文章编号：1003-0077（2011）00-0000-00

句法网与语义网的对比研究*

马丹[1]　赵怿怡[2]

（1.厦门大学 人文学院　福建 厦门 361005;2.厦门大学人文学院　福建 厦门 361005）

摘要：基于网络观的语言研究已经成为语言分析的趋势之一。但不同语言单位层级、不同语言单位关系的选取导致了语言网络主要参数的差异性特征。从词的同现网到句法网再到语义网所需要的语言学知识也逐步深化，本文旨在构建语义学理论支撑的语义网络，并把虚词纳入语义分析过程，分别以句法关系和语义关系作为联结，用 Cytoscape 构建了句法网和语义网。结果发现：语义网的直径，平均最短距离比句法网大、层级性比句法网差、聚集系数比句法网小，虚词节点“的”、“和”“个”等有可能是局部的中心节点。

关键词：句法网；语义网；虚词

中图分类号：TP391　　**文献标识码：**A

The comparative study of syntactic web and Semantic Web

MA Dan[1]　ZHAO Yiyi[2]

(1.College of Humanities,Xiamen University,Xiamen,Fujian 361005,China;

2. College of Humanities,Xiamen University,Xiamen,Fujian 361005,China)

Abstract: Language research based on the view of network has become one of the trends in language analysis. However, the selection of relations between different language units leads to differences in the main parameters of the language network from the co-occurrence network of words to the syntactic network to the semantic web site. The required linguistic knowledge is also gradually deepened. This paper aims to build a semantic network supported by the semantic theory, and incorporates the virtual words into the semantic analysis process. The syntax and semantic relations are used as a link, and the syntactic and semantic web results are constructed using Cytoscape. Findings: The diameter of the Semantic Web, the average shortest distance is larger than the syntactic net, the hierarchy is worse than the syntactic net, the clustering coefficient is smaller than the syntactic net, and the virtual word nodes "y", "and", "all", etc. may be local central nodes. .

Key Words: Syntactic web; Semantic Web; Function words

1 引言

语义分析指运用各种方法，学习与理解一段文本所表示的语义内容，任何对语言的理解都可以归为语义分析的范畴。语义分析又可进一步分解为词汇级语义分析、句子级语义分析以及篇章级语义分析。[1]本文关注句子级语义分析，句子级语义分析目标是分析整个句子所表达的语义内容和语义关系。研究发现，以往的语义分析主要关注句子中的实词之间的关系，“针对动态语料的语义依存分析若只考虑论元关系[2]，并不能充分实现句法分析到语义分析的转化，导致句法网络与语义网络产生不可解释的参数差异。”[3] 因此，本文把虚词纳入语义处理框架，从而实现从句法依存到语义依存的完全转换，进一步推动句法网络与语义网络的对比研究。本文基于同一文本以词为单位构建了句法网和语义网，试图从网络的整体数据和网络的局部节点讨论两者之间的差异。

---

* **收稿日期：**　　**定稿日期：**

基金项目：国家社会科学基金青年项目——基于同一文本的句法网络语义网络关系研究（NO.14CYY046）；厦门大学哲社科繁荣计划、两岸关系和平发展中心资助。

2 语言网络构建

句法网络是基于句法理论构建的语言网络。句法网络的构建是句法分析结果的直观反映。句法分析又可以分为短语结构句法分析和依存句法分析[4]。本文的句法分析依赖于依存关系[5]理论。刘海涛[6]认为依存句法分析比短语结构句法分析更容易发现句子中两词之间的关系。依存句法分析的前提是建立能够输入的依存句法树库。本文所用的树库是小型百科语料树库，我们采用表格的形式展示（参见表 1）。

表 1　依存树库示例

从属词序号	从属词	从属词词性	支配词序号	支配词	支配词词性	依存类型
1	人体	n	2	是	v	subj
2	是	v	0	ROOT	s	s
3	由	v	12	构成	v	adva
4	数以亿计	a	5	的	usde	dec
5	的	usde	11	细胞	n	atr
6	微小	a	10	的	usde	dec
7	而	cc	8	有	v	co
8	有	v	6	微小	a	c-dec
9	生命	n	8	有	v	obj
10	的	usde	11	细胞	n	atr
11	细胞	n	3	由	v	obj
12	构成	v	2	是	v	obj
13	的	um	2	是	v	esa

