wordnet整理

(2012-03-21 21:16:01)

转载▼

标签：

杂谈

分类： 一天一点积累

主要是复制黏贴的别人的

计算概念间的关联度

Ted Pedersen Siddharth Patwardhan Jason Michelizzi

摘要：

wordnet是一个免费的软件包，可以用来计算两个概念（或者word sense）之间的语义相似度，它提供了六种计算相似度 和三种计算概念之间的关联度方法，所有的这些方法都是基于wordnet这个词汇数据库。这些计算方法已经用Perl实现，算法模型的输入是两个概念，返回结果就是用来表示相似或者关联度的一个数量值。

概述：

计算两个概念相似的程度，主要是基于包含在层次结构“is－a”结构中的信息。比如：如果automobile和boat拥有同一个祖先vehicle，那么automobile 会被认为 与boat 的相似程度大于和tree的相似度。

自从wordnet把名词和动词组织成is－a这种层次关系，wordnet 就非常适合进行形似度的计算。在2.0版，有九个名词层次结构 包含80000个概念，还有554个动词层次结构包含13500个概念。

Wordnet中的is－a关系并不是跨词性的，所以，基于wordnet的相似度计算只是局限在名词之间（比如：cat 和dog）和动词对之间（比如：run和walk）。而wordnet也包含形容词和副词，这些词不能组织成is－a关系，所以这些词也就不能进行相似度的计算。

然而，概念之间除了通过相似度关联外，还可以通过很多种其他方法。比如：wheel 是car 的一部分，night是day的反面，snow是由water变成的，knife是用来切bread等等。Wordent也提供额外的联系 比如：has－part，is－made－of，is－an－attribute－of 等等。除此之外，每一个概念（或者word sense）都通过一段简短的定义或者注释加以描述。

计算关联的方法是基于这些额外的信息，并且这些方法可以运用的更广泛的概念对中。比如:可以跨越词性,然后可以达到使动词murder和名词gun相关联的程度。也可以计算不是is－a关系的词汇比如形容词 violent和harmful。

相似度计算

三种相似度的计算是基于概念之间路径长度的：lch（Leacock&Chodorow 1998）,wup(Wu&Palmer 1994),和path，其中lch方法是找到两个概念之间的最短路径，然后在is－a层次中找到最大路径长度。Wup方法是找到两个概念从最近的公共祖先节点到root 节点的路径长度。Path方法是两个概念之间最短路径长度的倒数。

另外，概念的信息内容来自SemCor的sense－tagged 集合。然而，也有现成的程序可以对没有标注的集合中的信息进行计算，比如Brown Corpus，the Penn Treebank, British Nation Corpus ,或者任意形式的普通文本。

WordNet::Similarity支持两种假定的根节点，这两种根节点可以存在也可以不存在，当存在时，一个根节点包含所有的名词节点，另一个包含所有的动词节点。这样便允许把相似度计算应用到任意的名词和动词对之间。如果两个根节点不存在，那么只有在同一个is－a层次中的词才能计算相似度。

计算关联度

有三种wordnet支持的计算关联度的方法分别是：hso（Hirst&St-Onge1998）,lesk(Banerjee&Pedersen 2003),和vector(Patwardhan 2003).hso方法是基于路径并且根据方向把关系分类。比如is－a关系是纵向的，has－part关系是横向的。最好能在两个概念之间找到一条路径，这条路径不能太长并且方向也不能改变的太频繁。

Wordnet中的每一个概念（或者word sense）都是通过一个短注释进行定义的。Lesk方法和Vector方法是把注释作为概念的唯一的标识方法。Lesk方法通过寻找和计算两个概念的注释的交叉部分，和在wordnet中同时与一个概念相连接的方法来计算相似度。

Vector方法建立一个由wordnet 注释组成的co－occurrence矩阵。Wordnet注释中每一个word都有一个关联向量。每一个注释都是由一个gloss vector的形式组成的，gloss vector 的内容通常是所有上下文向量的gloss中的词组成的。两个概念之间的关联度就是通过计算两个gloss向量的夹角余弦值。

运用 Wordnet::Similarity

Wordnet::Similarity 的实现具有Perl的面向对象的特征。利用Wordnet::QueryData 包（Rennie2000）来创建一个Wordnet的描述。已经有大量现成的方法实现现有的相似度测量方法。

当一个已有的方法要合并到Perl程序中，那种方法的对象必须通过new（）方法创建。然后通过getRelatedness()方法得出WorldNet中的一对sense的相似度以及关联度。

Wordnet::Similarity 也可以通过应用程序similarity.pl提供的命令行接口来访问。这种方式可以允许用户对于具体的一对概念(以word＃pos＃sense的形式给出)，利用某个方法交互式的操作。比如：car＃n＃3代表car 的第三个wordnet名词sense。另外还允许对于和一个word或者word＃pos的组合相关联的所有可能的sense进行详细的说明。出此之外，还有基于这种应用程序的web形式的接口。

我们忽略wordnet怎样运行，Wordnet::Similarity对于每一个不同的方法都提供了详细的跟踪并显示出各种诊断信息的具体说明。比如：对于依赖于路径长度的方法（lch，wup，path），跟踪显示出所有的两个概念之间的所有的路径。对于信息内容的追踪方法（res，lin，jcn）则包含了两个概念之间所有的路径以及最近的公共子概念。对于hso方法的追踪显示出通过wordnet找到的真正路径，而lesk则显示出了wordnet中两个概念相互重叠的gloss以及以及他们相近的关系。Vector方法显示了用于创建一个概念的gloss向量的所有word vectors。

我们已经把Wordnet::Similarity合并到通用的基于语义联系的词意消歧的方法中（Patwardhan,Banerjee,&Pdersen2003）。这种方法已经在SenseRelate 包实现了(http://senserelate.sourceforge.net). 这个算法的前提是一个word的sense可以通过查找它所有的sense中与它相邻的word的sense关联最大的那个而得到.我们现在正在摸索把相似度和关联度算法运用到评估一个text generation system 的组建选择上.

1）Synset类的定义：描述了WordNet 的pos.data文件中的同义词组。每一个Synset表达了一个概念，并且其包含了一组同义词。Synset被Pointers链接成相关概念的网络，这就是WordNet的Net；getTarget可以从这个链中抽取出目标；getPointer可以抽取出pointers自己本身。

A Synset, or synonym set, represents a line of a WordNet pos.data file. A Synset represents a concept, and contains a set of Words, each of which has a sense that names that concept (and each of which is therefore synonymous with the other words in the Synset). Synset‘s are linked by Pointers into a network of related concepts; this is the Net in WordNet. getTarget retrieves the targets of these links, and getPointer retrieves the pointers themselves.

（2）、A Pointer encodes a lexical or semantic relationship between Wordet entities. A lexical relationship holds between Words; a semantic relationship holds between Synsets. Relationships are directional: the two roles of a relationship are the source and target. Relationships are typed: the type of a relationship is a PointerType, and can be retrieved via getType.

（3）、An IndexWord represents a line of the pos.index file. An IndexWord is created retrieved or retrieved via DictionaryDatabase.lookupIndexWord(edu.gwu.wordnet.POS, java.lang.String), and has a lemma, a pos, and a set of senses, which are of type Synset.

（4）、POS：Instances of this class enumerate the possible major syntactic categories, or part‘s of speech. Each POS has a human-readable label that can be used to print it, and a key by which it can be looked up.

（5）、A Word represents the lexical information related to a specific sense of an IndexWord. Word‘s are linked by Pointers into a network of lexically related words. getTarget retrieves the targets of these links, and getPointer retrieves the pointers themselves.

（6）、A PointerTarget is the source or target of a Pointer. The target of a semantic PointerTarget is a Synset; the target of a lexical PointerTarget is a Word.

（7）、PointerType：Instances of this class enumerate the possible WordNet pointer types, and are used to label PointerTypes. Each PointerType carries additional information: a human-readable label, an optional reflexive type that labels links pointing the opposite direction, an encoding of parts-of-speech that it applies to, and a short string that represents it in the dictionary files.

Structure

The main relation among words in WordNet is synonymy, as between the words shut and close or car and automobile. Synonyms--words that denote the same concept and are interchangeable in many contexts--are grouped into unordered sets (synsets). Each of WordNet’s 117 000 synsets is linked to other synsets by means of a small number of “conceptual relations.” Additionally, a synset contains a brief definition (“gloss”注释) and, in most cases, one or more short sentences illustrating the use of the synset members. Word forms with several distinct meanings are represented in as many distinct synsets. Thus, each form-meaning pair in WordNet is unique.

Relations

The most frequently encoded relation among synsets is the super-subordinate relation (also called hyperonymy, hyponymy or ISA relation). It links more general synsets like {furniture, piece\_of\_furniture} to increasingly specific ones like {bed} and {bunkbed}. Thus, WordNet states that the category furniture includes bed, which in turn includes bunkbed; conversely, concepts like bed and bunkbed make up the category furniture. All noun hierarchies ultimately go up the root node {entity}. Hyponymy relation is transitive（传递性的） . WordNet distinguishes among Types (common nouns) and Instances (specific persons, countries and geographic entities). Thus, armchair is a type of chair, Barack Obama is an instance of a president. Instances are always leaf (terminal) nodes in their hierarchies.

Meronymy, the part-whole relation holds between synsets like {chair} and {back, backrest}, {seat} and {leg}. Parts are inherited from their superordinates: if a chair has legs, then an armchair has legs as well. Parts are not inherited “upward” （parts不能向上传递）as they may be characteristic only of specific kinds of things rather than the class as a whole: chairs and kinds of chairs have legs, but not all kinds of furniture have legs.

Verb synsets are arranged into hierarchies as well; verbs towards the bottom of the trees(troponyms：方式词a word that denotes a manner of doing something ) express increasingly specific manners characterizing an event, as in {communicate}-{talk}-{whisper}. The specific manner expressed depends on the semantic field; volume (as in the example above) is just one dimension along which verbs can be elaborated. Others are speed (move-jog-run) or intensity of emotion (like-love-idolize极端崇拜). Verbs describing events that necessarily and unidirectionally entail one another are linked: {buy}-{pay}, {succeed}-{try}, {show}-{see}, etc.

Adjectives are organized in terms of antonymy(反义词). Pairs of “direct” antonyms like wet-dry and young-old reflect the strong emantic contract of their members. Each of these polar adjectives in turn is linked to a number of “semantically similar” ones: dry is linked to parched炎热的, arid 干旱的, dessicated干燥 and bone-dry and wet to soggy透湿的；沉闷的, waterlogged, etc. Semantically similar adjectives are “indirect antonyms” of the contral member of the opposite pole. Relational adjectives ("pertainyms"--:meaning relating to or pertaining to) point to the nouns they are derived from (criminal-crime).

There are only few adverbs in WordNet (hardly, mostly, really, etc.) as the majority of English adverbs are straightforwardly derived from adjectives via morphological affixation (surprisingly, strangely, etc.)

Cross-POS relations

The majority of the WordNet’s relations connect words from the same part of speech (POS). Thus, WordNet really consists of four sub-nets, one each for nouns, verbs, adjectives and adverbs, with few cross-POS pointers. Cross-POS relations include the “morphosemantic”形态语义 links that hold among semantically similar words sharing a stem with the same meaning: observe (verb), observant (adjective) observation, observatory (nouns). In many of the noun-verb pairs the semantic role of the noun with respect to the verb has been specified: {sleeper, sleeping\_car} is the LOCATION for {sleep} and {painter}is the AGENT of {paint}, while {painting, picture} is its RESULT.

WordNet介绍和使用（转）

from http://blog.csdn.net/ictextr9/archive/2009/03/20/4008703.aspx

Wordnet是一个词典。每个词语(word)可能有多个不同的语义，对应不同的sense。而每个不同的语义（sense）又可能对应多个词，如topic和subject在某些情况下是同义的，一个sense中的多个消除了多义性的词语叫做lemma。例如，“publish”是一个word，它可能有多个sense：

1. (39) print,publish -- (put into print; "The newspaper published the news of the royalcouple's divorce"; "These news should not be printed")

2. (14) publish,bring out, put out, issue, release -- (prepare and issue for public distribution or sale; "publish a magazine or newspaper")

3. (4) publish,write -- (have (one's written work) issued for publication; "How manybooks did Georges Simenon write?"; "She published 25 books during herlong career")

在第一个sense中，print和publish都是lemma。Sense 1括号内的数字39表示publish以sense 1在某外部语料中出现的次数。显然，publish大多数时候以sense 1出现，很少以sense 3出现。

WordNet的具体用法

NLTK是python的一个自然语言处理工具，其中提供了访问wordnet各种功能的函数。下面简单列举一些常用功能：

得到wordnet本身：

from nltk.corpusimport wordnet

获得一个词的所有sense，包括词语的各种变形的sense：

wordnet.synsets('published')

[Synset('print.v.01'),

Synset('publish.v.02'),

Synset('publish.v.03'),

Synset('published.a.01'),

Synset('promulgated.s.01')]

得到synset的词性：

>>>related.pos

's'

得到一个sense的所有lemma：

>>>wordnet.synsets('publish')[0].lemmas

[Lemma('print.v.01.print'), Lemma('print.v.01.publish')]

得到Lemma出现的次数：

>>> wordnet.synsets('publish')[0].lemmas[1].count()

39

在wordnet中，名词和动词被组织成了完整的层次式分类体系，因此可以通过计算两个sense在分类树中的距离，这个距离反应了它们的语义相似度：

>>> x =wordnet.synsets('recommended')[-1]

>>> y =wordnet.synsets('suggested')[-1]

>>> x.shortest\_path\_distance(y)

0

形容词和副词的相似度计算方法：

形容词和副词没有被组织成分类体系，所以不能用path\_distance。

>>> a =wordnet.synsets('beautiful')[0]

>>> b =wordnet.synsets('good')[0]

>>>a.shortest\_path\_distance(b)

-1

形容词和副词最有用的关系是similar to。

>>> a =wordnet.synsets('glorious')[0]

>>>a.similar\_tos()

[Synset('incandescent.s.02'),

Synset('divine.s.06'),

……]

一。WordNet中的名词

\* 在WordNet 1.5版中包含了差不多80000名词 -- 60000词汇化的概念；其中许多都是collocation（搭配型词）。

\* WordNet跟其他传统词典的差别，主要不是在词义以及覆盖面方面，还是信息的组织方式的创新方面。

\* 传统的词典包括：拼写、发音、屈折变化形式、词源、派生形式、词性、定义以及不同意义的举例说明、同义词和反义词、特殊用法说明、临时用法等；

\* WordNet不包括发音、派生形态、词源信息、用法说明、图示举例等。WordNet尽量使词义之间的关系明晰并易于使用。

\* WordNet中的基础语义关系是synonymy（同义关系）。同义词集合（synset）构成了WordNet的基本建筑单位（building block）。Ravin（1992）已经开发了一些程序用于从同义词词林中抽取同义词集合。但WordNet的这类工作是手工进行的。

\* WordNet中的同义概念并不是指在任何语境中都具有可替换性。如果以这样的标准来衡量同义关系，语言中的同义词就少得很了。

\* {shot, pellet小球；[军] 小子弹（枪用）} 跟 {shot, injection} 之间没有同义关联，尽管两个synset中都有shot。

\* 大多数同义词集合（synset）有说明性的注释（explanatory gloss）相伴。这跟传统的词典情况类似。不过一个synset不等于词典中的一个词条。尤其是词典中的一个词条可能是个多义词（polysemous word），它就会包含多个解释，而一个synset只包含一个注释。

1 词汇层级（lexical hierarchy）

{robin, redbreast} @ -> {animal, animate\_being} @-> { organism, life\_form, living\_thing},

\* 一个名词通常只有一个直接上位词，因而编词典的人用这个上位词来定义该名词；一个名词通常不只一个下位词，因而编词典的人一般很少罗列这些下位词。

2 名词的25个基本类别为：

{act,activity}

{food}

{possession}

{animal,fauna}

动作行为

食物

所有物

动物

{group,grouping}

{process}

{artifact}

{location}

团体

过程

人工物

处所

{quantity,amout}

{attribute}

{motivation,motive}

{relation}

数量

属性

动机

关系

{body}

{natural\_object}

{shape}

{cognition,knowledge}

身体

自然物

外形

认知，知识

{natural\_phenomenon}

{state}

{communication}

{person,human\_being}

自然现象

状态

通信

人类

{substance}

{event,happening}

{plant,flora}

{time}

物质

事件

植物

时间

{feeling,emotion}情感

\* 这25类也可进一步概括为11个基本类

\* 由25个语义类形成的有关名词的25个元文件在语义层次上一般都是比较浅的。尽管对语义层次没有严格限制，但很少有超过10到12层的语义树，通常层次比较深的情况是由于专业词汇造成的，而不是日常语言中的用词。比如：

shetland pony @-> pony @-> horse @-> equid @-> odd-toed ungulate @-> placental mammal @-> mammal @-> vertebrate @-> chordate @-> animal @-> organism @-> entity

（共12层，其中6个层次是专业词汇，而非日常用词）

3 一些心理学假设（some psycholinguistic assumptions）

\* 尽管名词层级的一般结构是由上下位关系（hyponymy/hypernymy）产生的，但这些知识在人脑的词库中如何表示并不清楚。我们不妨假设概念的细节是由区别特征来刻画的。例如，知更鸟（robin）是一种鸟，有颜色，会唱歌，等等；显然，我们不仅要在robin和bird之间建立起上下位关系，还需要将“颜色”（color），歌唱（singing），飞（flying）等特征跟robin关联起来。此外，robin还需要从bird那里继承一些特征，比如恒温脊椎动物（warm-blooded vertebrate），有喙（beak），翅膀（wings），羽毛（feathers），孵蛋（lay eggs），等等。为了达到上述目的，至少要考虑以下三方面的特征：

（1）属性（attributes）: 是知更鸟属（redbreasted），恒温脊椎动物，

（2）部件（Parts）：beak, feathers, wings

（3）功能（functions）: sings, flies, lays eggs

尽管不同的区别特征应该被分别对待（比如attribute是形容词，部件是名词，功能是动词等），但大多数认知理论以同样的方式处理上述特征。给每个同义词集合（synset）配上诸如此类的特征集，就可以得到下面这样一种简单的对下位关系进行定义的方式：

如果同义词集合{ A }的所有特征被包含在同义词集合 { B }的特征集中，而不是相反，那么，{ B } 是 { A } 的下位概念（hyponym）。

\* 如果上下位关系依靠特征来定义，那么特征就成为非常重要的概念。对任何给定的同义词集合，用来定义它的特征必须是特定的，同时又必须是足够的。许多认知学家怀疑所有的词是否能如此容易地依靠定义性的特征来加以刻画。

\* 尽管大多数词汇学家和计算机科学家认为词汇层级是表示名词意义（nominal meaning）的一种自然的方式，但许多认知科学家对此表示怀疑。当然认知学者们并不是怀疑概念之间客观存在的上下位关系，而是对人脑词库中如何识别这些关系的认知过程表示疑惑。当人们去寻找不同的特征集合（list of feature）之间存在包含关系（inclusion relation）的经验证据时，上述问题就出现了。

\* 1969年，Collins 和 Quillian 报告说，人们证实 句子 “ A robin is a bird ”所需要的时间，比证实句子 “ A robin is an animal ” 所需要的时间 短。 于是他们主张，上述观察提供了心理学证据，证明词汇层级的距离的确对认知有影响 ： 在词汇层级上距离越远，就需要更多的时间来思考。这一可能性刺激出数量上相当可观的研究工作，不过一系列的问题最终导致人们对词汇层级的认知现实性的严重质疑。

\* 1981年，Smith 和 Medin非常谨慎地提出了反面意见。他们发现，感知“ a chicken is a bird ”所需的时间比感知 “ a robin（ ['rɔbin] n. 知更鸟）is a bird ”所需的时间长，尽管 chicken 和 robin 跟bird的分类关系是一样的。这个差异不是 robin 出现的频率比 chicken 高造成的。而是 作为鸟类的成员，robin比chicken更典型。在传统的经典分类理论中，典型性是没有地位的。

\* 在1980年代中期，对词汇层级的经典解释的攻击如此有诱惑力，以至于当时WordNet一出现，就被设想为有可能从根本上揭示为何词汇层级系统不是一个好的揭示词语意义的框架。事实上，WordNet的作者们很快就使自己确信，动词，形容词，副词都是以不同的方式组织的。不过，名词的层级结构似乎可以适应语言事实，尽管缺少好的解释。

\* 例如，句子“ A pistol is more dangerous than a rifle ”（手枪比步枪更危险）显示的意义是可以理解的。而“ A pistol is more dangerous than a gun ” （手枪比枪更危险）和“ A gun is more dangerous than a pistol ”就都不好理解（Bever and Rosenbaum 1970）。很显然，语言学上，比较句中两个比较项不能是上下位关系的两个词。而这条句法规则要发挥作用，必须基于给定的名词上下位关系知识基础上。

\* 再看一个例子，“ I gave him a good novel, but the book bored him ”容易理解，但“ I gave him a good novel, but the catsup bored him ”就让人费解。 因为词汇知识告诉我们，novel是book（有上下位关系），但不是catsup（番茄酱）。显然，要得出上述判断，也必须依赖名词之间的上下位关系。

\* 再一个例子，一些动词的搭配选择限制也表明名词上下位关系的重要性。比如动词“drink”的直接宾语可以是 beverage （饮料）的任何一个下位词。这也暗示有关名词的上下位关系的知识应该以一种人们能够快速访问和搜索到的方式存贮。

\* 由此，名物概念的层级组织方式似乎是大脑中的心智词典的一个必然特征。尽管它遭到认知理论的强烈攻击，但层级原则对名词词典来说确实重要。那么又如何看待典型性（typicality）呢？也许，典型性跟层级型二者可以并存。这就说，在名词的层级系统上，应该存贮更多的信息。除了一个名词在词汇层级系统上的位置，我们还需要知道更多有关这个名词的信息，但这并不意味着层级系统本身就不重要了。

