心智与计算, Volume 1 (2007), 142-146
文章编号：MC - 2007-011
收稿日期：2006-09-30
出版日期：2007-01

面向隐喻计算的语料库研究和建设

李剑锋，杨芸，周昌乐
（厦门大学艺术认知与计算实验室，福建 厦门 361005）
victorli@live.com

摘要：汉语隐喻计算化研究是中文自然语言处理领域的一个前沿课题，然而研究所需要的隐喻语料资源却极其有限，对隐喻计算研究的发展形成了一定的对影响。本文主要研究了面向隐喻计算的语料库的设计和构建方法，并设计构建了一个具有一万句规模 62 万字的汉语隐喻标注句库以及一个面向隐喻识别的汉语常用动词搭配库。这些资源的构建建立在对隐喻现象的深入分析基础之上，是隐喻计算模型研究的基础性资源，对汉语隐喻的认知计算有着积极意义。

关键词：隐喻计算；汉语隐喻句库；动词搭配语料库
中图分类号：TP18 **文献标识码：**A

Corpus Designing and Constructing for Metaphor Computation

LI Jian-feng, YANG Yun, ZHOU Chang-le
(Mind-Art-Computation Laboratory of Xiamen University, Xiamen 361005, China)

Abstract: The research of Chinese metaphor computation is one of the leading issues in the NLP field; however inadequate Chinese metaphor corpus limited development of it. This paper introduces the design and construction of corpus for metaphor computation; it constructs a Chinese metaphor corpus, which contains about 10,000 tagging sentences, 620,000 words, and a verb collocation corpus of common Chinese verbs. Construction of these corpuses is based on the in-depth analysis of the phenomenon of metaphor; they are the basic resources for the research of computational model of metaphor, and have a positive effect on Chinese cognitive computation of metaphor.

Key words: metaphor computation; Chinese metaphor corpus; verb collocation corpus

本项目受到国家自然科学基金资助（项目号：60373080）

隐喻是一种普遍的语言现象，是人类重要的认知思维方式。自然语言的隐喻性特征愈来愈受到学者的重视和认同，隐喻计算研究逐渐成为自然语言处理研究领域的一项重要的前沿性研究课题。国外对隐喻的计算模型的研究已经开展了一定的时间，而且取得了很多成果，其中就包括英语隐喻各种语料资源的建设，但是国内对汉语隐喻计算模型的研究还处于非常初期的阶段[1,2]，汉语隐喻语料资源，特别是面向计算的隐喻语料资源还非常缺乏，然而自然语言的研究离不开真实语料的支持，本文的主要工作就是设计并构建隐喻计算研究所需要的基础资源，为隐喻的研究做基础性工作，并为丰富汉语隐喻研究语料资源做一些工作。

1 汉语隐喻句库的构建

1.1 建库目的

隐喻的认知计算，离不开真实语料作为基础[3]，然而面向计算的汉语隐喻句库却极为贫乏。为满足隐喻分析和研究的需要，我们设计并构建了一个大规模汉语隐喻标注句库。

1.2 语料来源

隐喻句库的隐喻句均摘自与《读书》1980-1997 年间的文章语料库，共 47 万句，内容覆盖法律、政治、经济、历史人物等各个领域。我们采用人工判断的方式，以不超过 50 个词的句子为单位完成对《读书》中隐喻句子的判断与采集。

我们采用隐喻的广义定义，认为“隐喻并不能与我们理解的暗喻相等同，即隐喻更多指的是一种思维方式，当我们把某物 A 类比为 B 时，使用 B 范畴的词语来表述 A。”这样条件下的语句都归于隐喻范畴。

1.3 结构设计与标注

本隐喻句库采用文本方式进行存储，总共包含约一万个隐喻句，约 62 万字，涵盖了极为丰富的汉语隐喻现象，并具有可扩展性。

在生隐喻句库的基础上，进行句法分析和标注[4,5]，带有句法分析标注的隐喻句库通过哈尔滨工业大学信息检索研究室依存句法分析系统生成初步结果最后由人工校正来完成。这样，带有句法分析标注的隐喻句库中，每个隐喻句采用依存方式表示句子结构以及句内每个成分之间的关系[6]。

句库的结构设计如下：每一个隐喻句显示为三行，第一行是无标记的生隐喻句；第二行为句法成分的切分与词性标注后的隐喻句，句子中每个词及词性的前面加上序号，句子末尾的句号由标志<EOS>标识；第三行是隐喻句的依存句法关系。依存关系中，每个关系以一个依存对表示，依存对中的第一个词是核心，支配第二个词，如：“[2]公司_[1]我(ATT)”这个依存对表示“我”和“公司”存在依存关系 ATT，其中，“公司”是这个关系的核心成分，“我”依存于“公司”。依存对之间以两个 tab 制表符相隔，整个句子的依存中心单独列出，由标记(HED)标识，并与句尾句号构成一个依存对，如图 1 所示：