表中每一行代表的都是一个依存关系，本文采用的文本，共有 1000 个词，475 个依存关系。我们以词为节点，依存关系为边构建语言网络。

语义网络是介于句法和概念网络的中间层[7]，对于语义网络的特征研究有助于句法和语义之间的转换研究。语义网络的研究也是以依存理论为基础，本文所探讨的语义依存理论基于句法依存分析，是句法依存分析朝深度语义理解的进一步发展，并为构建语义网络提供理论支撑。不同的是，语义分析一般只对实词进行分析，不包含虚词。语义依存分析常常建立在句法依存分析的基础上，从句法分析到语义分析，虚词是否应该被保留呢？我们知道虚词在句子中起着经络的作用[8]。陈芯莹、刘海涛等[9]曾以新闻联播和实话实说为语料资源，探究虚词“的”“了”“在 p”在句法网络中的特征，发现“的”是网络的中心节点；“了”和“在 p”是局部网络节点去掉三个节点后，网络的平均度、网络密度、最大范围均有所降低，平均路径及直径增加。虚词在句法网络中作为中心节点或者局部中心节点存在，那么在语义网络中的作用呢？赵怿怡[3]在进行语义分析的时候，考虑到副词“不”会影响语义的表达，进行了保留。可见虚词会对语义分析产生影响，把虚词纳入语义分析的范畴，分析虚词在语义网络中的地位。

语义分析离不开讨论词的分类问题，这里我们主要讨论动词的分类问题。在句法分析中，动词按照形式分为助动词、系动词、趋向动词、不及物动词、小句宾语、双宾动词等。为了

满足语义分析的需要，我们参考陈昌来[10]对动词的分类及《汉语动词概述》[11]对动词进行了语义分类，分别为动作动词、存在动词、使令动词、趋向动词、心理动词、能愿动词、关系动词、先导动词。虚词等的标记参照刘海涛[6]的句法标注体系。原因是与句法分析中的词类标记保持一致，可以更直观地分析虚词在句法网络和在语义网络中的地位。语义标注的方法很多，这里不多赘述，本文主要参考陈昌来对语义角色的标注[12-16]、HowNet[17-18]的动态角色及哈工大 LTP[19]的语义角色标注。

依存关系转换成语言网络的方法，我们本文采用的是软件 Cytoscape[1],它是一个专注于开源网络可视化分析的平台，核心是提供基础的功能分布和网络查询，并依靠基本的数据形成可视化网络。它最先应用于生物学领域，显示分子之间的相互作用。[20]这里我们应用于依存网络的构建，展现各个语言单位之间的关系。Cytoscape 是以两个节点（source node,taget node）以及一个关系(interaction)为基础进行的网络构建[21]，这里支配词我们作为源节点（source node），被支配词作为（target node）来处理，关系就是两个词之间的支配关系。用表 1 中的支配词、被支配词、支配关系，可以转换成这样的语义网络（图 1 b）。本文句法关系的确定主要采用刘海涛[6]关于汉语依存关系的描述，利用 Cytoscape 同样可以构建如下网络：