\* WordNet的组织方式是提供语义关系的指针，而不是罗列特征。WordNet中的名词由同义词集合（synset）组成，synset之间有指针指示它们的语义关系（上下位关系）。在WordNet中，特征信息并不是以明确的显性形式来定义上下位关系或名词的层级关系。

4 WordNet中不包含的东西（some things not in WordNet）

\* WordNet中不包含这样的描述 ：企鹅不是会飞的鸟。蜘蛛不是昆虫。……等。

即 不包含“is not a (kind of)”这样的关系。

\* @->实际代表了不止一种语义关系，至少包括“is a kind of”和“is used as a kind of”两种关系。{chicken} @-> {bird} 是前者；{chicken} @-> {food}是后者。遗憾的是，WordNet中没有对上述区别加以描述。

5 整体部分关系（meronymy）包括三种情况：（WordNet中包含了这三种关系）

A是B的组成部分； beak / wing -> bird

A是B的成员； tree -> forest

A是B的构成材料。 aluminum （[ə'lju:minəm] n. 铝）-> plane

\* 在WordNet中，整体部分关系主要是在noun.body（身体部件名词）, noun.artifact（人工物），noun.quantity（数量名词）三个文件中。

\* 有时候，整体部分关系跟上下位关系会出现一些纠缠的现象。比如，{beak,bill,neb,nib}不仅是{bird}的部件，同时也是{jaw}的下位概念。而{jaw}又是{skull}的部件，是{skeletal\_structure}的下位概念。在下位关系和整体部分关系之间建立适合关系常常会产生的问题是，会有将部件名词放在名词层级系统中过高位置的倾向。比如，如果{wheel}是{vehicle}的部件，那么{sled}（雪橇）就不是vehicle了。在WordNet中，为此创建了一个中间概念，{wheeled\_vehicle}（带轮子的交通工具）。

\* 另一个重要的值得说明的问题是，“is a part of”通常用来测试整体部分关系，但这种测试方式并不总是可靠。比如，在许多例子中，传递性（transitivity）是受到限制的。Lyons(1977)注意到，“handle”是“door”的部件，“door”是“house”的部件，但如果我们说，“handle”是“house”的部件，就比较奇怪。Winston, Chaffin, Hermann(1987)举的例子更极端，我们可以说，“the branch is a part of the tree” 以及 “ the tree is a part of the forest ”，但不能因此导出，“the branch is a part of the forest ”，因为 “branch / tree ”之间的关系跟 “ tree / forest ”之间的关系不同。

6 Antonymy（反义关系）

\* 两个词构成反义关系的最强烈的心理学指示是，当给出其中一个词时，它的反义词（另一个词）通常就是那个最容易联想到的词。比如，当人们听到“victory”（胜利）时，最容易想到的就是“defeat”（失败），反之亦然。

\* 上述反义关系通常出现在所谓的降格形容词性名词(deadjectival noun)之间。比如名词“happiness”和“unhappiness”之间的反义关系来自反义形容词“happy”和“unhappy”之间的反义关系。这些 deadjectival noun都是{attribute}（属性）的下位概念。

\* 语义相反不是名词之间的基本组织关系，但这种关系确实存在，在WordNet中也有所反映。例如 [ { man } !-> { woman } ] 和 [ { woman } ! -> { man } ]都在源文件中存在。但这种反义关系不能被 { man } 和 { woman } 的下位概念继承。也许关于名词之间反义关系最有趣的是，两个反义名词一般总是有共同的直接上位概念。

7 属性和修饰语（attribute and modification）

\* 属性的值由形容词表达。例如，“size”和“color”是 “robin”（知更鸟）的两个属性。而知更鸟的size则由形容词“small”（小）来描述；color则由形容词“red”来描述。名词由此可以看作是属性词的论元。例如 SIZE (robin) = small, COLOR (robin) = red

\* 在WordNet中，{robin}和{red}之间不直接发生关联，而是通过{color}跟{red}之间发生关联。

\* friendliness（友好性）是dog的一个属性，因此我们可以说“ a dog can be friendly or unfriendly ”，但 “ stingy dog ”或“ shallow dog ”就只能解释为比喻用法。因为 generosity 和 depth 都不是 dog 的正常属性。

\* 形容词修饰成分在WordNet中扮演的主要角色可能是在搭配词或合成词中。比如 easy chair , electric chair, straight chair, high chair等，都比基本概念层级的名词 chair 更详细地刻画了 chair。

8 多义名词的相似意义/近义 similar meanings of polysemous nouns

\* Philip N. Johnson-Larid提出，如果一个词的两个意义是相似的，那么它们的下位词的意义也应该以同样的方式相似。例如：fish的两个意义，一个是动物（fish1），一个是食物（fish2）。fish的下位词perch,sole,bass等也都是如此。

\* 不过也有不同于上面情况的例外，比如 coral （珊瑚虫）是 animal 的下位概念之一。coral最后会形成 ocean reef （珊瑚）； 同时， coral 还有一个意思也是 food 的下位概念，是指 lobster roe （龙虾卵）。显然，这里的animal 和 food的下位概念之间的关系不是 动物 和 它的可食的肉之间的关系。因此，需要对这种例外情况加以特别说明。

\* 近义是最近才加入到WordNet中的关系。而且只在名词中得到了实施，这方面的工作仍在继续。

\* WordNet中区分了三种不同的近义情况：

（1） cousins （堂兄弟节点） 比如上面 例子 fish 的两个意思（概念）之间就是。

（2） sisters （姊妹节点） 比如 flounder 可以指不同的扁平鱼（比目鱼），在两种不同的指称意义下，它们的直接上位词是一样的（都是 flatfish ）。这种情况下，flounder的两个意义就是 sisters 近义。

（3） twins （孪生节点） 这是指两个同义词集合（synset）共享同样的词形式（word form），这两个同义词集合中的词形式应该有3个以上是共同的。比如，duo的一个意思是 musical group （二人音乐小组，或二重唱组合），另一个意思是 musical composition （音乐作品，二重唱音乐作品）。跟这两个意思相关的同义词集合中都包含{ duo, duet, duette }。五 WordNet中的动词

Collins英语词典中有43636个不同的名词和14190个不同的动词，动词的多义性比名词更突出。Collins词典中，名词平均意思为1.74，而动词是2.11。

（一） 动词在WordNet中的组织方式

1 将词库变成按语义组织的域 （breaking up the lexicon into semantic domains）。对动词进行分类，第一刀可能是分成event和state两类，而大多数动词可能属于前一类，这一类又进一步分成14个更专门的类：

verbs of motion

verbs of perception

verbs of contact

verbs of communication

verbs of competition竞争；比赛

verbs of change

verbs of cognition

verbs of consumption

verbs of creation

verbs of emotion

verbs of perception

verbs of possession

verbs of bodily care and functions

verbs referring to social behavior and interactions

（詹按：这里原文似有错误，出现了两个perception，有重复。另外，在文后所附动词语义基本分类中有weather/气象类，这里没有提到）

上面这个分类，部分基于Miller 和 Johnson-Laird 1976年的讨论，部分基于这种分类本身的合理性，因为这样分类似乎可以令得所有动词都有一个适合的类。当然，其他的研究人员可能有不同的分类方式，但是，不管是分类的数量还是类的名称，就WordNet中对动词词库进行分析的观念而言，都不是非常重要的。

上述这15个类是适合WordNet中所有动词的分类需求的。在多年发展之后，WordNet 1.5版大约包含11500个动词同义词集合。但是，必须强调，动词义类之间的界限是模糊的。比如许多动词无法毫不含糊地归入cognition类，也无法归入communication类。比如 wonder, speculate, confirm, judge等等。类似地，像 whistle这样的动词，在“The bullet whistled past him”中的whistle，既可归入sound emission类（发出声音），也可以归入运动类动词（motion），参见Atkins 和 Levin 1991的讨论。如果这样的动词被判定为是单义的，它们就不得不归属一个以上的语义类中。在WordNet中，是倾向于把这些动词处理为多义动词，如果一个动词确实可以归入两个语义类（semantic domain）。不过，一个动词归入那个具体的语义类，并不是件很要紧的大事，因为在WordNet中，一个动词的意义是通过它跟其他动词的关系以及同义词集合（synset）来表达的。

2 动词的第一级分类（起始类 unique begineer）：

Lyons(1977)提出一个分类集为：act, move, get, become, be, make。

Pulman(1983)的分类只有be和do，这实际是区分动作动词（activity）和状态动词（stative verb）。Jackendoff（1983）的分类是event和state。

首先，这些用来代表基本分类的动词几乎都是多义的。WordNet对do和be都进行了意义辨析，分别都至少得到12个意义（这个数字并不过分，一般的词典也是如此区分的）。当然其中的一些意义不适合作为基本分类使用，比如 “do my hair”或“do my room in blue”中的“do”已经是特定含义的动词了。其次，我们发现，我们设置的用来创建网络的，特别的语义关系，使得把这类抽象动词（如do）跟下一级语义类的动词如communicate和move链接起来不大容易了。例如，尽管communicate和chat之间存在上下位的层级关系，但do和communicate之间的同样的关系似乎就不太合适了。这两个概念似乎比communicate和chat之间的距离要远得多。同样的，move和run之间可以很容易地关联起来，但do和它的直接下位词move似乎就隔得很远。第三，似乎缺乏心理语言学的证据表明人们是把do跟行动动词如move关联起来的，而有证据显示人们是把move跟run联系在一起的（参见Chaffin, Fellbaum, 和Jenei 1994）。

采用Lyons(1977)和Pulman(1983)提出的初始分类似乎跟WordNet反映说话者词汇知识的愿望不能统一起来。因此，WordNet设置了更多的动词初始分类单元，达到14个类。

在一个单独的语义框架内，经常面对的情况是并非所有的动词都能组织到一个初始语义类概念之下。一些语义域只能被几个独立的树来表达。例如，运动动词（Motion verbs）有两个同形的顶级结点（top node）：move1和move2，分别表示位移类运动和无位移类运动。领属动词则可以上升到三个顶级结点：分别由三个同义词集合表达，即{give, transfer}， {take, receive} 和 {have, hold}。在很大程度上，它们的下位词就体现了使用该语言的社群对“领属”关系转移的编码方式，如“bequeath/遗赠, donate/捐赠, inherit/继承, usurp/篡夺, own/拥有, stock/进货 ”等。Communication动词由一级结点communicate打头，下面分成两个独立的树，分别表示语言交流和非语言的交流（gestural/手势语交流）。这两种情形在英语中没有相应的词汇化的体现。语言交流又被次分为口头交流和书面交流两种。其他语义类，例如身体动作类动词，由许多相互独立的语义类组成。这些语义类构成一个相互关联的语义场。如wash/洗, comb/梳, shampoo/洗发, make up/化妆, ache/疼痛, atrophy/萎缩。社会交互类动词（social interaction）也包含许多不同的语义子类，如政治类动词（elect, depose/免职），工作（hire/雇佣， subcontract/转包， strike/罢工），以及人际关系类动词（court/献殷勤, marry/婚嫁），等等。

3 动词的同义词集合（synset）

（1） 同义和近义（synonyms and near-synonyms）：

如果采用可替代性作为判断同义的标准，那么只有很少的动词是同义的。rise和fall可以选择抽象名词如“temperature”（温度）或“prices”（价格）等作为论元，但它们相近的同义词ascend和descend则不行。WordNet原则上避免把这样的动词放到一个同义词集合中。不过，由于许多显然属于同义的动词之间也存在类似的分布差别，因此注释和样例就显得很有帮助了，通过注释和例句，我们可以看出同义动词之间在使用上的差异。

（2）成语和比喻义（Idioms and Metaphors）：

像“kick the bucket”和“keep an eye on”这样的固定成语包含在适当的同义词集合（synset）中，那些有比喻意义（非字面含义）的动词也是如此，比如“die”可以用在汽车抛锚或计算机死机的情况下，因而die在WordNet中也被包含在break和break down所在的同义词集合中。有比喻义的动词不仅意义上跟相关动词相近（如die跟break和break down），而且用法上也有同样的句法特征（如die跟break和break down都是非宾格动词），并且一般可以用被动形式加以解释（如This task fell to me中的非宾格动词fall就可以解释为be assigned；The building went up中的go也可解释为be erected，等等）

（二） 动词之间存在词汇和语义关系的证据（Evidence for lexical and semantic relations among verbs）

1 动词语义存贮的心理语言学证据：

尽管有相当多的研究工作研究说话人对名词和形容词的语义记忆，但心理语言学家不太注意研究动词的语义记忆问题。动词的词联想数据是稀疏的。WordNet的设计基于例证关系，而不是适应不同句法类词语之间的直接链接。

除词联想数据外，替换错误也是研究词语和概念的心理组织的数据源。Garrett（1992）报告，说话人在一些动词的替换使用中出现错误情形，如ask和tell，go和come等。显然，这些动词对来自同一个语义范畴。这些替换错误被Garrett分类成反义对：start-stop, remember-forget, believe-doubt, ask - tell, preced - follow, fill - empty, love - hate, heard - said, taken - given。等等。这些动词对跟那些在刺激-反应实验中发现的联想动词对类似。（Palermo & Jenkins 1964, Chaffin, Fellbaum, and Jenei 1994）。在后一种心理实验中，反义是刺激和响应之间的一种主要关联。这类心理语言学证据表明，语义相反关系是说话人组织心理词库的一种重要方式。WordNet即包含了许多这样的动词对。

2 典型性和范畴成员关系判定的证据：

来自典型性和成员隶属程度判断的数据对词语和概念的研究也是一个很好的数据资源。不过这方面的研究同样也主要集中在名词上，例外是Pulman的研究工作（Pulman 1983）。他的主题是判断一些动词比另一些动词更典型地隶属于某个范畴。例如，murder是kill这个范畴的典型成员，隶属程度很高。survey则是look这个范畴的典型成员，相反，sacrifice和squint，之于kill和look，则都不太典型。这样的结果对心理词库建模是有帮助的。但是，Pulman的数据只局限于8个范畴，每个范畴涉及6个动词。

3 词典释义作为发现语义关系的一种启发式方法：

传统的词典在发现词语和概念之间联系方面提供了丰富的资源。名词通常用上位词来定义。词典也给出了动词的分类层级启示。例如，动词shuffle在韦氏词典（第三版）中被定义为，“to move or walk in a sliding, dragging manner without lifting the feet”； shout被定义为，“to utter in a loud voice.”这种类型的定义显示，一些动词是另一些更通用的动词经过一些修饰造成的。另一些动词不大好找上位词，就用它的否定方面来进行定义，例如，lose被定义为，“fail to sustain, maintain, or keep”。（这类似于词典在定义形容词时用“not X”这样的形式）。词典在定义许多动词时还采用“x while y-ing”这样的形式。这种定义有助于启示继承关系。许多有及物和开始状态的动词在词典中是以“to become or make X”的形式定义的（例如melt/使融化，被定义为，“to become or make liquid”）。在WordNet中，这两种意义是加以区分的，并通过CAUSE指针链接。

（三） 动词和同义词集合中的词汇与语义关系（Lexical and semantic relations among verbs and synsets）

1 继承/蕴涵（entailment）：

有些动词之间存在蕴涵关系，如snore（打鼾）蕴涵了sleep（睡觉）。“He is snoring”蕴涵了“He is sleeping”。动词之间的蕴涵关系很像名词中的meronymy（整体部分关系）。

人们接受上述事实，就可以认为这样的动词对是部分-整体关系的，如drive 是 ride的一部分；snoring 和 dreaming（做梦）是sleeping的一部分等。因为当一个人drive（驾驶）的时候，他必然是ride（乘）在交通工具上的。从时间性上讲，前一个动词必然在后一个动词发生的时段内。

（1） 下位关系（hyponymy among verbs）：

动词之间不同于名词之间的上下位关系，用来测试名词之间上下位关系的句子框架：An X is a Y，不大适合用来测试动词。Ambling is walking，To amble is kind of to walk或者Mumbling is talking都不被接受。

Fellbaum和Miller(1990)创造了一个术语troponymy方式关系来称说动词中的下位关系。动词之间的troponymy关系可以表示为公式：To V1 is to V2 in some particular manner. 这里Manner可以解释为多种语义维度，比如fight的下位词包括battle, war, tourney, joust, duel , feud等等，这些词都是从不同角度更具体地来刻画一个fight。

（2） 下位关系和蕴涵关系（troponymy and entailment）：

下位关系是一种特别的蕴涵关系。march是walk的下位动词，march同时也蕴涵walk。walk被包含在step中（即walk蕴涵step），但walk不是step的下位动词。snore蕴涵sleep，但snore也不是sleep的下位动词。

（3） 动词的分类树关系（verb taxonomies）：

动词层级分类体系通过下位关系实现，倾向于浅层分类。在大多情况下，分类层级不超过4层。例如：

Communicate - talk - [ babble / - mumble / - slur / - murmur / - bark ] - write

在动词分类层级体系中，动词下降一层，该层上的动词所能搭配的潜在论元（名词）的多样性也就相应地下降。例如figure（图形）或picture（照片）可以communicate和talk，甚至还可以deceive（欺骗）或lie（撒谎），但它们不能fib（撒小谎）或perjure（作伪证）了，只有人可以fib或perjure。有时候，把一个动词安置到单个的上位动词下非常困难。这跟名词的情形有些类似，例如piano在WordNet中既是一种“stringed instument”（弦乐器），同时也是一种“percussion instrument”（打击乐器）。就动词的例子而言，sprawl既可以是四肢分开地sit（坐），也可以是四肢分开地lie（躺）。Pulman(1983)提到，许多动词用于表示杀人的不同方式（比如hang, stangle, shoot, 等等），既可以是kill的方式，也可以是execute（行刑）的方式（即killing某人实际上是一种惩罚的形式，这样的惩罚是由社会提出加以制裁的）。人们希望语义上下位关系可以使得语义记忆模式显得经济一些，即下位概念的属性可以从上位概念继承。对动词的上下位语义层级分类，目标也同样如此。communicate（三价动词）的下位动词就从它们的上位词继承论元结构 -- 即communicator（发信者）, message（信息）, recipient（接收者）等。同样地，尽管speak的下位动词之间可能在方式上各不相同，但它们都共享speak的词义联想内容（声带振动，嘴唇和舌头的动作等等）

2 动词中的语义相反关系（Semantic opposition among verbs）：

动词词库的关系指针OPPOSITION实际上代表了复杂的若干种语义关系。例如，

converse是一种反义关系，这种反义关系没有共同的上位词或蕴涵动词，比如give/take; buy/sell; lend/borrow; teach/learn等等。这些词的共同之处在于出现在同样的语义场中，指称同样的行为，但各自联系的表层句子结构中的论元角色不同。

大多数反义动词是状态动词。这样的动词可以通过属性来表达各自之间的反义关系。如live/die; exclude/include; differ/equal; wake/sleep等等。

此外反义关系也经常出现在变化动词中，如lengthen/shorten; strengthen/weaken; prettify/uglify等等。

跟形容词中的反义关系类似，动词中的反义关系也存在标记问题（markedness）如，tie/untie和appear/disappear，前者是无标记的，后者是有标记的。

许多语义上相反的动词在分类层级上是姐妹关系（co-troponyms），比如rise/fall，walk/run等。

另有一些语义相反的动词对，共同蕴涵一个动词，如fail和succeed都蕴涵try，forget和remember都蕴涵know等。

3 致使语义关系（The cause relation）

这种关系连带两个动词概念，一个是因（如give），另一个是果（如have）。英语中词汇化的致使动词对的例子如：show - see（展现 - 看见）和 fell - fall （击倒 - 倒）等。

此外，WordNet包含从致使（causative）动词和及物动词，到相应的非致使动词和不及物动词的CAUSE语义关系指针。这主要出现在变化动词中。例如blacken, develop, break, shrink等。大多数非致使动词暗示了一个有生施事或无生的致使动因（如The glass door broke -- The storm/ The children broke the glass door）。少数动词只跟无生致使动因相关，如The wooden deck molded -- All that rain molded the wooden deck。比较：The house sitter molded the wooden deck是不合法的句子。致使动词也系统地出现在运动动词中，如bounce, roll, blow等等。（She blew a soap bubble in his face -- The soap bubble blew in his face）。Carter(1976)提到，致使关系是蕴涵关系中的一种特殊情形：如果V1必引起V2，则V1也就蕴涵V2。比如expel/驱逐和leave/离开，或bequeath/遗赠和own/拥有。跟所有的蕴涵关系一样，致使关系是单向的（unidirectional）。

下图显示了WordNet中全部4种不同的动词蕴涵关系。

（四） 多义性 （Polysemy）：

据Fellbaum(1990)，英语动词中的多义词大约是名词的两倍。那些高频动词（have, be, run, make, set, go, take等）的意义通常依赖于跟它们共现的名词。

1 多义性与下位关系：自我上下位关系 （Polysemy and Troponymy: Autohyponymy）

如behave既可以表示“conduct oneself”，又可表达“conduct oneself well”，后者（behave2）是前者（behave1）的下位动词。这就是所谓的自我上下位关系。

2 多义性与蕴涵：自我蕴涵关系（Polysemy and Entailment）

韦氏大词典第三版列有drive的两个义项如下：

to operate and steer a vehicle;

to have oneself carried in a vehicle

请看例句：The president drove on to the Capital.

