Technol Biomed 2002;6(2):102–8.

[4] Nelson SJ, Sheretz DD, Tuttle MS, Erlbaum MS. Using MetaCard: a HyperCard browser for biomedical knowledge sources. Proc Annu Symp Comput Appl Med Care 1990:151–4.

[5] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36:462–77.

[6] Cimino JJ. Auditing the unified medical language system with semantic methods. J Am Med Inform Assoc 1998;5(1):41–51.

[7] Greenacre MJ. Theory and applications of correspondence analysis. London: Academic Press; 1984.

[8] Herman I, Melancon G, Marshall MS. Graph visualization and navigation in information visualization: a survey. IEEE Trans Visual Comput Graphics 2000;6(1):24–43.

[9] Michailidis G, de Leeuw J. Data visualization through graph drawing. Computation Stat 2001;16(3):435–50.

[10] Hetzler B. Visual analysis and exploration of relationships. In: Green R, Bean CA, Myaeng SH, editors. The semantics of relationships: an interdisciplinary perspective. Boston: Kluwer Academic Publishers; 2002. p. 199–217.

[11] Tuttle MS, Cole WG, Sheretz DD, Nelson SJ. Navigating to knowledge. Methods Inf Med 1995;34(1-2):214–31.

[12] Hearst M, Karadi C. Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In: Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 1997:246–55.

[13] Li Q, Shilane P, Noy NF, Musen MA. Ontology acquisition from on-line knowledge sources. Proc AMIA Symp 2000:497–501.

[14] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003;4(1):2.

[15] Kamps J. Visualizing WordNet structure. In: Proceedings of the First Global WordNet Conference, 2002:182–6.

[16] Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp Biocomput 2000:529–40.

[17] Bormel JI, Ferguson LR. Visualization and analysis of co-occurrence and cross-tabulation data in medical research. Proc Annu Symp Comput Appl Med Care 1994:944–98.

[18] Good PI. Permutation tests: a practical guide to resampling methods for testing hypotheses. 2nd ed. New York: Springer; 2000.