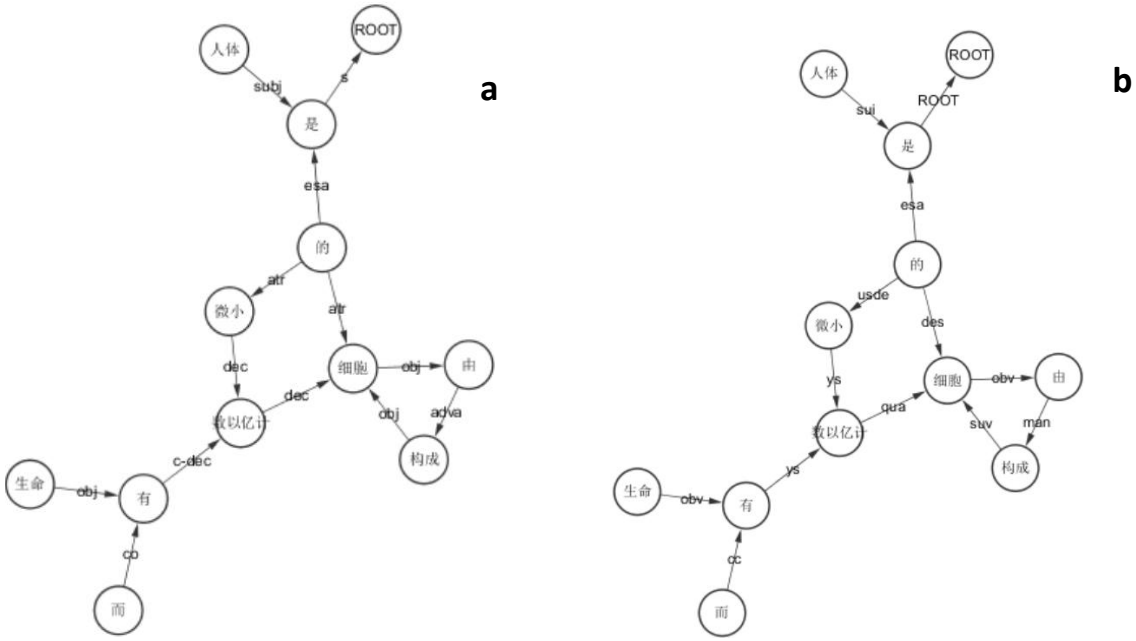


图 1 句法网络和语义网络

句法网（图 a）语义网（图 b）

图1是由12个节点构成的语义网和句法网，箭头表示各个节点之间的支配关系，箭头上标注的是各个节点之间的依存关系。对比图a和图b，虽然节点完全一致，但是由于构造方式的不同，结构存在较大差异，那么网络的参数是否也有较大的差异呢？

3 分析结果对比

我们以依存关系为边，词为节点，用 Cytoscape 软件构建了句法网和语义网。根据 Cytoscape 的数据对网络进行整体分析，这里为了更清晰地看到节点之间的依存关系，我们采用的是有向网络的分析方法。语言网络的对比主要从聚集系数[7]、最短路径[22]、平均相邻节点数、网络的层级性[23]、等方面进行考察。

[1] http://www.cytoscape.org

表 2 网络整体数据对比分析

语言网络	聚集系数	直径	最短路径	特征路径长度	平均相邻节点数	多边节点对
句法网	0.025	16	79383(35%)	5.310	3.651	66
语义网	0.018	23	62486(27%)	7.895	3.69	63

聚集系数 C（Clustering coefficient）是一种用来衡量网络聚类倾向或小集群形态的指标，设网络节点 i 有 k 条边和其他节点相连，那么该节点与这 Ki 个节点构成了一个子网络（集群）。而 K 条边连接的节点（k 个）之间最多可能存在的边的条数为 k(k-1)/2。如果将 Ei 看作是 ki 个节点之间实际存在的边数，那么 Ei 和 Ki 最多可有的边数之比就是节点 i 的聚集系数 Ci：

$$Ci=\frac{2Ei}{ki(ki-1)} \qquad (1)$$

那么整个网络的聚集系数 C 就是所有节点聚集系数 Ci 的平均值即：

$$C=\frac{1}{N}\sum_{i=1}^{N} Ci \qquad (2)$$

聚集系数是衡量网络集团化程度的标准，聚集系数越高说明各个节点之间的联系越紧密。由表 2，我们可以知道句法网的聚集系数比语义网高，直径比语义网的直径小。

最短路径 d 指的是网络中任意两点的最短路径，这里 Cytoscape 给出的是任意两个节点之间的最短路径数和最短路径在总路径中的百分比。句法网的最短路径数占 35%，语义网的最短路径数占 27%。

特征路径长度（平均路径长度）cpl 指任意两个节点的距离的平均值。设两个任意节点分别是 i,j，这两个任意节点之间的距离为 dij,网络的节点数为 N，则：

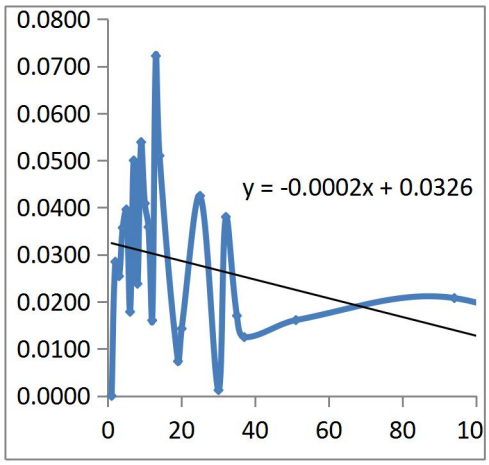
$$cpl=\frac{1}{\frac{1}{2}N(N+1)}\sum_{i\gg j} dij \qquad (3)$$