在这个例句中，drive的上述两个义项的意义区别就不再清晰了，因为总统可能是驾驶汽车的人，也可能只是乘客。再前一种情况下，“他驾驶”就蕴涵了他同时也是被交通工具承载的人。

再如：She sewed her dress 是一个歧义句。

sew有两个意思，一是fasten or join(fabric) by stitching（靠缝合使织品更紧凑或连到一起）；一是make or create by sewing（通过缝制生产衣服）。后一种意思蕴涵前一种意思。因此，如果这个例句是“她做了一件衣服”的意思，则这个例句蕴涵“她通过缝合使这些布片紧凑地组织到一起”的意思。

3 多义词与相反关系：自我反义关系 （Polysemy and Opposition: Autoantonymy）

例1. These manufacters now bone their shirt collars.（在领子中加领衬/硬领）

例2. She boned the turkey. （把火鸡的骨头抽去 - debone）

例3. The violinist strung his instrument. （给乐器加弦）

例4. Could you string these beans, please? （用线把豆子串起来）

Horn(1988)指出，在例1和例3中，动词选择“goal”（目标）论元；而在例2和例4中，动词选择的是“source”（来源）论元。同样形式的动词在不同的用法中构成自我反义关系。

例5. \* We strung the beads and the beans.

例6. \* She seeded the clouds and the grapes.

例5和例6之所以不合法，是因为在一个句子中，同时包含了上述两种论元（语义上不相容）。

（六） 动词词典（语义建模）的其他模式 （alternative models of the verb lexicon）

1 语义场理论 （semantic field）

从语义场理论的视角来看词典的结构，有些类似WordNet中的语义域。语义场分析方法的基础是这样一个信念，即一个给定语义场中的词的意义来自于它跟该语义场中其他词之间的相似和相反关系。但是，语义场分析跟WordNet中的情况还有所不同，语义场分析通常既考虑聚合关系（paradigmatic），又考虑组合关系（syntagmatic）。在WordNet中，关于一个词的选择优先性的信息不是网络结构的一部分，而更倾向于是跟一个同义词集合（synset）相伴随的说明性材料，这样的信息通常具体地标示什么样类型的名词是作为一个Synset中的动词的典型论元出现的。

2 脚本和框架分析（Schemata and Frame Analysis）

一些语言学家和词典学家认为纯粹的关系分析对描述说话人的动词词库表达是不够的。他们提议采用认知框架或者知识脚本来描述说话人的经验和信念，并提供表示词语意义的概念基础（例如Schank和Abelson, 1977）。

Fillmore和Atkins（1972）提议编纂一个基于框架的词典。在这样的词典中，词语意义跟认知结构或框架相连。比如商业交易框架（commercial transaction frame），其中包括{money, buyer, seller, goods}以及动词{buy, sell, charge, cost, etc.}。

尽管WordNet并没有把动词跟特定的名词关联起来，但它提供了每个动词的句法框架，来指示动词的确定数量的论元。但是，在这样的句法框架中的名词槽，目前既没有跟论旨角色（thematic role），也没有跟像buyer这样的语义范畴联系起来。

因此，WordNet可以在这方面有所加强，即在跟动词的每一个synset相伴随的框架中，可以将语义和论旨角色跟名词论元关联起来（link）。这样，原来处于反义关系（opposition relation）中的动词对，如buy和sell，就可以在名词论元的差别上显现出来。

3 义素分析（compositional analyses）

尽管通过组成成分的意义来建构整体的意义这一处理方式跟从关系角度来刻画词语意义不大相同，但二者之间仍有重合之处。WordNet中通过词语之间关系来刻画词语意义，其中也反映了义素分析的某些方面。比如在义素分析法看来，give这个动词的组成成分之一是CAUSE，而在WordNet中，这是通过give跟have这两个动词之间有CAUSE关系来体现的。

4 词汇从属理论（lexical subordination）

Levin 和 Rapoport（1988）指出，许多动词在语义上实际由其他动词组成。例如：

例1 brush the tangles out -- remove the tangles by brushing

例2 brush a hole in one's coat -- create a hole by brushing

例3 brush the coat clean -- cause to become clean by brushing

再比如：he nodded his assent （他点头表示同意）

在WordNet中，nod是一些基本动词的Troponym，比如nod的一个义项就是“express by nodding”。它的上位词是gesture（体态语），类似的词包括nod, shrug, wink等。

（七） 语义关系和句法规律性（Semantic relations and syntactic regularities）

对动词的语义和概念组成成分进行分析能够揭示出许多动词的句法属性。词汇概念结构相同的动词倾向于有共同的句法行为。Levin(1985, 1993)测试了大量的动词，语义上同类的英语动词，在句法模式上也有系统的对应性。

1 有区别的子树（distinguishing subtrees）

同一层级的动词的句法属性一般就是对这类动词进行次分类的基础。

例如：weave（编织）， mold（铸模）这两个词都是creation verb（创造类动词）。许多创造类动词都可以有下面这样的句式变换（材料/成品 交换句法位置）：

例1 She wove a rug from the black sheep's wool. （她用黑绵羊的毛织了一个毯子）

例2 She wove the black sheep's wool into a rug.

例3 They molded a head from the clay. （他们用粘土塑了一个人头）

例4 They molded the clay into a head.

有一些动词，像fabricate（编造）和compose（写），也是creation类动词中的成员，但不能像上面的动词那样变换句式。例如：

例5 The reporter fabricated a story out of the girl's account. （那个记者根据那个女孩所说的编出了一个故事）

例6 \* The reporter fabricated the girl's account into a stroy.

例7 She composed a quartet out of the old folk song. （她根据这首老民歌创作了一个四重奏）

例8 \* She composed the old folk song into a quartet.

在Fellbaum 和 Kegl (1988)对这些动词的讨论中，他们指出，有必要对创造类动词进行更精细的分类，一类是精神上的创造活动（如fabricate和compose），一类是对具体材料的加工（如weave和mold）。这样的区分就能解释这些动词在大多数情况下的系统差异。

WordNet注意到了上述区别，因而将creation类动词区分为“create from raw material”和“create mentally”两类，尽管这两个概念在英语中并没有词汇化。

2 动词在树结构中的位置的句法意义（syntactic reflexes of the verb's position within a tree structure）

从语义关系的角度观察动词同样也提供了理解动词句法性质的一条线索。Fellbaum和Kegl（1989）研究了英语中的一类动词（涉及到及物/不及物的变换）。例子如下：

例9 Mary ate a bag of pretzels. （玛丽吃了一袋椒盐脆饼干）

例10 Mary ate.

以往对这些动词的分析根据时态来进行解释。但是，对eat的下位动词的分析显示，可以将这些动词分为两种情况，一种是必须以及物用法出现，一种是总以不及物用法出现。前一类包括gobble， guzzle, gulp和 devour等（基本上都相当于汉语的“狼吞虎咽”），后一类包括dine（进餐）, graze（放牧，吃草等）, nosh（吃点心）, snack（吃零食）等。Fellbaum和Kegl提议，上述句法性质的差异并不仅仅是及物变换方面的不同，而是由语义激发的差别。他们认为，英语中有两个eat，分别占据语义网上的不同位置。不及物的eat的意思是“eat a meal”（吃饭）。它的下位词一部分来自名词，如dine, breakfast, picnic, feast等。另外一部分是munch, nosh, graze等。相比之下，及物的eat的意思是“ingest in some manner”。这种意义下，eat的下位词是指以特定方式eat的那些动词：gobble, gulp, devour，等等。

3 中间形式制约 （Restrictions on Middle Formation）

从语义分类树观察动词还能清楚地显示基于语义的特定的句法约束。

一个明显的结构是那些通常需要副词或副词短语出现才能站住的句子。比如：

例11 Her new novel sells \* ( fast / like hotcakes )

\* 她的新小说卖

她的新小说卖得很快

她的新小说很好卖

例12 This car drives \* ( easily / like a dream )

\* 这辆车开

这辆车很容易开

但是，有的情况下对副词出现与否就要求不高。比如：

例13 This vegetable microwaves (easily).

例14 This suitcase zips shut (in a flash).

对副词出现与否的限制要求可以用动词的语义层级来说明。下位动词对副词出现与否的要求相对宽松。例如：cook有许多下位动词，包括 fry, broil, braise, microwave等等。这些动词都表达了cook的某种特定方式。同样地，close的下位动词包括velcro, button, zip, snap等。这些动词都表达了以特定的方式来close。我们来看两个上位动词的例子：

例15 This vegetable cooks \* ( quickly ).

例16 This suitcase closes \* ( easily ).

在上面这两个句子中，副词都是必需的。因为其中的动词是“基础”层级（basic level）的动词，在人们的心理中，这样的动词本来就被期待是与特定的名词（它们的主语）相关联的。如果这些动词只是简单地来陈述它们的主语，就引不起听者特别的兴趣，因此，特别说明这些动词的具体方式（通过副词来表现）就是必要的，而且是自然的了。注意，这些句子的重音都落在副词上。

相比之下，如果动词是下位动词，副词性修饰成分就不是必需的了。因为这些动词本身就已经能给出新信息。microwave 意味着 “cook in a microwave oven”。 zip 意味着 “close with a zipper ”。

Ackerman和Goldberg (1996)注意到下面的例子，也可以从动词语义层级的角度来进行解释。他们注意到，一些来自动词过去分词的形式不太适合出现在名词前位置作限制性定语（attributive）

例17 \* a killed man

例18 \* a changed design

例19 \* a told secret

相比之下，如果动词在语义层级树上处在更下位的位置，它的过去分词形式就比较容易出现在名词前位置。

例20 a murdered man （一个被谋杀了的人）

例21 an altered design （一个被更改了的设计）

例22 a divulged secret （一个被泄露的秘密）

Ackerman和Goldberg得出结论说，这些不适合作修饰成分的动词，都没有提供足够的信息来限定它后面的中心动词，他们提出如下原则：

一个形容词性的动词过去分词如果来自基础层级的动词，就不适合作修饰语出现在名词前位置修饰限定中心名词，相反，如果动词是更下位的谓词，它的过去分词形式就容易起修饰作用。

附：

---------------------------

动词的基本语义类：

1. 身体动作动词（Verbs of Bodily Functions and Care）; 275个同义词集合

2. 变化动词（Verbs of Change）；约750个同义词集合

3. 通信动词（Verbs of Communication）； 710个以上的同义词集合

4. 竞争动词（Competition Verbs）；200个以上的同义词集合

5. 消费动词（Consumption Verbs）；130个同义词集合

6. 接触动词（Contact Verbs）；820个同义词集合

7. 认知心理动词（Cognition Verbs）；

8. 创造动词（Creation Verbs）；250个同义词集合

9. 运动动词（Motion Verbs）； 500个同义词集合

10. 情感心理动词（Emotion or Psych Verbs）；

11. 状态动词（Stative Verbs）；约200个同义词集合

12. 感知动词（Perception Verbs）；约200个同义词集合

13. 领属动词（Verbs of Possession）；约300个同义词集合

14. 社会交互（Verbs of Social Interaction）；约400个同义词集合

15. 气象动词（Weather Verbs）；约66个同义词集合

不同语义类的动词有非常不同的结构。其中一些可以进入到以上下位关系进行分类的框架中：这主要是“创造类动词、通信类动词、竞争类动词、接触类动词、运动类动词以及消费类动词（verbs of creation, communication, competition, contact, motion, and consumption）。状态动词和变化动词则有完全不同的结构，它们倾向于以反义和同义关系组织起来。跟形容词类似，它们也可以投射到两极对立的分类框架中。

四 WordNet中的修饰语

所有的语言都提供了修饰词语意义的方式，尽管在句法形式上各个语言的修饰手段不尽相同。英语中修饰成分的句法类主要是形容词和副词。形容词修饰名词；副词修饰名词之外的其他语言成分，包括动词，形容词，其他副词，小句或整句。

形容词的唯一功能是修饰名词（比如large chair中的large； comfortable chair中的comfortable）。此外，英语允许那些句法上属其他类的成分起到跟形容词同样的功能。比如名词起形容词功能：kitchen chair， barber chair等，或者动词的现在分词，过去分词，creaking chair, overstuffed chair等，介词词组也可以修饰名词，如chair by the window， 从句也可修饰名词，如 the chair that you bought at the auction。

WordNet 1.5包含了16428个形容词同义词集合，包括许多名词、分词（participles）、和介词词组。如前所述，这些成分通常都作为修饰性成分使用（比如home cooking, home office 中的home）。

WordNet把形容词区分为两类：描写性形容词（descriptive adjectives）和关系性形容词（relational adjectives）。前者如big、 beautiful、 interesting、 possible、 married、 ……；后者因其跟名词的关系而得名，如electrical engineer中的electrical实际跟名词electricity相关。

这两类形容词应该能代表英语形容词中的绝大多数，但并不是说就覆盖了所有的形容词。

（一）描写性形容词

1 反义（antonymy）

反义关系是描写性形容词之间的基本语义关系，心理学上的证据是，当一个人听到一个描写性形容词时，如果问关于这个形容词的最熟悉的词的第一反应是什么，首先反应出来的通常就是它的反义词。

如果认识到形容词是表达属性值的一种主要手段，而差不多所有的属性都有两极对立的值存在，就不难理解在描写性形容词的组织方式中，反义关系为何如此重要。例如，heavy的反义词是light，这两个词表示的是WEIGHT这个属性的两个相反的级值。

上述解释引发出两个相关问题：

（1）当两个形容词意义非常相近时，为什么它们的反义词不同？比如，heavy跟weighty的意义非常接近，但前者的反义词是light，后者的反义词是weightless？

（2）反义关系如此重要，但为什么许多描写性形容词似乎没有反义词呢？比如，仍以weight为例，ponderous（笨重）的反义词是什么？如果说ponderous的反义词是light，那当人们问light的反义词时，答案一定是heavy。

上面第一个问题造成的影响是，WordNet中用特定的箭头符号表示synset之间的语义关系，但反义关系似乎不能在两个synset之间建立起来。比如{heavy, weighty, ponderous}和{light, weightless, airy}之间如果用特定箭头表示反义关系似乎就不大合适。说英语的人对 heavy/weightless, ponderous/airy之间是否构成反义似乎倾向于表示怀疑。尽管概念上它们是反义的，但在词形式层面，似乎不是人们熟悉的那些反义词对形式。

于是有了下面这样的表示间接反义的方式：

反义关系 antonymy ，用 !-> 符号表示；

近义关系 similarity， 用 &-> 符号表示；

由此，就有在下图中显示的 moist &-> wet !-> dry 这样的语义关系。moist不直接跟dry构成反义，但通过图中的路径构成间接反义。

图1： 两极形容词结构

2 Gradation（等级/序关系）

大多数有关反义关系的讨论涉及到contrary （相对反义）和 contradictory （绝对反义）两种反义关系的区分，前者是可分程度的逐级变化，后者是两级对立的反义关系。“生”与“死”是绝对反义关系，二者是非此即彼的关系，即二者不能同真，也不能同假；“胖”与“瘦”是相对反义关系，二者不能同真，但可以同假。

下面是一个形容词序关系的一个例子

一个有等级的形容词可以被定义为能被程度副词very, decidedly, intensely, rather, quite, somewhat, pretty, extermely等等修饰的形容词（Cliff, 1959）。大多数形容词等级是通过形态变化规则（比较级，最高级）来实现的。

在形容词的synset之间用一种特定指示符号表示等级（序）关系并非难事，但估计在2500个以上的形容词类中不会超过2%的形容词是以这种方式组织的。既然等级关系在形容词概念组织中并不扮演中心角色，因此WordNet中并没有对等级关系加以描述。

3 标记性（markedness）

一些形容词表现出明显的倾向性。如汉语中的“高/矮”，其中“高”表示“高度”这个维度的方向（即缺省值），相比之下，“矮”就显得很突出，因而是“marked”（有标记的）。这可以比较“这座楼有30米高 -- \*这座楼有30米矮”看出。

4 多义性，选择优先（Polysemy and Selectional Preferences）

Justeson和Katz （1993）发现，多义形容词（比如old, right, short）的不同意思跟它同现的特定名词有关。例如，old的意义是“not young”时，它经常修饰“man”，而如果“old”的意义是“not new”时，它经常修饰像“house”之类的词。因此，形容词周围的名词可以用来对多义形容词进行排歧。Murphy和Andrew (1993)则对此持另一种观点。他们认为形容词是单义的，但有不同的外延（extension）。Murphy和Andrew断言，说话者是通过形容词所修饰的名词的意义来计算形容词的适合的意义的。他们还进一步反对这样的观点，即反义关系是两个形容词词形之间的关系。他们举例说，fresh的反义词将根据它是修饰shirt还是修饰bread而有所不同（译者注——修饰shirt时反义词是used；修饰bread时，反义词是stale）。上述观点也被Justeson和Katz（1993）采纳，他们指出不同的反义词可以用来区别多义形容词的不同义项。

（二） Reference-Modifying Adjectives

Bolinger（1967）首次指出reference-modifying 和 referent-modifying 形容词之间的区别。他指出，在“the former president”（前总统）这样的短语中，former是指所指对象的“总统身份”（president-hood），而不是指所指对象本身的某种性质；而在“my old friend”中，old可以解释为reference-modifying的形容词，这时是指朋友关系老（朋友本身也许是个年轻人）；old也可以解释为referent-modifying的形容词，这时是指朋友的年纪老（翻译成中文相当于“我的老伙计”）。由此，形容词old的两个意思有不同的反义词：作为reference-modifying意义时，反义词是recent / new ； 作为referent-modifying意义时，反义词是young。

reference-modifying形容词的功能很像副词，比如“my former teacher”意思是“he was formerly my teacher”；“the alleged killer”意思是“she is allegedly a killer”。这类形容词只能作为定语出现修饰名词，不能作谓语。可以比较“my friend is old.”，其中的old就已经不再是reference-modifying形容词了。

（三） Color Adjectives

英语中的颜色形容词同时也可以作为名词使用，但它们不是名形词，它们可以有比较级，最高级，可以名词化；不过这类形容词似乎没有直接和间接反义词。

只有一个颜色属性LIGHTNESS（亮度）的两个值是极性的：light/dark。

（四） 关系性形容词（relational adjectives）

关系性形容词是形容词中另一个大的开放类。

这种形容词只能出现在定语位置（attributive position）上。意义上跟一个名词非常相关。例如 fraternal twins （双卵双胞胎）/ fraternal 跟 brother意义相关；dental hygiene（牙齿卫生） / dental 跟 tooth相关。

有的名词既能被relational adjective修饰（作谓语受限），又能被描写性形容词修饰，例如musical instrument和musical child。前者是描写性的，指用在音乐方面的器具/乐器（an instrument used in music）；后者是关系性的，指有音乐天赋的小孩。类似地，“criminal law”跟“criminal behavior”不同。后一种情况下criminal是referent-modifying形容词，可以作谓语使用。

关系性形容词不能跟描写性形容词一道修饰中心名词，特别是两个形容词由连词连接时，比如“nervous and life-threatening disease”和“musical but not extraordinary talent”都是听起来比较怪的说法。（life-threatening nervous disease就是可以接受的说法，这表示关系形容词扮演着修饰性名词的作用）。与此相对，关系性形容词能跟修饰性名词联合起来修饰名词，比如 atom and nuclear bombs ; the Korean and Vietnam war。

关系性形容词大多数是来自希腊语或拉丁语，很少来自Anglo-Saxon（盎格鲁-萨克逊语）。英语词汇中通常有几个（同义的）形容词，来自不同语言中表达相同概念的名词。比如来自希腊语的rhinal（鼻腔）和来自盎格鲁萨克逊的nasal都跟名词nose有关。跟word相关的关系性形容词是verbal（来自拉丁语）和lexical（来自希腊语）。在多数情况下，这些同义词各自挑选它们自己的中心名词来修饰，在一个给定的上下文中不能被替换。比如 nasal passage / \* rhinal passage ; rhinal surgery / \* nasal surgery

相反地，一个关系性形容词有时候指向几个名词。比如：chemical 有两个意思跟两个名词一致：chemical （例子： chemical fertilizer / 化学肥料）和 chemistry （例子： chemical engineer / 化学工程师）

关系性形容词跟描写性形容词不一样的地方在于，关系性形容词不是跟一个属性相关（比如：没有跟关系形容词 criminal 和 musical 相对应的属性名词 criminality 和 musicality ），因而也不是作为属性值的形容词。

关系性形容词在概念上跟它相关的名词基本一样，只是词形式不同。

关系性形容词不指向它修饰的中心名词的属性（property）。这可以从关系性形容词没有相应的名词化形式看出。比如，nervous的描写性用法 the nervous person 可以变换结构为 the person's nervousness ， 但 nervous的关系性用法 the nervous disorder 没有平行的变换形式。

关系性形容词像名词，而不是描写性形容词，前者没有等级序关系，因此关系性形容词不受程度副词修饰（ \* the extremely atomic bomb ; \* the very baseball game ）， 而描写性形容词可以受程度副词修饰。

关系性形容词没有直接反义词，尽管它们常常可以跟 non- 连用。不过这样的形式不是表示一个属性取相反的属性值，而是表示“其他”（everything else）的意思。这类形容词有一种分类功能。在不少情况下，关系性形容词以前缀方式引入反面事物。比如：extracellular （细胞外的） vs. intracellular （细胞内的）。更常见的情况是，关系性形容词在跟特定中心名词搭配时，对事物进行分类，比如civil lawyer vs. criminal lawyer ； mechanical engineering vs. electrical engineering。

在WordNet中。一个关系性形容词的同义词集合包含一个相应的名词指针。例如：{stellar, astral, sidereal, noun.object: star}，表示关系性形容词stellar, astral, sidereal跟名词star相关。

不少形容词出现的句法位置是比较固定的，对此，WordNet也作了说明。并且这样的说明一般是针对特定形容词，而不是针对形容词的同义词集合的。例如：awake/asleep这两个形容词都只能出现在谓语位置，就以awake(p)这样的标记加以说明。