雷霆无情地怒吼，
[1]雷霆/n [2]无情/a [3]地/ui [4]怒吼/vg [5]，/wp [6]<EOS>/<EOS>
[3]地_[2]无情(DI)　　[4]怒吼_[1]雷霆(SBV)　　[4]怒吼_[3]地(ADV)　[6]<EOS>_[4]怒吼(HED)

图 1 句法分析示例
Fig.1 Example of syntactic parsing

句库的标注标准如下：词性标注使用的是 863 的标注体系的词性标准，包括 28 个词性标记；依存关系标记 24 个，如表 1 所示：

表 1 依存关系标记

Tab.1 Dependency relation tags

定中关系 ATT（attribute）	数量关系 QUN（quantity）
并列关系 COO（coordinate）	同位关系 APP（appositive）
前附加关系 LAD（left adjunct）	后附加关系 RAD（right adjunct）
动宾关系 VOB（verb-object）	介宾关系 POB（preposition-object）
主谓关系 SBV（subject-verb）	比拟关系 SIM（similarity）
核心 HED（head）	连谓结构 VV（verb-verb）
关联结构 CNJ（conjunctive）	语态结构 MT（mood-tense）
独立结构 IS（independent structure）	状中结构 ADV（adverbial）
动补结构 CMP（complement）	“的”字结构 DE
“得”字结构 DEI	“地”字结构 DI
“把”字结构 BA	“被”字结构 BEI
独立分句 IC（independent clause）	依存分句 DC（dependent clause）

1.4 隐喻句库辅助编辑工具

由于句库中标记繁多，而且依存结构以线性排列为主，使得句库依存结构的可读性不高，为了便于可视化查看隐喻句库中各隐喻句的句法依存关系结构并方便研究人员对句库中信息进行插入、删除以及修改等各项操作，我们还设计了基于本标注隐喻句库的“所见即可得”的编辑器，它具有很好的交互能力查看和修改功能，如图 2 所示。

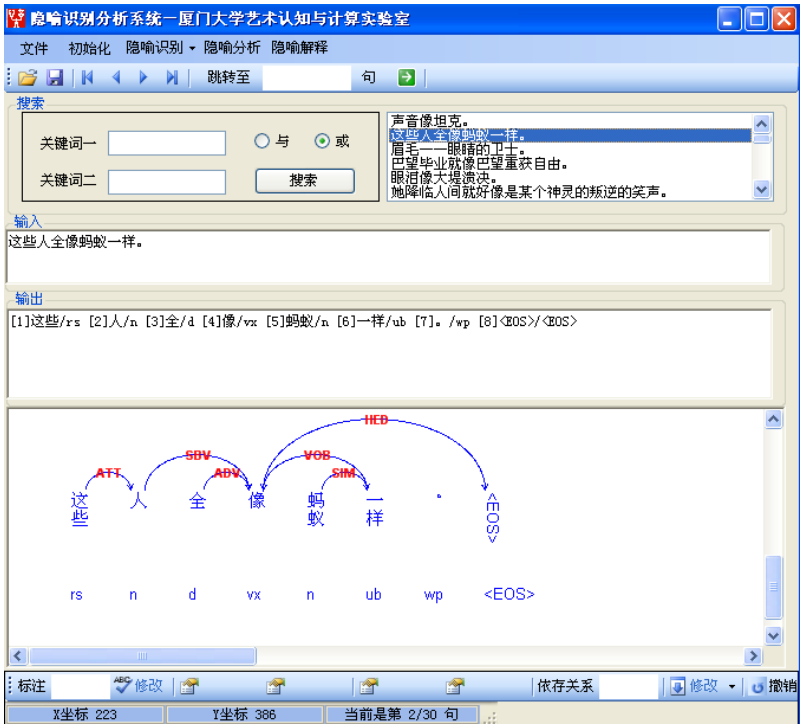

图 2 依存句法编辑器

Fig.2 Syntax dependence-relationship editor

依存句法标注隐喻句库及其辅助编辑工具，为隐喻句的分析和研究提供了丰富的资源和便捷的工具，是隐喻分析研究中的重要基础性资源。

2 常用汉语动词搭配语料库的建设

2.1 建库目的

常用动词搭配语料库[7]的设计出是面向隐喻句自动识别的需要，目前我们的隐喻自动识别研究需要用到浅层的汉语知识，如概念类别信息以及汉语语义搭配信息等，本项工作以动词为起点，探寻语义搭配信息库的一种可行的结构化设计方法，并应用于今后的各类相关资源的建设当中。

2.2 语料来源

动词搭配语料库的语料来源是《汉语常用动词搭配词典》(A Collocational Dictionary of Common Chinese Verbs，王砚农，焦庞颢，外语教学与研究出版社，以下简称《搭配词典》)。该词典共约 70 万字，收录了 1273 个常见动词的语法语义搭配信息。

《搭配词典》以每个动词为一个词条，以动词的意义为区分点，所涵盖的搭配关系包括了主要的动词用法，搭配关系包括主、宾、状、补四种。每一种语法语义搭配信息都被清晰地被分类和标识出来。