特征路径的长度与节点之间的距离有关系，无向网络节点之间的距离就是两点之间最短路径所包含的连线数，有向网络节点之间的距离是一个节点指向另外一个节点之间的距离，并且在相反方向上距离不同。若是网络看作有向网络，那我们发现语义网的特征路径长远大于句法网。

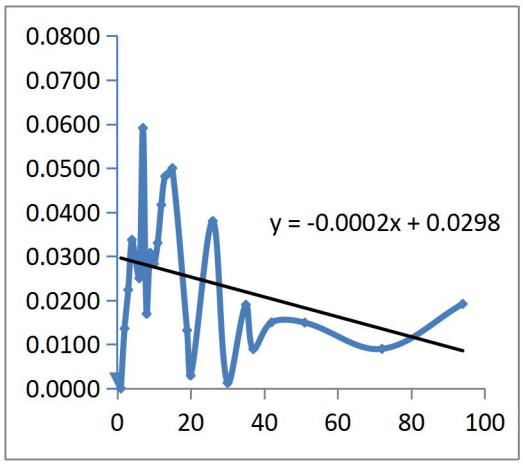
“网络的层级结构，可以用网络的聚集系数（Clustering coefficient）和节点度的相关性来表示，这种相关性 C（k）表示的是度为 k 的所有节点的平均聚集系数”[7]，计算公式如下：

$$\bar{C}(k)=\frac{1}{N_k}\sum_i C_i\ \delta_{k_i,k_j} \qquad (4)$$

其中N_k为节点度为 k 的所有节点总数，δ_{k_i,k_j}为克罗内克符号（Kronecker），当$k_i=k_j$的时候，即任意两个节点 i，j 的节点度相同，那么克罗内克符号的值就是 1，当$k_i\neq k_j$时，即两个节点的节点度不同的时候，那么符号的值就是 0（不执行求和）。在许多真实的网络中，如果节点度 k 变大，节点聚集系数 C（k）按照幂率衰减，那就说明网络的层级性比较明显，即低节点度节点，其邻节点互联的概率较大，而高节点度的节点，其邻节点互联概率较小。下面我们通过数据的计算对比句法网和语义网层级性明显程度的差异。



c



d

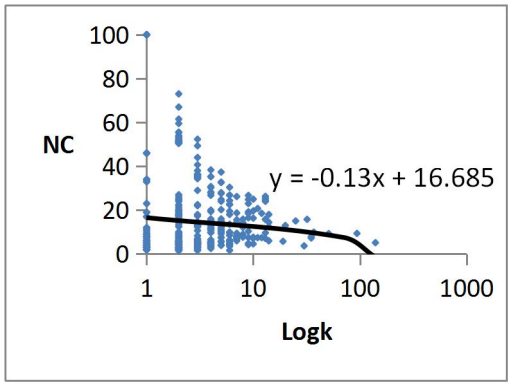
图 2 句法网（左）与语义网（右）节点度与聚集系数的相关性

其中横轴表示节点的度（k），句法网中节点最高的度为 94，语义网中节点最高的度也是 94，我们把最大节点度设为 100。纵轴表示平均聚集系数，计算方法为节点度相同的节点聚集系数的和/这些节点的个数。图中的拟合线表示图表的整体趋势，即是节点度 k 与节点度为 k 的平均聚集系数的相关性。