Bartning (1980) 观察到，如果形容词修饰的中心名词是由动词派生出来的，那么如果中心词指行动，形容词可以转化到谓语位置，如果指状态，形容词一般不能转化到谓语位置。比如 economic restructuring 可以转换为 the restructuring was economic ； 而 economic slump 则不能转换为 \* the slump is economic。

Bartning (1980）还进一步观察到，如果形容词跟中心名词之间有明显的语法关系，就不大能变换到谓语位置；如果没有明显的关系，一般可以变换到谓语位置。比如 presidential election （ president 是 elect 的结果/对象 ），二者之间有明显的关系，所以不能将 presidential 变换到谓语位置，不说 \* the election is presidential 。而在 manual labor （手工劳动）这个短语中，它的意思是 labor with/by hand ， manual 跟 labor之间没有显性的关系，因此可以说 this labor is manual。

形容词的反义是对词而言的，不是对概念而言的。比如large/small；big/little构成反义关系，而large跟little，big跟small并不构成反义关系，尽管从概念上讲，它们可以构成反义关系。large跟small；big跟little之间的反义关系则可以从语料中的搭配共现和联合关系反映出来。

关于WordNet的不成熟的想法可以追溯到20多年前，而这一想法开始逐渐具体化和清晰化则是1985年后才开始的。从85年开始，WordNet作为一个知识工程全面展开。不过，当时的WordNet和经过10多年后今天的WordNet还是很不一样的。

·  
这一工程最初的前提之一是"可分离性假设"（Separability  
hypothesis），即语言的词汇成分可以被离析出来并专门针对它加以研究。词汇编纂学的历史明确地告诉我们，在词语水平上可以得到有用的研究成果。词库（词典，lexicon）当然不是完全独立于其他语言成分的，但它的确是可以从其他成分中分离出来的。例如，尽管语音和语法知识在一个人的早年生活中就成型了，但词汇量却可以随着智力活动的不断积累而增加。这表明语言的不同成分涉及不同的认知过程。

· 另一个前提是"模式假设"（patterning  
hypothesis）：一个人不可能掌握他运用一种语言所需的所有词汇，除非他能够利用词义中存在的系统的模式和词义之间的关系。这种系统化的心智模式至少从柏拉图时代就成为一种进行推测的学问，现代语言学研究开始在自然语言的语义结构中识别这样的模式。但许多遵循这类路线的出色的研究工作在这一问题上碰到了困难。一个作者可能提出一种语义理论，并以20到50个英语单词为例来展示他的理论，而留下另外10万个单词让读者去做练习。

·  
第三个前提就是所谓的"广泛性假设"（comprehensiveness  
hypothesis）：计算语言学如果希望能像人那样处理自然语言，就需要像人那样储存尽可能多的词汇知识。

·  
建立包含词语意义描述的大规模词库的方式之一是基于语义成分分析的词汇语义学（componential  
lexical  
semantics）的方法（也可译为义素分析法）。这种方式把一个词的意义分析为更小的概念原子的组合。不过，定义一套概念原子却非易事。事实上，WordNet主帅George.A.Miller在1976年他与Philip  
N. Johnson-Laird合作的《Language and  
Perception》一书中还踌躇满志地探索义素分析的语义描写方法，但直到1985年，仍然没有能够出笼一个完整的定义清晰的清单，在上面列举出所有的概念原子。

·  
到1985年，许多认知心理学家和计算语言学家开始以"网"的形式来描述词语的意义。比如："桌子"（table）和"家具"（furniture）代表两个节点（node），而这两个节点之间有一个箭头（dart）来表示这样的命题：桌子是一种家具（a  
table is a kind of  
furniture），即"Is-A-KIND-OF"这样的语义关系。随着这方面研究的增多，越来越多的人自觉地意识到：除了利用语义成分（义素分析法）表示语义，还可以利用关系来表示语义（基于关系的词汇语义学relational  
lexical semantics），而且后者有可能替代前者。

·  
在WordNet的早期阶段，研究人员主要是在考虑用关系语义来描述词义的方式是否能够大规模地广泛使用，而不是仅仅停留在玩具式的演示水平上。到了研究人员确信这是可行的的时候，他们就编制了应用软件来把想法变成现实。实际上，在早期，Miller并没有关于构建一个大词库的完整想法。初步设想是识别由字符串组成的最重要的词节点，并探索其中的语义关系模式。当时的想法是，如果得到了正确的语义关系模式，词语的定义就能从中推理出来，因此，对于一个有关词义的关系网来说，词义的定义是多余的。

·  
在1978年的时候，Miller描述了一种"自动化词典"（automated  
dictionary）的想法。不过那时候他完全不知道该如何实现这种想法。由于Sloan基金会，Spencer基金会，IBM公司沃盛研究中心（Watson  
Research  
Center）的支持，Miller得以一直保持着他的想法，而没有中途放弃。到1984年的时候，Miller甚至在IBM  
PC机上做出了45个名词的小型语义网，他把这个小网叫做"word  
net"。Miller在IBM和Bellcore演示了这个示例成果。他在Bellcore的一帮好友，Lance  
Miller, Roy Byrd, Michael Lesk, Donald Walker, Robert Amsler,  
以及Stephen  
Hanson都鼓励他继续下去，并在技术上给予许多实际指导。

·  
Lesk邀请Miller参加了1985年11月在加拿大沃太卢（Waterloo）大学新牛津英语词典中心的第一次会议。Miller提交了一篇论文。Miller在那篇论文中解释这样的思想：  
我们可以使用同义词集合（synset）来代表词汇概念，并描述词汇矩阵，即在词的形式和意义之间建立起映射关系（mapping）。实际上，这正是在WordNet的发展中指导研究工作的主要思想。

·  
不过，也许Miller在那次会议上的报告所介绍的思想尚在其次，报告的标题反而意义更大。那个标题是由Lesk提议的，叫做：WordNet:  
A Dictionary  
Browser。（WordNet：一个词典浏览器）。在这里，WordNet被看作（设想为）是一个词典浏览器，是一个机器可读词典的辅助工具。而这样一个机器词典不是传统地按字母排序的，是基于意义的。

·  
也就是在WordNet开始成形的时候，Miller和他的一些志同道合的同事在普林斯顿合作发起了一个认知研究的计划（Program  
for Cognitive Studies）。Richard  
Cullingford从1983年到1985年在普林斯顿访问，他，Gilbert  
Herman，加上Miller说服Provost Neil  
Rudenstine提供了一台微型计算机。那台计算机也就成了普林斯顿认知科学实验室的心脏。有了设备在手，Miller又去说服海军研究办公室的Susan  
Chipman跟他们签合同来开发WordNet。Cullingford和Miller一起收到了军队研究所（Army  
Research  
Institute）的一个合同，开发一个词汇语义学的计算理论。1986年3月，James  
S.McDonnell基金向普林斯顿慷慨解囊，支持认知科学的研究工作。Marie  
BienKowski（Cullingford的研究生）和一个熟练的计算机专家，从1985年开始跟我们一道工作，不久就拿出了我们需要的软件。由此，WordNet真正成为普林斯顿新成立的认知科学实验室几项研究计划中的一个，并开始实际运作。

·  
用来创建WordNet的最重要的程序是所谓的Grinder（磨床）程序。Bienkowski在1986年用LISP语言写了Grinder的第一个版本。Dan  
Teibel在1987年用[**C语言**](http://lib.csdn.net/base/c)重写了这个程序。Antonio  
Romero在1989年又重写了一次。Randee  
Tengi从1991年开始负责管理该程序的所有这些版本。

· WordNet中的词来自不同的地方。Brown语料库、Laurence  
Urdang的同义反义小词典（1978）、Urdang修订的Rodale同义词词典（1978）、以及Robert  
Chapmand的第4版罗杰斯同义词词林（1977）等。1986年下半年，Miller得到海军研究与发展中心的Fred  
Chang的一个词表，Miller将Chang的词表跟WordNet已有的词表进行了比较，令人沮丧的结果是只有15%的重合词语，于是Miller把Chang的词表加入到WordNet中。1993年，Miller得到了Ralph  
Grishman和他在纽约大学的同事的一个词表，39143个词，这个词表实际上包含在著名的COMLEX词典中。这一次比较的结果是，WordNet中只包含了COMLEX中74%的词。于是Miller又把这个词表加入到WordNet中。

·  
随着词表长度的增加，组织工作的压力开始增加。第一步区分是通过句法范畴来进行的。我们创建了不同的文件用于存放名词、动词、形容词（1992年后加入副词）。但在每个句法范畴内，仍有太多的词语，如果没有进一步分类就难以把握。名词由于是开放的词类，因而问题最严重。

· 1987年春，Philip N.  
Johnson-Laird从英国剑桥应用心理学研究所来到普林斯顿大学访问。他发现WordNet中缺乏手段来区分形容词的意义在修饰不同名词时所发生的变化。他利用266对反义形容词（当时WordNet包含了这些反义形容词对），根据它们修饰名词的适合度来分出名词的次类。结果大约得到名词的25个次类。这些次类构成了名词的基础分类。而且这些不同次类的名词也相应地被分成不同的个别文件交给不同的研究人员去分头加工，从而从工程上推动了WordNet的进展。

· 1987年夏，Christiane  
Fellbaum加入到研究队伍中来，承担了对动词进行次分类的工作。

· 只有Kitty Miller负责的描写性形容词（descriptive  
adjective），从一开始就一直保持是一个大类，没有次范畴化，所有的这些形容词也保存在一个巨大的文件中。

· WordNet从一个简单的"词典浏览器"（dictionary  
browser）发展成一个自足的词汇[**数据库**](http://lib.csdn.net/base/mysql)（self-contained  
lexical  
database），主要的进步是从1989年年初开始的。当时SusanChipman不满于WordNet仅仅作为一个词汇浏览器而存在，要求研究小组开发一个工具。该工具可以在WordNet的基础上阅读一个文本，并报告文本中词语的各种信息。这一工具即所谓的"Word  
Filter"（词过滤器）。罕用的或不符合需要的词能够从小说文档中被过滤出去，而同时更常见的词语可以用来替代这些词。这个工作很快使我们意识到必须对词形的曲折变化进行处理。这使得我们处理了有关词形方面的一些问题，WordNet中仅包含词语的基本形式，如果文本中出现"ships"，WordNet就无法识别它。Richard  
Beckwith和Miceael  
Colon写了一个程序，叫做Morphy，可以识别出文本中的"ships"的词形式"ship"。到1989年9月，WordNet就可以处理文本中的词形变化，并在词库中找到相应的词语基本形式。

·  
上述工作导致另一个重要的进展，就是在文本中将词语跟WordNet中的相应意义关联起来。Claudia  
Leacock（1991年11月加入）和Brian  
Gustafson开发了一个界面--  
ConText，可以对文本进行预处理（实例化、词汇化、自动句法标注等），并跟WordNet一道显示目标词语（即对目标词语进行了语义标注）。1993年6月，Shari  
Landes负责把一整套的标注工具汇编到一起（包含一组标记），利用这样一套工具，可以尽最大可能产生一个无歧义的文本作为输出结果。对文本进行语义标注显著地提高了WordNet的覆盖面，即包括词汇覆盖面，也包括词义的覆盖面。

· WordNet的另一个重要变化是1989年春发生的，Antonio  
Romero修改了Grinder程序，使得它可以接受插入语定义（或理解为"注释"）作为同义词集合（synset）的一部分。随着WordNet中词语数量的增加，既要尽可能清楚地区别词义，同时又要保持同义词集合的纯粹性，就越来越困难了。换句话说，仅靠同义来定义词义是不够的。一开始，WordNet保持注释尽量得短。但一年后，注释就开始变得越来越长。注释数量也开始稳步增长。1989年4月，WordNet中有37409个同义词集合，没有注释；到1991年7月，WordNet包含44983个同义词集合，  
13688个注释（30%）；到1992年1月，WordNet包含49771个同义词集合，19382个注释（39%）；到1993年1月，WordNet包含61023个同义词集合，36880个注释（60%）；到1994年1月，WordNet中包含79542个同义词集合，58705个注释（74%）；到1995年1月，WordNet包括了91050个同义词集合，同时包含了75389个注释（占同义词集合数量的83%）。

·  
WordNet的研究人员一直把WordNet视作一个试验，而不是一个产品。因此当WordNet中的词语足够多的时候，他们就把WordNet向学术界免费公开了。Randee  
Tengi监督了WordNet一系列版本的发行。WordNet  
1.0版是1991年7月公布的；1.1版是在1991年8月公布的；1.2版是在1992年4月公布的；1.3版是在1992年12月公布的；1.4版是在1993年8月公布的；1.5版是在1995年3月公布的。目前加入WordNet邮件列表的个人和单位超过1000个。诸多研究人员下载了WordNet，并在各个不同的方面使用它。

· 对WordNet提供过资金支持的包括：海军研究室（Office  
of Naval Research），高级研究计划署（Advanced Research  
Projects Agency），James S.  
McDonnell基金，以及最近的语言学数据协作会（Linguistic  
Data Consortium）。

二 WordNet导言

·  
对于WordNet来说，10年后来清点清点得失似乎是合适的。每个参与其事的研究人员都真诚地感受到它的缺点，并且他们从未觉得这是一个"完工"了的项目。事实上，WordNet仍在继续发展中。

· "WordNet: An Electronic Lexical  
Database"一书分三部分，16章。第一部分从第1章到第4章，前3章分别介绍WordNet中的名词，形容词，动词，第4章介绍WordNet的设计细节及相关软件的情况（这主要是由普林斯顿大学认知科学实验室的研究人员写的）；第二部分和第三部分主要是由普林斯顿认知科学实验室之外的参加WordNet研究工作的研究人员撰写的。第5章和第6章描述了WordNet的改进；第7章从形式化的概念分析的角度描述了WordNet；第8到第16章讨论了WordNet的各种不同应用。

（一）计算机与词库（computers and lexicon）

·  
一个人即使不接受把人脑比作计算机的隐喻，也一定同意，计算机提供了一个良好的模式演练场，通过它，人们可以[**测试**](http://lib.csdn.net/base/softwaretest)各种关于人类认知能力的理论模型。

·  
越来越多的人认识到，一个大的词库对自然语言理解，[**人工智能**](http://lib.csdn.net/base/ai)的各方面研究都具有重要的价值。

·  
对大规模机器可读词典的需求同时也带来许多基础问题。首先是如何构造这样一个词库，是手工编制还是机器自动生成？第二，词典中应包含什么样的信息？第三，词典应如何设计，即信息如何组织，以及用户如何访问？实际上，这些问题涉及到词典的编纂方法，词典的内容，词典的使用方式这一系列非常基础的问题。

（二）构造词库数据库（constructing the lexical database）

·  构建词典的两种基本方式：自动获取 / 手工编制。

手工构建词典的优点之一是便于创建更为丰富的词条信息；其次是便于控制。

（三）WordNet的内容

· WordNet的描述对象包含compound（复合词）、phrasal  
verb（短语  
动词）、collocation（搭配词）、idiomatic  
phrase（成语）、word（单词），其中word是最基本的单位。

·  
WordNet并不把词语分解成更小的有意义的单位（这是义素分析法/componential  
analyses的方法）；WordNet也不包含比词更大的组织单位（如脚本、框架之类的单位）；由于WordNet把4个开放词类区分为不同文件加以处理，因而WordNet中也不包含词语的句法信息内容；WordNet包含紧凑短语，如bad  
person，这样的语言成分不能被作为单个词来加以解释。

·  
人们经常区分词语知识和世界知识。前者体现在词典中，后者体现在百科全书中。事实上二者的界限是模糊的。比如hit（"打"）某人是一种带有敌意的行为，这是百科知识；而hit跟strike（"击"）多多少少同义，并且hit可以带一个直接宾语论元，这是词语知识。但hit的直接宾语应该是固体（而不是像gas这样的气体），这是词语知识还是百科知识就界限模糊了。不过毫无疑问，要理解语言，这两部分知识是缺一不可的。Kay（1989）指出我们的大脑词库应该包含这两部分知识。但是百科知识太多难以驾驭，WordNet不试图包括百科知识。不过，在WordNet中，对于一些不常见的专业概念，比如不常见的植物和动物，词语知识和百科知识是融合在一起的。

（四）WordNet的设计（the design of WordNet）

·  
一般的词典都是按照单词拼写的正字法原则进行组织的。但如果为了获得词语意义信息的目的，通过词语语义属性来组织词典就更值得去做了。在线词典跟传统的纸张词典不同，允许使用者从不同的途径去访问词典信息。

·  
第一个以意义作为组织原则的词典是罗杰斯同义词词林（Roget's  
Thesaurus）。传统的词典是通过提供给用户关于词语的信息来帮助用户理解那些他们不熟悉的词的概念意义。WordNet既非传统词典，也非同义词词林。它混合了这两种类型的词典。

（五）作为同义词词林的WordNet （WordNet as a thesaurus）

·  
WordNet跟同义词词林相似的地方是：它也是以同义词集合（synset）作为基本建构单位进行组织的。用户脑子里如果有一个已知的概念，就可以在同义词集合中中找到一个适合的词去表达这个概念。

·  
但WordNet不仅仅是用同义词集合的方式罗列概念。同义词集合之间是以一定数量的关系类型相关联的。这些关系包括上下位关系、整体部分关系、  
继承关系等。

（六）作为一般词典的WordNet （WordNet as a dictionary）

·  
WordNet跟传统的词典相似的地方是它给出了同义词集合的定义以及例句。在同义词集合中包含对这些同义词的定义。对一个同义词集合中的不同的词，分别给出适合的例句来加以区分。

（七）WordNet中的关系 （relations in WordNet）

·  
不同句法词类中的语义关系类型也不同，比如尽管名词都动词都是分层级组织词语之间的语义关系，但在名词中，上下位关系是hyponymy关系，而动词中是troponymy关系；动词中的entailment（继承）关系有些类似名词中的meronymy（整体部分）关系。名词的meronymy关系下面还分出三种类型的子关系（见"WordNet中的名词"部分）。

（八）网球问题（the tennis problem）

·  
WordNet是基于同义性和反义（对义）性来描述词语和概念之间的各种语义关系类型的。由于WordNet的注意力不是在文本和话语篇章水平上来描述词和概念的语义，因此WordNet中没有包含指示词语在特定的篇章话题领域的相关概念关系。例如，WordNet中没有将racquet（网球拍）、ball（球）、net（球网）等词语以一定方式联系到一起。Roger  
Chaffin在一封私人信笺中，曾把这类问题称为"tennis  
problem"（网球问题），指的就是如何把racquet、ball、net、court  
game（场地比赛）；或者把physician（内科医生）跟hospital（医院）联系到一起。这对电子词典来说，是一个挑战。已经有一些相关的研究工作在探索如何从WordNet中包含的词汇和概念之间的语义关系，来推导出话题信息。Hirst和St-Onge描述了一种所谓的"词汇链"（lexical  
chain）的应用方法。"词汇链"是在基于名词的语义关系构成的上下文中的名词的序列。Al-Halimi和Kazman则在类似的基础上构造"词汇树"（lexical  
tree）来推导出话题信息。

（九）新的观点，改进，应用 （new perspectives,  
enhancements, and applications）

·  
许多WordNet的用户都对WordNet中缺乏跟语义处理的细节相匹配的句法信息而感到遗憾。的确，WordNet中几乎没有句法信息，因为它是作为一个语义知识库构建的。但是，对形容词的部分句法约束信息是包含在WordNet中的（考虑形容词跟中心名词的关系，以及形容词作为属性形容词作表语使用的情况）。句法对动词而言最为重要，对此，可以通过动词的名词论元、介词短语以及义素组成等不同来加以次范畴化（分出动词小类）。目前，WordNet的每个动词同义词集中包含了及物性和论元类型的基本信息，但有关这些论元的性质的细节就很少提到。知识工程以及推理方面的应用系统特别受益于动名间关系的信息。WordNet的一些用户依靠其他一些语法知识库，像COMLEX，来配合WordNet中的语义信息一道使用。事实上，有关动词的句法信息和语义信息的区分基本是人为的。Levin（1985，1993）已经收集了令人印象深刻的证据来说明动词的语义性质跟其句法行为之间的紧密联系。

（十）词语和它的上下文 （words and their contexts）

·  
为了提供词语的语境信息，普林斯顿（Princeton）认知科学实验室开发了一个语义检索工具（semantic  
concordance）--见《WordNet》一书第8章。该工具将文本和词库组成一个整体的数据库，从而使文本中的单词跟词库中合适的意义相关联。这样的语义检索工具，既可以看作是这样一个文本，其中的单词带有句法和语义信息的标注；也可以看作是一个词库，其中的词条都配有指示义项用法环境的例句。跟WordNet语义词库配合的文本是来自Brown语料库的语料（当代美国英语标准语料库）以及一个中短篇小说的全文（the  
complete text of a novella）。

（十一）意义排歧 （sense disambiguation）

·  
尽管我们很清楚，在确定的上下文中，说者赋予多义词确定的一个意义，但排歧的过程并不容易。对计算机而言，排歧需要多大的语境就是一个大问题。

·  
Leacock和Chodorow（见《WordNet》一书第11章）测试了对多义动词"serve"进行多义词歧义消解的不同策略。在三个试验中，他们发现，选择上下文的"窗口"大小为6个词比较适宜，所得结果最优；此外，当将上下文信息和WordNet中有关词语之间语义相似度的信息结合在一起使用时，排歧准确度最高。

（十二）信息检索 （information retrieval）

·  
意义排歧对许多应用来说都是关键因素，比如信息检索就是这样的应用领域。Voorhees（见《WordNet》一书第12章）解释说，要在大量文献中发现所需的文档，计算机就要在被查询词语和文档标题或摘要之间进行有效地匹配操作。Voorhees探讨了WordNet在词语匹配方面的效力，发现意义分辨方面的困难阻碍了有效利用WordNet中的语义信息。只有先依靠手工选择了概念，使得要查找的词语的意义已知，这种情况下，WordNet中的语义关系信息才对提高检索结果有帮助。