2.3 结构设计与标注

动词搭配类型共有四种，这里为了和依存句法标注隐喻句库相一致，搭配标记的设计采取如表 2 所示的标记来对应搭配词典中的搭配类型：

表 2 搭配类型

Tab.2 Collocation types

词典中的搭配类型	语料库中的搭配类型标记
[主~]	SBV
[~宾]	VOB
[~补]	CMP
[状~]	ADV

搭配库的结构设计如下：每一个动词以不超过 5 行的信息来描述：第一行为动词，并用一对大括号标记，第二行为主谓搭配，由[SBV]标记；第三行为述宾搭配，由[VOB]标记；第四行为动补搭配，由[CMP]标记；第五行为状中关系，由[ADV]标记。每行标记之后词紧跟每个语义上可与之搭配的词语，词语之间以“,”号相隔。由于动词具有及物或不及物的不同类型，因此不一定具有全部所列举的四种搭配情况，每个动词包括几种搭配则根据词典所提供的信息来进行设置，若不足四种搭配情况的，则按照 SBV，VOB，CMP 和 ADV 的先后进行排列。如图 3 所示：

{值得}
[SBV]很,也
[VOB]买,看,推广,学习,同情,重视,参观,研究,争吵,去写,赞许,怀疑,注意,考虑,深思,讨论,去听一听
[ADV]确实,真,特别,非常,一点也不,完全,根本不

图 3 汉语动词搭配语料库中文本示例

Fig.3 Example of text in verb collocation corpus

2.4 汉语动词搭配语料库辅助编辑工具

为配合搭配语料库的查看与编辑，我们也开发了与之相配套的专用检索工具，可以很方便的检索以及编辑修改一个动词在指定的搭配关系下常用搭配的词语，该工具对于对搭配语料库的扩充才作具有良好的性能。如图 4 所示：

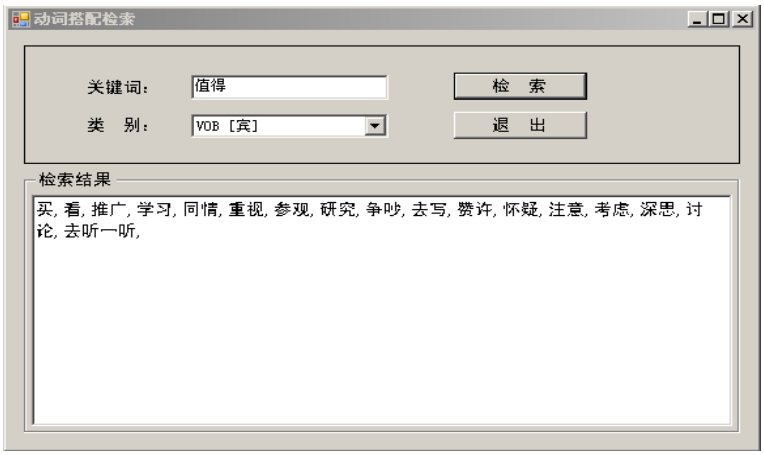

图 4 动词搭配检索工具

Fig.4 Search tool for verb collocation corpus

3 结语

在语言研究中，语料库方法是一种经验的方法，它能提供大量的自然语言材料，有助于研究者根据语言实际得出客观的结论，这种结论同时也是可观测和可验证的。在计算机技术的支持下，语料库方法对语言研究的许多领域产生了越来越多的影响。各种为不同目的而建立的语料库可以应用在词汇、语法、语义、语用、语体研究，社会语言学研究，口语研究，词典编纂，语言教学以及自然语言处理、人工智能、机器翻译、言语识别与合成等领域。

本文涉及并构建的两个语料库就是为了隐喻计算而专门建设的语料库，这两个语料库具有较大的规模，对于隐喻规律的认识和隐喻的自动识别的研究都具有重要的基础性意义。随着隐喻计算研究的发展我们将在现有语料资源的基础之上继续开展资源的构建与扩充工作，主要包括：1)扩充本文两个语料库的规模，特别是为了提高计算机隐喻自动识别的能力我们需要进一步扩大汉语常用动词搭配的规模以及设计构建除动词之外比如常用形容词等其他重要的搭配信息库。2)进一步设计隐喻信息的标注等。

参考文献：

[1] 周昌乐.心脑计算举要[M].北京：清华大学出版社,2003.
[2] 杨芸,周昌乐，等.基于机器理解的汉语隐喻分类研究初步[J].中文信息学报,2004，18(4):31-36.
[3] 黄昌宁,李娟子.语料库语言学[M].北京:商务印书馆,2002.
[4] 崔刚,盛永梅.语料库中语料的标注[J]. 清华大学学报:哲学社会科学版. 2000,1(15):89-94.
[5] Marcus M , Santorini B,et al. Building a Large Annotated Corpus of English : the Penn Treebank[J]. Computational Linguistics,1993, 19(2):313 - 330.
[6] 尤昉,李涓子,等.基于语义依存关系的汉语语料库的构建[J].中文信息学报, 2002,17(1):46-53.
[7] 白妙青,郑家恒.动词与动词搭配方法的研究[J].计算机工程与应用,2004.

作者简介：李剑锋（1983～），男，硕士研究生。