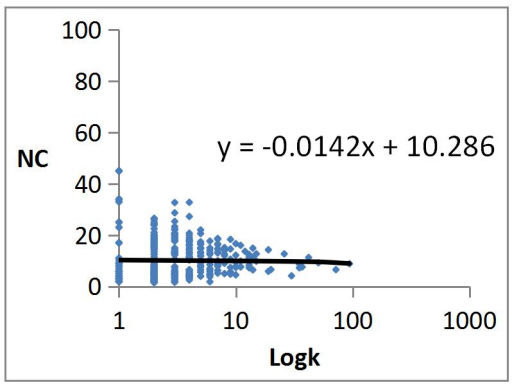
句法网(图 c)和语义网（图 d）点度与聚集系数之间的相关性都不是特别明显，但是两者的相关性一致。刘海涛[7]曾在统计语义网中节点度与聚集系数的相关性时发现，节点度为 1 的节点可能是导致网络层级性差的原因。虚词进入语义网络以后，语义网和句法网的层级性保持一致，可能是因为虚词的存在增强了语义网络的层级性。

节点的相关性表示一个节点的度与其相邻的节点度之间的相关性，我们可以用平均相邻节点度（KNN）来衡量网络节点之间的相关性。一般来说，如果在一个网络中，节点度数大（小）的节点常常与节点度数大（小）的节点连接，那么我们认为这个网络是正相关（assortativity）。相反，如果节点度大（小）的节点常常与节点度数小（大）的节点连接，那么这个网络就是负相关（disassortativity）[20]。

我们可以选择一个节点度为 k 的节点，然后统计这个节点与其相邻节点之间的相关性，如果随着 k 的变大，相邻的节点度也变大，则表明这个网络是正相关的；如果随着 k 的变大，相邻的节点度变小，则表明这个网络是负相关的；如果拟合线的斜率倾向于 0，则表示网络的节点间缺乏相关性。为了统计的方便，我们以节点度 k 为横轴，邻居节点的联通度（NC）为纵轴。可以用 Cytoscape 中对节点的邻居节点的平均度（Neigborhood connectivity）进行统计。



e



f

图 3 句法网和语义网节点度相关性

句法网（图 e）语义网（图 f）

句法网中拟合线的斜率为-0.13，这说明句法网中的节点与相邻节点的联通度成负相关的关系，语义网拟合线的斜率接近于 0，节点之间的相关性并不明显。在句法网中，实词和

虚词之间的关系紧密，节点与节点之间反映的的是实词与虚词之间的关系。语义网中，虚词只能充当被支配词，节点与邻居节点的相关性不强。我们有理由推断虚词是造成网络相关性差异的原因。

通过对网络整体性参数的对比分析，我们发现语义网和句法网在聚集系数、最短路径长、节点度的相关性都存在着差异，这说明网络构造方法的不同会对参数产生影响。

节点度是一个节点所拥有的连线（依存关系）的数量，如果把网络看作是一个简单无向图，那么节点的点度就是与其相邻的节点的数量。一个节点的点度就是对其领域的规模的大小的一种测量。高点度的节点往往位于网络的中心或者局部网络的中心。陈芯莹、刘海涛等[8]发现虚词“的”“了”“过”等虚词可能是句法网络的中心节点，那么这些虚词在语义网中是否也可能是中心节点呢？

表 3 句法网和语义网中高点度节点参数分析

	句法网				语义网			
节点	聚集系数	中介中心度	接近中心度	平均路径长	聚集系数	中介中心度	接近中心度	平均路径长
是	0.021	0.046	0.169	5.914	0.019	0.109	0.136	7.363
的	0.006	0.291	0.426	2.348	0.009	0	0.424	2.356
和	0.001	0	0.334	2.988	0.001	0	0.124	3.056
个	0.022	0.019	0.301	3.316	0.026	0.275	0.137	3.631

节点的聚集系数（云集系数）表示在该节点的邻点中，直接相连的邻点对，占所有邻点对的比例。它是衡量该节点与相邻节点之间的连通程度，反应节点之间关系的紧密度的参数。中介中心度（Betweenness Centrality）指在网络中所有节点之间的测地线[2]中，经过该节点的测地线所占的比例。一个节点在网络中起到多大的“中间”的作用就代表着这个节点在网络中占着多中心的位置。一个点度不高的节点也可能因为起着中介作用而成为网络中心或者局部中心。中介中心度和接近中心度都是用来测算节点在网络中的整体中心度。接近中心度（Closeness Centrality）指的是其他节点数除以该节点与其他节点的距离之和。总距离越大，接近度的值就越小。