（十三）语义关系与文本连贯性 （semantic relations and  
textual coherence）

· Hirst 和  
St-Onge（见《WordNet》第13章）也讨论了上下文的问题，尤其是一个连贯的文本是如何组成的。基于语篇是由意义相关的概念串联起来的假设，他们使用了"词汇链"（lexical  
chain）概念作为评估连贯性的一种方式。Hirst和St-Onge采用词汇链来检查文本中的用词错误情况（malapropism）。他们把用词错误定义为：一个词所对应的概念跟该词所在的文本中的其他词所对应的概念无关。利用评估一个词汇链中链接强度的方法，Hirst和St-Onge认为，文本中词语之间的语义距离越大，出现用词错误问题的可能性也越大。

·  
Al-Halimi和Kazman也对信息存贮，索引，检索等问题感兴趣（见《WordNet》第14章）。他们描述了一种自动对视频会议的脚本按照话题进行索引的方法（不是按照关键词索引），以及利用话题索引结果，通过匹配对脚本进行信息检索。Al-Halimi和Kazman将话题信息描述为"词汇树"（lexical  
tree）--这是对"词汇链"的一个修正。前者对后者的革新之一是考虑了不同的语义关系类型的信息相关性。

·  
Hirst和St-Onge指出，WordNet缺乏有关两个相关词之间语义距离的信息。他们举的例子是：more  
stew than steak（焖肉比牛排多），其中"more ...  
than"是一个格式，用来连接两个语义上相关的词语。在这个例子中，两个名词（stew和steak）分属6个同义词集合（synset），显然这无法反映出它们真实的语义距离。说英语的人知道"good  
person"（好人，圣人）的两个上下位概念之间语义上是非常相似的。这两个上下位概念分别是{saint,  
holy man, holy person, angel}，{plaster  
saint}，而且这两个概念之间的相似性与它们跟第三个下位概念之间的相似性不同。第三个下位概念是{square  
shooter, straight arrow}（正人君子）。

（十四）知识工程 （knowledge engineering）

·  
WordNet的诸多应用中，最具雄心壮志的也许是知识工程（见《WordNet》一书第15，16章）。

·  
Harabagiu和Moldovan（见《WordNet》一书第16章）指出，为常识推理建模需要一个扩展的知识库，其中包括数量巨大的概念和关系。WordNet提供了前者，但在关系方面不足以支持推理。他们的解决方案是对WordNet中的注释进行排歧，得到词语之间的更多关系，从而将WordNet中的注释转变为语义网络，其中包含不同词类之间的关系。他们举了一个例子：在hungry（饿）和refrigerator（冰箱）之间存在一个路径，因为这两个标记词在food（食物）这个节点上相撞，即通过food，可以把hungry和refrigerator联系到一起，从而用于常识推理。

三 WordNet中的名词

\* 在WordNet 1.5版中包含了差不多80000名词 --  
60000词汇化的概念；其中许多都是collocation（搭配型词）。

\*  
WordNet跟其他传统词典的差别，主要不是在词义以及覆盖面方面，还是信息的组织方式的创新方面。

\*  
传统的词典包括：拼写、发音、屈折变化形式、词源、派生形式、词性、定义以及不同意义的举例说明、同义词和反义词、特殊用法说明、临时用法等；

\*  
WordNet不包括发音、派生形态、词源信息、用法说明、图示举例等。WordNet尽量使词义之间的关系明晰并易于使用。

\*  
WordNet中的基础语义关系是synonymy（同义关系）。同义词集合（synset）构成了WordNet的基本建筑单位（building  
block）。Ravin（1992）已经开发了一些程序用于从同义词词林中抽取同义词集合。但WordNet的这类工作是手工进行的。

\*  
WordNet中的同义概念并不是指在任何语境中都具有可替换性。如果以这样的标准来衡量同义关系，语言中的同义词就少得很了。

\* {shot, pellet} 跟 {shot, injection}  
之间没有同义关联，尽管两个synset中都有shot。

\*  
大多数同义词集合（synset）有说明性的注释（explanatory  
gloss）相伴。这跟传统的词典情况类似。不过一个synset不等于词典中的一个词条。尤其是词典中的一个词条可能是个多义词（polysemous  
word），它就会包含多个解释，而一个synset只包含一个注释。

1 词汇层级（lexical hierarchy）

{robin, redbreast} @ -> {animal, animate\_being} @-> { organism,  
life\_form, living\_thing},

/\* @-> 可以读作"is a"或"is a kind of"。 \*/  
/\*  
跟@->相对的符号是"~->"，可以读作"subsume"包含。  
\*/

\*  
一个名词通常只有一个直接上位词，因而编词典的人用这个上位词来定义该名词；一个名词通常不只一个下位词，因而编词典的人一般很少罗列这些下位词。

2 名词的25个基本类别为：

{act,activity}  
{food}  
{possession}  
{animal,fauna}

动作行为  
食物  
所有物  
动物

{group,grouping}  
{process}  
{artifact}  
{location}

团体  
过程  
人工物  
处所

{quantity,amout}  
{attribute}  
{motivation,motive}  
{relation}

数量  
属性  
动机  
关系

{body}  
{natural\_object}  
{shape}  
{cognition,knowledge}

身体  
自然物  
外形  
认知，知识

{natural\_phenomenon}  
{state}  
{communication}  
{person,human\_being}

自然现象  
状态  
通信  
人类

{substance}  
{event,happening}  
{plant,flora}  
{time}

物质  
事件  
植物  
时间

{feeling,emotion}

情感

\* 这25类也可进一步概括为11个基本类

\*  
由25个语义类形成的有关名词的25个元文件在语义层次上一般都是比较浅的。尽管对语义层次没有严格限制，但很少有超过10到12层的语义树，通常层次比较深的情况是由于专业词汇造成的，而不是日常语言中的用词。比如：

shetland pony @-> pony @-> horse @-> equid @-> odd-toed ungulate @->  
placental mammal @-> mammal @-> vertebrate @-> chordate @-> animal @->  
organism @-> entity  
（共12层，其中6个层次是专业词汇，而非日常用词）

3 一些心理学假设（some psycholinguistic assumptions）

\*  
尽管名词层级的一般结构是由上下位关系（hyponymy/hypernymy）产生的，但这些知识在人脑的词库中如何表示并不清楚。我们不妨假设概念的细节是由区别特征来刻画的。例如，知更鸟（robin）是一种鸟，有颜色，会唱歌，等等；显然，我们不仅要在robin和bird之间建立起上下位关系，还需要将"颜色"（color），歌唱（singing），飞（flying）等特征跟robin关联起来。此外，robin还需要从bird那里继承一些特征，比如恒温脊椎动物（warm-blooded  
vertebrate），有喙（beak），翅膀（wings），羽毛（feathers），孵蛋（lay  
eggs），等等。为了达到上述目的，至少要考虑以下三方面的特征：

（1）属性（attributes）:  
是知更鸟属（redbreasted），恒温脊椎动物，

（2）部件（Parts）：beak, feathers, wings

（3）功能（functions）: sings, flies, lays eggs

尽管不同的区别特征应该被分别对待（比如attribute是形容词，部件是名词，功能是动词等），但大多数认知理论以同样的方式处理上述特征。给每个同义词集合（synset）配上诸如此类的特征集，就可以得到下面这样一种简单的对下位关系进行定义的方式：

如果同义词集合{ A }的所有特征被包含在同义词集合 {  
B }的特征集中，而不是相反，那么，{ B } 是 { A }  
的下位概念（hyponym）。

\*  
如果上下位关系依靠特征来定义，那么特征就成为非常重要的概念。对任何给定的同义词集合，用来定义它的特征必须是特定的，同时又必须是足够的。许多认知学家怀疑所有的词是否能如此容易地依靠定义性的特征来加以刻画。

\*  
尽管大多数词汇学家和计算机科学家认为词汇层级是表示名词意义（nominal  
meaning）的一种自然的方式，但许多认知科学家对此表示怀疑。当然认知学者们并不是怀疑概念之间客观存在的上下位关系，而是对人脑词库中如何识别这些关系的认知过程表示疑惑。当人们去寻找不同的特征集合（list  
of feature）之间存在包含关系（inclusion  
relation）的经验证据时，上述问题就出现了。

\* 1969年，Collins 和 Quillian 报告说，人们证实 句子 " A  
robin is a bird "所需要的时间，比证实句子 " A robin is  
an animal " 所需要的时间 短。  
于是他们主张，上述观察提供了心理学证据，证明词汇层级的距离的确对认知有影响  
：  
在词汇层级上距离越远，就需要更多的时间来思考。这一可能性刺激出数量上相当可观的研究工作，不过一系列的问题最终导致人们对词汇层级的认知现实性的严重质疑。

\* 1981年，Smith 和  
Medin非常谨慎地提出了反面意见。他们发现，感知" a  
chicken is a bird "所需的时间比感知 " a robin is a bird  
"所需的时间长，尽管 chicken 和 robin  
跟bird的分类关系是一样的。这个差异不是 robin  
出现的频率比 chicken 高造成的。而是  
作为鸟类的成员，robin比chicken更典型。在传统的经典分类理论中，典型性是没有地位的。

\*  
在1980年代中期，对词汇层级的经典解释的攻击如此有诱惑力，以至于当时WordNet一出现，就被设想为有可能从根本上揭示为何词汇层级系统不是一个好的揭示词语意义的框架。事实上，WordNet的作者们很快就使自己确信，动词，形容词，副词都是以不同的方式组织的。不过，名词的层级结构似乎可以适应语言事实，尽管缺少好的解释。

\* 例如，句子" A pistol is more dangerous than a rifle  
"（手枪比步枪更危险）显示的意义是可以理解的。而"  
A pistol is more dangerous than a gun "  
（手枪比枪更危险）和" A gun is more dangerous than a pistol  
"就都不好理解（Bever and Rosenbaum  
1970）。很显然，语言学上，比较句中两个比较项不能是上下位关系的两个词。而这条句法规则要发挥作用，必须基于给定的名词上下位关系知识基础上。

\* 再看一个例子，" I gave him a good novel, but the book bored  
him "容易理解，但" I gave him a good novel, but the catsup  
bored him "就让人费解。  
因为词汇知识告诉我们，novel是book（有上下位关系），但不是catsup（番茄酱）。显然，要得出上述判断，也必须依赖名词之间的上下位关系。

\*  
再一个例子，一些动词的搭配选择限制也表明名词上下位关系的重要性。比如动词"drink"的直接宾语可以是  
beverage  
（饮料）的任何一个下位词。这也暗示有关名词的上下位关系的知识应该以一种人们能够快速访问和搜索到的方式存贮。

\*  
由此，名物概念的层级组织方式似乎是大脑中的心智词典的一个必然特征。尽管它遭到认知理论的强烈攻击，但层级原则对名词词典来说确实重要。那么又如何看待典型性（typicality）呢？也许，典型性跟层级型二者可以并存。这就说，在名词的层级系统上，应该存贮更多的信息。除了一个名词在词汇层级系统上的位置，我们还需要知道更多有关这个名词的信息，但这并不意味着层级系统本身就不重要了。

\*  
WordNet的组织方式是提供语义关系的指针，而不是罗列特征。WordNet中的名词由同义词集合（synset）组成，synset之间有指针指示它们的语义关系（上下位关系）。在WordNet中，特征信息并不是以明确的显性形式来定义上下位关系或名词的层级关系。

4 WordNet中不包含的东西（some things not in WordNet）

\* WordNet中不包含这样的描述  
：企鹅不是会飞的鸟。蜘蛛不是昆虫。......等。  
即 不包含"is not a (kind of)"这样的关系。

\* @->实际代表了不止一种语义关系，至少包括"is a kind  
of"和"is used as a kind of"两种关系。{chicken} @-> {bird}  
是前者；{chicken} @->  
{food}是后者。遗憾的是，WordNet中没有对上述区别加以描述。

5  
整体部分关系（meronymy）包括三种情况：（WordNet中包含了这三种关系）

A是B的组成部分； beak / wing -> bird  
A是B的成员； tree -> forest  
A是B的构成材料。 aluminum -> plane

\*  
在WordNet中，整体部分关系主要是在noun.body（身体部件名词）,  
noun.artifact（人工物），noun.quantity（数量名词）三个文件中。

\*  
有时候，整体部分关系跟上下位关系会出现一些纠缠的现象。比如，{beak,bill,neb,nib}不仅是{bird}的部件，同时也是{jaw}的下位概念。而{jaw}又是{skull}的部件，是{skeletal\_structure}的下位概念。在下位关系和整体部分关系之间建立适合关系常常会产生的问题是，会有将部件名词放在名词层级系统中过高位置的倾向。比如，如果{wheel}是{vehicle}的部件，那么{sled}（雪橇）就不是vehicle了。在WordNet中，为此创建了一个中间概念，{wheeled\_vehicle}（带轮子的交通工具）。

\* 另一个重要的值得说明的问题是，"is a part  
of"通常用来测试整体部分关系，但这种测试方式并不总是可靠。比如，在许多例子中，传递性（transitivity）是受到限制的。Lyons(1977)注意到，"handle"是"door"的部件，"door"是"house"的部件，但如果我们说，"handle"是"house"的部件，就比较奇怪。Winston,  
Chaffin, Hermann(1987)举的例子更极端，我们可以说，"the  
branch is a part of the tree" 以及 " the tree is a part of the  
forest "，但不能因此导出，"the branch is a part of the  
forest "，因为 "branch / tree "之间的关系跟 " tree /  
forest "之间的关系不同。

6 Antonymy（反义关系）

\*  
两个词构成反义关系的最强烈的心理学指示是，当给出其中一个词时，它的反义词（另一个词）通常就是那个最容易联想到的词。比如，当人们听到"victory"（胜利）时，最容易想到的就是"defeat"（失败），反之亦然。

\*  
上述反义关系通常出现在所谓的降格形容词性名词(deadjectival  
noun)之间。比如名词"happiness"和"unhappiness"之间的反义关系来自反义形容词"happy"和"unhappy"之间的反义关系。这些  
deadjectival noun都是{attribute}（属性）的下位概念。

\*  
语义相反不是名词之间的基本组织关系，但这种关系确实存在，在WordNet中也有所反映。例如  
[ { man } !-> { woman } ] 和 [ { woman } ! -> { man }  
]都在源文件中存在。但这种反义关系不能被 { man } 和  
{ woman }  
的下位概念继承。也许关于名词之间反义关系最有趣的是，两个反义名词一般总是有共同的直接上位概念。

7 属性和修饰语（attribute and modification）

\* 属性的值由形容词表达。例如，"size"和"color"是  
"robin"（知更鸟）的两个属性。而知更鸟的size则由形容词"small"（小）来描述；color则由形容词"red"来描述。名词由此可以看作是属性词的论元。例如  
SIZE (robin) = small, COLOR (robin) = red

\*  
在WordNet中，{robin}和{red}之间不直接发生关联，而是通过{color}跟{red}之间发生关联。

\*  
friendliness（友好性）是dog的一个属性，因此我们可以说"  
a dog can be friendly or unfriendly "，但 " stingy dog "或"  
shallow dog "就只能解释为比喻用法。因为 generosity 和  
depth 都不是 dog 的正常属性。

\*  
形容词修饰成分在WordNet中扮演的主要角色可能是在搭配词或合成词中。比如  
easy chair , electric chair, straight chair, high  
chair等，都比基本概念层级的名词 chair  
更详细地刻画了 chair。

8 多义名词的相似意义/近义 similar meanings of polysemous  
nouns

\* Philip N.  
Johnson-Larid提出，如果一个词的两个意义是相似的，那么它们的下位词的意义也应该以同样的方式相似。例如：fish的两个意义，一个是动物（fish1），一个是食物（fish2）。fish的下位词perch,sole,bass等也都是如此。

\* 不过也有不同于上面情况的例外，比如 coral  
（珊瑚虫）是 animal 的下位概念之一。coral最后会形成  
ocean reef （珊瑚）； 同时， coral 还有一个意思也是  
food 的下位概念，是指 lobster roe  
（龙虾卵）。显然，这里的animal 和  
food的下位概念之间的关系不是 动物 和  
它的可食的肉之间的关系。因此，需要对这种例外情况加以特别说明。

\*  
近义是最近才加入到WordNet中的关系。而且只在名词中得到了实施，这方面的工作仍在继续。

\* WordNet中区分了三种不同的近义情况：

（1） cousins （堂兄弟节点） 比如上面 例子 fish  
的两个意思（概念）之间就是。  
（2） sisters （姊妹节点） 比如 flounder  
可以指不同的扁平鱼（比目鱼），在两种不同的指称意义下，它们的直接上位词是一样的（都是  
flatfish ）。这种情况下，flounder的两个意义就是 sisters  
近义。  
（3） twins （孪生节点）  
这是指两个同义词集合（synset）共享同样的词形式（word  
form），这两个同义词集合中的词形式应该有3个以上是共同的。比如，duo的一个意思是  
musical group  
（二人音乐小组，或二重唱组合），另一个意思是  
musical composition  
（音乐作品，二重唱音乐作品）。跟这两个意思相关的同义词集合中都包含{  
duo, duet, duette }。

9 小 结

\* 在有关WordNet的更早期描述（Beckwith et al. 1991; Miller  
[et.al](http://et.al/),  
1990）中，研究人员曾经主张WordNet是基于心理语言学原则的，就如同牛津英语词典基于历史原则一样。不过这个主张并没有结出一如当初人们期望的那样的果实。事实上，WordNet在很大程度上被心理语言学家忽视了。相反，计算语言学家则对它表示出更大的兴趣。

\*  
不过，WordNet中有关名词概念的层级描述，以及关于名词间同义，反义，上下位关系，整体部分关系等的描述，确实在解释语言行为方面发挥着作用。但心理语言学家并不把WordNet中的这些描写看作是可以用来揭示人类语言心理基础的基本原则。也许他们认为更为重要的是解释更为基本的认知过程。

\*  
计算语言学家当然不这么看，只要有助于处理自然语言，甚至在某种程度上达到理解的水平，一棵关于词义概念的层级树就是有用的（心理语言学家关心的恰恰是词汇层级树上不那么明显的那些关系）。

\*  
WordNet中的名词库的开发因此更多的是受到计算语言学方面的潜在应用的推动，胜过来自认知心理学理论的推动。也许，这个结果是应该能够预见到的。正如一个基于历史原则编纂的词典，实际上对历史研究的贡献微乎其微一样。

附：一个名词语义网络的示例

四  WordNet中的修饰语

所有的语言都提供了修饰词语意义的方式，尽管在句法形式上各个语言的修饰手段不尽相同。英语中修饰成分的句法类主要是形容词和副词。形容词修饰名词；副词修饰名词之外的其他语言成分，包括动词，形容词，其他副词，小句或整句。

形容词的唯一功能是修饰名词（比如large  
chair中的large； comfortable  
chair中的comfortable）。此外，英语允许那些句法上属其他类的成分起到跟形容词同样的功能。比如名词起形容词功能：kitchen  
chair， barber  
chair等，或者动词的现在分词，过去分词，creaking chair,  
 overstuffed chair等，介词词组也可以修饰名词，如chair by  
the window， 从句也可修饰名词，如 the chair that you bought  
at the auction。

WordNet  
1.5包含了16428个形容词同义词集合，包括许多名词、分词（participles）、和介词词组。如前所述，这些成分通常都作为修饰性成分使用（比如home  
cooking, home office 中的home）。

WordNet把形容词区分为两类：描写性形容词（descriptive  
adjectives）和关系性形容词（relational  
adjectives）。前者如big、 beautiful、 interesting、 possible、  
married、  
......；后者因其跟名词的关系而得名，如electrical  
engineer中的electrical实际跟名词electricity相关。

这两类形容词应该能代表英语形容词中的绝大多数，但并不是说就覆盖了所有的形容词。

（一）描写性形容词

1 反义（antonymy）

反义关系是描写性形容词之间的基本语义关系，心理学上的证据是，当一个人听到一个描写性形容词时，如果问关于这个形容词的最熟悉的词的第一反应是什么，首先反应出来的通常就是它的反义词。

如果认识到形容词是表达属性值的一种主要手段，而差不多所有的属性都有两极对立的值存在，就不难理解在描写性形容词的组织方式中，反义关系为何如此重要。例如，heavy的反义词是light，这两个词表示的是WEIGHT这个属性的两个相反的级值。

上述解释引发出两个相关问题：

（1）当两个形容词意义非常相近时，为什么它们的反义词不同？比如，heavy跟weighty的意义非常接近，但前者的反义词是light，后者的反义词是weightless？

（2）反义关系如此重要，但为什么许多描写性形容词似乎没有反义词呢？比如，仍以weight为例，ponderous（笨重）的反义词是什么？如果说ponderous的反义词是light，那当人们问light的反义词时，答案一定是heavy。

上面第一个问题造成的影响是，WordNet中用特定的箭头符号表示synset之间的语义关系，但反义关系似乎不能在两个synset之间建立起来。比如{heavy,  
weighty, ponderous}和{light, weightless,  
airy}之间如果用特定箭头表示反义关系似乎就不大合适。说英语的人对  
heavy/weightless,  
ponderous/airy之间是否构成反义似乎倾向于表示怀疑。尽管概念上它们是反义的，但在词形式层面，似乎不是人们熟悉的那些反义词对形式。

于是有了下面这样的表示间接反义的方式：

反义关系 antonymy ，用 !-> 符号表示；  
近义关系 similarity， 用 &-> 符号表示；

由此，就有在下图中显示的 moist &-> wet !-> dry  
这样的语义关系。moist不直接跟dry构成反义，但通过图中的路径构成间接反义。

图1： 两极形容词结构

2 Gradation（等级/序关系）

大多数有关反义关系的讨论涉及到contrary  
（相对反义）和 contradictory  
（绝对反义）两种反义关系的区分，前者是可分程度的逐级变化，后者是两级对立的反义关系。"生"与"死"是绝对反义关系，二者是非此即彼的关系，即二者不能同真，也不能同假；"胖"与"瘦"是相对反义关系，二者不能同真，但可以同假。

下面是一个形容词序关系的一个例子

一个有等级的形容词可以被定义为能被程度副词very,  
decidedly, intensely, rather, quite, somewhat,  pretty,  
extermely等等修饰的形容词（Cliff,  
1959）。大多数形容词等级是通过形态变化规则（比较级，最高级）来实现的。

在形容词的synset之间用一种特定指示符号表示等级（序）关系并非难事，但估计在2500个以上的形容词类中不会超过2%的形容词是以这种方式组织的。既然等级关系在形容词概念组织中并不扮演中心角色，因此WordNet中并没有对等级关系加以描述。

3 标记性（markedness）

一些形容词表现出明显的倾向性。如汉语中的"高/矮"，其中"高"表示"高度"这个维度的方向（即缺省值），相比之下，"矮"就显得很突出，因而是"marked"（有标记的）。这可以比较"这座楼有30米高  
-- \*这座楼有30米矮"看出。

4 多义性，选择优先（Polysemy and Selectional Preferences）

Justeson和Katz （1993）发现，多义形容词（比如old, right,  
short）的不同意思跟它同现的特定名词有关。例如，old的意义是"not  
young"时，它经常修饰"man"，而如果"old"的意义是"not  
new"时，它经常修饰像"house"之类的词。因此，形容词周围的名词可以用来对多义形容词进行排歧。Murphy和Andrew  
(1993)则对此持另一种观点。他们认为形容词是单义的，但有不同的外延（extension）。Murphy和Andrew断言，说话者是通过形容词所修饰的名词的意义来计算形容词的适合的意义的。他们还进一步反对这样的观点，即反义关系是两个形容词词形之间的关系。他们举例说，fresh的反义词将根据它是修饰shirt还是修饰bread而有所不同（译者注--修饰shirt时反义词是used；修饰bread时，反义词是stale）。上述观点也被Justeson和Katz（1993）采纳，他们指出不同的反义词可以用来区别多义形容词的不同义项。

（二） Reference-Modifying Adjectives

Bolinger（1967）首次指出reference-modifying 和  
referent-modifying 形容词之间的区别。他指出，在"the  
former  
president"（前总统）这样的短语中，former是指所指对象的"总统身份"（president-hood），而不是指所指对象本身的某种性质；而在"my  
old  
friend"中，old可以解释为reference-modifying的形容词，这时是指朋友关系老（朋友本身也许是个年轻人）；old也可以解释为referent-modifying的形容词，这时是指朋友的年纪老（翻译成中文相当于"我的老伙计"）。由此，形容词old的两个意思有不同的反义词：作为reference-modifying意义时，反义词是recent  
/ new ； 作为referent-modifying意义时，反义词是young。

reference-modifying形容词的功能很像副词，比如"my former  
teacher"意思是"he was formerly my teacher"；"the alleged  
killer"意思是"she is allegedly a  
killer"。这类形容词只能作为定语出现修饰名词，不能作谓语。可以比较"my  
friend is  
old."，其中的old就已经不再是reference-modifying形容词了。

（三） Color Adjectives

英语中的颜色形容词同时也可以作为名词使用，但它们不是名形词，它们可以有比较级，最高级，可以名词化；不过这类形容词似乎没有直接和间接反义词。

只有一个颜色属性LIGHTNESS（亮度）的两个值是极性的：light/dark。

（四） 关系性形容词（relational adjectives）

关系性形容词是形容词中另一个大的开放类。

这种形容词只能出现在定语位置（attributive  
position）上。意义上跟一个名词非常相关。例如  
fraternal twins （双卵双胞胎）/ fraternal 跟  
brother意义相关；dental hygiene（牙齿卫生） / dental 跟  
tooth相关。

有的名词既能被relational  
adjective修饰（作谓语受限），又能被描写性形容词修饰，例如musical  
instrument和musical  
child。前者是描写性的，指用在音乐方面的器具/乐器（an  
instrument used in  
music）；后者是关系性的，指有音乐天赋的小孩。类似地，"criminal  
law"跟"criminal  
behavior"不同。后一种情况下criminal是referent-modifying形容词，可以作谓语使用。

关系性形容词不能跟描写性形容词一道修饰中心名词，特别是两个形容词由连词连接时，比如"nervous  
and life-threatening disease"和"musical but not extraordinary  
talent"都是听起来比较怪的说法。（life-threatening  
nervous  
disease就是可以接受的说法，这表示关系形容词扮演着修饰性名词的作用）。与此相对，关系性形容词能跟修饰性名词联合起来修饰名词，比如  
atom and nuclear bombs ;  the Korean and Vietnam war。

关系性形容词大多数是来自希腊语或拉丁语，很少来自Anglo-Saxon（盎格鲁-萨克逊语）。英语词汇中通常有几个（同义的）形容词，来自不同语言中表达相同概念的名词。比如来自希腊语的rhinal（鼻腔）和来自盎格鲁萨克逊的nasal都跟名词nose有关。跟word相关的关系性形容词是verbal（来自拉丁语）和lexical（来自希腊语）。在多数情况下，这些同义词各自挑选它们自己的中心名词来修饰，在一个给定的上下文中不能被替换。比如  
  nasal passage  /  \* rhinal passage ; rhinal surgery / \* nasal surgery

相反地，一个关系性形容词有时候指向几个名词。比如：chemical  
有两个意思跟两个名词一致：chemical （例子： chemical  
fertilizer / 化学肥料）和 chemistry （例子： chemical  
engineer / 化学工程师）

关系性形容词跟描写性形容词不一样的地方在于，关系性形容词不是跟一个属性相关（比如：没有跟关系形容词  
criminal 和 musical 相对应的属性名词 criminality 和  
musicality ），因而也不是作为属性值的形容词。

关系性形容词在概念上跟它相关的名词基本一样，只是词形式不同。

关系性形容词不指向它修饰的中心名词的属性（property）。这可以从关系性形容词没有相应的名词化形式看出。比如，nervous的描写性用法  
the nervous person 可以变换结构为 the person's nervousness ，  
但 nervous的关系性用法 the nervous disorder  
没有平行的变换形式。

关系性形容词像名词，而不是描写性形容词，前者没有等级序关系，因此关系性形容词不受程度副词修饰（  
\* the extremely atomic bomb ; \* the very baseball game ），  
而描写性形容词可以受程度副词修饰。

关系性形容词没有直接反义词，尽管它们常常可以跟  
non-  
连用。不过这样的形式不是表示一个属性取相反的属性值，而是表示"其他"（everything  
else）的意思。这类形容词有一种分类功能。在不少情况下，关系性形容词以前缀方式引入反面事物。比如：extracellular  
（细胞外的） vs. intracellular  
（细胞内的）。更常见的情况是，关系性形容词在跟特定中心名词搭配时，对事物进行分类，比如civil  
lawyer  vs. criminal lawyer  ； mechanical engineering vs. electrical  
engineering。

在WordNet中。一个关系性形容词的同义词集合包含一个相应的名词指针。例如：{stellar,  
astral, sidereal, noun.object: star}，表示关系性形容词stellar,  
astral, sidereal跟名词star相关。

不少形容词出现的句法位置是比较固定的，对此，WordNet也作了说明。并且这样的说明一般是针对特定形容词，而不是针对形容词的同义词集合的。例如：awake/asleep这两个形容词都只能出现在谓语位置，就以awake(p)这样的标记加以说明。

Bartning (1980)  
观察到，如果形容词修饰的中心名词是由动词派生出来的，那么如果中心词指行动，形容词可以转化到谓语位置，如果指状态，形容词一般不能转化到谓语位置。比如  
economic restructuring 可以转换为 the restructuring was economic  
； 而 economic slump 则不能转换为 \* the slump is economic。

Bartning  
(1980）还进一步观察到，如果形容词跟中心名词之间有明显的语法关系，就不大能变换到谓语位置；如果没有明显的关系，一般可以变换到谓语位置。比如  
presidential election （ president 是 elect 的结果/对象  
），二者之间有明显的关系，所以不能将 presidential  
变换到谓语位置，不说 \* the election is presidential 。而在  
manual labor （手工劳动）这个短语中，它的意思是 labor  
with/by hand ， manual 跟  
labor之间没有显性的关系，因此可以说 this labor is  
manual。

形容词的反义是对词而言的，不是对概念而言的。比如large/small；big/little构成反义关系，而large跟little，big跟small并不构成反义关系，尽管从概念上讲，它们可以构成反义关系。large跟small；big跟little之间的反义关系则可以从语料中的搭配共现和联合关系反映出来。

五  WordNet中的动词

Collins英语词典中有43636个不同的名词和14190个不同的动词，动词的多义性比名词更突出。Collins词典中，名词平均意思为1.74，而动词是2.11。

（一） 动词在WordNet中的组织方式

1 将词库变成按语义组织的域 （breaking up the lexicon into  
semantic domains）

为了将英语动词词库组织成一个关系网络，将词库分成语义域来组织似乎是一个好办法，这有两个理由。第一，它提供了一个对英语词库中数千多义动词的初步的，基于语义的组织方式；第二，对这方面素有研究的研究人员已经指出通过语义和词汇关系联系起来的那些词语通常都属于同一个语义域（semantic  
domain）。

对动词进行分类，第一刀可能是分成event和state两类，而大多数动词可能属于前一类，这一类又进一步分成14个更专门的类：

verbs of motion  
verbs of perception  
verbs of contact  
verbs of communication  
verbs of competition  
verbs of change  
verbs of cognition  
verbs of consumption  
verbs of creation  
verbs of emotion  
verbs of perception  
verbs of possession  
verbs of bodily care and functions  
verbs referring to social behavior and interactions

（詹按：这里原文似有错误，出现了两个perception，有重复。另外，在文后所附动词语义基本分类中有weather/气象类，这里没有提到）

上面这个分类，部分基于Miller 和 Johnson-Laird  
1976年的讨论，部分基于这种分类本身的合理性，因为这样分类似乎可以令得所有动词都有一个适合的类。当然，其他的研究人员可能有不同的分类方式，但是，不管是分类的数量还是类的名称，就WordNet中对动词词库进行分析的观念而言，都不是非常重要的。

上述这15个类是适合WordNet中所有动词的分类需求的。在多年发展之后，WordNet  
1.5版大约包含11500个动词同义词集合。但是，必须强调，动词义类之间的界限是模糊的。比如许多动词无法毫不含糊地归入cognition类，也无法归入communication类。比如  
wonder, speculate, confirm, judge等等。类似地，像  
whistle这样的动词，在"The bullet whistled past  
him"中的whistle，既可归入sound  
emission类（发出声音），也可以归入运动类动词（motion），参见Atkins  
和 Levin  
1991的讨论。如果这样的动词被判定为是单义的，它们就不得不归属一个以上的语义类中。在WordNet中，是倾向于把这些动词处理为多义动词，如果一个动词确实可以归入两个语义类（semantic  
domain）。不过，一个动词归入那个具体的语义类，并不是件很要紧的大事，因为在WordNet中，一个动词的意义是通过它跟其他动词的关系以及同义词集合（synset）来表达的。

2 动词的第一级分类（起始类 unique begineer）：

Lyons(1977)提出一个分类集为：act, move, get, become, be,  
make。  
Pulman(1983)的分类只有be和do，这实际是区分动作动词（activity）和状态动词（stative  
verb）。Jackendoff（1983）的分类是event和state。

首先，这些用来代表基本分类的动词几乎都是多义的。WordNet对do和be都进行了意义辨析，分别都至少得到12个意义（这个数字并不过分，一般的词典也是如此区分的）。当然其中的一些意义不适合作为基本分类使用，比如  
"do my hair"或"do my room in  
blue"中的"do"已经是特定含义的动词了。其次，我们发现，我们设置的用来创建网络的，特别的语义关系，使得把这类抽象动词（如do）跟下一级语义类的动词如communicate和move链接起来不大容易了。例如，尽管communicate和chat之间存在上下位的层级关系，但do和communicate之间的同样的关系似乎就不太合适了。这两个概念似乎比communicate和chat之间的距离要远得多。同样的，move和run之间可以很容易地关联起来，但do和它的直接下位词move似乎就隔得很远。第三，似乎缺乏心理语言学的证据表明人们是把do跟行动动词如move关联起来的，而有证据显示人们是把move跟run联系在一起的（参见Chaffin,  
Fellbaum, 和Jenei 1994）。

采用Lyons(1977)和Pulman(1983)提出的初始分类似乎跟WordNet反映说话者词汇知识的愿望不能统一起来。因此，WordNet设置了更多的动词初始分类单元，达到14个类。

在一个单独的语义框架内，经常面对的情况是并非所有的动词都能组织到一个初始语义类概念之下。一些语义域只能被几个独立的树来表达。例如，运动动词（Motion  
verbs）有两个同形的顶级结点（top  
node）：move1和move2，分别表示位移类运动和无位移类运动。领属动词则可以上升到三个顶级结点：分别由三个同义词集合表达，即{give,  
transfer}， {take, receive} 和 {have,  
hold}。在很大程度上，它们的下位词就体现了使用该语言的社群对"领属"关系转移的编码方式，如"bequeath/遗赠,  
donate/捐赠, inherit/继承, usurp/篡夺, own/拥有, stock/进货  
"等。Communication动词由一级结点communicate打头，下面分成两个独立的树，分别表示语言交流和非语言的交流（gestural/手势语交流）。这两种情形在英语中没有相应的词汇化的体现。语言交流又被次分为口头交流和书面交流两种。其他语义类，例如身体动作类动词，由许多相互独立的语义类组成。这些语义类构成一个相互关联的语义场。如wash/洗,  
comb/梳, shampoo/洗发, make up/化妆, ache/疼痛,  
atrophy/萎缩。社会交互类动词（social  
interaction）也包含许多不同的语义子类，如政治类动词（elect,  
depose/免职），工作（hire/雇佣， subcontract/转包，  
strike/罢工），以及人际关系类动词（court/献殷勤,  
marry/婚嫁），等等。

3 动词的同义词集合（synset）

（1） 同义和近义（synonyms and near-synonyms）：

如果采用可替代性作为判断同义的标准，那么只有很少的动词是同义的。rise和fall可以选择抽象名词如"temperature"（温度）或"prices"（价格）等作为论元，但它们相近的同义词ascend和descend则不行。WordNet原则上避免把这样的动词放到一个同义词集合中。不过，由于许多显然属于同义的动词之间也存在类似的分布差别，因此注释和样例就显得很有帮助了，通过注释和例句，我们可以看出同义动词之间在使用上的差异。

（2）成语和比喻义（Idioms and Metaphors）：

像"kick the bucket"和"keep an eye  
on"这样的固定成语包含在适当的同义词集合（synset）中，那些有比喻意义（非字面含义）的动词也是如此，比如"die"可以用在汽车抛锚或计算机死机的情况下，因而die在WordNet中也被包含在break和break  
down所在的同义词集合中。有比喻义的动词不仅意义上跟相关动词相近（如die跟break和break  
down），而且用法上也有同样的句法特征（如die跟break和break  
down都是非宾格动词），并且一般可以用被动形式加以解释（如This  
task fell to me中的非宾格动词fall就可以解释为be  
assigned；The building went up中的[**Go**](http://lib.csdn.net/base/go)也可解释为be  
erected，等等）

（二） 动词之间存在词汇和语义关系的证据（Evidence  
for lexical and semantic relations among verbs）

/\*  
在WordNet之前，已经有不少研究工作是针对名词之间的语义网络的，但很少有人对动词之间的关系及其组织下工夫进行研究的。不过也有例外，即Evens（1988）的研究工作和Melcuk（1984）的研究。他们把名词间语义关系的研究拓展到动词的语义关系研究上。但是，他们的研究工作涉及的关系范围太广，不仅包括语义关系，还包含形态和句法关系。此外，这些语义网络也不涉及到为说话人的词汇组织方式建模的研究工作。\*/

1 动词语义存贮的心理语言学证据：

尽管有相当多的研究工作研究说话人对名词和形容词的语义记忆，但心理语言学家不太注意研究动词的语义记忆问题。动词的词联想数据是稀疏的。WordNet的设计基于例证关系，而不是适应不同句法类词语之间的直接链接。

除词联想数据外，替换错误也是研究词语和概念的心理组织的数据源。Garrett（1992）报告，说话人在一些动词的替换使用中出现错误情形，如ask和tell，go和come等。显然，这些动词对来自同一个语义范畴。这些替换错误被Garrett分类成反义对：start-stop,  
remember-forget, believe-doubt, ask - tell, preced - follow, fill -  
empty, love - hate, heard - said, taken -  
given。等等。这些动词对跟那些在刺激-反应实验中发现的联想动词对类似。（Palermo  
& Jenkins 1964, Chaffin, Fellbaum, and Jenei  
1994）。在后一种心理实验中，反义是刺激和响应之间的一种主要关联。这类心理语言学证据表明，语义相反关系是说话人组织心理词库的一种重要方式。WordNet即包含了许多这样的动词对。

2 典型性和范畴成员关系判定的证据：

来自典型性和成员隶属程度判断的数据对词语和概念的研究也是一个很好的数据资源。不过这方面的研究同样也主要集中在名词上，例外是Pulman的研究工作（Pulman  
1983）。他的主题是判断一些动词比另一些动词更典型地隶属于某个范畴。例如，murder是kill这个范畴的典型成员，隶属程度很高。survey则是look这个范畴的典型成员，相反，sacrifice和squint，之于kill和look，则都不太典型。这样的结果对心理词库建模是有帮助的。但是，Pulman的数据只局限于8个范畴，每个范畴涉及6个动词。

3 词典释义作为发现语义关系的一种启发式方法：

传统的词典在发现词语和概念之间联系方面提供了丰富的资源。名词通常用上位词来定义。词典也给出了动词的分类层级启示。例如，动词shuffle在韦氏词典（第三版）中被定义为，"to  
move or walk in a sliding, dragging manner without lifting the  
feet"； shout被定义为，"to utter in a loud  
voice."这种类型的定义显示，一些动词是另一些更通用的动词经过一些修饰造成的。另一些动词不大好找上位词，就用它的否定方面来进行定义，例如，lose被定义为，"fail  
to sustain, maintain, or  
keep"。（这类似于词典在定义形容词时用"not  
X"这样的形式）。词典在定义许多动词时还采用"x  
while  
y-ing"这样的形式。这种定义有助于启示继承关系。许多有及物和开始状态的动词在词典中是以"to  
become or make  
X"的形式定义的（例如melt/使融化，被定义为，"to  
become or make  
liquid"）。在WordNet中，这两种意义是加以区分的，并通过CAUSE指针链接。

（三） 动词和同义词集合中的词汇与语义关系（Lexical  
and semantic relations among verbs and synsets）

1 继承/蕴涵（entailment）：

有些动词之间存在蕴涵关系，如snore（打鼾）蕴涵了sleep（睡觉）。"He  
is snoring"蕴涵了"He is  
sleeping"。动词之间的蕴涵关系很像名词中的meronymy（整体部分关系）。

人们接受上述事实，就可以认为这样的动词对是部分-整体关系的，如drive  
是 ride的一部分；snoring 和  
dreaming（做梦）是sleeping的一部分等。因为当一个人drive（驾驶）的时候，他必然是ride（乘）在交通工具上的。从时间性上讲，前一个动词必然在后一个动词发生的时段内。

（1） 下位关系（hyponymy among verbs）：

动词之间不同于名词之间的上下位关系，用来测试名词之间上下位关系的句子框架：An  
X is a Y，不大适合用来测试动词。Ambling is walking，To  
amble is kind of to walk或者Mumbling is talking都不被接受。

Fellbaum和Miller(1990)创造了一个术语troponymy来称说动词中的下位关系。动词之间的troponymy关系可以表示为公式：To  
V1 is to V2 in some particular manner.  
这里Manner可以解释为多种语义维度，比如fight的下位词包括battle,  
war, tourney, joust, duel ,  
feud等等，这些词都是从不同角度更具体地来刻画一个fight。

（2） 下位关系和蕴涵关系（troponymy and entailment）：

下位关系是一种特别的蕴涵关系。march是walk的下位动词，march同时也蕴涵walk。walk被包含在step中（即walk蕴涵step），但walk不是step的下位动词。snore蕴涵sleep，但snore也不是sleep的下位动词。

（3） 动词的分类树关系（verb taxonomies）：

动词层级分类体系通过下位关系实现，倾向于浅层分类。在大多情况下，分类层级不超过4层。例如：

Communicate - talk - [ babble / - mumble / - slur / - murmur / - bark ]  
- write

在动词分类层级体系中，动词下降一层，该层上的动词所能搭配的潜在论元（名词）的多样性也就相应地下降。例如figure（图形）或picture（照片）可以communicate和talk，甚至还可以deceive（欺骗）或lie（撒谎），但它们不能fib（撒小谎）或perjure（作伪证）了，只有人可以fib或perjure。有时候，把一个动词安置到单个的上位动词下非常困难。这跟名词的情形有些类似，例如piano在WordNet中既是一种"stringed  
instument"（弦乐器），同时也是一种"percussion  
instrument"（打击乐器）。就动词的例子而言，sprawl既可以是四肢分开地sit（坐），也可以是四肢分开地lie（躺）。Pulman(1983)提到，许多动词用于表示杀人的不同方式（比如hang,  
stangle, shoot,  
等等），既可以是kill的方式，也可以是execute（行刑）的方式（即killing某人实际上是一种惩罚的形式，这样的惩罚是由社会提出加以制裁的）。人们希望语义上下位关系可以使得语义记忆模式显得经济一些，即下位概念的属性可以从上位概念继承。对动词的上下位语义层级分类，目标也同样如此。communicate（三价动词）的下位动词就从它们的上位词继承论元结构  
-- 即communicator（发信者）, message（信息）,  
recipient（接收者）等。同样地，尽管speak的下位动词之间可能在方式上各不相同，但它们都共享speak的词义联想内容（声带振动，嘴唇和舌头的动作等等）