在表 3 中，句法网中 “的”的聚集系数、中介中心度均比语义网高，说明“的”的中心地位在句法网中更为明显。但值得一提的是 “的”在语义网中虽然不起“中间”作用，但是接近中心度与句法网基本持平，平均路径也很短，我们有理由认为“的”在语义网中的中心地位比较突出。

“和”作为连词出现在两个网络中，节点度、接近中心度较高，平均路径长比较短，这说明“和”很有可能作为局部中心节点存在。“和”在两个网络中的参数基本保持一致性，这说明“和”在网络中的地位并不受网络构建方式的影响。

量词“个”在语义网和在句法网中，地位大体相同。在句法网中中介中心度比较高，这说明在句法网络中，“个”的“中间”作用更为突出。从平均路径上看，两个网络中“个”的平均路径都很短，这说明“个”可能是处在网络中心附近的节点。

“是”在两个网络中都具有很高的点度、入度和出度，聚集系数也较高，这说明“是”在两个网络中的地位重要，并且与邻居节点的连通性很好。但是两个网络中“是”的接近中

[2]无向网络中，两个节点之间的距离，就是两点之间最短路径所含的连线数。平均最短路径又称测地线。

心度都很小，平均路径也很长，这说明“是”不可能处于网络的中心节点，可能作为局部中心节点存在。

为了验证节点在网络中的地位，我们统计了观察剔除节点以后网络特征的变化。这里主要从平均度（average degree）、网络的中心度（network centralization）、特征路径长（characteristic path length）、孤立节点数（isolated nodes）几个方面讨论。

平均度指的是每个节点平均具有的节点度数。计算方法是各个节点的度数之和与节点数之比。

网络中心度指整个网络的中心化程度，中心度在各个节点之间的变异越大，网络就越中心化，也就是说节点中心度的变异越大，网络的中心化程度就越高。

特征路径长又叫平均路径长，指任意两点之间的平均最短路径，计算方法见公式（3）。

孤立节点指的是节点度为 0 的节点。这里是去节点之后产生的孤立节点。

表 4 去节点之后网络参数分析

句法网 / 语义网		边数	平均度	中心度	特征路径长	孤立节点
节点	完整	1000	4.2105	0.204	3.929	0
的	去“的”	860	3.9241	0.089	4.422	8
是	去“是”	912	4.0211	0.203	4.076	5
和	去“和”	970	4.1561	0.205	4.042	0
个	去“个”	965	4.1455	0.205	3.948	2
的	去“的”	928	4.0111	0.086	4.344	0
是	去“是”	912	4.0211	0.097	4.291	5
和	去“和”	970	4.1561	0.099	4.266	0
个	去“个”	965	4.1456	0.099	4.156	2

表 4 显示，去掉“的”之后的，句法网和语义网的平均度、中心度明显下降，特征路径长变长，产生了 8 个孤立节点。去“的”之后，语义网络的中心度变小了，但是变化程度远远低于句法网去“的”之后。原因是“的”在句法网中接近中心度更高，去掉之后，各个节点之间的差异性会变小，但是在语义网中，“的”的接近中心度不高，对各个节点之间的差异影响不大。“的”去掉之后，语义网的特征路径长变大，中心度降低，这说明“的”在语义网中虽然不占据中心位置，但仍然与其他节点保持着联系。去掉虚词“的”导致语义网的参数发生了变化，这说明“的”在语义网中的重要作用。

去掉“是”之后，两个网络的平均度，中心度和密度均降低了，平均路径都增加了，网络直径都保持不变，孤立节点数都是 5。节点“是”在句法网中的中心度降低了 0.4%，语义网中降低了 1%，这说明去掉“是”以后，语义网中节点之间的差异在语义网中变得更小，“是”在语义关系连接中具有更强的中心节点功能。这很可能说明节点“是”在语义网中比在句法网中更占据中心的位置，当然这需要更大的数据库来验证这一结论。

剔除节点“和”之后，两个网络的中心度都变大，句法网中增大幅度为 0.4%，语义网增大了 1%，网络的中心度变大，说明“和”在两个网络中都不处于中心节点的位置，但是节点“和”在语义网中的重要性要弱于在句法网。