2  动词中的语义相反关系（Semantic opposition among  
verbs）：

动词词库的关系指针OPPOSITION实际上代表了复杂的若干种语义关系。例如，

converse是一种反义关系，这种反义关系没有共同的上位词或蕴涵动词，比如give/take;  
buy/sell; lend/borrow;  
teach/learn等等。这些词的共同之处在于出现在同样的语义场中，指称同样的行为，但各自联系的表层句子结构中的论元角色不同。

大多数反义动词是状态动词。这样的动词可以通过属性来表达各自之间的反义关系。如live/die;  
exclude/include; differ/equal; wake/sleep等等。

此外反义关系也经常出现在变化动词中，如lengthen/shorten;  
strengthen/weaken; prettify/uglify等等。

跟形容词中的反义关系类似，动词中的反义关系也存在标记问题（markedness）如，tie/untie和appear/disappear，前者是无标记的，后者是有标记的。

许多语义上相反的动词在分类层级上是姐妹关系（co-troponyms），比如rise/fall，walk/run等。

另有一些语义相反的动词对，共同蕴涵一个动词，如fail和succeed都蕴涵try，forget和remember都蕴涵know等。

3  致使语义关系（The cause relation）

这种关系连带两个动词概念，一个是因（如give），另一个是果（如have）。英语中词汇化的致使动词对的例子如：show  
- see（展现 - 看见）和 fell - fall （击倒 - 倒）等。

此外，WordNet包含从致使（causative）动词和及物动词，到相应的非致使动词和不及物动词的CAUSE语义关系指针。这主要出现在变化动词中。例如blacken,  
develop, break,  
shrink等。大多数非致使动词暗示了一个有生施事或无生的致使动因（如The  
glass door broke -- The storm/ The children broke the glass  
door）。少数动词只跟无生致使动因相关，如The wooden  
deck molded -- All that rain molded the wooden deck。比较：The  
house sitter molded the wooden  
deck是不合法的句子。致使动词也系统地出现在运动动词中，如bounce,  
roll, blow等等。（She blew a soap bubble in his face -- The soap  
bubble blew in his  
face）。Carter(1976)提到，致使关系是蕴涵关系中的一种特殊情形：如果V1必引起V2，则V1也就蕴涵V2。比如expel/驱逐和leave/离开，或bequeath/遗赠和own/拥有。跟所有的蕴涵关系一样，致使关系是单向的（unidirectional）。

下图显示了WordNet中全部4种不同的动词蕴涵关系。

（四） 多义性 （Polysemy）：

据Fellbaum(1990)，英语动词中的多义词大约是名词的两倍。那些高频动词（have,  
be, run, make, set, go,  
take等）的意义通常依赖于跟它们共现的名词。

1 多义性与下位关系：自我上下位关系 （Polysemy and  
Troponymy: Autohyponymy）

如behave既可以表示"conduct oneself"，又可表达"conduct  
oneself  
well"，后者（behave2）是前者（behave1）的下位动词。这就是所谓的自我上下位关系。

2 多义性与蕴涵：自我蕴涵关系（Polysemy and Entailment）

韦氏大词典第三版列有drive的两个义项如下：

to operate and steer a vehicle;  
to have oneself carried in a vehicle

请看例句：The president drove on to the Capital.

在这个例句中，drive的上述两个义项的意义区别就不再清晰了，因为总统可能是驾驶汽车的人，也可能只是乘客。再前一种情况下，"他驾驶"就蕴涵了他同时也是被交通工具承载的人。

再如：She sewed her dress 是一个歧义句。

sew有两个意思，一是fasten or join(fabric) by  
stitching（靠缝合使织品更紧凑或连到一起）；一是make  
or create by  
sewing（通过缝制生产衣服）。后一种意思蕴涵前一种意思。因此，如果这个例句是"她做了一件衣服"的意思，则这个例句蕴涵"她通过缝合使这些布片紧凑地组织到一起"的意思。

3 多义词与相反关系：自我反义关系 （Polysemy and  
Opposition: Autoantonymy）

例1. These manufacters now bone their shirt  
collars.（在领子中加领衬/硬领）  
例2. She boned the turkey. （把火鸡的骨头抽去 - debone）  
例3. The violinist strung his instrument. （给乐器加弦）  
例4. Could you string these beans, please?  
（用线把豆子串起来）

Horn(1988)指出，在例1和例3中，动词选择"goal"（目标）论元；而在例2和例4中，动词选择的是"source"（来源）论元。同样形式的动词在不同的用法中构成自我反义关系。

例5. \* We strung the beads and the beans.  
例6. \* She seeded the clouds and the grapes.

例5和例6之所以不合法，是因为在一个句子中，同时包含了上述两种论元（语义上不相容）。

（五） 测试WordNet的语义模式的心理合理性（Testing the  
psychological validity of the wordnet model）

Chaffin，  
Fellbaum，和Jenei（1994）完成了几项试验，目的是测试WordNet中动词关系的心理现实性（psychological  
reality）。

首先，Chaffin，  
Fellbaum和Jenei想看看未经语言学训练的人是否能确定许多动词对之间的不同关系。被试者需要完成涉及四种蕴含关系（troponymy，  
proper inclusion， backward presupposition，  
cause）的三种不同任务。在一个类比任务的测试中，48个被试首先看到1个动词对（verb  
pair）--作为种子对，并被要求从另外6个动词对中选出1对来，而这一对中的成员之间的关系应该跟种子动词对中两个动词之间的关系相同。差不多一半的回答跟WordNet中的分类不一致。

在第二个分类试验（sorting  
experiment）中，另外一组12个被试被要求对四种关系的例子进行整理分类。这个任务允许给被试更多自由来显示他们自己的分类。一个层级聚类的分析显示，被试确定的四种主要的关系类，跟WordNet对四种Entailment的分类是大致相当的。

最后一个试验的对象是有40名被试的第三组人，他们拿到的是显示不同Entailment关系的动词对，要求他们造一个推论句（rationale  
sentence），来详细说明动词对之间的关系。结果显示，被试同意WordNet中关系的分类，尽管很少有被试造出的具体句子（用来描述这些关系的句子）是重合的。在这个实验中，对Troponymy关系有最高的认同度。而且Troponymy在第一个类比试验中也是被试给出答案最准确的，在第二个分类试验中也是一致性最高的。在所有三个实验中，基本都达到了最低限度的一致性。proper  
inclusion关系和presupposition关系似乎在直觉上低于troponymy关系和cause关系。

（六） 动词词典（语义建模）的其他模式 （alternative  
models of the verb lexicon）

1  语义场理论 （semantic field）

从语义场理论的视角来看词典的结构，有些类似WordNet中的语义域。语义场分析方法的基础是这样一个信念，即一个给定语义场中的词的意义来自于它跟该语义场中其他词之间的相似和相反关系。但是，语义场分析跟WordNet中的情况还有所不同，语义场分析通常既考虑聚合关系（paradigmatic），又考虑组合关系（syntagmatic）。在WordNet中，关于一个词的选择优先性的信息不是网络结构的一部分，而更倾向于是跟一个同义词集合（synset）相伴随的说明性材料，这样的信息通常具体地标示什么样类型的名词是作为一个Synset中的动词的典型论元出现的。

2  脚本和框架分析（Schemata and Frame Analysis）

一些语言学家和词典学家认为纯粹的关系分析对描述说话人的动词词库表达是不够的。他们提议采用认知框架或者知识脚本来描述说话人的经验和信念，并提供表示词语意义的概念基础（例如Schank和Abelson,  
1977）。

Fillmore和Atkins（1972）提议编纂一个基于框架的词典。在这样的词典中，词语意义跟认知结构或框架相连。比如商业交易框架（commercial  
transaction frame），其中包括{money, buyer, seller,  
goods}以及动词{buy, sell, charge, cost, etc.}。

尽管WordNet并没有把动词跟特定的名词关联起来，但它提供了每个动词的句法框架，来指示动词的确定数量的论元。但是，在这样的句法框架中的名词槽，目前既没有跟论旨角色（thematic  
role），也没有跟像buyer这样的语义范畴联系起来。

因此，WordNet可以在这方面有所加强，即在跟动词的每一个synset相伴随的框架中，可以将语义和论旨角色跟名词论元关联起来（link）。这样，原来处于反义关系（opposition  
relation）中的动词对，如buy和sell，就可以在名词论元的差别上显现出来。

3  义素分析（compositional analyses）

尽管通过组成成分的意义来建构整体的意义这一处理方式跟从关系角度来刻画词语意义不大相同，但二者之间仍有重合之处。WordNet中通过词语之间关系来刻画词语意义，其中也反映了义素分析的某些方面。比如在义素分析法看来，give这个动词的组成成分之一是CAUSE，而在WordNet中，这是通过give跟have这两个动词之间有CAUSE关系来体现的。

4  词汇从属理论（lexical subordination）

Levin 和  
Rapoport（1988）指出，许多动词在语义上实际由其他动词组成。例如：

例1 brush the tangles out -- remove the tangles by brushing  
例2 brush a hole in one's coat -- create a hole by brushing  
例3 brush the coat clean -- cause to become clean by brushing

再比如：he nodded his assent （他点头表示同意）

在WordNet中，nod是一些基本动词的Troponym，比如nod的一个义项就是"express  
by  
nodding"。它的上位词是gesture（体态语），类似的词包括nod,  
shrug, wink等。

（七） 语义关系和句法规律性（Semantic relations and  
syntactic regularities）

对动词的语义和概念组成成分进行分析能够揭示出许多动词的句法属性。词汇概念结构相同的动词倾向于有共同的句法行为。Levin(1985,  
1993)测试了大量的动词，语义上同类的英语动词，在句法模式上也有系统的对应性。

1  有区别的子树（distinguishing subtrees）

同一层级的动词的句法属性一般就是对这类动词进行次分类的基础。

例如：weave（编织），  
mold（铸模）这两个词都是creation  
verb（创造类动词）。许多创造类动词都可以有下面这样的句式变换（材料/成品  
交换句法位置）：

例1 She wove a rug from the black sheep's wool.  
（她用黑绵羊的毛织了一个毯子）

例2 She wove the black sheep's wool into a rug.

例3 They molded a head from the clay.  
（他们用粘土塑了一个人头）

例4 They molded the clay into a head.

有一些动词，像fabricate（编造）和compose（写），也是creation类动词中的成员，但不能像上面的动词那样变换句式。例如：

例5 The reporter fabricated a story out of the girl's account.  
（那个记者根据那个女孩所说的编出了一个故事）

例6 \* The reporter fabricated the girl's account into a stroy.

例7 She composed a quartet out of the old folk song.  
（她根据这首老民歌创作了一个四重奏）

例8 \* She composed the old folk song into a quartet.

在Fellbaum 和 Kegl  
(1988)对这些动词的讨论中，他们指出，有必要对创造类动词进行更精细的分类，一类是精神上的创造活动（如fabricate和compose），一类是对具体材料的加工（如weave和mold）。这样的区分就能解释这些动词在大多数情况下的系统差异。

WordNet注意到了上述区别，因而将creation类动词区分为"create  
from raw material"和"create  
mentally"两类，尽管这两个概念在英语中并没有词汇化。

2  动词在树结构中的位置的句法意义（syntactic reflexes  
of the verb's position within a tree structure）

从语义关系的角度观察动词同样也提供了理解动词句法性质的一条线索。Fellbaum和Kegl（1989）研究了英语中的一类动词（涉及到及物/不及物的变换）。例子如下：

例9 Mary ate a bag of pretzels.  
（玛丽吃了一袋椒盐脆饼干）

例10 Mary ate.

以往对这些动词的分析根据时态来进行解释。但是，对eat的下位动词的分析显示，可以将这些动词分为两种情况，一种是必须以及物用法出现，一种是总以不及物用法出现。前一类包括gobble，  
guzzle, gulp和  
devour等（基本上都相当于汉语的"狼吞虎咽"），后一类包括dine（进餐）,  
graze（放牧，吃草等）, nosh（吃点心）,  
snack（吃零食）等。Fellbaum和Kegl提议，上述句法性质的差异并不仅仅是及物变换方面的不同，而是由语义激发的差别。他们认为，英语中有两个eat，分别占据语义网上的不同位置。不及物的eat的意思是"eat  
a meal"（吃饭）。它的下位词一部分来自名词，如dine,  
breakfast, picnic, feast等。另外一部分是munch, nosh,  
graze等。相比之下，及物的eat的意思是"ingest in some  
manner"。这种意义下，eat的下位词是指以特定方式eat的那些动词：gobble,  
gulp, devour，等等。

3  中间形式制约 （Restrictions on Middle formation）

从语义分类树观察动词还能清楚地显示基于语义的特定的句法约束。

一个明显的结构是那些通常需要副词或副词短语出现才能站住的句子。比如：

例11 Her new novel sells \* ( fast / like hotcakes )  
    \* 她的新小说卖  
      她的新小说卖得很快  
      她的新小说很好卖

例12 This car drives \* ( easily / like a dream )  
    \* 这辆车开  
      这辆车很容易开

但是，有的情况下对副词出现与否就要求不高。比如：

例13 This vegetable microwaves (easily).

例14 This suitcase zips shut (in a flash). /\*  
这里shut时表示结果的形容词 \*/

对副词出现与否的限制要求可以用动词的语义层级来说明。下位动词对副词出现与否的要求相对宽松。例如：cook有许多下位动词，包括  
fry, broil, braise,  
microwave等等。这些动词都表达了cook的某种特定方式。同样地，close的下位动词包括velcro,  
button, zip,  
snap等。这些动词都表达了以特定的方式来close。我们来看两个上位动词的例子：

例15 This vegetable cooks \* ( quickly ).

例16 This suitcase closes \* ( easily ).

在上面这两个句子中，副词都是必需的。因为其中的动词是"基础"层级（basic  
level）的动词，在人们的心理中，这样的动词本来就被期待是与特定的名词（它们的主语）相关联的。如果这些动词只是简单地来陈述它们的主语，就引不起听者特别的兴趣，因此，特别说明这些动词的具体方式（通过副词来表现）就是必要的，而且是自然的了。注意，这些句子的重音都落在副词上。

相比之下，如果动词是下位动词，副词性修饰成分就不是必需的了。因为这些动词本身就已经能给出新信息。microwave  
意味着 "cook in a microwave oven"。 zip 意味着 "close with  
a zipper "。

Ackerman和Goldberg  
(1996)注意到下面的例子，也可以从动词语义层级的角度来进行解释。他们注意到，一些来自动词过去分词的形式不太适合出现在名词前位置作限制性定语（attributive）

例17 \*  a killed man

例18 \*  a changed design

例19 \* a told secret

相比之下，如果动词在语义层级树上处在更下位的位置，它的过去分词形式就比较容易出现在名词前位置。

例20 a murdered man （一个被谋杀了的人）

例21 an altered design （一个被更改了的设计）

例22 a divulged secret （一个被泄露的秘密）

Ackerman和Goldberg得出结论说，这些不适合作修饰成分的动词，都没有提供足够的信息来限定它后面的中心动词，他们提出如下原则：

一个形容词性的动词过去分词如果来自基础层级的动词，就不适合作修饰语出现在名词前位置修饰限定中心名词，相反，如果动词是更下位的谓词，它的过去分词形式就容易起修饰作用。

附：

---------------------------

动词的基本语义类：

1. 身体动作动词（Verbs of Bodily Functions and Care）;  
275个同义词集合  
2. 变化动词（Verbs of Change）；约750个同义词集合  
3. 通信动词（Verbs of Communication）；  
710个以上的同义词集合  
4. 竞争动词（Competition Verbs）；200个以上的同义词集合  
5. 消费动词（Consumption Verbs）；130个同义词集合  
6. 接触动词（Contact Verbs）；820个同义词集合  
7. 认知心理动词（Cognition Verbs）；  
8. 创造动词（Creation Verbs）；250个同义词集合  
9. 运动动词（Motion Verbs）； 500个同义词集合  
10. 情感心理动词（Emotion or Psych Verbs）；  
11. 状态动词（Stative Verbs）；约200个同义词集合  
12. 感知动词（Perception Verbs）；约200个同义词集合  
13. 领属动词（Verbs of Possession）；约300个同义词集合  
14. 社会交互（Verbs of Social  
Interaction）；约400个同义词集合  
15. 气象动词（Weather Verbs）；约66个同义词集合

不同语义类的动词有非常不同的结构。其中一些可以进入到以上下位关系进行分类的框架中：这主要是"创造类动词、通信类动词、竞争类动词、接触类动词、运动类动词以及消费类动词（verbs  
of creation, communication, competition, contact, motion, and  
consumption）。状态动词和变化动词则有完全不同的结构，它们倾向于以反义和同义关系组织起来。跟形容词类似，它们也可以投射到两极对立的分类框架中。

WordNet词库和搜索软件的设计与实施

Design and Implementation of the WordNet Lexical Database

Richard Beckwith, George A. Miller, Randee Tengi

（Doubtfire草译整理于2001年6月，仅供参考。欲对WordNet有更准确地了解，请细读原文）

词典编纂者必须既关心词典的内容，也关心如何去呈现内容，当呈现内容的媒介从纸变成电脑屏幕的时候，就需要投入更多的关心了。传统的纸上词典经过多年出版发行的实践，已经形成相当标准的模式（Vizetelly,  
1915），至于电子词典，大概还是摸着石头过河的阶段吧。事实上，计算机技术本身也在日新月异地变化之中。电子词典可以凭借计算机技术的支持以多种相对不受限制的方式呈现在读者面前。这些不同方式的利弊高下目前也仍然还是人们辩论的话题，并没有定论。由于这种不确定性，电子词典的编纂者就应该投入更多的精力去关注词典的呈现方式。

WordNet将是这方面工作的一个先例。研究人员的努力使得WordNet的词库实践富于成效。不过这并不代表现有的办法就是最终的标准。我们希望读者不仅仅注意到目前我们所采用的方式的不足之处，同时更应该多去思考改进之道。

人们对WordNet的第一个印象很可能是它是一个在线义类词典（online  
thesaurus）。同义词集合（sets of  
synonyms）确实是WordNet词库的基石，也是WordNet构成一个义类词典的根本所在。当简短的注释（short  
gloss）加入到同义词集合（synonym  
set）中时，WordNet就好像成了一部补充了同义词作为交叉参考的在线词典（Calzolari  
1998）。但实际上，WordNet包含的信息远不止此。作为将英语使用者的英语词汇知识加以模式化的一种尝试，WordNet给出了词语形式（word  
form）与同义词集合（synonym set,  
即synset）之间关系的许多详细信息。如何将这种关系结构呈现给读者所引起的问题，超出了传统的词典编纂所考虑的问题的范围。

在WordNet的开发过程中，将开发工作分成两个相对独立的任务为我们带来了便利。这两个相对独立的任务，跟传统的词典编写和印刷有大致的类似性。一个任务是编写源文件（包含基本的词汇资料）--这些文件的内容是WordNet词库的实体；另一个任务是开发一系列计算机程序，这些程序可以处理源文件，并最终产生出可以在用户面前呈现的词典内容。

WordNet系统自然地分成了四部分：（1）WordNet词典编纂人员的源文件；（2）将这些源文件转成WordNet词汇数据库的软件；（3）WordNet词汇数据库；（4）用于访问这些数据库的一套软件工具；WordNet系统是在一个Sun-4工作站网络环境中开发的。软件工具的编程环境是C语言、Unix工具、外部命令解释脚本等。到目前为止，WordNet已经可以在如下计算机环境下运行：Sun-3,  
DEXstation, NeXT，IBM PC及兼容机，Macintosh。

接下来我们将讨论WordNet的设计和实施的总的特点。"WordNet参考手册"（WordNet  
Reference  
Manual）是一组描述WordNet系统的细节的文件，尤其是对用户界面和文件格式进行了描述。这两方面合在一起，可以提供一个关于WordNet系统的相当全面的观察视角。

1  熟悉度指数（Index of Familiarity）

关于大脑词库的最著名的也是最重要的心理语言学事实之一是，人们对一些词语比另一些词语更熟悉。对一个词语的熟悉度在许多方面会有所表现：阅读速度，理解速度，易于回忆，使用概率，等等。这些方面的影响如此普遍地存在，以至于那些希望研究词语其他性质的实验者，即便付出极大的努力，也很难将不同词语的熟悉度程度视作一样。换言之，词库的初衷是反映心理语言学原则，如果在词库中忽略词语的熟悉度在上述表现上的差异，将是不可想象的。

为将词语熟悉度的差异反映到WordNet中，我们给每个词形式（Word  
form）关联了一个熟悉度的句法标记指数。这个指数并不反映熟悉度差异的所有结果  
-- 一些理论家可能会要求给每个关系附加上权重  
--  
但全部结果的精确信息并不容易得到。目前的熟悉度指数还只是第一步的工作。

使用频率通常被认为是熟悉度的最好体现。那些扮演着重要的句法角色的封闭类词语是使用频率极高的词语，不过，甚至在开放类词语中，使用频率上也存在着较大差异  
--  
使用频率通常被假定为跟熟悉度的差异相关，或者干脆就用前者来解释后者。词频数据在一些技术文献中可以查到，但是，对于WordNet这样规模的词库来说，原有的词频数据还是不够的。Thorndike和Lorge(1994)出版了基于500万词文本语料库的统计结果的词频表，不过他们只报道了3万常用词的结果。此外，他们对词的定义是两个空格间的字符串，因此他们对同形异义字（homograph）的统计是不可靠的，比如他们的结果无法说明lead这个词作为名词和作为动词出现的频率有什么差别。Francis和Kucvera(1982)用他们自己的句法类标记来标明词语的词性，不过他们报告的结果仅仅是从包含1,014,000个单词的文本中得到的结果（含有50400个词形，其中包括许多专有名词）--  
因此这个结果对反映非常用词的频度是不够的。（通常的语速为120词/分钟，因此100万词大约相当于140个小时的话语，或者一个人两周所说的话）

幸运的是，我们还有另外的办法来表示熟悉度。Zipf(1945)的研究表明，词语出现的频率跟多义性是相关的。平均来说，频度越高的词语，在词典中也就有越多的不同意义。心理语言学一项令人感到有趣的发现（Jastrezembski，1981）是，多义性似乎预示了人们访问大脑词库的时间，就好像一个词的频度所能起到的作用那样。

因此，WordNet不用词语的出现频度来指示熟悉度，而用多义性来反映熟悉度。词语义项数可以从一部在线词典中得到。如果那些不在这部词典中出现的词语被指派熟悉度指数值为0，对于词典中收录的词语，则根据词语的义项数来指派熟悉度指数（比如1、2、3、......等等），那么，这样的数值就可以为各个词类中的每一个词指派一个。因此，对于WordNet中的每个词形式，都用一个整数值来记录该词形式（作为名词、动词、形容词、副词使用时）的义项数。WordNet所依据的词典是Collins英语词典。

下面表1显示了熟悉度指数的用处。

Table 1 Hypernyms of bronco and their index  
values（表1：bronco的上位词和义项指数）/\*  
詹按：下表中义项数有些跟用wordnet浏览器查询时的结果是不相符的  
\*/

如果要查bronco（野马）的上位词（superodinate），WordNet可以回应给用户答案：表1中的一系列上位概念(hypernym)。现在，如果这些词都带有熟悉度指数（按照义项数计算）--那么，那些值为0，1的词语就可以被滤去，因为这些词都属于专业术语类的词语，这样，bronco的上位概念就简单地包含：

bronco @-> pony @-> horse @-> animal @-> organism @-> entity

这个缩短了的链更接近人们期望的答案。显然，词语的熟悉度信息是有用的，在人们选择措辞的时候对此会有所体会。用户可以通过在WordNet层级中检查多义性来寻找那些更为人们所熟悉的词语。

如果熟悉度指数可以指派到词的义项上，而不仅仅是指派给词形式上，那么，WordNet对人类语义知识的模拟会显得更真实。比如，名词tie作为{tie,  
necktie}意义使用，远远比作为{tie, tie  
beam}更常用，但目前这两种意义下，tie的熟悉度指数都是一样的13。/\*  
詹按：查wordnet 1.6版浏览器结果

2  词典编纂者的源文件（Lexicographers' Source Files）

WordNet的源文件是由词典学家编写的。这些文件是词汇语义学分析的产物：多种词汇关系和语义关系被用来表示词汇知识的组织方式。词形式（word  
form）和词义(word  
meaning)是这些文件中可见的两个基本构件。词形式以规范的词形表示；词义以同义词集合（synset）表示。词汇关系是两个词形式之间的关系；语义关系是两个词义之间的关系。

WordNet将名词、动词、形容词、副词都组织到synset中，并且进一步根据句法类和其他组织原则分配到不同的源文件中。副词保存在一个文件中，名词和动词根据语义类组织到不同的文件中。形容词分为两个文件（descriptive形容词和relational形容词）。附录A给出了这些源文件的名称。

每个源文件都包含了一个词类的synset的一个列表。每个synset由这些内容组成：同义词形式，关系指针，以及其他一些信息。三种指针表示的关系包括（但不限于）：hypernymy/hyponymy（上下位），entailment（继承），meronymy/holonymy（部分-整体/材料-实体/成员-集体）。多义词是那些出现在一个以上synset中的词，因此代表了不止一个概念。一个词典编纂者常常在一个synset中加入文本注释（textual  
gloss），通常可以提供有关这个synset中同义词及其用法的一些信息。在显示的时候，文本注释包含在数据库中，可以由查询软件显示。评注则在synset之外，由圆括号标示，不包含在数据库中。

描写性形容词组织成一些聚类（cluster），代表了一些属性的一系列的值，从一端到另一端。因此，每个形容词聚类都有两部分（偶尔也有三部分），每部分由一个反义词形式对（antonymous  
pair of word forms）作为中心，叫做中心同义词集合（head  
synset）。大多数中心同义词集合有一个或多个卫星同义词集合相随，每个卫星同义词集合代表了一个概念，跟中心同义词集合所代表的概念比较接近。读者可以通过车轮这个形象来想象这个结构，中间的轴（hub）代表了中心同义词集合，围绕中轴的轮辐就是卫星同义词集合。两个或多个车轮根据逻辑上的反义关系连接起来，反义关系相当于两个车轮间的横轴。

Grinder软件工具将词典编纂者编写的源文件转换成WordNet数据库。在这个过程中，它检查文件的语法是否正确，解析（resolve）关系指针，然后生成WordNet数据库，供检索软件和其他研究工具软件使用。

3  词形式（Word forms）

在WordNet中，一个词形式是由一个单个词或一串由下划线连起来的"词"来表示的。一个"串词"指一个搭配词，表示一个单一的概念，比如名词搭配词：fountain\_pen（自来水笔）

在词典编纂者编写的源文件中，一个词形式可能增添了额外的信息，这对正确的处理和解释源文件数据是必需的。用整数表示的义项号添加到词形式后，用于确定一个词形式的准确意思，如果这个词形式在一个文件中出现一次以上的话。此外，还在圆括号中包含了一个句法标记（syntactic  
marker），补充在这样的形容词形式后，这些形容词的特点是，它们出现的位置相对于它们所修饰的名词而言是固定的。WordNet中的每一个词形式靠这些信息来标示：词形本身，句法类，语义类，义项号。这些信息成为所谓的索引关键字"key"，唯一的来识别数据库中的一个词形式。

4  关系指针（Relational Pointers）

关系指针代表了一个synset中的词形式跟另一个synset中的词形式之间的关系（既包括词汇关系，也包括语义关系）。词汇关系存在于关系性形容词和它们相关的名词的之间，以及副词和导出这些副词的形容词之间。语义关系存在于形容词和跟这些形容词所表示的值对应的属性名词之间。名词属性和表示这些属性的值的形容词之间的语义关系也被编码了。上述关系就是目前WordNet中包括的从一个句法类到另一个类的指针。反义关系也是词汇关系。同义关系是暗含在同义词集合中的。表2归纳了这些关系指针（按照句法类列举）。

Table 2 WordNet Relational Pointers（表2：WordNet的关系指针）

Noun  
Verb  
Adjective  
Adverb

Antonym !  
Antonym !  
Antonym !  
Antonym !

Hyponym ~  
Troponym ~  
Similar &  
Derived from \

Hypernym @  
Hypernym @  
Relational Adj. \

Meronym #  
Entailment \*  
Also See

Holonym %  
Cause >  
Attribute =

Attribute =  
Also See

Meronymy关系可以通过在"meronym指针"后添加字符来进一步明确是哪一种类型的Meronymy关系。在WordNet中，用"p"表示"部分-整体"关系；用"s"表示"材料-实体"关系；用"m"表示"成员-集体"关系。Holonymy关系也以同样的方式进一步加以明确，每种关系都表示跟Meronymy关系相对相反的一种关系（就好像上位关系跟下位关系那样）。

许多指针是相互对称的（reflexive），如果一个synset包含一个指针，指向另一个synset，那么后一个synset也应该包含一个相应的指针指向前一个synset。下面表3中列出了WordNet中的相互对称关系指针。如果满足相互对称关系的一方没有出现，Grinder程序可以自动地产生。

Table 3 Reflexive Pointers（表3：相互对称指针）

Pointer  
Reflect

Antonym  
Antonym

Hyponym  
Hypernym

Hypernym  
Hyponym

Holonym  
Meronym

Meronym  
Holonym

Similar to  
Similar to

Attribute  
Attribute

词典编纂者可以选择两种方式中的一种来给出关系指针。如果一个指针是用来表示两个synset之间的关系--语义关系--该指针可以跟在synset中的词形之后输入。Hypernymy总是连接一个synset和另一个synset的，它是语义关系的例子。词典编纂者也可以将一个词形式跟一些关系指针包含在一个方括号中（[  
...  
]），来定义词形式之间的词汇关系。关系形容词就是以这种方式进入，来表示关系形容词跟相应的名词之间的关系的。

5  动词的句子框架（Verb Sentence Frames）

每个动词synset包含了一个动词框架的清单，来展示跟这个synset中动词用法有关的简单句句型。附录B是动词简单句句型框架的清单。

6 同义词集合的形式句法结构定义（Synset Syntax）

源文件中满足如下形式定义的字符串被当作是synset（不过，下面只是简单的示意，而不是完整的形式描述，正式的规范在"WordNet参考手册"的wninput(5)中可以找到）。

[1] 每个synset以 { 开始；

[2] 每个synset以 } 结束；

[3]  
每个synset中包含一个或多个词形式，之间以逗号分隔；

[4]  
为表示语义关系，词形式后面可以跟上关系指针，具体形式是：一个词形式后面是逗号，然后是一个关系指针符号；

[5]  
对动词synset来说，"框架"（frame）是由逗号隔开的，跟在所有的关系指针之后。

[6]  
为表示词汇关系，一个词形式后面可以跟上[4]或[5]中的基本元素形式，放在方括号  
[ ... ]中。

[7]  
为表示形容词聚类，聚类的每个部分（中心同义词集合，可以选择跟上卫星同义词集合），靠一条仅包含连字符的短线跟其他部分隔开。每个完整的聚类包含在一个方括号中。

7 档案系统（Archive System）

词典编纂者的源文件保存在一个档案管理系统中。该档案管理系统基于Unix修改控制系统（RCS）--  
一个用于文本文件的多重修改管理工作的系统。这个档案系统的建立基于几个理由：（1）允许WordNet数据库的任何一个版本的重构；（2）保存对源文件进行修改的历史记录；（3）阻碍人们对同一个文件进行有冲突的修改；（4）确保产生最新的WordNet数据库。这个档案管理系统中的程序是Unix外部脚本，可以一定方式封装RCS命令，使得词典编纂人员在友好的用户界面下对源文件进行操作。

reserve（预留）命令是从档案中抽取一个或多个给定文件最近修改的部分，并且在用户操作该文件期间对文件上锁保护。review（检查）命令从档案中抽取一个或多个给定文件最近修改的部分，用于检查，此时该文件并没有上锁。为防止修改，review文件时不允许进行写操作。restore（恢复）命令校验一个被预留的文件的完整性，并将该文件返回给档案管理系统。release（释放）命令用于解除一个文件的锁。当词典编纂者决定不应该将对文件的修改返回给档案管理系统时，就可以使用release命令。whose命令用于了解文件是否当前预留的文件，如果是，则显示预留该文件的用户名。

8  将源文件转换成WordNet词库的Grinder工具（Grinder  
Utility）

Grinder是将词典源文件转换成数据库形式的工具。Grinder由几个选项来控制对输入文件的操作。为建立一个完整的WordNet数据库，所有的词典源文件必须同时进行处理。Grinder同时也是一个校验工具，确保词典源文件在用restore命令返回到档案管理系统时的一致性、形式合法性。

9  实施（Implementation）

Grinder是多遍编译器（multi-pass  
compiler），C语言编程。第一遍使用一个用yacc和lex写的分析器（parser），检查输入文件的文法是否正确，并建立被分析后synset的内部表示。接下来的编译过程都针对词典数据的内部表示操作。第一遍编译尝试尽可能地发现形式句法错误和结构错误。源文件中的形式句法错误是跟规范不符的输入，结构错误指的是由于某种原因无法"解析"（resolve）的关系指针。通常这些错误是词典编纂者的打字错误造成的，比如将一个指针指向一个不存在的文件，或者引用一个多义词时没有明确义项号。第一遍编译无法确定多个指针指向多个文件中的结构错误，因为这多个文件不是一起处理的。作为校验工具使用就如同restore命令所做的那样，只运行第一遍编译过程。

第二遍编译过程中，Grinder解析所有的语义关系和词汇关系指针。为此，每个synset中的指针依次被检查，每个指针的目标（synset或synset中的词形式）则被找出来。源文件中的指针被解析为一个指向内部[**数据结构**](http://lib.csdn.net/base/datastructure)的入口，这个入口标明了指针所指目标的"位置"。碰到相互/反身指针（reflexive  
pointer）的情况，就搜索目标指针的synset。如果找到，代表反身指针的数据结构就被修改，来表示目标--源指针的"位置"。如果没有找到这样的反身指针，Grinder自动产生一个带有所有相关信息的指针。

随后的一遍编译遍历词形式表，指派一个多义词的义项号。对每个词形式按照句法类不同来分配义项号。

Grinder的最后一遍编译生成WordNet数据库。

10  内部表示（Internal Representation）

词汇数据的内部表示是一个相关链接表形式的网络。词典源文件被分析过（parsed）后，形成一个关于词形式的哈希表（hash  
table）。小写字母字符串作为关键字；如果词形式不是小写字母，就保留下来作为数据结构的一部分，以便包含在数据库文件中。当分析程序处理输入文件时，它调用函数来创建一个synset中的词形式、指针以及动词句型框架的数据结构。一旦一个完整的synset被分析，就有一个数据结构被创建出来表示该synset。输入文件中所有的synset作为一个单一的链接表来维护。Grinder程序对该数据结构的每"遍"（Pass）访问，既可以通过synset的链接表访问，也可以通过词形式的哈希表访问。一个synset的列表（每个synset明确地表明了每个词形式的意义）用于解析指针，并且生成数据库的索引文件。

11  WordNet数据库（WordNet Database）

对每个句法类，有两个文件来代表WordNet数据库 --  
索引文件（index.pos）和数据文件（data.pos） （这里  
pos是part of speech的缩写，代表 noun, verb, adj 或 adv  
）--  
实际文件名可能由于计算机系统平台不同而有所差别。数据库是ASCII码格式存贮的，人和机器都可读，对于那些希望利用这个数据库支持自己的应用程序的人来说，非常方便访问。每个索引文件是WordNet中一个句法类中所有词形式按照字母顺序排列的表。每个数据文件包含了来自词典编纂者的源文件的所有词汇数据（按句法类分开），同时有经过解析的关系指针来表示词语在数据文件中的地址。

索引和数据文件是相关的。在一个索引文件中，每个记录有一部分是一个或多个字节的偏移量（offset），每个偏移量指明一个synset在数据文件中的起始地址。检索synset或其他信息的第一步通常情况下是在一个或多个索引文件中搜索一个词形式，以获得包含这个词形式的所有synset的数据文件地址。每个地址都是一个按字节计的偏移量（指向数据文件），表示synset信息的起始位置。属于一个单个synset的信息的编码方式可参见下面13小节"数据文件"部分的介绍。

尚书数据库结构的一个缺点是，尽管所有文件都是ASCII码的，因而是可以编辑的，并且理论上也是可以可扩展的，但实际上这些几乎是不可能的。Grinder的主要功能之一是计算synset在数据文件中的地址。编辑任何一个数据库文件都可能造成错误的偏移字节量，从而造成搜索错误。目前，建构一个WordNet数据库需要使用Grinder同时处理所有的词典源文件。

下面将简要描述索引文件和数据文件，非常概要地介绍WordNet数据库的形式结构、规范、和组织方式。更详细的信息请参考手册"WordNet  
Reference Manual"中的wndb(5)页。

12  索引文件（Index Files）

索引文件中的词形式是小写的，不管它们原先在词典源文件中取什么形式。索引文件按照ASCII码字符集排序，可以利用二分法快速搜索。

每个索引文件都以几行包含版权申明、版本号、使用许可协议等信息的文字开始，之后是数据行。每行数据包含如下信息：来自在线词典（Collins英语词典）的一个词的义项数信息；包含这个词的所有synset中使用的关系指针类型的列表（这个信息是用于检索的）；一个偏移字节量的索引列表，每个偏移字节量指向相应的数据文件中的一个synset中的词形式。每行数据以行尾字符作为结束标志。

13  数据文件（Data Files）

数据文件中包含的信息对应着词典源文件中定义的synset，以及解析为data.pos文件中的字节偏移量的指针。

每个数据文件都以几行包含版权申明、版本号、使用许可协议等信息的文字开始，之后是一个所有输入文件名的列表。之后是数据行。每行数据包含了synset的信息，以及由Grinder程序附加的信息（可以用于检索或其他软件）。每行数据也是以行尾字符作为结束标志。在数据文件中，一个synset里的词形式对应着词典源文件中的词语规范写法。

每行的第一部分信息是偏移字节量（即synset的地址）。这个信息多少有些是冗余的，因为几乎所有的从一个数据文件中读取synset的计算机程序都知道synset的偏移量，但这个信息仍然是有用的，比如使用像grep这样的unix工具来跟踪synset和指针时，因为有了这个信息，就不需要使用复杂的软件了。此外，这个信息还提供了一个synset的唯一"关键字"（key），可以供用户程序使用。在这部分信息之后是一个整数值，对应着synset所在文件在一个文件名列表中的位置。这个信息可以用于检索软件在显示synset时标记该synset所在的源文件名，同时也有助于区别义项。在这个信息之后是一个词表，关系指针，以及动词句型框架。数据行的最后一段是文本注释部分。这部分是可选的（即可有可无的）。

关系指针由几段信息来表示。指针符号打头，之后是目标synset的地址，以及它的句法类（这个信息对那些指向不同句法类的指针是必需的），之后的一个片断用于区分当前指针是词汇关系指针还是语义关系指针。如果是词汇关系指针，这个片断就指示哪些词形式是该指针在源synset和目标synset中涉及到的词形式。如果是语义关系指针，这个片断的值就是0。

14  查询词汇信息（Retrieving Lexical Information）

为了让用户可以访问数据库中的信息，需要为用户提供一个界面。这个界面可以使最终用户查找WordNet数据库，并通过一个基于Windows的工具或者以命令行方式来显示查询结果。这一点使得WordNet这样一个在线词典跟传统的印刷出版物形式的辞典非常不同。即便是一般传统词典的在线版本，WordNet仍然跟它有所不同，传统词典的在线版本（电子化）的信息是以固定格式存贮的，根据需要来加以显示。WordNet的信息以面向计算机的方式存贮，对于一般读者而言，反而是不可读的。因此，用户界面为用户提供了多种途径检索和显示词典信息。不同的界面可以提供给不同需求的用户。但所有界面都基于同一个词汇数据库。

WordNet的用户界面可以有很多形式。标准界面是一个X  
Windows应用程序。该程序可以在几种不同的计算机[**操作系统**](http://lib.csdn.net/base/operatingsystem)平台上运行。支持微软Windows操作系统和苹果Macintosh操作系统的版本也已经开发出来。尽管如此，以命令行方式访问WordNet数据库仍然是需要的。一些用户没有窗口环境，此外，外部脚本程序以及其他一些程序也可能围绕命令行界面来编写。

搜索过程也是跟搜索请求的具体类型无关的。第一步是检索索引文件中的索引入口。索引入口包含了地址信息，即被检索词语所在的synset在数据文件中的地址。然后，每一个synset就作为结果回应检索请求。因为每个包含了被检索词的synset中还包含指向数据文件中其他synset的指针，这些synset也可能是检索结果，需要被显示出来。

WordNet的用户界面和其它软件工具依赖一个函数库。这样一个相当全面的函数库已经提供了出来。可以支持搜索和检索，词形处理，以及其他功能。附录C包含了对这些函数的简要介绍。

findtheinfo( )函数的输入变元是一个词形式（word  
form），词类，和搜索类型；findtheinfo(  
)调用一个低级函数在索引文件中寻找相应的入口，对每一个义项，调用适合的函数跟踪跟搜索类型相应的指针。大多数都是由traceptrs(  
)函数完成的，但特定的函数不一定适合标准的层级搜索。

上述通用搜索和检索[**算法**](http://lib.csdn.net/base/datastructure)在实现WordNet的用户界面时是以几种不同方式起作用的。搜索类型根据词类不同而有所区别，但对应着上面表2中列出的关系指针。层级搜索可能由所有的关系指针来完成，除了反义关系和"参见"（also  
see）关系。此外，调用findtheinfo(  
)函数可以搜索多义项信息、动词句型框架信息、或名词同位信息（那些有共同上位的词语也作为搜索字串被搜索到）。

搜索功能无法完成词形处理（morphological  
operations），因此调用findtheinfo(  
)函数的同时需要调用morphstr(  
)函数来将搜索字串（单词）翻译成一个或多个基础形式（base  
form）。

15  X Windows界面（X Windows Interface）

16  数据库搜索（Searching the Database）

17  选项（Options）

18  结果输出（Output）

19  词语形态处理（Morphy）

20  例外词表（Exception Lists）

WordNet中每个词类（副词除外）都包含一个例外词表。这个词表包含词语的不规则词形变化，因此无法以通用的算法方式来处理。这个例外词表的每一行都包含一个含有屈折形式的单词，之后是这个单词的一个或多个基础形式。这个词表以字母顺序排序，采用二分法搜索。

21  单字词（Single Words）

22  搭配词（Collocations）

23  含连字符的词（Hyphenation）

24  进一步的工作（Future Work）

因为许多名词搭配词包含了介词，例如"line of  
products"（"产品路线、产品线"）这个词，因此有必要开发一个跟处理动词类似的算法来处理名词。在目前的框架下，如果Morphy分析了"lines  
of products"，这个搜索串就成为"line of  
product"，而"line of  
product"在WordNet词库中不存在。Morphy应该在这方面有所改进。当它分析的是基础形式时，应该能够保留该基础形式中的"屈折成分"。