去“个”之后，两个网络的平均度下降，中心度和特征路径长均变大，产生了两个孤立节点。网络的中心度是网络中各个节点之间的差异程度，差异越大，中心度越高。去掉“个”之后中心度变大，说明网络节点之间的差异变大，网络的集中度变高，也就是“个”在两个网络中的存在影响了网络的集中度。

4 总结

把虚词纳入语义分析的范畴，用同一文本构建语言网络，是从句法依存分析到语义理解的进一步发展。虚词只具有功能性意义，但是却会对语义分析产生影响。通过对语义网和句法网的参数分析发现，虚词“的”“个”“和”在语义网中同样具有着重要地位。本文研究的意义在于讨论虚词在语义网中的地位，初步研究句法到语义完全转换。本研究还会在此基础上进一步扩大语料，探究更多虚词在语言网络中的作用。本文在建立包含虚词的语义处理框架之后，对网络进行了对比分析，以求进一步推动从句法到语义之间的完全转换研究。

参考文献

[1]人工智能中的语义分析技术及其应用[J]. 软件和集成电路, 2017(04):42-47.

[2]袁毓林. 一套汉语动词论元角色的语法指标[J]. 世界汉语教学, 2003, (03):24-36+2.

[3]赵怿怡, 刘海涛. 语言同现网、句法网、语义网的构建与比较[J]. 中文信息学报, 2014, 28(05):24-31+65.

[4]刘海涛. 依存语法的理论与实践[M]. 北京：科学出版社，2009.

[5]Tesnire,L.Elments de la syntaxe structurale[M].Paris :Klincksieck,1959

[6]刘海涛, 赵怿怡. 基于树库的汉语依存句法分析[J]. 模式识别与人工智能, 2009, 22(01):17-21.

[7]HaiTao Liu　Statistical properties of Chinese semantic networks[J]. Chinese Science Bulletin 2009,54(16): 2781-2785

[8]陆俭明. 现代汉语语法研究教程（第三版）[M]. 北京：北京大学出版社. 2005

[9]陈芯莹，刘海涛. 汉语句法网络的中心节点研究[J]. 科学通报，2011，56（10）：735-740.

[10]陈昌来. 现代汉语动词的句法语义属性研究[M]. 上海：学林出版社，

[11]范晓. 汉语动词概述[M]. 上海：上海教育出版社. 1987 , 2002. 6:77-90

[12]刘海涛. 汉语句法网络的复杂性研究[J]. 复杂系统与复杂性科学, 2007(04):38-44.

[13]陈昌来. 论现代汉语句子的语义结构[J]. 烟台师范学院学报(哲学社会科学版), 2000(01):67-72+77.

[14]陈昌来. 论现代汉语句子语义结构中的施事[J]. 吉安师专学报, 1999(02):69-77.

[15]陈昌来. 论汉语句子语义结构中的受事[J]. 吉安师专学报, 2000(01):30-37+47.

[16]陈昌来. 论语义结构中的与事[J]. 语文研究, 1998(02):23-28.

[17]董振东　董强　知网简介 http://www.keenage.com/zhiwang/c_zhiwang.html

[18]董振东, 董强. 知网和汉语研究[J]. 当代语言学, 2001(01):33-44+77.

[19]刘挺, 车万翔, 李正华. 语言技术平台[J]. 中文信息学报, 2011, 25(06):53-62.

[20]Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821–7826

[21]Michael Smoot. Network visualization and data analysis using Cytoscape [A].中国化学会计算机化学专业委员会. 第十届全国计算(机)化学学术会议论文摘要集

[22]（荷）诺伊，（斯洛文）姆尔瓦，（斯洛文）巴塔盖尔吉著；林枫译. 蜘蛛：社会网络分析技术[M]. 北京：世界图书出版公司北京公司，2012. 10.

[23]Ravasz E,Somera A L,Mongru D A,et al. Hierarchical organization of modularity in metabolic networks.science, 2002, 297:1551—1555

作者简介：作者一马丹（1989——），女，硕士主要研究领域为应用语言学，语言复杂网络。Email: 1250242179@qq.com；**作者二**赵怿怡（1982——），女，博士，副教授、硕士生导师，主要研究领域为应用语言学、依存语法、语言复杂网络。Email: zhaoyiyi@xmu.edu.cn；